ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

MODELS FOR THE ANALYSIS OF CONTINUOUS OUTCOME LONGITUDINAL DATA WITH SPECIAL EMPHASIS ON MIXED MODELS

By

Chrisovaladis C. Malesios

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2002

ISBN: 960-8287-42-1

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

MODELS FOR THE ANALYSIS OF CONTINUOUS OUTCOME LONGITUDINAL DATA WITH SPECIAL EMPHASIS ON MIXED MODELS

By

Chrisovaladis C. Malesios

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
March 2003

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΟΝΤΕΛΑ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΕΧΩΝ ΔΙΑΜΗΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΙΔΙΑΙΤΕΡΗ ΕΜΦΑΣΗ ΣΤΑ ΜΙΚΤΑ ΜΟΝΤΕΛΑ

Χρυσοβαλάντης Χ. Μαλέσιος

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Μάρτιος 2003

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfilment of
the requirements for the degree of
Master of Science

MODELS FOR THE ANALYSIS OF CONTINUOUS OUTCOME LONGITUDINAL DATA WITH SPECIAL EMPHASIS ON MIXED MODELS

Chrisovaladis C. Malesios

Supervisor:
Dr. E. Xekalaki
Professor

External examiner:
N. Balakrishnan,
Professor, Department
of Mathematics & Statistics,
McMaster University,
CANADA

Athens, March 2003

Michael Zazanis, Associate Professor
Director of the Graduate Program

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Evdokia Xekalaki for her help and guidance during the writing of this dissertation. I would also like to thank the Ph.D. student George Makatis for his encouragement, and everyone else who helped in anyway to complete this thesis.

VITA

I was born in Athens in 1973. I finished High School in 1990. I became a student in the University of Athens, department of Mathematics, in September 1990. I graduated in September 1996. In September 1999 I was accepted as a post-graduate student in the International Master Program in Statistics of the Athens University of Economics and Business.

ABSTRACT

Chrisovaladis Malesios

MODELS FOR THE ANALYSIS OF CONTINUOUS OUTCOME LONGITUDINAL DATA WITH SPECIAL EMPHASIS ON MIXED MODELS

March 2003

The aim of the present thesis is to provide a comprehensive review of the existing work on the statistical analysis of continuous outcome longitudinal data. In particular, we provide in detail, an extensive and unified overview of all available modelling strategies and inferential procedures for the statistical analysis of continuous response longitudinal data. We begin by reviewing approaches that are considered to be classical approaches to longitudinal data modelling, namely univariate analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) for longitudinal data, and then proceed to more recently developed approaches for the modelling of (continuous-type) longitudinal data.

One of the newest developments is the incorporation of mixed models to longitudinal data analysis. This is based on the mixed model methodology initially developed in the animal breeding field, and results in very general and flexible models for handling continuous-type longitudinal data. We illustrate in as much detail as possible the implementation of the general linear mixed model (GLMM) in longitudinal data analysis, by reviewing a great part of the existing literature on the subject. Besides the important issue of parameter estimation (for both fixed and random parameters of the model), we treat the question of modeling of the covariance/correlation structure of longitudinal data and discuss the ways that have been considered in the literature on how the latter may be incorporated into the analysis by means of the GLMM. Special emphasis is put on describing the various computational procedures existing in the literature, for estimating the unknown variance parameters of the model (e.g. iterative algorithms such as Newton-Raphson and Expectation-Maximization).

While the majority of work on methods for analysis of continuous longitudinal data has been focused on response data that are linear in their parameters, a significant number of applications required utilization of Non-linear Mixed Effects (NLME) models. Here, we review the proposed methodology for analysing longitudinal data via NLME models, and present the most important and widely applied estimation procedures.

Finally, we provide a review on the most popular plotting techniques for representing longitudinal data (e.g. parallel plot, Draftman's display), since graphical representations may assist significantly in statistical modeling analysis by revealing useful features of the data.

ΠΕΡΙΛΗΨΗ

Χρυσοβαλάντης Μαλέσιος

ΜΟΝΤΕΛΑ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΕΧΩΝ ΔΙΑΜΗΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΙΔΙΑΙΤΕΡΗ ΕΜΦΑΣΗ ΣΤΑ ΜΙΚΤΑ ΜΟΝΤΕΛΑ

Μάρτιος 2003

Σκοπός της παρούσας διατριβής είναι να δώσει μια περιεκτική παρουσίαση του υπάρχοντος έργου πάνω στη στατιστική ανάλυση των συνεχών διαμήκων δεδομένων. Συγκεκριμένα, παρέχεται μία εκτενής και ενοποιημένη σύνοψη των διαθέσιμων στρατηγικών μοντελοποίησης και των διαδικασιών συμπερασματολογίας για τη στατιστική ανάλυση των διαμήκων δεδομένων συνεχούς απόκρισης. Ξεκινώντας με την ανάλυση προσεγγίσεων οι οποίες θεωρούνται κλασσικές για τη μοντελοποίηση διαμηκών δεδομένων, δηλαδή τη μονομεταβλητή ανάλυση διακύμανσης (ANOVA) και τη πολυμεταβλητή ανάλυση διακύμανσης (MANOVA), προχωρούμε στην παρουσίαση περισσότερο πρόσφατων προσεγγίσεων για τη μοντελοποίηση των συνεχών διαμήκων δεδομένων.

Μια απο τις νεότερες αναπτύξεις είναι η ενσωμάτωση των Μικτών μοντέλων στην ανάλυση των διαμήκων δεδομένων. Αυτή βασίζεται στη μεθοδολογία των Μικτών μοντέλων, η οποία αναπτύχθηκε αρχικώς στο πεδίο της ζωικής αναπαραγωγής, και η οποία έχει ως αποτέλεσμα πολύ γενικά και ευέλικτα μοντέλα όσον αφορά τη διαχείριση συνεχών διαμήκων δεδομένων. Δίνεται μια αναλυτική περιγραφή της υλοποίησης του Γενικού Γραμμικού Μικτού Μοντέλου (GLMM) στην ανάλυση διαμήκων δεδομένων καλύπτοντας ένα μεγάλο μέρος της υπάρχουσας

βιβλιογραφίας πάνω στο θέμα. Πέραν του σημαντικού ζητήματος της εκτίμησης των παραμέτρων (σταθερών και τυχαίων), ασχολούμεθα επίσης με τη μοντελοποίηση της δομής της συνδιασποράς/συσχετίσεως των διαμήκων δεδομένων και τους τρόπους με τους οποίους η δομή αυτή μπορεί να ενταχθεί στην ανάλυση, μέσω του GLMM. Επιπρόσθετα, ιδιαίτερη έμφαση έχει δοθεί στην περιγραφή των διαφόρων υπολογιστικών διαδικασιών (π.χ. επαναληπτικοί αλγόριθμοι όπως οι αλγόριθμοι Newton-Raphson και Expectation-Maximization), για την εκτίμηση των αγνώστων παραμέτρων διακύμανσης του μοντέλου.

Παρότι η πλειονότης του έργου πάνω σε μεθόδους για την ανάλυση συνεχών διαμηκών δεδομένων έχει εστιασθεί σε δεδομένα τα οποία είναι γραμμικά όσον αφορά τις παραμέτρους τους, ένας σημαντικός αριθμός εφαρμογών απαίτησε τη χρησιμοποίηση μοντέλων Μη Γραμμικών Μικτών Παραγόντων (NLME). Εδώ, καλύπτουμε την μέχρι σήμερα προταθείσα μεθοδολογία για την ανάλυση διαμήκων δεδομένων μέσω των NLME μοντέλων, και παρουσιάζουμε τις πιο σημαντικές και ευρέως εφαρμόσιμες διαδικασίες εκτίμησης.

Τέλος, αναφερόμαστε στις πιο δημοφιλείς τεχνικές σχεδιασμού για διαμήκη δεδομένα (π.χ. parallel plot, Draftman's display), λόγω της σημαντικής βοήθειας που αυτά τα γραφήματα προσφέρουν στη στατιστική ανάλυση αποκαλύπτοντας χρήσιμα χαρακτηριστικά των δεδομένων.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Longitudinal data is a special type of multivariate data, consisting of time sequences of measurements (usually called repeated measurements), counts or categorical responses taken from one or more experimental units or subjects. In recent years, statisticians have put a great amount of effort into developing suitable statistical models for the analysis of this type of data. One of the main reasons for such a wide interest on studying longitudinal data in the last years is partially due to the fact that the modern technology and the recent developments of statistical software have greatly reduced cost and effort.

Over the years, a number of various approaches for representing longitudinal data in terms of a statistical model have been developed. Specifically, for discrete-type data such as binary and count data Generalized Estimating Equations (GEE) models are mainly implemented, while for continuous responses that are assumed to be normally distributed usually ones uses analysis of variance (ANOVA) for repeated measurements, multivariate analysis of variance (MANOVA) for repeated measurements or mixed model methodology for longitudinal data.

The aim of this thesis is to provide a comprehensive review of the existing work on the statistical analysis of continuous outcome longitudinal data, giving enough detail to serve as a useful reference to the interested reader. The material of the thesis is organized in six Chapters, the first of which is a short introduction and overview of the remaining chapters.

Chapter 2 introduces the reader to the nature of longitudinal data as well as of repeated measures data, two terms closely related to each other (Section 2.2) and discusses their main differences with more general types of data associated with them, such as general multivariate data or time series data (Section 2.3).

One of the basic and distinguishing characteristics of longitudinal data is that of covariance, especially exhibited by data on a specific unit. As is explained, this is quite natural to occur since longitudinal data consist of repeated measurements on the same units over time, and therefore it is expected to have covariation between the data collected on the same unit (Section 2.4). A detailed presentation of two classical methods used for the analysis of longitudinal data, the classical univariate analysis of variance (ANOVA) and the multivariate analysis of variance (MANOVA) is also made (Sections 2.5.1 and 2.5.2).

In particular, as concerns the ANOVA method, we present the ANOVA table and the F-tests based on that table for testing significance of fixed/random effects of the model, a method to estimate the fixed effects parameters, as well as a routine way to estimate the variance components of the ANOVA model. Of particular interest is the issue of advantages and most importantly the disadvantages of the above methods. The main disadvantage of these two methods (ANOVA, MANOVA) in the case of longitudinal studies is that they can be applied only under extremely special circumstances. For instance, the ANOVA method can be implemented only in balanced longitudinal designs (where we have same number of measurements for all units, taken at exactly the same times). Furthermore, in order for this method to provide valid inferences a very restrictive structure for the variance-covariance matrix of the within-units observations must be assumed (the compound symmetric structure). A discussion on this issue is presented in section 2.5.1.

On the other hand, MANOVA models do not require a specific (hence restrictive) structure for the within-units variance-covariance matrix, but instead assume a complete arbitrary structure. This turns out to be undesirable and unattractive too since, although one can cope with the problem of assuming the specific (and rather unrealistic for longitudinal data) structure of compound symmetry, one faces the additional problem of estimating a large number of unknown parameters (the variance components of the unstructured variance-covariance matrix of within-units observations).

For the latter reasons more recent approaches for the modelling of (continuous-type) longitudinal data have been developed, with most popular among them the *general linear mixed effects model for longitudinal data* [*Laird and Ware (1982)*, *Harville (1977)*], which is based on the theory of the general linear mixed model. Due to the latter relation, a review of the general linear mixed model (GLMM) theory is made in Chapter 3. In particular, a presentation of the general linear model (GLM) theory (Section 3.2) is made with specific emphasis on parameter estimation, i.e. ML, least squares and best linear unbiased estimation (BLUE) of the fixed effects of the model. The GLMM can be considered as an extension of the GLM, that introduces another source of randomness in the model except the random error, namely that of the random effects. Hence, a comparison between the two models (GLM and GLMM) is quite useful, especially in showing how the method of best linear unbiased estimation (BLUE) is extended to the so called method of best linear unbiased prediction (BLUP) in the case of estimating random effects of the model. It should be noted that again BLUP concerns estimators of the (realized values of) random effects, but according to a convention that has somehow been developed, the statistics estimating fixed effects are referred to as estimators, while those estimating random effects are referred to as predictors.

Section 3.3.1 presents the GLMM, and Section 3.3.3 gives the estimators of fixed effects and random effects of the model (BLUE and BLUP respectively). The mixed model equations (Section 3.3.4), due to *Henderson (1950)*, are of great importance since they provide the above mentioned BLUE and BLUP with less computational efforts, compared to classical methods of derivation, as the one of Section 3.3.1. In what follows, we calculate the solutions of Henderson's mixed model equations (MME in abbreviation), and prove that indeed these two solutions are equivalent to BLUE and BLUP. In the sequel, we concentrate our attention on the estimation of the variance components (or variance parameters) of the GLMM. (Section 3.3.4). Two close-related methods are considered, maximum likelihood (ML) and restricted maximum likelihood (REML) since both methods are very popular in the field of variance component estimation. REML is usually preferred over ML mainly due to that the former corrects the drawback of the latter to produce biased (downwards) estimates of variance components, by maximizing (with respect to the variance components) only the portion of the likelihood that does not depend on the fixed effects (Section 3.3.4.2).

The subject of Chapter 4 is exploratory data analysis, where we attempt to review the basic plots and graphical procedures used in longitudinal data analysis. The most popular plot for representing longitudinal data is the parallel plot (Section 4.2). The necessity that led statisticians to adopt the parallel plot as a visualization tool for longitudinal data is mainly the insufficiency of other classical graphical techniques such as the scatterplot to present multivariate data (in fact, the usefulness of scatterplots proves significant only in two-dimensional planes). Since longitudinal data usually consist of (a large number) of repeated measurements on different units, each set of repeated measures of each unit can be considered to be a high-dimensional point. Thus, in essence, the presentation of longitudinal data is the plotting of those high-dimensional points corresponding to each unit in a single plot. The scatterplot, as previously noted, is effective mostly in presenting 2-dimensional data points. The main difference between the scatterplot and the parallel plot, is that while the scatterplot preserves orthogonality between axes, in the parallel plot the axes are drawn parallel to each other. Of interest is the relation between the two plots, proven to be a geometrical coordinate transformation, in the context of projective geometry (Section 4.2.1). The utility of the parallel plot in the statistical field, as a tool for displaying and revealing useful information about longitudinal data and repeated measures data is demonstrated in section 4.2.2.

Chapter 4 concludes with another practical and useful exploratory graphical tool for longitudinal data, the so called Draftman's display (Section 4.3). The Draftman's display is the most common way for checking visually the, within-subject, covariance structure of a typical balanced longitudinal data set.

Chapter 5 is devoted to the presentation of the General Linear Mixed Model (GLMM) for longitudinal data, also known and as the 'Laird-Ware' model. In sections 5.2.1 and 5.2.2 the 'Laird-Ware' model is defined and the various forms under which it is met in the statistical literature are presented. Estimation methodology for the fixed-effects vector of the model as well as available methods for predicting its random effects are the subject of section 5.3.

An important advantage of mixed-model methodology is that it permits the (possible) covariation between measures on the same unit/subject to be incorporated into the statistical model. Various covariance structures can be adopted in order to model this, possibly existing, covariation of each subject's response vector. Section 5.4 focuses on this issue, and thoroughly reviews the most representative covariance structures, by covering a wide range of the available choices from a completely unstructured covariance pattern to more complex covariance patterns borrowed from time series analysis.

The issue of estimating the model's (unknown) variance components (already considered in Chapter 4) is again discussed, this time for the specific case of the 'Laird-ware' model (Section 5.5). An inherent difficulty with the estimation of both fixed-effects parameters and variance components in mixed model methodology (and accordingly with the Laird-Ware model), is the insufficiency to come up with closed-form solutions for the estimators of fixed effects and variance components since expressions for fixed effects estimates involve the unknown variance components and vice-versa. To this end, numerical iterative techniques [such as the expectation-maximization (EM) algorithm (Section 5.5.1) or the Newton-Raphson (NR) algorithm and variations (Section 5.5.3)], must be employed in order to overcome this problem. The implementation of iterative schemes such as the above provides us with estimates of fixed-effects parameters and variance parameters simultaneously, using a unified procedure. In both cases, the procedure requires initial values of the parameters and using information on the slope of the likelihood surface the current estimates are moved in a direction that increases the log-likelihood of the data. The iterations continue until a satisfactory degree of convergence is reached. Our review on the specific subject includes ML/REML estimation via the EM algorithm (Sections 5.5.2.1 and 5.5.2.3, respectively), as well as ML/REML estimation via the NR algorithm (Sections 5.5.5.1 and 5.5.5.2, respectively). Also section 5.5.6 provides a description of a variant of NR, namely the Fisher scoring algorithm, very often used to compute estimates of fixed-effects and variance components of the Laird-Ware model. Moreover, the comparison of the two principal numerical algorithms (i.e., the EM and the Newton-Raphson algorithms) is attempted in subsection 5.5.7. Tests of hypotheses associated with the fixed-effects are of great practical importance in mixed-model analysis longitudinal data. Section 5.6 focuses on the presentation of the most commonly used tests, namely the Wald test statistic (subsection 5.6.1), the likelihood ratio test (subsection 5.6.2) and the F-test (subsection 5.6.3).

A slightly modified approach to mixed-effects modeling of longitudinal and repeated measures data came from *Diggle (1988)* who developed a parametric model that suggests an alternative specification for the (within-subject) error term's variance-covariance matrix compared to the Laird-Ware model, and in this sense may be viewed as an extension of the GLMM for longitudinal data (cf. Section 5.7). The semivariogram (*Matheron, 1963*), a very important graphical technique for visualization and validation of the covariance structure of the within-subject data, devised by *Diggle (1988)* is the topic of subsection 5.7.2. Finally, some remarks on available software for the statistical analysis of longitudinal data via linear mixed model methodology are offered in section 5.8.

While most of the developed theory on mixed effects modeling of (continuous) longitudinal data has focused on data where each subject's response is assumed to be linear in both the fixed effects and the random effects (GLMM for longitudinal data), there are often situations where longitudinal data are inherently nonlinear with respect to a given response function, say $f(\cdot)$. Data of this type are common in many applications, for example, pharmacokinetic and pharmacodynamic studies. As a consequence, the general linear mixed model does not seem appropriate enough to describe the relationship between individual response vectors and the unknown parameters of interest. Thus, naturally, the need for developing more general models that allow for the mean response function to be nonlinear in the parameters became apparent. To this end, a great deal of attention (parallely to the GLMM) has also been given to nonlinear mixed effects models for longitudinal and repeated measures data. In Chapter 6, we present an overview of nonlinear mixed effects (NLME in abbreviation) models and the associated estimation procedures currently used for the analysis of continuous response longitudinal data. A very important statistical challenge in the developing of nonlinear mixed model methodology is to circumvent the problem caused by the fact that random effects enter the model in a nonlinear fashion, making hence the evaluation of the full data likelihood function an extremely difficult task. To date, several methods have been developed in dealing with this significant statistical challenge. While nonparametric, semiparametric and Bayesian methods have been proposed, the principal approach until now has been the approximation of the likelihood in a fully parametric context. Special emphasis is placed on describing in as much detail the most salient of these methods, appeared in the literature. In particular, Chapter 6 is organized as follows. Section 6.1 consists of a brief introduction to NLME models. In section 6.2 we present the NLME model. Section 6.3 presents the widely-used approximation method of *Sheiner* and *Beal (1980)*, based on a first-order Taylor series expansion. Ideas similar to the approach of Sheiner and Beal have been used by *Lindstrom* and *Bates (1990) (*cf. Section 6.4) In the sequel (Section 6.5), the Laplacian approximation method, which has proven to be close-connected to the Lindstrom and Bates approximate estimation method is treated. Approximations to the log-likelihood function based on Gaussian quadrature rules are discussed in section 6.6. We then address nonlinear model formulation and parameter estimation in a non- and semi-parametric framework (Section 6.7). The considerable literature on Bayesian approaches for treating nonlinear models for longitudinal data is the topic of section 6.8. Finally, available commercial packages for the implementation of NLME models are discussed in section 6.9.

Chapter 2

Longitudinal Data

2.1 Introduction

One of the main interests of statistical science is to draw conclusions for some population, finite or infinite. (In most occasions the populations are finite, except for some situations where we are confronted with finite populations of very large number of elements, or situations where for the convenience of the statistical inference we assume that the population is infinite). The best way to derive information about the population of interest would of course be the collection and examination of the entire set of elements of the specific population. Of course, in practice this is not always feasible. Very often data are only available for a subset of the population, or the collection of the elements of the entire population is simply impossible due to the large size of the specific population. In such cases, the practicing statistician collects information, by using only this small group (or subset) of the complete population, so that the desired conclusions can be sufficient and representative of the population under study. Such a given subset of population elements is generally called a *sample*, and the methods and principles for the collection and analysis of data of (finite) populations constitute the branch of Statistics known as: ‘*Sample Survey Methods’*, or ‘*Sampling’*. Sample Surveys are nowadays widely accepted as a means of providing (statistical) data on an extensive range of subjects for both research and administrative purposes, and have become a powerful tool for research in a wide area of applications, such as industry, education, social psychology, sociology and many more.

The objective of sample design is how to select the part of the population to be included in the survey. For this purpose, various techniques have been developed, each subject to the specific population and its associated characteristics (e.g. simple random sampling, systematic sampling, stratified sampling, multistage sampling, cluster sampling etcetera). The need for sampling from larger and more widespread populations led statisticians to move from relatively simple techniques (e.g. simple random sampling), towards more complex techniques. The most commonly applied types of survey designs are listed below:

- *Simple cross-sectional:* Cross-sectional is the design in which each subject is measured or evaluated on only one occasion. Thus, we have only one round of data, collected at a specific time point. The cross-sectional design may be regarded as the simplest survey design, lacking however the availability to investigate and evaluate changes over time.

- *Stratified:* The entire population is categorized into subgroups (strata) and a random selection is made within each of the strata. The number and categories (or strata) selected, depend on the needs of the survey. This design provides greater assurance that each type of category is adequately represented and can be analyzed.

- *Multi-subject:* Data are collected using the same sample on the same survey on a number of different topics. In this design, the survey results are added on to a survey that has already been designed and whose sample has already been selected. Designs of this nature find practical application in cases for example where we have only a few questions to ask, and therefore it could be appropriate to include these questions in an already existing survey. The great disadvantage of a multi-subject design is that by using it we are lacking the opportunity to select a survey design that fits exactly to our needs.

- *Longitudinal:* In general terms, a longitudinal design is a design in which each subject (or unit) is measured (or observed) on multiple occasions over a period of time. The specific design allows the tracking, and analysis of changes over a period of time. Longitudinal studies have gained a great interest in the social sciences, such as sociology, psychology and medical science. One of the main reasoning for this vast interest on longitudinal studies in the last years is partially due to that the modern technology and recent developments of statistical software has greatly reduced the cost and effort. Three are the basic types of longitudinal survey, namely:

- *Trend studies*
- *Cohort studies*
- *Panel studies*

Trend studies formulate a type of longitudinal survey that uses different individuals in order to conduct a study over time; a cohort study on the other hand, while still uses different individuals over time, is focusing on the same group of people. By definition, it is a study that uses the same specific population each time, but uses different samples. Finally, a panel study is a type of longitudinal survey that studies the same people over time.

The purpose of the current thesis is to review, in an extensive manner, the literature associated with the statistical analyses of (continuous response) longitudinal data of panel forms. Data of this type have proven to be very important in medical research and in many other disciplines, and various techniques have been devised for their analysis. Our intention is to provide a coverage of all available approaches for handling and analyzing such data, ranging from classical (cf. univariate analysis of variance; multivariate analysis of variance) to more recently developed (cf. mixed model methodology).

2.2 Longitudinal Data/Repeated Measures Data

By the term 'repeated measures', we refer to data with multiple observations on the same sample element (or unit). Essentially, data of this kind are Panel Designs where the data are collected for the same sampled elements on each round, and are in contrast to cross-sectional panels where observations are taken at only one fixed point in time.

In most cases, the multiple observations are taken over time, but they could be over space for example. Exactly this essence of the repeated measurement occasion (e.g. time or space), defines a slightly departure between these two terms; longitudinal and repeated measures data. So, we can define longitudinal data, as data in which the repeated measurement occasion is strictly time. That is, although repeated measurement most often takes place over time, this is not the only way measurements may be taken repeatedly on the same unit. In other words, longitudinal data consist of time sequences of measurements, counts or categorical responses taken from one or more experimental units or subjects. In this manner, longitudinal data can be viewed as a subset of repeated measures data.

Any data set in which subjects are measured repeatedly over time can be described as longitudinal data. Seeking for a more formal and a firmer statistically based definition, we have:

Definition 2.1: Longitudinal data *(and consequently repeated measures data), are multivariate observations on m units (or elements), with n_i (repeated) measurements on the ith unit and total number of measurements for all units: $n = \sum_{i=1}^{m} n_i$. $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in_i})^t$ is the $(n_i \times 1)$ vector of the repeated observations on subject i, and observations y_{ij} on each subject i, are taken at times t_{ij} with $\mathbf{t}_{ij} = (t_{i1}, t_{i2}, \ldots.., t_{in_i})^t$. The vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots.., \mathbf{y}_m)^t$ represents the entire set of measurements on the m units.*

It is a straightforward consequence of the above definition that:

$$y_{ij} = \textit{the jth measurement taken on unit } i$$

and also:

$t_{ij} =$ *the time at which the jth measurement on unit i was taken*

This vector notation has been proven quite helpful, since by this way we can refer and also manipulate the entire set of the repeated measurements of a single unit as a unique mathematical entity and consequently summarize and display in an elegant form observations on each unit. In fact, it is standard in longitudinal studies to think of all the data from a particular unit together, so that the complex relationships over time may be summarized. It is worth noting that vector notation and more generally matrix notation is very popular for summarizing longitudinal data. This is indeed the case in the literature, particularly when discussing some of the newer methods, where matrix algebra is a basic tool not only for the presentation of longitudinal data but also in the development and analysis of the various statistical models available until now, which try to describe these data.

The number of measurements on each unit i is denoted by n_i. This implies that one can have different numbers of observations on the various units. Quite often on the other hand, especially for designed experiments, each subject has the same number of observations taken at the same set of time points. Thus, for this special case we have $n_i = n$ for all i and $t_{ij} = t_j$ for all i and $j = 1, 2, \ldots.., n$. Such data is generally referred to as *balanced longitudinal data.* Otherwise the data are *unbalanced.* It happens that although balanced data is usually the aim, unbalanced longitudinal data do arise for a variety of reasons; occasionally the data are unbalanced or incomplete by design; in general however, the main reason for unbalanced data in a longitudinal study is the occurrence of missing values, in the sense that intended measurements are not taken, are lost or are otherwise unavailable.

2.3 Longitudinal Data and Associations with other types of Multivariate Data

2.3.1 Longitudinal Data/Multivariate Data

Longitudinal data can be considered without question as multivariate data. In fact, they are a special form of multivariate data, in the sense that by their definition, certain restrictions are introduced in the collection of such data, restrictions that are not incorporated in the general multivariate data definition.

To be more specific, Multivariate data refers to the case where the same unit is measured *on more than one* outcome variable, for example one variable could be the height of an individual, another the weight or blood-pressure, all of them measured at the same time on each individual in the study. By considering the repeated measurements on each subject as response vectors, the first noticeable difference in compare to longitudinal data, is that here the vectors consist of measures on different variables, and consequently the measurements are incomparable to each other.

Longitudinal responses on the other hand, are quite different. Here, the response vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t$ of the ith unit consists of repeatedly measuring the same quantity (same variable), hence only a single variable is measured.

2.3.2 Longitudinal Data/Time Series

A question that naturally arises from the definition of longitudinal data is whether such data relate to the familiar time series data. Indeed, such kind of data contain elements of multivariate data and time series data. In fact, it is not easy to draw a distinct boundary between longitudinal data and time series data. In general, time series methods are more appropriate to analyze long series of data. A single time series consists of one variable measured for one object on at least say twenty and often many more occasions (e.g. consecutive years), whereas longitudinal data rarely contain ten observations per subject but consist of many subjects.

The distinction between these two data types relates more to the analysis methods, with forecasting and econometric methods applied to time series data and multivariate, categorical and cross sectional analyses for longitudinal data, as *Kloesgen (1999)* notices.

2.4 Covariance: The Distinguishable Feature of Longitudinal Data

The primary distinguishing feature of longitudinal data is the assumption/restriction of covariance incorporated in the within-subject observations. Hence, in the statistical analysis of longitudinal data is addressed the issue of covariation between the repeated measures on the same unit. But what exactly has driven statisticians to introduce this restriction in the analysis of longitudinal data? The answer is quite simple; if we consider as $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t$ the vector of responses for unit i, collected over time, it is natural to be concerned about the possibility of correlation existing among them.

For example, consider a clinical trial, where repeated measurements (most often with respect to time) are collected from patients in order to examine the effect of a new treatment on a disease process over time. This is a typical example of longitudinal data where the units are the individuals (patients) who participate in the study, and the data consist of measurements of some characteristic of the patients taken at multiple time points (a period or a visit). Unquestionably, it is much more realistic to assume that the measurements on the same patient are more or less related to each other, than to consider otherwise.

In other words, it is rather plausible to expect for two consecutive measurements of the same individual to 'vary' together. Hence, a large for example value for the one measurement suggests that the other value could be large too. Respectively, small values of the first measurement suggests small value for the second, too. As is well known, a measure of how two random variables 'vary' together is the covariance. Formally, if

y_i, y_j are two random variables that follow probability distributions with means μ_i, μ_j respectively, then we define the covariance between y_i and y_j as:

$$Cov\left(y_i, y_j\right) = E\left[\left(y_i - \mu_i\right)\left(y_j - \mu_j\right)\right]. \tag{2.1}$$

Since the repeated measurements y_{ij} of longitudinal data and the way they occur, is attributed to some probability model, they can be regarded as random variables, and consequently a measure of association between them, such as the covariance, can be defined.

As mentioned previously, if we denote by y_{ij}, $y_{ij'}$ $\left(j \neq j^{'}\right)$ the measurements within the same individual i, it is reasonable to think of some association (dependence) between them. Hence, we believe that these measurements tend to 'vary' together, or in other words they are positive together or negative together, and as a consequence of this, the product $\left(y_{ij} - \mu_{ij}\right)\left(y_{ij'} - \mu_{ij'}\right)$ in the covariance formula will be positive for most pairs of values. Thus we expect:

$$Cov\left(y_{ij}, y_{ij'}\right) \neq 0. \tag{2.2}$$

In an analogous manner, we have no reason to expect any kind of association between measurements of two different subjects. For convenience, we will denote as y_{ij}, $y_{i'j'}$ $\left(i \neq i^{'}, j \neq j^{'}\right)$ two random measurements of two different individuals. If y_{ij} and $y_{i'j'}$ are considered unrelated, then 'large' y_{ij} are very likely to happen with 'small' $y_{i'j'}$ and vice versa. This means that the deviations $\left(y_{ij} - \mu_{ij}\right)$ and $\left(y_{i'j'} - \mu_{i'j'}\right)$ will be positive and negative in no real systematic way. As a consequence of this, the product $\left(y_{ij} - \mu_{ij}\right)\left(y_{i'j'} - \mu_{i'j'}\right)$ may be negative or positive with no special tendency, and the average $E\left[\left(y_{ij} - \mu_{ij}\right)\left(y_{i'j'} - \mu_{i'j'}\right)\right]$ is likely to be zero. So we expect that if y_{ij} is an observation from individual (patient) i, and $y_{i'j'}$ is an observation from individual $i^{'} \neq i$, then:

$$Cov\left(y_{ij}, y_{i'j'}\right) = 0. \tag{2.3}$$

2.5 Statistical Modeling in Longitudinal Data

There are a number of various approaches for representing longitudinal data in terms of a statistical model, with the characteristic of most of them being regression-based models. In thick lines we can distinguish the statistical modeling analysis of longitudinal data in the following major categories:

- analysis of variance (ANOVA) for repeated measurements
- multivariate analysis of variance (MANOVA) for repeated measurements
- generalized estimating equation (GEE) models[1]
- mixed-effects models

While GEE and mixed-effects models are considered as recent developments, ANOVA and MANOVA models are considered to be classical approaches for analyzing longitudinal data, since the first analyses of such data were based upon these models. As a consequence of the long-term implementation and application of these models, and also due to the simplicity and connection with the familiar analysis of variance techniques, anova and manova methods and models for longitudinal data have become quite popular, and are often adopted by default, sometimes even without proper attention to the validity of the assumptions that these models take in account. In the following lines, we present briefly the outlying principles of the two methods and right afterwards their advantages and disadvantages are discussed in detail.

2.5.1 Univariate Analysis of Variance

The analysis of variance method for longitudinal data is applicable only in the special case where the data are under the restrictive balanced form; that is, where the vector of

[1] GEE models (initially developed for longitudinal data), provide a regression framework for analyzing correlated data that are not necessarily assumed to be normal, instead are usually correlated discrete-type data such as binary and count data.

the ith subject's measurements is given by $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})^t$, hence the responses of each individual (unit) occur at the same n times $\mathbf{t} = (t_1, t_2, ..., t_n)^t$ for all units, with no deviations from these times or missing values for any unit. This specific structure of the data, where we have m units and n measurements on these units, with each round of measurements collected at the same time, resembles that of a randomized block design. The role of factors here are played by the individual and time effects, considered random and fixed respectively. The levels of the individual factor are the individuals taken part in the longitudinal study, and correspondingly the levels of the time factor are the specific time points that the responses were obtained. Thus a classical analysis of variance model seems as a plausible way to proceed with the analysis (in fact the analysis of longitudinal data via the analysis of variance model is identical to the analysis of the split-plot design, considering the particular in this occasion, respect to time structure).

A common analysis of variance (ANOVA) model is described formally, by the following equation:

$$y_{ij} = \mu + \tau_j + b_i + \varepsilon_{ij} \qquad (i = 1, ..., m)\,, (j = 1, ..., n) \tag{2.4}$$

where:

y_{ij}: is the jth observation on the ith unit (response)

μ: is the overall mean

τ_j: is the time effect

b_i: is the individual effect, and

ε_{ij}: is a random error.

Examining the terms placed on the right hand side of the previous stated model (i.e. μ , $\tau_j, b_i, \varepsilon_{ij}$), we deduce that we have a model containing both fixed effects (τ_j), and random effects (b_i, ε_{ij}). This classification of effects as either fixed or random [see, e.g. *Searle (1971)* for a thorough discussion on fixed/random effects], appears to be a straightforward decision, due to the following reasoning; first of all, the individuals (subjects) considered in longitudinal studies, in an analogous way to every other statistical modeling procedure,are treated as a random sample from the (entire) population of subjects. The population is assumed to have zero mean and constant variance, say σ_b^2. The zero mean value is assumed for simplicity, and is a standard issue for analysis of variance models, since by using a linear transformation, we could define: $\mu^* = \mu + E(b_i)$ and $b_i^* = b_i - E(b_i)$. Then we can write the model (2.4) as:

$$y_{ij} = \mu^* + b_i^* + \tau_j + \varepsilon_{ij} \equiv \mu + b_i + \tau_j + \varepsilon_{ij} \tag{2.5}$$

where now for the equivalent model (2.5), we have $E(b_i^*) = 0$. Consequently, the b_i's which denote the effect (deviation from the overall mean) on the response y_{ij} due to the ith individual, are considered to be random variables from the same population. Moreover, especially in the case where the responses y_{ij} are continuous, the normal (Gaussian) distribution is often the most suitable one for modeling the components of the analysis of variance model. Thus, for the random effect b_i , representing the deviation caused by the fact that y_{ij} measurement is measured on the ith particular subject, we may assume:

$$b_i \sim N\left(0, \sigma_b^2\right) \tag{2.6}$$

Consider now the model term ε_{ij}. As previously noted, ε_{ij} represents an error term that serves the purpose of measuring the within-units variation, since responses vary not only between individuals but also because of variation within each individual, for example due to measurement error. In a similar way to the random effect b_i, we may think that measurement errors ε_{ij} come from a population of all possible measurement errors, and consequently can be thought of as random variables following a normal distribution with zero mean and variance σ_e^2. Thus, we may write:

$$\varepsilon_{ij} \sim N\left(0, \sigma_e^2\right) \tag{2.7}$$

Assumption (2.7) indicates constant variance for the random term ε_{ij} of model (2.4) and can be considered as a special case of the more general assumption $\varepsilon_{ij} \sim N\left(0, \sigma_{ej}^2\right)$

(by $Var\left(\varepsilon_{ij}\right)=\sigma_{ej}^{2}$ we assume different variances at each time point j), but since (2.7) is the most commonly applied case, we use this instead.

Worthwhile mentioning is the fact that we used different names to describe the two random terms in the model; **random effect** for the individual effect b_i, and **random error** for the measurement effect ε_{ij}. This was done on purpose, since in analysis of variance models, the term random effect is customary to describe a model component that addresses the among-unit variation, while the term random error is usually used to describe the model component that addresses the within-unit variation (*Davidian, 2000*).

The time effect τ_j on the other hand, denotes the effect of the jth time point ($j=1,.....,n$) on the response y_{ij}. This effect of time is considered to be fixed, since the measurements are collected at specific time points and the researchers' concentration is fixed upon just these times (the specific factor levels included in the study), and no others, meaning that inferences to be drawn are about the specific times that the data are collected.

Finally, random terms b_i and ε_{ij} are assumed to all be mutually independent (this represents the view that errors in taking measurements are of similar magnitude, regardless of the magnitudes of the individual deviations b_i associated with the units on which the observations are made). Furthermore, b_i's are assumed independent to each other, and the same stands for the fixed parameters τ_j.

2.5.1.1 The Covariance Matrix V of the Observations y_{ij}

The analysis of variance model (2.4) finds application to balanced data, where each subject i ($i=1,.....,m$) is measured at occasions j ($j=1,.....,n$). Due to this, we have $m \cdot n$ as the total number of observations. In the current section we are going to determine the covariance matrix, denoted by $\mathbf{V}$, of the total mn observations y_{ij}. More specific, we will show that $\mathbf{V}$, is in general, a block-diagonal matrix (i.e. a matrix that its off diagonal elements are zero, and its diagonal elements are submatrices), with a particular feature shared by these submatrices, that of having a specific form, known as compound symmetric form.

As an initial step, let us define the variance-covariance matrix (some authors use the term covariance matrix), of a random vector (a vector where its elements are random variables) $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = (y_1, y_2, ..., y_n)^t$. (the superscript t denotes the transpose).

Definition 2.2: *For a random vector* $\mathbf{y} = (y_1, y_2, ..., y_n)^t$ *we define as* **variance matrix** *or* **variance-covariance matrix V** *of* **y** *the matrix:*

$$\mathbf{V} = Var(\mathrm{y}) = E\left\{(\mathrm{y} - \boldsymbol{\mu})(\mathrm{y} - \boldsymbol{\mu})^t\right\} =$$

$$= \begin{pmatrix} E(y_1 - \mu_1)^2 & E(y_1 - \mu_1)(y_2 - \mu_2) & ...... & E(y_1 - \mu_1)(y_n - \mu_n) \\ E(y_2 - \mu_2)(y_1 - \mu_1) & E(y_2 - \mu_2)^2 & ...... & E(y_2 - \mu_2)(y_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(y_n - \mu_n)(y_1 - \mu_1) & E(y_n - \mu_n)(y_2 - \mu_2) & ...... & E(y_n - \mu_n)^2 \end{pmatrix} =$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \cdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix},$$

where:

$$\boldsymbol{\mu} = \begin{pmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix},$$

and for $j, k = 1, ..., n$ *we define:* $Var(y_j) = \sigma_j^2$ *and* $Cov(y_j, y_k) = \sigma_{jk}$.

As already mentioned in Section 2.3, the covariance between different measurements of the same subject is considered in general, non zero, while we assume that observations from different subjects have no apparent associations and therefore their covariance is taken to be zero.

We will see that this general principle is conveyed successively into the analysis of variance model (2.4), considering of course the model assumptions already stated. Before proceeding with the determination of the covariance matrix of the $m \cdot n$ (balanced) observations of the longitudinal study, we remind the analysis of variance model, which is expressed as :

$$y_{ij} = \mu + \tau_j + b_i + \varepsilon_{ij} \qquad (i = 1, ..., m), (j = 1, ..., n)$$

where: $b_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ constitute the random terms of the model, and τ_j is considered fixed. Data of this balanced longitudinal model consist of m vectors $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})^t = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{pmatrix}$, where each vector $\mathbf{y}_i$ is of $(n \times 1)$ dimension, including the measurements on subject $i, (i = 1, ..., m)$. By stacking together all the $\mathbf{y}_i$ vectors, one on top of the other, a single $(nm \times 1)$ vector, say $\mathbf{y}$, is formed which contains the entire set of all repeated measurements on all subjects i. Hence,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n} \\ \vdots \\ \vdots \\ y_{m1} \\ y_{m2} \\ \vdots \\ y_{mn} \end{pmatrix}$$

Essentially, and in accordance to Definition 2.2, we have to determine the variance-covariance matrix of vector $\mathbf{y}$, and this is achieved by finding the elements of $\mathbf{V}$, that is the variances as long as the covariances between all the y_{ij} measurements. To begin with, the variance of all y_{ij} measurements is the same across all subjects and is calculated as follows:

$$
\begin{aligned}
Var\left(y_{ij}\right) &= Cov\left(y_{ij}, y_{ij}\right) \\
&= E\left[\left(y_{ij} - \mu_{y_{ij}}\right)^2\right] \underset{from\ 2.4}{=} E\left[\left(\mu + b_i + \tau_j + \varepsilon_{ij} - \mu - \tau_j\right)^2\right] \qquad (2.8) \\
&= E\left(b_i^2 + \varepsilon_{ij}^2 + 2b_i\varepsilon_{ij}\right) = \sigma_b^2 + \sigma_e^2.
\end{aligned}
$$

[In the above calculations we used $E\left(b_i^2\right) = Var\left(b_i\right) = \sigma_b^2$, $E\left(\varepsilon_{ij}^2\right) = Var\left(\varepsilon_{ij}\right) = \sigma_e^2$ since $E\left(b_i\right) = 0$, $E\left(\varepsilon_{ij}\right) = 0$ and $E\left(b_i\varepsilon_{ij}\right) = 0$, since b_i, ε_{ij} are independent].

While the variance of each measurement y_{ij} is constant for all subjects, this is not the case for the covariances between the measurements of vector $\mathbf{y}$. The between-subjects measurements covariances differentiate from the within-subjects measurements covariances. For a formal verification, take y_{ij}, $y_{ij'}$ to be two measurements on the same subject i. Then:

$$
\begin{aligned}
& Cov\left(y_{ij}, y_{ij'}\right) = \\
& = E\left[\left(b_i + \tau_j + \varepsilon_{ij}\right)\left(b_i + \tau_{j'} + \varepsilon_{ij'}\right)\right] - E\left(b_i + \tau_j + \varepsilon_{ij}\right) E\left(b_i + \tau_{j'} + \varepsilon_{ij'}\right) \qquad (2.9) \\
& = \ldots\ldots = E\left(b_i^2\right) + \tau_j\tau_{j'} - \tau_j\tau_{j'} = E\left(b_i^2\right) = \sigma_b^2.
\end{aligned}
$$

[We have substituted y_{ij} and $y_{ij'}$ by $y_{ij} - \mu$ and $y_{ij'} - \mu$ since

$Cov\left(y_{ij}, y_{ij'}\right) = Cov\left(y_{ij} - \mu, y_{ij'} - \mu\right)$ as a consequence of the known covariance property:

$Cov\left(aX + b, cY + d\right) = acCov\left(X, Y\right)$ where a, b, c, d are constant real numbers and X, Y are random variables]. In an analogous way, the covariance between two measurements y_{ij} and $y_{i'j}$ that belong to two different subjects i, i' is calculated and can be shown that:

$$
Cov\left(y_{ij}, y_{i'j}\right) = 0 \qquad (2.10)
$$

Alternatively, we can summarize the above calculations in a unified manner, by using the Kronecker delta δ_{ij}, taking values $\delta_{ij} = 1$ for $i = j$ and 0 for $i \neq j$. If y_{ij} and $y_{i'j'}$ denote two measurements on different individuals i, i' at different time points j, j' , then:

$$\begin{aligned}
&Cov\left(y_{ij}, y_{i'j'}\right) = \\
&= E\left[\left(b_i + \tau_j + \varepsilon_{ij}\right)\left(b_{i'} + \tau_{j'} + \varepsilon_{i'j'}\right)\right] - E\left(b_i + \tau_j + \varepsilon_{ij}\right) E\left(b_{i'} + \tau_{j'} + \varepsilon_{i'j'}\right) \\
&= ..... = E\left(b_i b_{i'}\right) + E\left(\varepsilon_{ij}\varepsilon_{i'j'}\right) = \delta_{ii'}\sigma_b^2 + \delta_{ii'}\delta_{jj'}\sigma_e^2 = \\
&= \left\{ \begin{array}{c} \sigma_b^2 + \sigma_e^2,\ for\ i = i', j = j' \\ \sigma_b^2,\ for\ i = i', j \neq j' \\ 0,\ for\ i \neq i', j \neq j' \end{array} \right\}
\end{aligned}$$

Equations (2.8), (2.9), (2.10) provide us with all the necessary information needed to construct the variance-covariance matrix, $\mathbf{V}_i$, of the ith subject's vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})$ and also the variance-covariance matrix $\mathbf{V} = Var(\mathbf{y})$ of vector $\mathbf{y}$ of all measurements which is:

$$\mathbf{V} = \left(\begin{array}{ccc|ccc|c|ccc}
\sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\
\hline
0 & 0 & 0 & \sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \vdots & \ddots & \vdots & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 & \ddots & 0 & \cdots & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & \ddots & \ddots & 0 & \cdots & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2
\end{array}\right) \qquad (2.11)$$

To comment on the structure of variance-covariance matrix $\mathbf{V}$, we can say that consists of a diagonal series of m in total blocks (submatrices) of order $(n \times n)$:

$$\mathbf{V}_i = Var\left(\mathrm{y}_i\right) = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_e^2 \end{pmatrix} \qquad (2.12)$$

corresponding to the variance-covariance matrix of $\mathbf{y}_i$, the vector of measurements on each subject i $(i = 1, ..., m)$, each one of which has $\sigma_b^2 + \sigma_e^2$ as its diagonal elements and σ_b^2 as its off-diagonal elements. Observing the form of the submatrices, we notice that its main diagonal elements are equal to each other, and additionally its off-diagonal are also equal. This specific structure possessed by a variance-covariance matrix is known as compound-symmetry structure and correspondingly the matrix of this form is called a compound symmetric matrix.

Dividing each element of variance-covariance matrix $\mathbf{V}_i$ by the product of the square roots of the corresponding diagonal elements, in this case $\sigma_b^2 + \sigma_e^2$, we obtain the corresponding correlation matrix of $\mathbf{V}_i$:

$$\mathbf{R}_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, \qquad (2.13)$$

where $\rho = \sigma_b^2 / \sigma_b^2 + \sigma_e^2$. (Observing correlation matrix $\mathbf{R}_i$ one can easily deduce that it also has equal diagonal elements as well as equal off-diagonal elements, hence is also a compound symmetric matrix). Correlation coefficient ρ of $\mathbf{R}_i$, as a correlation between different measurements on the same individual, is called the **intraclass correlation coefficient** [see, e.g., *Hand* and *Crowder (1990)*, page 27]. It ranges from 0 to 1, taking the value $\rho = 0$ if and only if $\sigma_b^2 = 0$,corresponding to the null hypothesis H_0 of (2.17) described in the next section, indicating that no individual effects are present. If this is the case, random term b_i can be omitted from model (2.4). The case of $\rho = 1$ occurs when $\sigma_e^2 = 0$. This case does not occur in practice since it implies that there is no random (measurement) error and no intrasubject variability; however, it is possible for the intraclass correlation to be close to 1 indicating that the within-subject variation σ_e^2 is very small compared to the between-subject variation σ_b^2.

Finally notice that each of the $\mathbf{V}_i$ submatrices can be rewritten as:

$$
\begin{aligned}
&\begin{pmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 \end{pmatrix} = \sigma_b^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \\
&= \sigma_b^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \sigma_b^2 \mathbf{1}_n \mathbf{1}_n^t + \sigma_e^2 \mathbf{I}_n = \\
&= \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n, \qquad (2.14)
\end{aligned}
$$

where $\mathbf{1}_n$ is a vector consisting of n 1's, $\mathbf{I}_n$ is a $(n \times n)$ identity matrix and $\mathbf{J}_n$ is a $(n \times n)$ matrix of 1's. The benefit of this alternative notation will become evident later on, in the compound symmetry and sphericity section. A direct consequence of (2.14) is that the variance-covariance matrix $\mathbf{V}$ of vector $\mathbf{y}$ can now take the form

$$
\mathbf{V} = \begin{pmatrix} \mathbf{V}_i & 0 & \cdots & 0 \\ 0 & \mathbf{V}_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{V}_i \end{pmatrix} = \begin{pmatrix} \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n & 0 & \cdots & 0 \\ 0 & \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n \end{pmatrix}.
$$

2.5.1.2 The Analysis of Variance Procedure

The analysis of variance procedure for a (balanced) longitudinal study, from a computational point of view, is similar to that of a two-factor ANOVA experiment. As a consequence, under the assumptions that model (2.4) is correct and that the observations are normally distributed, it is possible to show that the usual F-ratios (also called Mean Square ratios) constructed using the principles of analysis of variance in order to test hypotheses on the parameters of the model, are still valid. Validity, as is well known, implies that these F-ratios have sampling distributions that are exact F distributions, central if the null hypotheses discussed above are true. The request of a correct model basically concerns the compound symmetry assumption presented in the previous section, yielding that each data vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in})^t$ exhibits the compound symmetry covariance structure of (2.12) (in fact, as it will be shown later on, the F-ratios are valid, under an even more general covariance structure of y_i, the type H covariance structure).

The analysis of variance procedure is summarized in the following table, known as analysis of variance table.

Table 2.1

Analysis of Variance for a Balanced Longitudinal Study

source of variation	*d.f.*	*sum of squares(SS)*	*mean squares (MS)*	*expected mean squares*
subject	$m-1$	$SS_B = n \sum_{i=1}^{m} (\overline{y}_{i.} - \overline{y}_{..})^2$	$MS_B = SS_B / m-1$	$E(MS_B) = \sigma_e^2 + n\sigma_b^2$
time	$n-1$	$SS_T = m \sum_{j=1}^{n} \left(\overline{y}_{.j} - \overline{y}_{..}\right)^2$	$MS_T = SS_T / n-1$	$E(MS_T) =$ $= \sigma_e^2 + m \sum \tau_j^2 / n-1$
error	$(m-1)(n-1)$	$SS_E =$ $= \sum_{i=1}^{m} \sum_{j=1}^{n} \left(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}\right)^2$	$MS_E = SS_E / (m-1)(n-1)$	$E(MS_E) = \sigma_e^2$
total	$mn-1$	$SS_{Total} = \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{..})^2$		

In the above table, $y_{..}$ represents the grand total of all the measurements y_{ij}, and $\overline{y}_{..}$ represent the grand mean of all y_{ij}. Similarly, $y_{i.}$ represent the total of the measurements under the ith subject and $\overline{y}_{i.}$ is the average of the n measurements under the ith subject.

Symbolically, we have:

$$
\begin{aligned}
y_{..} &= \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} \; and \; \overline{y}_{..} = \frac{y_{..}}{mn} \\
y_{i.} &= \sum_{j=1}^{n} y_{ij} \; and \; \overline{y}_{i.} = \frac{y_{i.}}{n} \\
y_{.j} &= \sum_{i=1}^{m} y_{ij} \; and \; \overline{y}_{.j} = \frac{y_{.j}}{m}
\end{aligned}
$$

Two are the hypotheses of main interest in model (2.4). One of the them is the hypothesis of zero time effects, that can be stated formally via the following hypothesis testing:

$$
\begin{aligned}
& H_0 : \; \tau_j = 0 \; for \; all \; j \\
vs & \\
& H_1 : \; at \; least \; one \; \tau_j \neq 0
\end{aligned}
\tag{2.15}
$$

An equivalent way to write the above hypothesis is in terms of the n time means μ_j $(j = 1, ..., n)$. We are interested in testing the equality of the n means, that is:

$$
\begin{aligned}
& H_0 : \; \mu_1 = \mu_2 = ... = \mu_n \; for \; all \; j \\
vs & \\
& H_1 : \; \mu_j \neq \mu_{j'} \; for \; at \; least \; one \; pair \; (j, j')
\end{aligned}
\tag{2.16}
$$

In order to test the null hypothesis H_0, the ratio (test statistic) $F_T = \frac{MS_T}{MS_E}$ is calculated. By Cochran's theorem [see, e.g., *Montgomery (1997)*], the sums of squares SS_T, SS_B, SS_E, and SS_{Total} are independently distributed chi-square random variables with $n-1$, $m-1$, $(n-1)(m-1)$ and $nm-1$ degrees of freedom respectively. Thus the ratio F_T,under the null hypothesis follows an F (*snedecor*) distribution with $n-1$ and $(n-1)(m-1)$ degrees of freedom. Observing the expected mean squares of Table 2.1, one can see that MS_E is an unbiased estimator of σ_e^2, since $E(MS_E) = \sigma_e^2$. Also, under the null hypothesis, MS_T is an unbiased estimator of σ_e^2, too. However, if the null hypothesis is false, the expected value of MS_T, $E\left(MS_T\right)$ is greater than σ_e^2. So, under the alternative hypothesis, the expected value of the numerator of the test statistic is greater than the expected value of the denominator and this suggests the rejection of H_0 on values of the test statistic that are too large.

The critical region K, that leads to the rejection of the null hypothesis at a significance level α is given by:

$$K: \; F_T > F_{n-1,(n-1)(m-1)}\left(\alpha\right),$$

or equivalently using the p-value approach, if the probability of observing a value of the test statistic large or larger than the estimated F_T is less than α, we are led to the rejection of H_0.

Similarly working, we are in position to conduct another useful test, the one that checks the importance of the main effects of subjects. As in the case of every other random effect, testing hypotheses concern the effects of the specific subjects participating in the study, is meaningless, so instead we test the following hypothesis:

$$\begin{array}{c} H_0 : \sigma_b^2 = 0 \\ vs \\ H_1 : \sigma_b^2 > 0 \end{array} \tag{2.17}$$

The appropriate test statistic now is the ratio $F_B = \frac{MS_B}{MS_E}$, distributed as F with $m-1$ and $\left(m-1\right)\left(n-1\right)$ degrees of freedom, under the null hypothesis. Thus, the null hypothesis H_0 is rejected, at a (predetermined) significance level α, if:

$$F_B > F_{m-1,(n-1)(m-1)}\left(\alpha\right),$$

or equivalently, if the probability that one would observe a value of the test statistic as large or larger than the estimated F_B if H_0 were true, is less than α, H_0 is rejected.

2.5.1.3 Estimation of Fixed Effect Parameters τ_j

Thus far, we have being mostly occupied with the statistical analysis of variance model (2.4) and the associated hypotheses, concerning the significances of fixed effects of time (τ_j) and individual random effects (b_i). The specific hypotheses were made possible to check by forming each time appropriate F-tests. Later on, we are going to see that as long as the assumptions about the distributional form and variance-covariance structure of each subject's i data vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in})^t$ [that $\mathbf{y}_i \sim N_n(\boldsymbol{\mu}_i, \mathbf{V}_i)$ and $\mathbf{V}_i = \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n$] hold, then the F-ratios constructed to perform the tests will indeed have sampling distributions that are F-distributions under the null hypotheses of concern, and consequently the F-tests will provide valid inferences. Another important issue, of analogous interest as with the fixed and random effects F-tests, is that of estimating the fixed effects parameters, τ_j, of the univariate analysis of variance model (2.4).

The estimation of τ_j's may be performed using standard estimating procedures of the general linear model theory (see Section 3.2), and one of these standard approaches is illustrated in the following lines. For this, we consider again the ANOVA model:

$$y_{ij} = \mu + \tau_j + b_i + \varepsilon_{ij} \qquad (i = 1, \ldots.., m), (j = 1, \ldots.., n).$$

The above model can be written in the alternative form:

$$y_{ij} = \mu + \tau_j + u_{ij}, \tag{2.18}$$

where the term u_{ij} comprises both random effects b_i and random error ε_{ij}, 'covering' in this way all sources of random variation. Now, let $\mathbf{u}_i$ denoting the vector $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{in})^t$, and as already given, let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in})^t$ be the vector of repeated measurements on subject i. One can easily prove (working similarly to 2.8 and 2.9 equations) that the variance-covariance matrix of $\mathbf{y}_i$ has again the same compound symmetric structure of (2.12) and the vector of all measurements $\mathbf{y}$ has variance-covariance matrix given by (2.11). Following *Hand* and *Crowder (1996)*, observe that model (2.18) can be written in the matrix form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{u}, \tag{2.19}$$

where $\mathbf{y}$ is, as before, denotes the $(nm \times 1)$ vector of all measurements, $\mathbf{X}$ is the $(nm \times n)$ design matrix, $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_n)^t$ is the vector of fixed parameters that must be estimated and finally $\mathbf{u} = (\mathbf{u}_1^t, \mathbf{u}_2^t, ..., \mathbf{u}_m^t)$ is the $(nm \times 1)$ vector of random terms. Since, the variance-covariance matrix of vector $\mathbf{y}$ is the unknown block diagonal matrix $\mathbf{V}$, of (2.11) (the elements of $\mathbf{V}$ consist of the unknown variances σ_b^2, σ_e^2), if we denote by $\hat{\mathbf{V}}$ the estimated matrix $\mathbf{V}$, then by using one of the well-known estimation methods of generalized linear models, (e.g. generalized least squares, maximum likelihood) we obtain the parameter estimates

$$\tilde{\boldsymbol{\tau}} = \left(\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{y}. \tag{2.20}$$

2.5.1.4 Estimation of the Variance Components of V

Along with the estimation of the parameters in univariate analysis of variance models, i.e. the estimation of the (fixed) parameters τ_j of model (2.4), we are usually interested in estimating the variance components of the model (being σ_b^2 and σ_e^2 for the particular model). Further, in deriving the estimators of time effects τ_j, given in a vector form $\tilde{\boldsymbol{\tau}} = \left(\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{y}$, we have already emphasized that due to the fact that the variance-covariance matrix $\mathbf{V}$ is unknown ($\mathbf{V}$ is formed from zeros and the unknown variances σ_b^2, σ_e^2), we are urged to replace it by an estimate of $\mathbf{V}$, say $\hat{\mathbf{V}}$. Henceforth, what remains in suspense is the estimation of $\mathbf{V}$ and in particular the estimation of variances σ_b^2 and σ_e^2, usually referred to as the **variance components**.

Generally, the feature that differentiates the estimation of variance components is the form of the longitudinal design of interest. More specific, the chosen method of variance component estimation is depending upon whether the design is balanced or unbalanced. For balanced designs, like the ANOVA model of (2.4), the estimation relies almost exclusively on one method (contrary to unbalanced data models, such as the linear mixed effects model discussed in Chapter 3, where there have been developed several methods for variance component estimation). The general outline of the method is, after calculating the mean squares (MS) of the model and deriving the expected values of the mean squares $[E(MS)]$, to equate these expected mean squares to their calculated (observed) values. Since the obtained equations will always be linear functions of the variance components, then solving for the latter we are in position to derive the corresponding estimators.

This method of estimating variance components (for balanced data) is known as the 'analysis of variance method', because it makes use of the lines in the analysis of variance Table. Concentrating on the specific ANOVA model $y_{ij} = \mu + \tau_j + b_i + \varepsilon_{ij}$, and its corresponding analysis of variance Table (Table 2.1), the estimated variances $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ are calculated by equating the (observed) mean squares to their expected values, obtaining the two equations

$$\left\{ \begin{array}{c} MS_E = E(MS_E) \\ MS_B = E(MS_B) \end{array} \right\}. \tag{2.21}$$

Now, $E(MS_E) = \sigma_e^2$ simply tells us that MS_E is an (unbiased) estimator of σ_e^2, hence we can write $\hat{\sigma}_e^2 = MS_E$. In a similar way, it is $\hat{\sigma}_e^2 + n\hat{\sigma}_b^2 = MS_B$, and system (2.21) becomes

$$\left\{ \begin{array}{c} MS_E = \hat{\sigma}_e^2 \\ MS_B = \hat{\sigma}_e^2 + n\hat{\sigma}_b^2 \end{array} \right\}.$$

The solutions to the above system are simply the estimators $\hat{\sigma}_e^2$, $\hat{\sigma}_b^2$ of the variance components σ_e^2 , σ_b^2, given by

$$\left\{ \begin{array}{c} \hat{\sigma}_e^2 = MS_E \\ and \\ \hat{\sigma}_b^2 = (MS_B - MS_E)/n \end{array} \right\}.$$

Remark 2.1: *Although in the analysis of variance procedures involving the calculations of F-tests, the use of those tests was founded upon normality assumptions, this is not the case with the analysis of variance method for estimating variance components, since the expected values of mean squares do not use these normality assumptions (this is because the expected values apply to any distributions that have zero means and finite variances, hence the analysis of variance method of estimation can therefore be used regardless of distributional properties). Furthermore, estimators of variance components derived by the analysis of variance method for balanced data (as the above $\hat{\sigma}_e^2$ and $\hat{\sigma}_b^2$), are always unbiased (for proof see Searle, 1971). In addition, as it has been showed by Graybill and Hultquist (1961), these estimators are also minimum variance quadratic unbiased. This means that among all estimators of a variance component, say σ^2, which are both quadratic functions of the observations and unbiased, those derived by the analysis of variance method have the smallest variance. In spite of those 'good' properties, ANOVA method for variance components estimation suffers from an important disadvantage; estimates obtained by the analysis of variance method can unfortunately take negative values. When this happens, several possibilities for dealing with this issue have been proposed, no one of them being quite satisfactory though. There is the option to accept the obtained negative estimate as an evidence that the true value of the corresponding component is zero, hence one can just replace the negative estimate with zero. Interpreting a negative estimate as indication of a wrong model is another possible course of action. Finally, there is the option of using, instead of ANOVA method for balanced data, other estimation procedures, e.g. Bayes estimators [for more on this subject see Tiao and Box (1967), Federer (1968), Hill (1965,1967)].*

2.5.1.5 Compound Symmetry and Sphericity

In previous sections, we saw how the univariate analysis of variance procedure is applied to longitudinal studies. Also, by using the familiar variance-covariance matrix notation, we defined the covariance structure of both within and between subjects' repeated measurements in (2.11), making a special reference to the specific structure that each submatrix $\mathbf{V}_i$ of $\mathbf{V}$ shares, known as compound symmetric structure.

Each (sub)matrix $\mathbf{V}_i$ stands for the variance-covariance matrix of the within-subjects measurements and as already demonstrated is of the form:

$$\begin{pmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2+\sigma_e^2 \end{pmatrix}$$

This compound symmetry assumption is a straightforward consequence of the fact that for two measurements $y_{ij}, y_{ij'}$ on the same subject i (within unit measurements), it is easily proved that:

$$\begin{aligned} Cov\left(y_{ij}, y_{ij'}\right) &= \sigma_b^2 \\ & \qquad\qquad \forall\ j, j' \ \ \left(j \neq j'\right). \\ Var\left(y_{ij}\right) &= \sigma_b^2+\sigma_e^2 \end{aligned} \tag{2.22}$$

The significance of the compound symmetry structure is implied by the fact that under this (rather restrictive) assumption, the F-tests (for the time-related terms) in the univariate analysis of variance procedure are valid. In fact, compound symmetry can be considered as a special case of a more general condition, called '*sphericity*' or '*circularity*', which still has the advantage of providing valid F-tests if we assume that the within-subjects covariance matrices share the sphericity property. Sphericity (denoted by ε and sometimes referred to as circularity), is one of the few distinguishable assumptions of repeated measures ANOVA in comparison to the classical analysis of variance, and the need for its introduction was caused by the special feature of longitudinal data, where the dependence of the within-subjects observations must be taken in account.

It is well known that in classical analysis of variance with fixed effects, two are the basic assumptions that must be satisfied in order to obtain valid inferences concerning model parameters via the F-tests. Specifically, the assumptions are that the errors are normally and independently distributed, with zero mean and **constant** variance σ_e^2. As a consequence, the variance-covariance matrix of all observations is a diagonal matrix, with its main diagonal consisting of constant variances σ_e^2. Similarly, in experiments where additionally to the random error term ε_{ij}, another source of variation (random factor) is considered, the assumption of constant variance is modified and hence rather than assuming constant variances σ_e^2 for all measurements y_{ij} of the experiment, we assume equality of variances between all levels of the random factor. It is common to refer to this constant variances assumption as the **homogeneity of variance** assumption. In analysis of variance applied to repeated measures data (and consequently to longitudinal data), although variance-covariance matrix $\mathbf{V}$ is no longer diagonal, the usual F-tests are still valid under the compound symmetry condition assumed for the within-subject variance-covariance matrix $\mathbf{V}_i$.

In what follows we illustrate in detail exactly why the compound symmetric structure of $\mathbf{V}_i$ is a sufficient (but not necessary) condition for the validity of the F-tests for mean comparisons (associated with the levels of the time-related factors) in analysis of variance for longitudinal data. In order to demonstrate the previous claim, the necessity of reminding some basic ideas of the classical experimental design analysis of variance arises.

Many important comparisons in analysis of variance, between level means of a factor may be made using **contrasts** (see, e.g. *Montgomery, 1997*). Consider for example model (2.4), and assume that, the relative to this model, null hypothesis (2.15), of equal time-factor means is rejected. Rejection of this hypothesis indicates that some (or even all) of the means μ_j are different from the others, but exactly which pairs of means are different (or equal) is not a provided information from the specific hypothesis testing. For circumstances like this, where we seek to make comparisons of pairs of means, or other partial comparisons of interest to the researcher, implementation of contrasts comes to our rescue.

In general, contrasts are linear combinations formed by the multiplication of suitable chosen vectors with vectors consisting of the level totals of the specific factor of interest. More specific, consider a factor with α different levels, each level consisting of n measurements. A contrast, concerning level means of the factor, is the product of the transpose of a suitably chosen vector $\mathbf{c} = (c_1, c_2, ..., c_\alpha)^t$ and the vector $\mathbf{y} = (y_{1\cdot}, y_{2\cdot}, ..., y_{\alpha\cdot})^t$, where the elements $y_{i\cdot}$ of $\mathbf{y}$ denote the summation of all measurements within the ith level of the factor. The contrast is thus:

$$\mathbf{c}^t\mathbf{y} = (c_1, c_2, ..., c_\alpha)\begin{pmatrix} y_{1\cdot} \\ y_{2\cdot} \\ \vdots \\ y_{\alpha\cdot} \end{pmatrix} = \sum_{i=1}^{\alpha} c_i y_{i\cdot} \tag{2.23}$$

with the restriction that $\sum_{i=1}^{\alpha} c_i = 0$. The elements c_i of vector $\mathbf{c}$ are called coefficients. The sum of squares for any contrast is:

$$SS_{contrast} = \frac{\left(\sum_{i=1}^{\alpha} c_i y_{i\cdot}\right)^2}{n \sum_{i=1}^{\alpha} c_i^2}, \tag{2.24}$$

and has a single degree of freedom. In the case of an unbalanced design, that is a design where the number of measurements is not the same for each level of the factor, the contrast sum of squares of (2.24) becomes:

$$SS_{contrast} = \frac{\left(\sum_{i=1}^{\alpha} c_i y_{i\cdot}\right)^2}{\sum_{i=1}^{\alpha} n_i c_i^2}, \tag{2.25}$$

where n_i denotes the number of measurements within the ith level $(i = 1, 2, ..., \alpha)$. Generally, a contrast is tested by comparing its sum of squares to the error mean square.

The statistic (essentially the ratio of contrast sum of squares and error mean square) would be distributed as F with 1 and $\alpha\,(n-1)$ degrees of freedom.

A very important special case of the above defined contrast, is that of the *orthogonal contrasts.* Two contrasts, $\mathbf{c}^t\mathbf{y}$, $\mathbf{d}^t\mathbf{y}$, with $\mathbf{c} = (c_1, c_2, \ldots.., c_\alpha)^t$ and $\mathbf{d} = (d_1, d_2, \ldots.., d_\alpha)^t$ respectively, are called orthogonal if:

$$\mathbf{c}^t\mathbf{d} = 0 \Leftrightarrow \sum_{i=1}^{\alpha} c_i d_i = 0. \qquad (2.26)$$

The importance of orthogonal contrasts lies on the fact that tests performed on orthogonal contrasts are independent.

In order to outline the construction of a contrast, the well known test for the equality of means between two independent populations, by its familiarity and simplicity, proves to be a suitable example. Suppose we have two independent samples: $y_1, y_2, \ldots.., y_r$ and $y_{r+1}, y_{r+2}, \ldots.., y_s$, coming from the two corresponding populations, and let $\overline{y}_1$, $\overline{y}_2$ be the two sample means. For convenience in the calculations we consider the simplest case, where the two samples come from a $N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$ distribution, respectively. Moreover a common known variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ is assumed for the two samples. Applying the test (usually called z-test), we are able to test the null hypothesis:

$$H_0 : \mu_1 = \mu_2 \qquad (2.27)$$

of the equality of the two population means, μ_1 and μ_2, against the (two-sided) alternative hypothesis:

$$H_1 : \mu_1 \neq \mu_2. \qquad (2.28)$$

For this purpose the z_0-statistic:

$$z_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sigma\sqrt{\frac{1}{r} + \frac{1}{s}}}$$

is computed, where r, s denote the sample sizes. The null hypothesis (2.27) is rejected, in a (predetermined) confidence level α, if:

$$z_0 < -z_{\alpha/2} \quad or \quad z_0 > z_{\alpha/2},$$

where, in general, z_α is a value defined such that: $\Pr(Z \geq z_\alpha) = \alpha$. Z denotes the standard normal distribution [The motivation for constructing z_0 is directly related to the following: since $\overline{y}_1 \sim N\left(\mu_1, \frac{\sigma^2}{r}\right)$, $\overline{y}_2 \sim N\left(\mu_2, \frac{\sigma^2}{s}\right)$, then due to the independence of the two populations, $\overline{y}_1 - \overline{y}_2 \sim N\left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{r} + \frac{1}{s}\right)\right]$ and consequently if $H_0 : \mu_1 = \mu_2$ were true, $\overline{y}_1 - \overline{y}_2 \sim N\left(0, \sigma^2\left(\frac{1}{r} + \frac{1}{s}\right)\right)$ or $\frac{\overline{y}_1 - \overline{y}_2}{\sigma\sqrt{\frac{1}{r} + \frac{1}{s}}} \sim N(0,1)$].

Let us see now, how the above hypothesis (2.27) can be recomposed, in an alternative form. First, observe that the null hypothesis (2.27) is equivalent to:

$$H_0^{'} : \ \mu_1 - \mu_2 = 0. \tag{2.29}$$

Now, let us consider the vectors $\boldsymbol{\mu} = \left(\mu_1 \mathbf{1}_r^t, \mu_2 \mathbf{1}_s^t\right)^t$ and $\mathbf{c}^t = \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right)$, where $\mathbf{1}_r = (1, 1, ..., 1)^t$ is a vector consisting of r 1's, and $\mathbf{1}_s = (1, 1, ..., 1)^t$ is a vector consisting of s 1's. The product $\mathbf{c}^t\boldsymbol{\mu}$ produces a scalar, since $\mathbf{c}^t$ and $\boldsymbol{\mu}$ are of $1 \times (r+s)$ and $(r+s) \times 1$ dimensions, respectively. More specifically:

$$\begin{aligned} \mathbf{c}^t\boldsymbol{\mu} &= \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right)' \left(\mu_1 \mathbf{1}_r^t, \mu_2 \mathbf{1}_s^t\right)^t = \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right) \begin{pmatrix} \mu_1 \mathbf{1}_r \\ \mu_2 \mathbf{1}_s \end{pmatrix} \\ &= \mu_1 \frac{\mathbf{1}_r^t \mathbf{1}_r}{r} - \mu_2 \frac{\mathbf{1}_s^t \mathbf{1}_s}{s} = \mu_1 \frac{r}{r} - \mu_2 \frac{s}{s} = \mu_1 - \mu_2. \end{aligned} \tag{2.30}$$

As an obvious result, the hypothesis $H_0^{''} : \mathbf{c}^t\boldsymbol{\mu} = 0$ is an equivalent representation of hypothesis (2.27). In order to construct a statistic for this hypothesis equivalent to the statistic z_0, this time by forming a suitable contrast, we proceed as follows. Initially, consider the (scalar) product $\mathbf{c}^t\mathbf{y}$, where $\mathbf{y} = (y_1, y_2, ..., y_r, y_{r+1}, ..., y_s)^t$ denotes the vector that comprises all measurements from both samples. This product gives:

$$\begin{aligned}
\mathbf{c}^t\mathbf{y} &= \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right)\begin{pmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_s \end{pmatrix} = \left(\frac{1}{r}, \ldots, \frac{1}{r}, -\frac{1}{s}, \ldots, -\frac{1}{s}\right)\begin{pmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_s \end{pmatrix} \\
&= \frac{y_1 + y_2 + \ldots + y_r}{r} - \frac{y_{r+1} + y_{r+2} + \ldots + y_s}{s} = \overline{y}_1 - \overline{y}_2. \qquad (2.31)
\end{aligned}$$

Thus $\mathbf{c}^t\mathbf{y}$ is an alternative way of writing the difference of sample means, $\overline{y}_1 - \overline{y}_2$. As a next step, we have to find an alternative presentation of the variance of the sample mean difference $\mu_1 - \mu_2$, $\sigma^2\left(\frac{1}{r} + \frac{1}{s}\right)$. Working similarly to the above calculations, we can easily verify that:

$$\begin{aligned}
\mathbf{c}^t\mathbf{c} &= \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right)\begin{pmatrix} \frac{1}{r}\mathbf{1}_r^t \\ -\frac{1}{s}\mathbf{1}_s^t \end{pmatrix} = \frac{1}{r^2}\mathbf{1}_r^t\mathbf{1}_r + \frac{1}{s^2}\mathbf{1}_s^t\mathbf{1}_s \\
&= \frac{1}{r^2}(1, 1, \ldots.., 1)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \frac{1}{s^2}(1, 1, \ldots.., 1)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{r^2}r + \frac{1}{s^2}s \\
&= \frac{1}{r} + \frac{1}{s} \Rightarrow \sigma^2\mathbf{c}^t\mathbf{c} = \sigma^2\left(\frac{1}{r} + \frac{1}{s}\right). \qquad (2.32)
\end{aligned}$$

Combining (2.30), (2.31) and (2.32) we finally arrive at the position to reformulate the previous stated assumption $\overline{y}_1 - \overline{y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{r} + \frac{1}{s}\right)\right)$ as $\mathbf{c}^t\mathbf{y} \sim N\left(\mathbf{c}^t\boldsymbol{\mu}, \sigma^2\mathbf{c}^t\mathbf{c}\right)$. Hence, under H_0'' $(\mathbf{c}^t\boldsymbol{\mu} = \mathbf{0})$, it is:

$$\mathbf{c}^t\mathbf{y} \sim N\left(\mathbf{0}, \sigma^2\mathbf{c}^t\mathbf{c}\right) \Rightarrow \frac{\mathbf{c}^t\mathbf{y}}{\left(\sigma^2\mathbf{c}^t\mathbf{c}\right)^{1/2}} \sim N(\mathbf{0}, 1) \Rightarrow$$

$$\Rightarrow \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\sigma^2\mathbf{c}^t\mathbf{c}} \sim X_1^2. \tag{2.33}$$

[for the above, we have used that if $X \sim N\left(\mu, \sigma^2\right)$ a normal random variable, then $Z = \frac{X-\mu}{\sigma} \sim N\left(0,1\right)$ and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2 \sim x_1^2$, where x_1^2 denotes the chi-square distribution with 1 degree of freedom].

Summarizing, so far we have shown that an alternative way of expressing the null hypothesis $H_0 : \mu_1 = \mu_2$ is via the hypothesis $H_0^{''} : \mathbf{c}^t\boldsymbol{\mu} = \mathbf{0}$, and also that an alternative to $\overline{y}_1 - \overline{y}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{r} + \frac{1}{s}\right)\right)$ which has been used for the construction of z_0-statistic is given by $\mathbf{c}^t\mathbf{y} \sim N\left(\mathbf{c}^t\boldsymbol{\mu}, \sigma^2\mathbf{c}^t\mathbf{c}\right)$. The issue that remains in suspense is the one of showing that this alternative representation discussed so far is in the form of a contrast. Indeed, since $\mathbf{c}^t\mathbf{y}$ can be rewritten as:

$$\begin{aligned}
\mathbf{c}^t\mathbf{y} &= \left(\frac{1}{r}\mathbf{1}_r^t, -\frac{1}{s}\mathbf{1}_s^t\right)\begin{pmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_s \end{pmatrix} = \left(\frac{1}{r}, \ldots\ldots, \frac{1}{r}, -\frac{1}{s}, \ldots\ldots, -\frac{1}{s}\right)\begin{pmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_s \end{pmatrix} \\
&= \frac{y_1 + \ldots + y_r}{r} - \frac{y_{r+1} + \ldots + y_s}{s} = \frac{1}{r}y_{1.} - \frac{1}{s}y_{2.} = c_1y_{1.} + c_2y_{2.} = \\
&= \sum_{i=1}^{2} c_iy_{i.} \qquad (2.34)
\end{aligned}$$

where $c_1 = \frac{1}{r}$, $c_2 = -\frac{1}{s}$ are the contrast coefficients, then in accordance to (2.23), $\mathbf{c}^t\mathbf{y}$ constitutes a contrast. If we consider the two-sample comparison as a special case of an unbalanced design, the corresponding sum of squares of contrast $\mathbf{c}^t\mathbf{y}$ is, according to

(2.25):

$$SS_{contrast} = \frac{\left(\sum\limits_{i=1}^{2} c_i y_{i\cdot}\right)^2}{\sum\limits_{i=1}^{2} n_i c_i^2} = \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{rc_1^2 + sc_2^2} = \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{r\frac{1}{r^2} + s\frac{1}{s^2}}$$

$$= \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\frac{1}{r} + \frac{1}{s}} \underset{(2.27)}{=} \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\mathbf{c}^t\mathbf{c}}. \tag{2.35}$$

As mentioned previously, a contrast is tested by forming the ratio $\frac{SS_{contrast}}{MS_{Error}}$, which under the null hypothesis is distributed as an F with 1 and $\alpha(n-1)$ degrees of freedom. The calculations given above, showed that $\frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\sigma^2\mathbf{c}^t\mathbf{c}} \sim x_1^2$ and $SS_{contrast} = \frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\mathbf{c}^t\mathbf{c}}$. Combining these results it seems rather reasonable to use as the numerator of $\frac{SS_{contrast}}{MS_{Error}}$ the ratio $\frac{\left(\mathbf{c}^t\mathbf{y}\right)^2}{\sigma^2\mathbf{c}^t\mathbf{c}}$ since not only includes the $SS_{contrast}$ but additionally follows a chi-square distribution with 1 degree of freedom.

A natural extension of the hypothesis (2.27) that compares the two population means μ_1 and μ_2 could be a hypothesis that involves more than one comparison, such as:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \tag{2.36}$$

that involves three samples of r, s and l sizes corresponding to the three populations. Testing the above hypothesis by the formulation of a contrast proves to be a little more complex task compared to the previous example. Working similarly to the pairwise comparison $H_0 : \mu_1 = \mu_2$ it can be shown that (2.36) is equivalent to

$$H_0^{'} : \mathbf{c}_1^t\boldsymbol{\mu} = \mathbf{c}_2^t\boldsymbol{\mu} = 0, \tag{2.37}$$

where $\mathbf{c}_1$, $\mathbf{c}_2$ are suitable chosen vectors and $\boldsymbol{\mu} = \left(\mu_1\mathbf{1}_r^t, \mu_2\mathbf{1}_s^t, \mu_3\mathbf{1}_l^t\right)^t$. In order to construct a combined F-test, that will test both of the partial hypotheses $\mathbf{c}_1^t\boldsymbol{\mu} = 0$, $\mathbf{c}_2^t\boldsymbol{\mu} = 0$ simultaneously, we have to form a ratio with the numerator being a sum (with suitable weights) of the contrasts squares $\left(\mathbf{c}_1^t\mathbf{y}\right)^2$ and $\left(\mathbf{c}_2^t\mathbf{y}\right)^2$. One basic restriction that arises now is that the contrasts $\mathbf{c}_1^t\mathbf{y}$, $\mathbf{c}_2^t\mathbf{y}$ must satisfy the orthogonality condition (2.26) since by this way we achieve the required independence between the partial comparisons. Hence, the contrast vectors $\mathbf{c}_1$, $\mathbf{c}_2$ must be chosen such that $\mathbf{c}_1^t\mathbf{c}_2 = 0$. The above considerations can be extended out in the general case. Suppose that the hypothesis to be tested is expressed as

$$H_0:\ \mathbf{c}_j^t\boldsymbol{\mu} = 0\ for\ j = 1, ..., \nu.$$

Then, assuming that the $\mathbf{c}_j$'s have been *orthonormalized* (by this is meant that $\mathbf{c}_j$'s have been modified, if necessary, to be mutually orthogonal, that is $\mathbf{c}_k^t\mathbf{c}_j = 0$ for $k \neq j$, and moreover are normalized satisfying $\mathbf{c}_j^t\mathbf{c}_j = 1\ \forall j$).

At last now, having illustrated the concept of a contrast through the relatively simple examples such the ones of comparing two and three population means respectively, we are ready to proceed (taking advantage of the contrast notion discussed above) with showing that assuming a compound symmetric form for the variance-covariance matrix $\mathbf{V}_i$ of the within-subjects measurements is a sufficient condition for the validity of the analysis of variance F-tests. For this reason, let $\mathbf{y} = (y_1, y_2, ..., y_n)^t$ be the vector of n measurements on a subject and furthermore assume that the (within-subjects) variance-covariance matrix $\mathbf{V}_i$ of $\mathbf{y}$ can be written as:

$$Var\left(\mathbf{y}\right) = \mathbf{V}_i = \sigma_b^2 \mathbf{1}_n \mathbf{1}_n^t + \sigma_e^2 \mathbf{I}_n = \sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n, \tag{2.38}$$

where $\mathbf{J}_n$ is a $n \times n$ matrix of 1's and $\mathbf{I}_n$ is a $(n \times n)$ identity matrix. Thus, according to (2.14), $\mathbf{V}_i$ is a compound symmetric matrix. Consider now a hypothesis about $\boldsymbol{\mu} = \mathbf{E}\left(\mathbf{y}\right) = \left(E\left(y_1\right), E\left(y_2\right), ..., E\left(y_n\right)\right) = \left(\mu_1, \mu_2, ..., \mu_n\right)$. As already described, by using a standard procedure, it is possible to express the specific hypothesis as $H:\ \mathbf{c}_j^t\boldsymbol{\mu}$ =0 for $j = 1, ..., v$. As previously, we think of forming a ratio of which both numerator (the combined contrast's sum of squares) and denominator will follow a chi-square distribution (F-ratio). More specific, for the F-numerator, the (combined) sum of squares is constructed by adding the (weighted) $\left(\mathbf{c}_j^t\mathbf{y}\right)^2$, and as in the example of the comparison of three means, we need independent chi-squares. Independence of contrasts $\mathbf{c}_l^t\mathbf{y}$ and $\mathbf{c}_k^t\mathbf{y}$ $\forall$ $l, k \in \{1, \ldots.., v\}$, $(l \neq k)$ is obtained if[2]:

$$\begin{aligned}
& Cov\left(\mathbf{c}_l^t\mathbf{y}, \mathbf{c}_k^t\mathbf{y}\right) = 0 \Rightarrow \mathbf{c}_l^t Cov\left(\mathbf{y}, \mathbf{y}\right) \mathbf{c}_k = 0 \Rightarrow \\
\Rightarrow \;& \mathbf{c}_l^t Var\left(\mathbf{y}\right) \mathbf{c}_k = 0 \Rightarrow \mathbf{c}_l^t \Sigma_i \mathbf{c}_k = 0 \underset{from\ (2.34)}{\Rightarrow} \\
\Rightarrow \;& \mathbf{c}_l^t\left(\sigma_b^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n\right) \mathbf{c}_k = 0 \Rightarrow \mathbf{c}_l^t\left(\sigma_b^2 \mathbf{1}_n \mathbf{1}_n^t + \sigma_e^2 \mathbf{I}_n\right) \mathbf{c}_k = 0 \Rightarrow \\
\Rightarrow \;& \sigma_b^2\left(\mathbf{c}_l^t \mathbf{1}_n\right)\left(\mathbf{1}_n^t \mathbf{c}_k\right) + \sigma_e^2\left(\mathbf{c}_l^t \mathbf{c}_k\right) = 0 \Rightarrow \\
\Rightarrow \;& \sigma_b^2\left(\mathbf{c}_l^t \mathbf{1}_n\right)\left(\mathbf{c}_k^t \mathbf{1}_n\right) + \sigma_e^2\left(\mathbf{c}_l^t \mathbf{c}_k\right) = 0 \underset{\text{since } \sigma_b^2, \sigma_e^2 > 0}{\Rightarrow} \\
\Rightarrow \;& \mathbf{c}_l^t \mathbf{c}_k = 0 \textit{ and at least one of } \mathbf{c}_l^t \mathbf{1}_n, \mathbf{c}_k^t \mathbf{1}_n = 0.
\end{aligned}$$

Hence, in words, if $\mathbf{V}_i$ is a compound symmetric matrix (i.e. the n repeated measurements exhibit equal variances and have equal covariances) and the $\mathbf{c}_j^t\mathbf{y}$'s are orthogonal contrasts ($\mathbf{c}_l^t\mathbf{c}_k = 0$), then the $\mathbf{c}_j^t\mathbf{y}$'s are independent $N\left(\mathbf{c}_j^t\boldsymbol{\mu}, \sigma^2\right)$ (since due to the normalization: $\mathbf{c}_j^t\mathbf{c}_j = 1$) and consequently the F-numerator for testing H can be validly constructed, producing a valid F-ratio as *Hand* and *Crowder (1990)* point out.

The following definitions will assist our efforts in explaining as well as possible the concept of sphericity and its association with compound symmetry.

Definition 2.3: *The variance-covariance matrix* $\mathbf{V}$ *of a (random) vector* $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^t$ *following a n-variate normal distribution, is said to satisfy the sphericity (or* ***circularity****) condition if-f is of the form:*

$$\mathbf{V} = \sigma^2 \mathbf{V}_0,$$

where $\mathbf{V}_0$ *is a fixed (known) positive definite matrix*[3] *and* σ^2 *unknown.*

Remark 2.2: *Assuming a variance-covariance matrix to be of the form* $\mathbf{V} = \sigma^2\mathbf{V}_0$ *is*

[2] We know that if X,Y independent random variables, then Cov(X,Y)=0 (due to Cov(X,Y)=E(XY)-E(X)E(Y)). In the special case of the normal distribution the converse is also true, that is if X,Y uncorrelated then X,Y are independent.

[3] Suppose $\mathbf{A}$ is a square symmetric matrix and $\underline{\mathbf{x}}$ is a non-zero column vector. Then $\underline{\mathbf{x}}^t\mathbf{A}\underline{\mathbf{x}}$ is called a quadratic form. The quadratic form and the matrix $\mathbf{A}$ are defined to be positive definite if $\underline{\mathbf{x}}^t\mathbf{A}\underline{\mathbf{x}} > 0$.

equivalent to assume that $\mathbf{V} = \sigma^2\mathbf{I}$ *(with* $\mathbf{I}$ *being an identity matrix) as we can transform* $y_1, y_2, \ldots.., y_n$ *to* $z_1, z_2, \ldots.., z_n$ *by* $z_j = \mathbf{G}x_j$ *where* $\mathbf{G}$ *is a matrix such that* $\mathbf{G}\mathbf{V}_0\mathbf{G}^t = \mathbf{I}$*. Hence,* $\mathbf{V} = \sigma^2\mathbf{V}_0$ *is equivalent to assume that we have a set of* n *independent random variables* $z_1, z_2, \ldots.., z_n$ *with a common variance* σ^2*. Thus:*

Definition 2.4: *The variance-covariance matrix* $\mathbf{V}$ *of a (random) vector* $\mathbf{Y}$*=(*y_1*,*y_2*,…* *following a n-variate normal distribution, is said to satisfy the* ***sphericity*** *condition if-f is of the form:*

$$\mathbf{V} = \sigma^2\mathbf{I},$$

where $\mathbf{I}$ *is a* $(n \times n)$ *identity matrix and* σ^2 *unknown.*

Definition 2.5: *A* $(n \times n)$ *matrix* $\mathbf{M}$ *is said to be of* Type H *if it may be written in the following form:*

$$\mathbf{M} = \begin{pmatrix} \lambda + 2\alpha_1 & \alpha_1 + \alpha_2 & \cdots & \alpha_1 + \alpha_n \\ \alpha_2 + \alpha_1 & \lambda + 2\alpha_2 & \cdots & \alpha_2 + \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_n + \alpha_1 & \alpha_n + \alpha_2 & \cdots & \lambda + 2\alpha_n \end{pmatrix}, \qquad (2.39)$$

where λ *and* α'*s are constant numbers.*

It may be shown (see *Huynh* and *Feldt, 1970*), that as long as (within-subject) variance-covariance matrix $\mathbf{V}_i$ of each ith subject's data vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^t$ shares the above so-called Type H form, the ANOVA F-tests already discussed are valid. The Type H condition required for the validity of the univariate ANOVA tests is also known as the **Huynh-Feldt (H-F) condition**, and is mathematically less stringent compared to the compound-symmetry condition (i.e. equal variances and covariances of within-subject variance-covariance matrix).

Observe that Type H structure implies equality of variances of differences for all pairs of responses assumed to be correlated (i.e. the within-subject responses). Indeed, let us assume a response vector $\mathbf{y}_i$ that summarizes the measurements for the ith subject, and exhibits a variance-covariance matrix of Type H such as the one given in (2.39). Then, all possible differences $y_{ij} - y_{ij'}$ $\left(j \neq j^{'}\right)$ are equally variable, since:

$$\begin{aligned} Var\left(y_{ij} - y_{ij'}\right) &= Var\left(y_{ij}\right) + Var\left(y_{ij'}\right) - 2Cov\left(y_{ij}, y_{ij'}\right) \underset{(2.39)}{=} \\ &= \lambda + 2\alpha_j + \lambda + 2\alpha_{j'} - 2\left(\alpha_j + \alpha_{j'}\right) \\ &= 2\lambda = cons\tan t. \end{aligned}$$

The importance of a variance-covariance matrix of a multivariate response vector $\mathbf{y}_i$ of Type H lies in the fact that, whenever the response vector follows a (multivariate) normal distribution, sphericity is equivalent to equality of variances of differences of the components of the vector, a property shared by matrices of Type H as already shown. Thus, as long as data vectors $\mathbf{y}_i$, $(i = 1, 2, ..., m)$ are multivariate normal with common variance-covariance matrix of the form (2.39), sphericity condition is satisfied and the usual analysis of variance F-tests are valid.

In the above, we emphasized the importance of the sphericity condition as a necessary and sufficient condition for providing valid F-tests in the univariate analysis of variance for balanced longitudinal studies. It is, thus, of great interest to be able to test whether or not variance-covariance matrices of response vectors in ANOVA analyses satisfy the sphericity condition (or alternatively, under the normality assumption, the Type H condition). A test of the hypothesis that a matrix satisfies the sphericity condition was derived by *Mauchly (1940)*, and is henceforth known as the Mauchly's test of sphericity. It is based on a statistic of the following form (see, e.g. *Hand* and *Crowder; 1996*):

$$W = \frac{\left|\mathbf{C}\mathbf{V}_n\mathbf{C}^t\right|}{\left|tr\left(\mathbf{C}\mathbf{V}_n\mathbf{C}^t\right) \big/ n - 1\right|^{n-1}},$$

where $\mathbf{V}_p$ denotes the pooled within-subject sample variance-covariance matrix, $\mathbf{C}$ denotes a matrix of $n - 1$ orthogonal contrasts, and n is the number of the repeated observations on each subject $(j = 1, 2, .., n)$. The statistic W (as originally shown by *Mauchly, 1940* for $n = 2$), has approximately a x^2 with $(n-2)(n+1)\big/2$ degrees of freedom.

Further contributions in connection to the W criterion are given by *Nagarsenker* and *Pillai (1972, 1973); Mathai* and *Rathie (1970)* and *Khatri* and *Srivastava (1971)* among others.

2.5.2 Multivariate Analysis of Variance

Multivariate analysis of variance [see, e.g. *Johnson* and *Wichern (1992)*], in simple words, is just an analysis of variance (ANOVA) with several dependent variables. Equivalently, we can say that it is a multivariate method that extends univariate methods in cases where the response is not just a scalar, but is considered to be a vector. In such circumstances, univariate hypotheses testings for comparisons, as the familiar t-test for a single comparison, or the one-way ANOVA for multiple comparisons, have no practical usage, and generalizations of those methods known as multivariate analysis of variance (MANOVA) methods have developed. The key concept of those methods is that they consider an observation to be an entire vector, instead of a single scalar response.

At this point, we have to note that in general, multivariate analysis of variance methods were created with the scope of analyzing a broader category of data than longitudinal data, namely multivariate data, where the responses are not necessarily come from the same variable. However, since longitudinal data can be viewed as a special case of multivariate data , where the measurements y_{ij} of each subject i refer to the same unique variable, the general theory of multivariate analysis of variance is easily conveyed to longitudinal studies.

Before turning our attention on how MANOVA is adjusted and fitted in the longitudinal studies analyses, it would be useful to present the basic features of the method in the situation of implementation to a general multivariate problem. In multivariate analysis of variance (MANOVA), unlike univariate analysis of variance (ANOVA), the responses are not measurements on the same single variable, but instead could come from multiple dependent interval variables. Essentially, it is a 'generalized' ANOVA where now the place of the single dependent variable is occupied by two or more dependent variables.

Hence, while ANOVA is testing for differences in means of the (single) interval dependent variable for various categories of the categorical independent variables, MANOVA tests for differences in the vector of means of the multiple interval dependents for various categories of the independents.

Multiple dependent variables make MANOVA much more complicated to perform compared to ANOVA, due to the fact that multiple dependent variables are usually not independent of each other. As a consequence, the multivariate analysis of variance tests can no longer based only on the sum of squares between and within groups, as in ANOVA. In fact, the sums of squares for between and within for the one dependent variable in the ANOVA case must be replaced now with a matrix containing the sums of squares for each one of the dependent variables as well as their cross-products. Determinants of those matrices are then used as measures of the overall variance in each matrix.

To illustrate implementation of MANOVA in the longitudinal data setting let us now consider the following situation; suppose we have a collection of repeated measurements on i subjects $(i = 1, 2, ..., m)$, collected at times j $(j = 1, 2, ..., n)$. We should remind here that this setting corresponds to a collection of balanced longitudinal data. Further suppose that the m subjects are randomized into k different groups $(k = 1, 2, ..., q)$. By denoting $\mathbf{y}_{ik} = (y_{i1k}, y_{i2k}, ..., y_{ink})^t$ the data vector that contains all observations on the ith subject that belongs in the kth group, we can construct the following model:

$$\mathbf{y}_{ik} = \boldsymbol{\tau} + \gamma_k + \varepsilon_{ik}, \ \ (i = 1, 2, ..., m), \ \ (k = 1, 2, ..., q), \tag{2.40}$$

where

- $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_n)^t$ is the $(n \times 1)$ vector associated with time,
- γ_k is the $(n \times 1)$ vector effect for the population from which the kth group of subjects was drawn, and
- ε_{ik} is a $(n \times 1)$ vector of errors.

Under normality assumptions, we can assume the response $\mathbf{y}_{ik}$ (which is now a vector, rather than a scalar) to satisfy:

$$\mathbf{y}_{ik} \sim N_n\left(\boldsymbol{\mu}_k, \mathbf{V}_i\right), \tag{2.41}$$

where $\boldsymbol{\mu}_k$ is the mean response vector for group k and $\mathbf{V}_i$ is the variance-covariance matrix of vector $\mathbf{y}_{ik}$, assumed to be the same for each group. The essential difference now, in comparison to a ANOVA model, is that the variance-covariance matrix $\mathbf{V}_i$ is not of a particular structure (e.g. the compound symmetric structure assumed by ANOVA methods), but instead is assumed to be completely unstructured which means that the only knowledge we have about $\mathbf{V}_i$ is that it is a $(n \times n)$ symmetric matrix of the form:

$$\mathbf{V}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}. \tag{2.42}$$

Thus, while ANOVA assumes the very restrictive compound symmetric structure, multivariate analysis of variance takes the entirely opposite direction assuming very little about the nature of the covariance structure of the data vector $\mathbf{y}_{ik}$.

In the sequel we will concentrate on how we can utilize multivariate analysis of variance theory to test hypotheses of interest, associated with the above model. For instance, the interest could be focused on comparing the different groups. Let us suppose that we have a null hypothesis stating that there is no difference in the means of the dependent variables (this reduces to only one variable in the longitudinal data situation) for the different groups. Formally, this is equivalent to testing the following (null) hypothesis:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \dots = \boldsymbol{\mu}_q, \ \textit{for every } k = 1, 2, \dots, q \tag{2.43}$$

against the alternative hypothesis H_1 which states that H_0 is not true. To test the above hypothesis, what is required is the construction of a table (usually called the MANOVA table), essentially being a straightforward generalization of the univariate analysis of variance table. The only complication is that the entries of the MANOVA table, due to that there are more than one variables now present, can no longer be single sums of squares as was the case with the entries of a univariate analysis of variance table. In fact, we also have to take into account the sums of products between the different variables. To comprise both sums of squares and sums of products between variables, we use as entries in the MANOVA table matrices whose rows and columns are indexed by the variables. For example, the (ij)th element of such matrix would be the sum of products relating to the ith and jth variables. Similarly, the ith diagonal element of the matrix would be the sum of squares for the ith variable. Matrices of this form are usually called the **sums of squares and cross-products matrices** (SS&CP), and their purpose in the MANOVA table is similar to that of the (scalar) sums of squares in the ANOVA table. Table 2.2 corresponds to the MANOVA table, associated with hypothesis (2.43).

Table 2.2

Multivariate Analysis of Variance for a Balanced Longitudinal Study

source of variation	*SS&CP*	*d.f.*
among groups	$\mathbf{Q}_H = \sum_{k=1}^{q} m_k \left(\overline{\mathbf{y}}_{\cdot k} - \overline{\mathbf{y}}_{\cdot\cdot}\right)\left(\overline{\mathbf{y}}_{\cdot k} - \overline{\mathbf{y}}_{\cdot\cdot}\right)^t$	$q-1$
within groups	$\mathbf{Q}_E = \sum_{k=1}^{q} \sum_{i=1}^{m_k} \left(\mathbf{y}_{ik} - \overline{\mathbf{y}}_{\cdot k}\right)\left(\mathbf{y}_{ik} - \overline{\mathbf{y}}_{\cdot k}\right)^t$	$m-q$
total	$\mathbf{Q}_H + \mathbf{Q}_E = \sum_{k=1}^{q} \sum_{i=1}^{m_k} \left(\mathbf{y}_{ik} - \overline{\mathbf{y}}_{\cdot\cdot}\right)\left(\mathbf{y}_{ik} - \overline{\mathbf{y}}_{\cdot\cdot}\right)^t$	$m-1$

Observe that the entries in the MANOVA table are now matrices, and not scalars as in a usual ANOVA table. Specifically, matrix $\mathbf{Q}_H$ stands for the between groups SS&CP matrix, and accordingly $\mathbf{Q}_E$ for the within groups SS&CP matrix. Matrices $\mathbf{Q}_H$, $\mathbf{Q}_E$ involve except from the already defined data vector $\mathbf{y}_{ik}$, the vectors $\overline{\mathbf{y}}_{\cdot k} = \left(\overline{y}_{\cdot 1k}, \overline{y}_{\cdot 2k}, \ldots, \overline{y}_{\cdot nk}\right)^t$ which comprises all sample means $\overline{y}_{\cdot jk}$ $(j = 1, 2, \ldots n)$, and $\overline{\mathbf{y}}_{\cdot\cdot} = \left(\overline{y}_{\cdot 1\cdot}, \overline{y}_{\cdot 2\cdot}, \ldots, \overline{y}_{\cdot n\cdot}\right)^t$ which stands for the vector that comprises all $\overline{y}_{\cdot 1\cdot}$'s $(j = 1, 2, \ldots n)$. Further, m_k denotes the subjects belonging in group k. Due to that the MANOVA table consists entirely of matrices, it is no longer possible to construct a simple F-ratio based only on the sum of squares between and within groups (as is usual in ANOVA procedures) to test for group differences. Hence, to test H_0 other alternative statistics have been proposed, statistics that are mainly based on comparing the 'magnitude' of the SS&CP matrices, namely $\mathbf{Q}_H$ and $\mathbf{Q}_E$. Among them, three are the most widely applied; **Wilks' lambda, Pillai's trace** and **Roy's greatest root**.

Wilks' lambda is the most common, traditional multivariate test where there are more than two groups formed. As already stated, the proposed statistic is a measure of the difference between the groups means, and for the specific hypothesis is given by the following:

$$Wilks'\ lambda = \frac{\mid \mathbf{Q}_E \mid}{\mid \mathbf{Q}_H + \mathbf{Q}_E \mid}. \tag{2.44}$$

As one notices, the statistic is making use of the determinants of $\mathbf{Q}_H$ and $\mathbf{Q}_E$ rather than the SS&CP matrices themselves, reducing in this way the matrices to single scalars. One rejects H_0 for small values of lambda. Roy's greatest root (or Roy's largest eigenvalue) on the other hand, is described as being the largest eigenvalue of matrix $\mathbf{Q}_H\mathbf{Q}_E^{-1}$ (i.e. the largest root of $\mid \mathbf{Q}_H\mathbf{Q}_E^{-1} - \lambda\mathbf{I} \mid = \mathbf{0}$). Finally, the Pillai's trace (or the Pillai-Bartlett trace) is defined to be the trace of matrix $\mathbf{Q}_H\left(\mathbf{Q}_H + \mathbf{Q}_E\right)^{-1}$.

Although Pillai's trace has been found to be the most robust among the three tests (see *Olson, 1976*), in general none of the above statistics been shown to be significantly superior to the others. In fact, many software packages for MANOVA provide all three of the statistics, leaving the user to choose among them.

2.5.3 Advantages/Disadvantages of Classical Methods for Longitudinal Data

ANOVA and MANOVA models for longitudinal data, share specific advantages, as well as important disadvantages; in favor of them is the fact that the specific methods are based on well-understood and well-developed methods, hence have become more approachable to the average practicing statistician and to the researchers of other disciplines, compared to other, relatively recent modeling techniques which are still under continuous development and modifications. Furthermore, almost all of the widely used statistical packages contain facilities for the implementation of those methods, by this way freeing the hands of practitioners and providing them with software tools that have become standard and commonly available.

Certain disadvantages of these methods on the other hand, are equally important and reduce their applicability and validity on longitudinal data models. First of all, the ANOVA model for longitudinal data requires a very restrictive assumption about the associations (correlation) among observations on the same unit (or individual). To be more specific, it assumes a compound symmetric variance-covariance matrix $\mathbf{V}_i$ for the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in})^t$ of responses within each individual i. If this assumption is correct then the ANOVA method will provide valid inferences. However if the assumption of compound symmetry does not hold, application of ANOVA method may lead to erroneous conclusions. The problem about the compound symmetric form of the variance-covariance matrix assumed by ANOVA, is that in fact compound symmetry is not at all a realistic structure for covariance modeling of longitudinal data. MANOVA models on the other hand, in contrast to the ANOVA models, do not assume a specific structure for the covariance matrix, such as the compound symmetric form, but instead use an arbitrary covariance matrix, of a completely unstructured form. That is, the only knowledge we have about $\mathbf{V}_i$ is that is of the form:

$$\mathbf{V}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}, \tag{2.45}$$

and is the same for all subjects i. Thus, the advantage of not assuming a specific covariance structure in fact turns out to become a drawback for the analysis, since the number of parameters to be estimated increases significantly. In summary, when the covariance matrix was assumed to have the compound symmetric structure, the entire matrix depended on only two parameters. In contrast, in the MANOVA model, where no structure is assumed, the matrix depends on $\frac{n(n+1)}{2}$ parameters. Unfortunately now, under this consideration, a great many more parameters are required to describe how observations within the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in})^t$, vary and covary.

In addition to the restrictions that the two methods impose as concerns the modeling of data's covariance structure, another major disadvantage shared by both ANOVA and MANOVA methods is their failure to handle unbalanced longitudinal data. When the data are unbalanced with possibly different numbers of observations of each subject i, it is not possible to think in terms of ANOVA or MANOVA analysis. These deficiencies of classical methods inevitably forced statisticians to turn their attention towards other, newer methods, with less restrictive assumptions. An interesting and challenging problem was to find more realistic statistical models, models that do not rely on the assumption of compound symmetry, but can handle different covariance structures, each time depending on the specific problem and data under study. One of these models is the *General Linear Mixed-Effects* model for longitudinal data.

Concluding however, it should be noted that despite the development of new approaches both univariate and multivariate analysis of variance have been and still are very popular to analyzing repeated measures and longitudinal data, in some disciplines. Behavioural sciences and Psychology in particular have made extensive use of the ANOVA and MANOVA. There is a large number of articles and books appeared in the literature that discuss extensively the implementation of the latter approaches to repeated measures data. Representative examples are *Hand* and *Crowder (1990), Hand* and *Crowder (1996)*, *Rouanet* and *Lepine (1970)*, *O'Brien* and *Kaiser (1985), Hertzog* and *Rovine (1985),* and *Hand* and *Taylor (1986).*

Chapter 3

The General Linear **Mixed Model**

3.1 Introduction

Linear models have received great attention both in theory and in practice. From the theoretical point of view they are mathematically tractable, and in practical applications of wide variety they have proved to be of great value. The techniques of linear regression analysis find applications in almost every field of study, including social sciences, physical and biological sciences, business and technology, and the humanities.

Most types of statistical analysis based on linear models of single variables are included in the following categories:

- Simple and multiple linear regression
- Analysis of variance (ANOVA)
- Analysis of covariance (ANCOVA)
- Mixed model analysis of variance

The first three categories of the above linear models are considered as special cases of the **General Linear** Model (GLM in abbreviation) since all these can be written in the form of GLM [see, e.g. *Nelder* and *Wedderburn (1972)*]. In fact, every fixed-effects model that is linear in the parameters is called a general linear model. Its great importance lies in its broadness, since that within the GLM theory the whole spectrum of methods for analyzing one continuous response variable (Y) and multiple explanatory variables (X_i) is covered.

On the other hand, Mixed model analysis is a relatively recent type of statistical analysis, based on linear regression models. In particular, it applies to research involving factors whose levels can be controlled by the researcher (fixed) as well as factors whose levels are beyond the researcher's control (random effects). The "*mixed Model*" term is attributed to *Eisenhart (1947)*, who codified much of the material relating to linear models into three models. His Model **I** (fixed effects) and Model **II** (random effects) were given at that time extensive discussion. The mixed model, which was a combination (a mixture) of Model **I** and Model **II** was introduced as well, but was given relatively little discussion. These three models-Model **I**, Model **II** and the Mixed model-have generally provided the framework within which the bulk of applied linear model methods have been developed. Nowadays, it is established to refer to Model **I** as the fixed effects model and Model **II** as the random effects model. The genetics was the science field on which the first steps for the development of mixed model analysis were based. Especially in animal breeding, where the prediction and estimation of unobservable genetic parameters is an issue of great significance, mixed model analysis and especially mixed linear models have found the suitable soil for their implementation. This was mainly due to the fact that data arising in animal genetics are usually not balanced, henceforth methods developed for balanced data (as the ANOVA method described in Chapter 2), were not suitable any more. The basic challenge of genetics and animal breeding data was not only the formulation of a suitable model for such data (a linear model with many fixed environmental and many random genetic factors), but also the estimation of the associated variance components arising from such models. Animal breeding scientists, and especially *Henderson C.R.*, developed the initial results of mixed linear models, results that proved to be the foundation for nearly all applications of analyses for mixed linear models.

3.2 The General Linear Model (GLM)

The above discussion shows that the **General Linear Mixed Model** (**GLMM**) is not directly related to the GLM. In fact, the GLMM can be considered as an extension of the GLM viewed from the perspective that GLM is essentially a fixed effects model (contains only the vector of fixed parameters, $\mathbf{b}$), while GLMM, as already stated, contains both fixed and random effects.

Nevertheless, we present in the following lines the basic results of GLM theory (especially, the results associated with the estimation of $\mathbf{b}$) not only for the purpose of a complete and adequate illustration of GLM model, but mainly due to the fact that a parallel comparison of both GLM and GLMM results will be quite enlightening in detecting the similarities as well as the differences between the two models.

The general linear model is described by the equation:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where

- $\mathbf{y}$ is a $(n \times 1)$ vector of responses y_i
- $\mathbf{X}$ is a $(n \times p)$ known matrix (design matrix, incidence matrix or model matrix)
- $\mathbf{b}$ is a $(p \times 1)$ vector of all the unknown populations parameters b_j $(j = 1, ..., p)$
- $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector that contains the random errors ε_i $(i = 1, ..., n)$

Generally, it is assumed that $\boldsymbol{\varepsilon}$ follows a n-variate normal distribution (n is the total number of observations) with zero mean vector and variance-covariance matrix $\mathbf{V}$, i.e. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{V})$.

Remark 3.1: *The feature that distinguishes the general linear model* $\mathbf{y} = \mathbf{X}\mathbf{b} + \varepsilon$ *for the various special cases of linear models analysis (i.e. ANOVA, ANCOVA, simple and multiple linear regression) is in essence the type of matrix* $\mathbf{X}$*. Indeed, an ANOVA model can be constructed by using as elements of* $\mathbf{X}$ *dummy variables so that the* X_i *explanatory variables represent the model's factors and interactions. When* $\mathbf{X}$ *consists of observed values of the* X_i*'s we fall in the simple and multiple linear regression model. Finally, the case where some of the elements of matrix* $\mathbf{X}$ *are observed* X_i*'s and others are dummy variables, hence representing a model that combines both regression (simple or multiple) as well as linear models involving factors and interactions is the generally known covariance analysis.*

3.2.1 Estimation of Fixed Effects b

Three are the most commonly used approaches to the statistical estimation of parameters (**b**) for the GLM, namely:

- the Method of Least Squares
- the Method of Maximum Likelihood and
- the best linear unbiased estimator of **b** (BLUE)

An important distinction between the methods of least squares and maximum likelihood is that the former can be used without making any assumptions about the distribution of the response vector **y**, beyond specifying its expectation and possibly its variance-covariance matrix. On the contrary, it is not possible to derive maximum likelihood (ML) estimators without first specifying the distribution of the response vector **y**, since ML method requires the knowledge of the probability function of **y**. In summary, the basic results of the two methods as well as the derivation of BLUE are shown below (the notation of estimators is according to *Searle, 1971*).

3.2.1.1 The Least Squares Method

Two are the most popular and familiar variations of this estimation procedure. Both of them are based upon the general idea of minimizing sums of squares, with the distinguishing feature between them being the variance-covariance structure of the vector of responses, $\mathbf{y}$.

Ordinary Least Squares: As already stated, in order to derive least squares estimators all that one needs is vector's $\mathbf{y}$ expected value and variance-covariance matrix. The ordinary least squares method is based on the rather simple assumption that $\boldsymbol{\varepsilon}$ has zero mean $[E(\boldsymbol{\varepsilon}) = \mathbf{0}]$, and all the elements of $\boldsymbol{\varepsilon}$ are uncorrelated with one another with the same variance, σ_{ε}^{2}, so that the variance-covariance matrix of $\boldsymbol{\varepsilon}$ is:

$$Var(\boldsymbol{\varepsilon}) = \sigma_{\varepsilon}^{2}\mathbf{I}_n,$$

with $\mathbf{I}_n$ the $(n \times n)$ identity matrix. Thus, in accordance we have $E(\mathbf{y}) = \mathbf{Xb} + E(\boldsymbol{\varepsilon}) = \mathbf{Xb}$, and $Var(\mathbf{y}) = \sigma_{\varepsilon}^{2}\mathbf{I}_n$. The ordinary least squares (OLS) estimator of $\mathbf{b}$, usually denoted by $\mathbf{b}^{o}$, is chosen to be the value of $\mathbf{b}$ that minimizes the sum of squares of observations from their expected values, that is the sum of squares of the error terms: $\sum\limits_{i=1}^{n} \varepsilon_i^2 = \sum\limits_{i=1}^{n} [y_i - E(y_i)]^2 = \sum\limits_{i=1}^{n} (y_i - \mu_i)^2$ (where n denotes again the total number of responses). Notice that $\sum\limits_{i=1}^{n} (y_i - \mu_i)^2$ can be equivalently rewritten in matrix notation as:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \mu_i)^2 &= (y_1 - \mu_1)^2 + \ldots + (y_n - \mu_n)^2 \\
&= (y_1 - \mu_1, \ldots, y_n - \mu_n) \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix} = \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix}^{t} \begin{pmatrix} y_1 - \mu_1 \\ \vdots \\ y_n - \mu_n \end{pmatrix} \\
&= [\mathbf{y} - E(\mathbf{y})]^{t} [\mathbf{y} - E(\mathbf{y})] = (\mathbf{y} - \mathbf{Xb})^{t} (\mathbf{y} - \mathbf{Xb}) = \boldsymbol{\varepsilon}^{t}\boldsymbol{\varepsilon},
\end{aligned}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)^t$ the response vector, and $E(\mathbf{y}) = (\mu_1, \mu_2, ..., \mu_n)^t$. Hence the OLS estimator of $\mathbf{b}$ can be equivalently obtained by the minimization of the vector product $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{Xb})^t (\mathbf{y} - \mathbf{Xb})$. To achieve minimization we differentiate $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}$ with respect to the elements of $\mathbf{b}$ and equate $\partial(\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}) / \partial\mathbf{b}$ to zero. The resulting equations obtained are:

$$\mathbf{X}^t\mathbf{X}\mathbf{b}^o = \mathbf{X}^t\mathbf{y}. \tag{3.2}$$

Equations (3.2) are known as the **Normal Equations**. When matrix $\mathbf{X}^t\mathbf{X}$ is nonsingular[1], thus securing the existence of the inverse $(\mathbf{X}^t\mathbf{X})^{-1}$, the symbol $\tilde{\mathbf{b}}$ is used in place of $\mathbf{b}^o$, and the solution to the normal equations (OLS estimator) is given by:

$$\tilde{\mathbf{b}}_{OLS} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}. \tag{3.3}$$

In the case where $\mathbf{X}^t\mathbf{X}$ is not of full column rank, hence singular and its inverse does not exist, the **generalized inverse** of $\mathbf{X}^t\mathbf{X}$, namely $(\mathbf{X}^t\mathbf{X})^-$ is used instead satisfying $\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{X} = \mathbf{X}^t\mathbf{X}$. Then the corresponding solution of the normal equations is:

$$\mathbf{b}^o = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{y}. \tag{3.4}$$

The notation $\mathbf{b}^o$, instead of $\tilde{\mathbf{b}}$, in the above equation is used in order to emphasize that $\mathbf{b}^o$ is **only** a solution to normal equations and **not** an (OLS) estimator of $\mathbf{b}$. This is because although $\mathbf{b}^o$ is an estimator of some expression (not $\mathbf{b}$), this expression depends entirely upon which generalized inverse $(\mathbf{X}^t\mathbf{X})^-$ is used in obtaining $\mathbf{b}^o$.

Remark 3.2: *The generalized inverse (or g-inverse) of a $(m \times n)$ matrix $\mathbf{A}$, denoted as $\mathbf{A}^-$, is every $(n \times m)$ matrix satisfying the relation*

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{3.5}$$

[1] A square matrix $\mathbf{A}$ is said to be nonsingular ór regular if $|\mathbf{A}| \neq \mathbf{0}$. Otherwise $\mathbf{A}$ is said to be singular. The following are equivalent: $\mathbf{A}$ nonsingular$\Leftrightarrow |\mathbf{A}| \neq \mathbf{0} \Leftrightarrow \mathbf{A}^{-1}$ exists.

The generalized inverse of a given matrix $\mathbf{A}$ *is not unique. There is an infinite number of matrices that satisfy* (3.5). *Generalized inverse finds applications in cases where the (regular) inverse* $\mathbf{A}^{-1}$ *of* $\mathbf{A}$ *cannot be defined. This is the case where either* $m \neq n$ *hence* $\mathbf{A}$ *is not a square matrix, or* $m = n$ *but* $\mid \mathbf{A} \mid = \mathbf{0}$ *(since OLS estimator requires the inverse of matrix* $\mathbf{X}^t\mathbf{X}$ *which is a* $(p \times p)$ *square matrix, the inverse* $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ *may not always exist, e.g. when* $\mid \mathbf{X}^t\mathbf{X} \mid = \mathbf{0}$*, and the only way to come up with an OLS estimator is by using a generalized inverse).*

The problem that arises now is due to the fact that although the generalized inverse $\left(\mathbf{X}^t\mathbf{X}\right)^{-}$ *of* $\mathbf{X}^t\mathbf{X}$ *always exists, as mentioned above is not unique and as a consequence the OLS estimator* $\mathbf{b}^o$ *depends clearly upon the specific choice of* $\left(\mathbf{X}^t\mathbf{X}\right)^{-}$.

Generalized Least Squares: The ordinary least squares estimators of (3.3) was based on the assumption $Var\left(\varepsilon\right) = \sigma_\varepsilon^2\mathbf{I}_n$. If instead of assuming this special variance-covariance structure for ε we consider the more general situation of

$$Var\left(\varepsilon\right) = \mathbf{V}$$

for some positive definite symmetric matrix $\mathbf{V}$, then the generalized least squares (G.L.S.) estimator of $\mathbf{b}$ (also known as weighted least squares estimator), is obtained by minimizing:

$$\left[\mathbf{y}-E\left(\mathbf{y}\right)\right]^t \mathbf{V}^{-1}\left[\mathbf{y}-E\left(\mathbf{y}\right)\right] \equiv \left(\mathbf{y}-\mathbf{X}\mathbf{b}\right)^t \mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\mathbf{b}\right), \tag{3.6}$$

with respect to $\mathbf{b}$. Minimization of (3.6) yields the **generalized least squares equations**

$$\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.7}$$

with the solution

$$\tilde{\mathbf{b}}_{GLS} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.8}$$

being the generalized least squares (GLS) estimator of $\mathbf{b}$.

3.2.1.2 The Method of Maximum Likelihood

For obtaining the above (ordinary and generalized) least squares estimators of fixed effects vector $\mathbf{b}$, no specific assumption was made about the form of the distribution of the vector of random errors $\boldsymbol{\varepsilon}$, and consequently about the distributional form of the response vector $\mathbf{y}$. Yet, for the maximum likelihood estimator of $\mathbf{b}$ it is necessary to make some assumption about the distribution. Most times this assumption states that $\boldsymbol{\varepsilon}$ is normally distributed, with zero mean and variance-covariance matrix $\mathbf{V}$, i.e., $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{V})$. As a consequence the response vector $\mathbf{y}$ is distributed as:

$$\mathbf{y} \sim N_n(\mathbf{Xb}, \mathbf{V}).$$

The corresponding probability density function (p.d.f.) of the normally distributed vector $\mathbf{y}$ is well-known to be:

$$f(\mathbf{y}; \mathbf{Xb}, \mathbf{V}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{V} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{Xb})^t \mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})\right\}. \tag{3.9}$$

To derive the M.L. estimator of $\mathbf{b}$, the likelihood of the sample of observations must be maximized with respect to $\mathbf{b}$. Now, for vector $\mathbf{y}$ the likelihood function $L(\mathbf{Xb}, \mathbf{V}; \mathbf{y})$ is algebraically the same as the p.d.f. $f(\mathbf{y}; \mathbf{Xb}, \mathbf{V})$ in (3.9) (the change in the sequence of symbols between f and L is made to emphasize that while the weight in f is on the response variable vector $\mathbf{y}$, the emphasis in L is on the parameters considering the $\mathbf{y}$ to be the fixed observations).

Maximizing $L(\mathbf{Xb}, \mathbf{V}; \mathbf{y})$ with respect to $\mathbf{b}$ is equivalent, due to that ln is an increasing function, to maximizing $\ln L(\mathbf{Xb}, \mathbf{V}; \mathbf{y})$. For this reason the equation $\partial(\ln L)/\partial\mathbf{b} = \mathbf{0}$ is solved and the solution, which is the M.L. estimator of $\mathbf{b}$ is:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.10}$$

the same as the generalized least squares estimator. By assuming now (as in the OLS estimation) a less general covariance structure for $\boldsymbol{\varepsilon}$, i.e. $\mathbf{V} =\sigma_{\varepsilon}^{2}\mathbf{I}_n$ then the ML estimator of $\mathbf{b}$ simplifies to

$$\hat{\mathbf{b}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y},$$

the ordinary least squares estimator. The only difference now, is the distributional assumption of $\boldsymbol{\varepsilon}$ which is considered to be multivariate normal, hence $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0},\sigma_{\varepsilon}^{2}\mathbf{I}_n\right)$.

3.2.1.3 The Best Linear Unbiased Estimator (BLUE)

In the situation of least squares estimation of the fixed effects $\mathbf{b}$, the estimators $\tilde{\mathbf{b}}_{OLS}$ and $\tilde{\mathbf{b}}_{GLS}$ were derived by making specific assumptions about the expected value and variance-covariance matrix of $\boldsymbol{\varepsilon}$. Furthermore, on assuming normality for $\boldsymbol{\varepsilon}$ it was made possible to obtain the maximum likelihood estimators $\hat{\mathbf{b}}$ (identical to $\tilde{\mathbf{b}}_{OLS}$ and $\tilde{\mathbf{b}}_{GLS}$). In this section, we will present an estimator-known as best linear unbiased estimator-that does not require any assumptions at all about the moments of $\boldsymbol{\varepsilon}$ or its distributional form. Instead, two of the most widely applied criteria on the investigation of estimators in mathematical statistics are utilized: the unbiasedness criterion and the criterion of minimum variance of the estimator. The specific estimation procedure is based upon the *Gauss-Laplace-Markov* theory of linear estimation, using a generalized version of the Gauss-Markov theorem (for more details on the topic we refer to *Harville, 1976*). By this approach, it is possible to estimate not just the fixed effects vector $\mathbf{b}$, but in general any linear function of $\mathbf{b}$, $\mathbf{t}^t\mathbf{b}$ with $\mathbf{t}$ any column vector [the only restriction about the vector $\mathbf{t}$ is on its dimension, that has to be in compliance with the dimension of $\mathbf{b}$. Hence, since vector $\mathbf{b}$ was chosen to be of $(p \times 1)$ dimension, $\mathbf{t}$ must be also a $(p \times 1)$ vector, resulting a well-defined scalar $\mathbf{t}^t\mathbf{b}$].

The three characteristics (with the two of them essentially being criteria for obtaining an optimum estimator) of the estimator, described in its definition, lead to its derivation. We now describe analytically how this derivation is accomplished. The technique of b.l.u. estimation proceeds as follows: Initially, we take the estimator of $\mathbf{t}^t\mathbf{b}$ to be the linear function $\boldsymbol{\lambda}^t\mathbf{y}$, of the vector of observations $\mathbf{y}$ (this is where the term linear in the b.l.u.e.

definition corresponds). As a next step, we restrict the set of estimators of the form $\boldsymbol{\lambda}^t\mathbf{y}$, by imposing the unbiasedness criterion for $\boldsymbol{\lambda}^t\mathbf{y}$. The estimator $\boldsymbol{\lambda}^t\mathbf{y}$ of $\mathbf{t}^t\mathbf{b}$, is said to be an unbiased estimator if-f:

$$\begin{aligned} E\left(\boldsymbol{\lambda}^t\mathbf{y}\right) = \mathbf{t}^t\mathbf{b} \Rightarrow \boldsymbol{\lambda}^t E\left(\mathbf{y}\right) = \mathbf{t}^t\mathbf{b} \\ \underset{E(\mathbf{y})=\mathbf{X}\mathbf{b}}{\Rightarrow} \boldsymbol{\lambda}^t\mathbf{X}\mathbf{b} = \mathbf{t}^t\mathbf{b}\ \forall\mathbf{b} \Rightarrow \boldsymbol{\lambda}^t\mathbf{X} = \mathbf{t}^t. \end{aligned} \tag{3.11}$$

Taking transposes on both sides of $\boldsymbol{\lambda}^t\mathbf{X} = \mathbf{t}^t$, for reasons that will become evident in a while, we obtain

$$\mathbf{X}^t\boldsymbol{\lambda} = \mathbf{t}. \tag{3.12}$$

Having in mind (3.12), we additionally require of $\boldsymbol{\lambda}^t\mathbf{y}$ to be the 'best' estimator of $\mathbf{t}^t\mathbf{b}$; 'best' means that in the class of linear, unbiased estimators of $\mathbf{t}^t\mathbf{b}$, the best is to be the one that has minimum variance [i.e. a minimum variance unbiased estimator (M.V.U.E.), being linear at the same time]. This criterion will assist in deriving the $\boldsymbol{\lambda}^t$ of estimator $\boldsymbol{\lambda}^t\mathbf{y}$. Indeed, let $Var\left(\mathbf{y}\right) = \mathbf{V}$. Then,

$$Var\left(\boldsymbol{\lambda}^t\mathbf{y}\right) = \boldsymbol{\lambda}^t Var\left(\mathbf{y}\right)\boldsymbol{\lambda} = \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda},$$

and in order for $\boldsymbol{\lambda}^t\mathbf{y}$ to be 'best', that is to have the minimum variance between all unbiased linear estimators $\boldsymbol{\lambda}^t\mathbf{y}$, $\boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda}$ must be minimum. This minimization is well-executed using the Lagrangian technique[2], taking $2\boldsymbol{\theta}^t$ as a vector of lagrange multipliers and $\mathbf{X}^t\boldsymbol{\lambda} - \mathbf{t}$ of (3.12) to be the constraint. Hence, we minimize the quantity

$$Q = \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} - 2\boldsymbol{\theta}^t\left(\mathbf{X}^t\boldsymbol{\lambda} - \mathbf{t}\right)$$

[2] A technique, devised by the French mathematician J.L. Lagrange, for calculating the optimal value of a variable, subject to some constraint, in order to maximize or minimize another variable.

with respect to the elements of $\boldsymbol{\lambda}^t$ and $\boldsymbol{\theta}^t$. This is accomplished by solving the system:

$$\left\{\begin{array}{c} \frac{\partial Q}{\partial \theta}=0 \\ \frac{\partial Q}{\partial \lambda}=0 \\ \mathrm{X}^t\boldsymbol{\lambda}=\mathrm{t} \end{array}\right\}. \tag{3.13}$$

Solving $\partial Q/\partial\boldsymbol{\theta}=\mathbf{0}$ we get $\mathbf{X}^t\boldsymbol{\lambda}=\mathbf{t}$, [note that $\mathbf{X}^t\boldsymbol{\lambda}=\mathbf{t}$ is equivalent to $\boldsymbol{\lambda}^t\mathbf{X}=\mathbf{t}^t$ since $\mathbf{X}^t\boldsymbol{\lambda}=\mathbf{t}\Leftrightarrow(\mathbf{X}^t\boldsymbol{\lambda})^t=\mathbf{t}^t\Leftrightarrow\boldsymbol{\lambda}^t\mathbf{X}=\mathbf{t}^t$]. Furthermore, $\partial Q/\partial\boldsymbol{\lambda}=\mathbf{0}$ gives $\mathbf{V}\boldsymbol{\lambda}=\mathbf{X}\boldsymbol{\theta}$, or after solving for $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}=\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\theta}$. Taking transposes on both sides, it is $\boldsymbol{\lambda}^t=\boldsymbol{\theta}^t\mathbf{X}^t\mathbf{V}^{-1}$ $[(\mathbf{V}^{-1})^t=\mathbf{V}^{-1}$, since $(\mathbf{V}^{-1})^t=(\mathbf{V}^t)^{-1}$ and $\mathbf{V}^t=\mathbf{V}]$. Substituting $\boldsymbol{\lambda}^t$ to $\boldsymbol{\lambda}^t\mathbf{X}=\mathbf{t}^t$, we obtain $\boldsymbol{\theta}^t\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}=\mathbf{t}^t\Rightarrow\boldsymbol{\theta}^t=\mathbf{t}^t(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}$. Now, replacing $\boldsymbol{\theta}^t$ in $\boldsymbol{\lambda}^t=\boldsymbol{\theta}^t\mathbf{X}^t\mathbf{V}^{-1}$ we can derive a solution for the $\boldsymbol{\lambda}^t$, of the estimator $\boldsymbol{\lambda}^t\mathbf{y}$, that does not contain any more the unknown vector $\boldsymbol{\theta}$, and is:

$$\boldsymbol{\lambda}^t=\mathrm{t}^t\left(\mathrm{X}^t\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\mathrm{X}^t\mathrm{V}^{-1}. \tag{3.14}$$

Hence, the b.l.u.e. of $\mathbf{t}^t\mathbf{b}$, $\boldsymbol{\lambda}^t\mathbf{y}$ is given by

$$\boldsymbol{\lambda}^t\mathbf{y}=\mathbf{t}^t\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathrm{X}^t\mathrm{V}^{-1}\mathrm{y}. \tag{3.15}$$

It can be shown using (3.15) (for the proof see, e.g. *Searle, 1971)* that the b.l.u.e. of fixed effects vector $\mathbf{b}$,denoted by $\mathbf{b}_{BLUE}$, is:

$$\mathrm{b}_{BLUE}=\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.16}$$

identical to the generalized least squares and maximum likelihood estimators, respectively [assuming $\boldsymbol{\varepsilon}\sim\mathbf{N}_n(\mathbf{0},\mathbf{V})$].

3.3 General Linear Mixed Model (GLMM)

In spite of its great popularity and applicability, the already discussed GLM of equation (3.1) has also an important disadvantage; the problem is that this model allows only one source of randomness, the random error term ε. The general linear mixed model (GLMM) removes this restriction, by allowing for other error structures except ε, error structures that are popularly known as 'random effects' factors or simply random effects. Under this perspective, GLMM can be considered of being an extension of GLM. This model has received increasing attention, mainly due to its wide applicability and ease of interpretation. Within this framework, one is generally interested in inference procedures and estimation of the parameters of the GLMM. There is a considerable literature on the subject of estimation, which is usually based on maximum likelihood or restricted maximum likelihood (*Patterson* and *Thompson, 1971*) for the fixed effects and the variance parameters, with best linear unbiased prediction (BLUP) for the random effects of the model.

3.3.1 The Model Equation

The *General Linear Mixed Effects Model* is described by the following equation (see, e.g. *Harville, 1977)*:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \varepsilon, \tag{3.17}$$

where

- $\mathbf{y}$ is a $(n \times 1)$ vector of random variables whose observed realizations are the responses

- $\mathbf{X}$ is a $(n \times p)$ matrix of known coefficients that relates observations to fixed effects

- $\mathbf{b}$ is a $(p \times 1)$ vector of unobservable parameters (the fixed effects)

- $\mathbf{Z}$ is a $(n \times q)$ matrix of known coefficients that relates observations to random effects

- $\mathbf{u}$ is a $(q \times 1)$ vector of unobservable parameters (the random effects) and finally

- $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of unobservable random errors

By definition we take the random terms $\mathbf{u}$, $\boldsymbol{\varepsilon}$ of the above model to have zero expectations:

$$E\left(\mathbf{u}\right) = \mathbf{0},\ E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0} \tag{3.18}$$

and variance-covariance matrices:

$$Var\left(\mathrm{u}\right) = \mathbf{D},\ Var\left(\boldsymbol{\varepsilon}\right) = \mathbf{R}, \tag{3.19}$$

where $\mathbf{D}$, $\mathbf{R}$ are considered known, positive definite matrices. Further, we also define:

$$Var\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \mathrm{D} & 0 \\ 0 & \mathrm{R} \end{pmatrix}. \tag{3.20}$$

From this assumption it follows that the random vectors $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are independent. Taking advantage of (3.18), (3.19), (3.20) we can easily determine the expectation and the variance-covariance matrix of the third random term of the model, the response vector $\mathbf{y}$. We have:

$$\begin{aligned} E\left(\mathrm{y}\right) &= E\left(\mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}\right) = E\left(\mathbf{Xb}\right) + E\left(\mathbf{Zu}\right) + E\left(\boldsymbol{\varepsilon}\right) \\ &= \mathbf{Xb} + \mathbf{Z}E\left(\mathbf{u}\right) + E\left(\boldsymbol{\varepsilon}\right) = \mathbf{Xb} + \mathbf{Z0} + \mathbf{0} \\ &= \mathbf{Xb} \end{aligned} \tag{3.21}$$

and

$$Var\left(\mathbf{y}\right) = \mathbf{V} = Var\left(\mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}\right) = Var\left(\mathbf{Zu} + \boldsymbol{\varepsilon}\right)$$

$$\begin{aligned} &= \mathbf{Z}Var\left(\mathbf{u}\right)\mathbf{Z}^{t} + Var\left(\boldsymbol{\varepsilon}\right) + \mathbf{Z}Cov\left(\mathbf{u},\boldsymbol{\varepsilon}\right) + Cov\left(\boldsymbol{\varepsilon},\mathbf{u}\right)\mathbf{Z}^{t} \\ &= \mathbf{ZDZ}^{t} + \mathbf{R}. \end{aligned} \tag{3.22}$$

3.3.2 Estimation of Fixed Effects/Prediction of Random Effects

All estimation methods of section 3.2.1 (Least squares, maximum likelihood, best linear unbiased estimation via the Gauss-Markov theorem), were concerned with the estimation of $\mathbf{b}$, a (vector) parameter that comprises all fixed effects of the GLM: $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$. For the more complex mixed-effects model $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$ the issue of estimation becomes somewhat more complicated, due to the extra random parameter $\mathbf{u}$ introduced into the model.

While, at least from a statistical point of view, estimating fixed effects $\mathbf{b}$ (or functions of $\mathbf{b}$) is considered to be a problem of major importance, in contrast the problem of making inferences about a realized or sample value of random vector $\mathbf{u}$ did not received analogous attention from statisticians. The reasons that lead statisticians to "neglect" estimation of $\mathbf{u}$ (and estimation of random effects, in general), are obvious in a way; for the classical (frequentist) statistical school of thought, the main distinction between fixed and random effects lied on the fact that the effects are random when we are not interested in their specific individual values. Hence, estimation of random effects never seemed to be of any practical usefulness. Instead, what has been established as a standard procedure is the estimation of variance components, i.e. the elements (variances and covariances) of the random term's variance-covariance matrices. Motivation for this is simple; rather than making inferences about specific realized values of a random variable, which essentially are just a small sample from a (finite or most usually infinite) population, it is much more useful to direct the efforts in drawing conclusions about the population's variation.

In specific situations though, realized values of $\mathbf{u}$ are important; such situations often occur in animal breeding applications, where (linear) combinations of $\mathbf{b}$ and $\mathbf{u}$ correspond to the breeding values of individual animals, and the primary objective of the statistical analysis is to evaluate these same animals as candidates for some future breeding program. In this context, estimates of $\mathbf{u}$ has been used extensively in order to decide which animals are best, in some sense.

As concern terminology now, it is of interest to note that it is typical in the literature to meet the term "predictor", rather than "estimator", when referring to random effects. It has become common practice to estimate fixed effects and to predict random effects. Although not of much practical importance, this terminology issue has gone under long discussion, and the question of which of the two terms is the correct one is still under debate (see *Robinson, 1991*). Nevertheless, from now on we are going to use the term prediction when referring to the estimation of random effects $\mathbf{u}$. Finally, note that for deriving the following results, model's $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$ variance-covariance structure is considered to be known (i.e. $\mathbf{D}$, $\mathbf{R}$ and consequently $\mathbf{V}$ are known).

3.3.2.1 Estimation of Fixed Effects b

The methods used for deriving estimates for the fixed-effects parameter vector $\mathbf{b}$ of the GLM $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$, apply also for the estimation of $\mathbf{b}$ in the GLMM. Hence, once again, widely applied procedures such as maximum likelihood and best linear unbiased estimation are considered to be the most suitable in order to form adequate expressions for the estimates of $\mathbf{b}$. Consequently, the analytic expressions of both the ML and the best linear unbiased estimator (BLUE) of the fixed effects vector $\mathbf{b}$ for the GLMM are exactly identical to those presented in section 3.2.1, concerning the GLM, thus:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}, \tag{3.23}$$

expresses the ML estimator of $\mathbf{b}$, and is identical to the best linear unbiased estimator:

$$\mathbf{b}_{BLUE} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}. \tag{3.24}$$

3.3.2.2 Prediction of Random Effects $\mathbf{u}$

Estimation of fixed effects involves various estimation methods of equal importance and applicability (Least squares, maximum likelihood, BLUE). In contrast, for the prediction of the random effects $\mathbf{u}$ (partially due to the non-acceptance of the frequentists), there has not been shown an early analogous interest for developing an estimation method. But what mainly animal breeders needed were predictors of random effects, such as breeding values. The answer to their problem was best linear unbiased prediction[3] (BLUP), which until now remains the most widely used procedure for random-effects prediction in mixed model analysis. It should be mentioned at this point, that as was the case with best linear unbiased estimation of fixed effects, the results concerning BLUP estimates are also obtained by use of the Gauss-Markov theorem (*Harville, 1976, 1977*) this time extended to include the estimation of random effects.

The Best Linear Unbiased Predictor (BLUP): The current subsection deals with methods associated with the derivation of best linear unbiased predictors for the random effects vector $\mathbf{u}$. We have already stated that best linear unbiased prediction (or BLUP), is in general a method of estimating random effects, originally developed for ranking and selection in the contexts of animal breeding and genetics. A vast literature on the derivation of BLUP estimates for the random effects vector $\mathbf{u}$, (denoted similarly to the BLUE notation as $\mathbf{u}_{BLUP}$), in the general linear mixed model exists since a large number of statisticians, most usual of different statistical wiewpoints (i.e. Classical or Bayesian schools of thought), have been concentrated on this issue. Among them, *Henderson (1949,1950)*, *Goldberger (1962)* and *Searle (1971, 1995)* have contributed the most elegant and applicable derivations of BLUP, from a Classical perspective, while Bayesian derivations of BLUP were given by *Dempfle (1977)* and *Lindley* and *Smith (1972)*. One of the most popular derivations is presented below, a relatively recent derivation provided by *Searle (1995)*, which is very similar to that of $\mathbf{b}_{BLUE}$ of the fixed effects vector $\mathbf{b}$, in the general

[3] The best linear unbiased predictors are also known as empirical Bayes estimators.

linear model (3.1). Before proceed in presenting the method, it should be noticed that it is a rather general procedure, estimating not just $\mathbf{u}$, but linear combinations of $\mathbf{u}$ and $\mathbf{Xb}$, namely of the form

$$\mathbf{w} = \mathbf{t}_1^t\mathbf{Xb} + \mathbf{t}_2^t\mathbf{u}, \tag{3.25}$$

yielding $\mathbf{u}_{BLUP}$ of $\mathbf{u}$ as a special case. Also, similarly to BLUE, no distributional assumptions concerning the random parts of the model are required.

Deriving the BLUP in a **Manner Similar** to **Deriving the BLUE:** As already mentioned, we want to predict the function $\mathbf{w} = \mathbf{t}_1^t\mathbf{Xb} + \mathbf{t}_2^t\mathbf{u}$, involving both $\mathbf{b}$ and $\mathbf{u}$. First of all, we take the estimator of $\mathbf{w}$ to be linear in $\mathbf{y}$, i.e. to be of the form $\boldsymbol{\lambda}^t\mathbf{y}$. Now, as in the best linear unbiased estimator case, we take advantage of the same criteria used there, the one of unbiasedness and minimum variance. As so, we have (for unbiasedness to hold):

$$\begin{aligned}
&E\left(\boldsymbol{\lambda}^t\mathbf{y}\right) = \mathbf{t}_1^t\mathbf{Xb} + \mathbf{t}_2^t\mathbf{u} \Rightarrow E\left(\boldsymbol{\lambda}^t\mathbf{y} - \mathbf{t}_1^t\mathbf{Xb} - \mathbf{t}_2^t\mathbf{u}\right) = 0\\
&\Rightarrow \boldsymbol{\lambda}^t E\left(\mathbf{y}\right) - \mathbf{t}_1^t\mathbf{Xb} - \mathbf{t}_2^t E\left(\mathbf{u}\right) = 0 \Rightarrow \boldsymbol{\lambda}^t\mathbf{Xb} - \mathbf{t}_1^t\mathbf{Xb} = 0\\
&\Rightarrow \boldsymbol{\lambda}^t\mathbf{Xb} = \mathbf{t}_1^t\mathbf{Xb} \Rightarrow \left(\boldsymbol{\lambda}^t\mathbf{Xb}\right)^t = \left(\mathbf{t}_1^t\mathbf{Xb}\right)^t \Rightarrow \mathbf{b}^t\mathbf{X}^t\boldsymbol{\lambda} = \mathbf{b}^t\mathbf{X}^t\mathbf{t}_1 \ \forall\ \mathbf{b}\\
&\Rightarrow\ \mathbf{X}^t\boldsymbol{\lambda} = \mathbf{X}^t\mathbf{t}_1.
\end{aligned} \tag{3.26}$$

Next, we choose $\boldsymbol{\lambda}$ to minimize (for the 'best' to hold):

$$\begin{aligned}
&Var\left(\boldsymbol{\lambda}^t\mathbf{y} - \mathbf{t}_1^t\mathbf{Xb} - \mathbf{t}_2^t\mathbf{u}\right) = Var\left(\boldsymbol{\lambda}^t\mathbf{y} - \mathbf{t}_2^t\mathbf{u}\right)\\
&= Var\left(\boldsymbol{\lambda}^t\mathbf{y}\right) + Var\left(\mathbf{t}_2^t\mathbf{u}\right) - Cov\left(\boldsymbol{\lambda}^t\mathbf{y}, \mathbf{t}_2^t\mathbf{u}\right) - Cov\left(\mathbf{t}_2^t\mathbf{u}, \boldsymbol{\lambda}^t\mathbf{y}\right)\\
&= \boldsymbol{\lambda}^t Var\left(\mathbf{y}\right)\boldsymbol{\lambda} + \mathbf{t}_2^t Var\left(\mathbf{u}\right)\mathbf{t}_2 - \boldsymbol{\lambda}^t Cov\left(\mathbf{y}, \mathbf{u}\right)\mathbf{t}_2 - \mathbf{t}_2^t Cov\left(\mathbf{u}, \mathbf{y}\right)\boldsymbol{\lambda}\\
&= \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - \boldsymbol{\lambda}^t Cov\left(\mathbf{y}, \mathbf{u}\right)\mathbf{t}_2 - \mathbf{t}_2^t Cov\left(\mathbf{u}, \mathbf{y}\right)\boldsymbol{\lambda}.
\end{aligned}$$

Since $Cov\left(\mathbf{y}, \mathbf{u}\right) = Cov\left(\mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}, \mathbf{u}\right) = Cov\left(\mathbf{Zu}, \mathbf{u}\right) = \mathbf{Z}Cov\left(\mathbf{u}, \mathbf{u}\right) = \mathbf{Z}Var\left(\mathbf{u}\right) = \mathbf{ZD}$ and similarly $Cov\left(\mathbf{u}, \mathbf{y}\right) = ... = Cov\left(\mathbf{u}, \mathbf{Zu}\right) = Cov\left(\mathbf{u}, \mathbf{u}\right)\mathbf{Z}^t = Var\left(\mathbf{u}\right)\mathbf{Z}^t = \mathbf{DZ}^t$, the above becomes [taking advantage of the fact that $\mathbf{t}_2^t\mathbf{DZ}^t\boldsymbol{\lambda}$ is a scalar, hence

$\left(\mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\boldsymbol{\lambda}\right)^t = \mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\boldsymbol{\lambda}$]:

$$\begin{aligned}
&\boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - \boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\boldsymbol{\lambda} \\
&= \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - \boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \left(\mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\boldsymbol{\lambda}\right)^t \\
&= \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - \boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}\mathbf{t}_2 - \boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}^t\mathbf{t}_2 \\
&\underset{\mathbf{D}^t=\mathbf{D}}{=} \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - 2\boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}\mathbf{t}_2.
\end{aligned}$$

To minimize the latter, we once again implement the Lagrange theory, taking $2\boldsymbol{\theta}^t$ to be a vector of lagrange multipliers, and (3.26) to be the constraint. Thus, we minimize:

$$Q = \boldsymbol{\lambda}^t\mathbf{V}\boldsymbol{\lambda} + \mathbf{t}_2^t\mathbf{D}\mathbf{t}_2 - 2\boldsymbol{\lambda}^t\mathbf{Z}\mathbf{D}\mathbf{t}_2 + 2\boldsymbol{\theta}^t\left(\mathbf{X}^t\boldsymbol{\lambda} - \mathbf{X}^t\mathbf{t}_1\right).$$

Working similarly to the derivation of $\mathbf{b}_{BLUE}$ for the GLM, after the calculations this minimization yields:

$$\begin{aligned}
&\left(\mathbf{t}_1^t\mathbf{X}\mathbf{b} + \mathbf{t}_2^t\mathbf{u}\right)_{BLUP} = \boldsymbol{\lambda}^t\mathbf{y} \\
&= \mathbf{t}_1^t\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y} + \mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\right]\mathbf{y} \qquad (3.27) \\
&= \mathbf{t}_1^t\mathbf{X}\mathbf{b}_{BLUE} + \mathbf{t}_2^t\mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{BLUE}\right).
\end{aligned}$$

The above identity for $\left(\mathbf{t}_1^t\mathbf{X}\mathbf{b} + \mathbf{t}_2^t\mathbf{u}\right)_{BLUP}$ is true for any $\mathbf{t}_1^t$, $\mathbf{t}_2^t$. Therefore, by taking $\mathbf{t}_1^t = \mathbf{0}$ and $\mathbf{t}_2^t$ to be successive rows of $\mathbf{I}_q$ (for $\mathbf{u}$ having q elements) we are able to obtain the $\mathbf{u}_{BLUP}$ as a special case of (3.27) given by:

$$\mathbf{u}_{BLUP} = \mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{BLUE}\right), \qquad (3.28)$$

or due to the equivalence between $\mathbf{b}_{BLUE}$ and $\hat{\mathbf{b}}$,

$$\mathbf{u}_{BLUP} = \mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right).$$

The obtained best linear unbiased predictors $\mathbf{u}_{BLUP}$ are best in the sense that they minimize the sampling variance, linear in the sense that they are linear functions of vector $\mathbf{y}$, and unbiased in the sense that $E\left(\mathbf{u}_{BLUP}\right) = \mathbf{u}$ holds.

3.3.3 Henderson's Mixed Model Equations

The calculation of the $\mathbf{b}_{BLUE}$ and $\mathbf{u}_{BLUP}$ solutions of equations (3.24) and (3.28) respectively, requires the inverse of the variance-covariance matrix of the responses vector $\mathbf{y}$, namely $\mathbf{V}$. Generally, this is a matrix of order equal to the number of elements in $\mathbf{y}$, hence for the GLMM of equation (3.17) since $\mathbf{y}$ is a $(n \times 1)$ vector, consequently $\mathbf{V}$ is a $(n \times n)$ matrix. An obvious problem with these solutions is that in most linear mixed model, unbalanced data situations, with the number of elements in $\mathbf{y}$ being very large ($\mathbf{y}$ could contain many hundreds or even thousands of observations), the computation of the inverse of $\mathbf{V}$, $\mathbf{V}^{-1}$ can be quite difficult.

For this kind of situations, other methods for the derivation of BLUE and BLUP of fixed and random effects had to be developed, methods that would not require the inversion of the dispersion matrix $\mathbf{V}$. *Henderson (1949, 1950)* and *Henderson et al. (1959)*, were the first who solved this problem, by developing a set of two equations, defined as:

$$\begin{aligned} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} &= \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)\tilde{\mathbf{u}} &= \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} \end{aligned} \tag{3.29}$$

The above equations are *Henderson's* Mixed Model Equations (usually mentioned under the acronym MME), and $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{u}}$ obtained by solving these equations are referred to as the 'mixed model solutions'. In fact, as it will be shown in the sequel, the 'mixed model solutions' $\tilde{\mathbf{b}}$, $\tilde{\mathbf{u}}$ of MME are identical to the $\mathbf{b}_{BLUE}$, $\mathbf{u}_{BLUP}$, thus what essentially MME offer is a practical method of obtaining $\tilde{\mathbf{b}} = \mathbf{b}_{BLUE}$ and $\tilde{\mathbf{u}} = \mathbf{u}_{BLUP}$. In other words, although MME are nothing more than just one of several computing tools for the derivation of $\mathbf{b}_{BLUE}$ and $\mathbf{u}_{BLUP}$, they proved to be the most useful among these computing algorithms, having the advantage of requiring the less possible computational efforts..

Notice that the MME equations can be alternatively written in a matrix form, as the single equation:

$$\begin{pmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\mathrm{b}} \\ \tilde{\mathrm{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{3.30}$$

Notably, MME are most frequently presented by the matrix form of (3.30), since this matrix representation provides us with a method of jointly obtaining $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{u}}$, by calculating the vector $\left(\tilde{\mathbf{b}},\, \tilde{\mathbf{u}}\right)^t$. The basic advantage of MME when compared with other methods for deriving $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{u}}$ is, as noted previously, the absence of $\mathbf{V}^{-1}$ within them. On the other hand, one might argue that there is now real gain, since we still have to calculate two inverses, $\mathbf{R}^{-1}$ and $\mathbf{D}^{-1}$ in order to get the estimates. This is not right though, since $\mathbf{R}^{-1}$ and $\mathbf{D}^{-1}$ are much easier to obtain, compared with $\mathbf{V}^{-1}$. To see that let us consider the dimensionality of the dispersion matrices $\mathbf{R}$ and $\mathbf{D}$, compared to the dimensionality of $\mathbf{V}$. Recall that $\mathbf{X}$ and $\mathbf{Z}$ are of dimensions $(n \times p)$ and $(n \times q)$ respectively. Consequently, $\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}$ is a $(p \times p)$ matrix, $\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}$ is a $(p \times q)$ matrix, $\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}$ is a $(q \times p)$ matrix and $\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1}$ a $(q \times q)$ matrix. Hence, the matrix

$$\begin{pmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{pmatrix},$$

(often called coefficient matrix), is a $(p+q) \times (p+q)$ dimensional matrix and to obtain $\left(\tilde{\mathbf{b}},\, \tilde{\mathbf{u}}\right)^t$ estimates requires finding the inverse (or generalized inverse, in case the regular inverse does not exist) of this $(p+q) \times (p+q)$ matrix. The latter task is much easier than finding the inverse of an $(n \times n)$ matrix, such as $\mathbf{V}$, since p and q are the number of fixed and random effects parameters and usually it is $p \ll n$ and $q \ll n$. This suggests that the coefficient matrix is of considerably less order than $\mathbf{V}$, and therefore less computational efforts are required for its inversion, compared to the inversion of $\mathbf{V}$.

3.3.3.1 The Derivation of the MME

Henderson (1950) described the estimates obtained from the MME (3.29) as being "joint maximum likelihood estimates". Only later though, he discovered [see *Henderson (1973)*] that what he actually did was not a maximum likelihood estimation, but a joint probability density function (pdf) maximization, instead.

In this context, *Henderson* assumed that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are normally distributed vectors, with $\mathbf{u} \sim N_q(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{R})$, where $\mathbf{u}$, $\boldsymbol{\varepsilon}$ are once again the random terms of the GLMM:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}.$$

Consequently to the above assumptions, we have that the response vector $\mathbf{y}$ is distributed as $\mathbf{y} \sim N_n(\mathbf{Xb}, \mathbf{V})$, with $\mathbf{V} = \mathbf{ZDZ}^t + \mathbf{R}$ being the dispersion matrix of $\mathbf{y}$. *Henderson* derived the MME, not by maximizing the likelihood function of the data vector $\mathbf{y}$:

$$L(\mathbf{Xb}, \mathbf{V}; \mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{V} \mid^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{Xb})^t \mathbf{V}^{-1} (\mathbf{y} - \mathbf{Xb}) \right\},$$

but instead he maximized the joint density of $\mathbf{y}$ and $\mathbf{u}$, $f(\mathbf{y}, \mathbf{u})$, which can be written with the aid of the definition of conditional distribution as

$$f(\mathbf{y}, \mathbf{u}) = f(\mathbf{u}) f(\mathbf{y} \mid \mathbf{u}). \qquad (3.31)$$

Now, since the probability density function $f(\mathbf{u})$ of $\mathbf{u} \sim N_q(\mathbf{0}, \mathbf{D})$ is given by

$$f(\mathbf{u}; \mathbf{0}, \mathbf{D}) = \frac{1}{(2\pi)^{\frac{q}{2}} \mid \mathbf{D} \mid^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \mathbf{u}^t \mathbf{D}^{-1} \mathbf{u} \right\},$$

and the probability density function $f(\mathbf{y} \mid \mathbf{u})$ of $\mathbf{y} \mid \mathbf{u} \sim N_n(\mathbf{Xb} + \mathbf{Zu}, \mathbf{R})$ is given by

$$f(\mathbf{y} \mid \mathbf{u}; \mathbf{Xb} + \mathbf{Zu}, \mathbf{R}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{R} \mid^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{Xb} - \mathbf{Zu}) \right\},$$

then, using (3.31), the function $f\left(\mathbf{y},\mathbf{u}\right)$ being maximized is succinctly written as:

$$
\begin{aligned}
& f\left(\mathbf{y},\mathbf{u}\right) = f\left(\mathbf{u}\right) f\left(\mathbf{y} \mid \mathbf{u}\right) = \\
= \ & \frac{1}{\left(2\pi\right)^{\frac{n+q}{2}}\left(\mid \mathbf{D} \mid\mid \mathbf{R} \mid\right)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[\mathbf{u}^t\mathbf{D}^{-1}\mathbf{u}+\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)^t\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)\right]\right\}.
\end{aligned}
$$

As usual, instead of maximizing $f\left(\mathbf{y},\mathbf{u}\right)$, we equivalently maximize $\ln f\left(\mathbf{y},\mathbf{u}\right)$, given by:

$$
\begin{aligned}
& \ln f\left(\mathbf{y},\mathbf{u}\right) = \\
= \ & -\frac{\left(n+q\right)}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\mid \mathbf{D} \mid\mid \mathbf{R} \mid\right)-\frac{1}{2}\mathbf{u}^t\mathbf{D}^{-1}\mathbf{u}-\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)^t\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Z}
\end{aligned}
$$

In order to obtain the MME, what remains is to take the partial derivatives of $\ln f\left(\mathbf{y},\mathbf{u}\right)$ with respect to $\mathbf{b}$, $\mathbf{u}$ and equate them to zero. For the calculations of the first-order partial derivatives that follow we are making use of the general result of matrix derivation[4] (see, e.g. *Harville, 1997*), stating that:

$$
\frac{\partial\left(\mathbf{x}^t\mathbf{A}\mathbf{x}\right)}{\partial\mathbf{x}} = 2\mathbf{A}\mathbf{x}, \tag{3.32}
$$

where $\mathbf{x}$ a random vector and $\mathbf{A}$ known, symmetric matrix.

Hence, regarding (3.32) and since differentiating the first three parts of $\ln f\left(\mathbf{y},\mathbf{u}\right)$ with respect to $\mathbf{b}$ gives zero, it is:

$$
\begin{aligned}
\frac{\partial \ln f\left(\mathbf{y},\mathbf{u}\right)}{\partial\mathbf{b}} &= -\frac{1}{2}\frac{\partial\left[\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)^t\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)\right]}{\partial\mathbf{b}} \\
&= -\frac{1}{2}\left[-2\mathbf{X}^t\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right)\right] = \mathbf{X}^t\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{Xb}-\mathbf{Zu}\right) \\
&= \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y}-\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}-\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u},
\end{aligned}
$$

[4] For $\mathbf{A}$ a general known matrix, the partial derivative of the quadratic form $\mathbf{x}^t\mathbf{A}\mathbf{x}$ with respect to $\mathbf{x}$ is given in matrix notation by $\partial\left(\mathbf{x}^t\mathbf{A}\mathbf{x}\right)/\partial\mathbf{x} = \left(\mathbf{A}+\mathbf{A}^t\right)\mathbf{x}$. In the special case where $\mathbf{A}$ is symmetric, the above result simplifies to result 3.31.

and equating to zero we get

$$\begin{aligned}&\mathbf{X}^t\mathbf{R}^{-1}\mathbf{y}-\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}-\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}=\mathbf{0}\\ &\Rightarrow \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}+\mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}=\mathbf{X}^t\mathbf{R}^{-1}\mathbf{y}.\end{aligned}$$

This last equation as we easily notice is the first of the mixed model equations of (3.29). To obtain the second mixed model equation we simply have to calculate the partial derivative of $\ln f(\mathbf{y},\mathbf{u})$ with respect to the random vector $\mathbf{u}$, and equate the result to zero. Indeed,

$$\begin{aligned}\frac{\partial \ln f(\mathbf{y},\mathbf{u})}{\partial \mathbf{u}} &= -\frac{1}{2}\frac{\partial\left(\mathbf{u}^t\mathbf{D}^{-1}\mathbf{u}\right)}{\partial \mathbf{u}}-\frac{1}{2}\frac{\partial\left[(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})^t\,\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})\right]}{\partial \mathbf{u}}\\ &= -\frac{1}{2}\left(2\mathbf{D}^{-1}\mathbf{u}\right)-\frac{1}{2}\left[-2\mathbf{Z}^t\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})\right]\\ &= -\mathbf{D}^{-1}\mathbf{u}+\mathbf{Z}^t\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b}-\mathbf{Z}\mathbf{u})\\ &= -\mathbf{D}^{-1}\mathbf{u}+\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y}-\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}-\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}.\end{aligned}$$

Equating now the last term to zero, results to the second MME:

$$\begin{aligned}&\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y}-\mathbf{D}^{-1}\mathbf{u}-\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}-\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}=\mathbf{0}\\ &\Rightarrow \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}+\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{u}+\mathbf{D}^{-1}\mathbf{u}=\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y}\\ &\Rightarrow \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\mathbf{b}+\left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1}\right)\mathbf{u}=\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y}.\end{aligned}$$

Thus far, we have stated the form of *Henderson's* mixed model equations, and presented a method for deriving them. What remains in suspense, is to calculate the estimates $\tilde{\mathbf{b}},\tilde{\mathbf{u}}$ of MME and moreover to verify that indeed these estimates are the $\mathbf{b}_{BLUE}$ and $\mathbf{u}_{BLUP}$, respectively. The above is the subject of the following section.

3.3.3.2 Estimates $\tilde{\mathrm{b}},\tilde{\mathrm{u}}$ of Mixed Model Equations

To obtain the mixed model solution $\tilde{\mathbf{u}}$ of the random vector $\mathbf{u}$, one simply has to take the second equation of (3.29) (the second mixed model equation), and solve with respect to $\tilde{\mathbf{u}}$ as follows:

$$\begin{aligned}
&\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)\tilde{\mathbf{u}} = \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} \\
&\Rightarrow \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)\tilde{\mathbf{u}} = \mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right) \\
&\Rightarrow \tilde{\mathbf{u}} = \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right) \\
&\Rightarrow \tilde{\mathbf{u}} = \mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right),
\end{aligned} \qquad (3.33)$$

where $\mathbf{T} = \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}$. Accordingly, the MM solution $\tilde{\mathbf{b}}$ of $\mathbf{b}$, can be found by replacing in the first mixed model equation the obtained estimate $\tilde{\mathbf{u}}$ of (3.33), and solving for $\tilde{\mathbf{b}}$, as shown below:

$$\begin{aligned}
&\mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{u}} = \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\
&\Rightarrow \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\left[\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right)\right] = \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\
&\Rightarrow \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\left(\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} - \mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}}\right) = \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\
&\Rightarrow \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} + \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\
&\Rightarrow \mathbf{X}^t\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}^t\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{y}.
\end{aligned} \qquad (3.34)$$

Assuming that the inverse of matrix $\mathbf{X}^t\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{X}$ exists, the MM solution is given by:

$$\tilde{\mathbf{b}} = \left[\mathbf{X}^t\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{X}\right]^{-1}\mathbf{X}^t\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{y}.$$

3.3.3.3 Equivalence of MME Estimates and BLUE/BLUP

For the general linear mixed model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$, the BLUE of $\mathbf{b}$ and BLUP of $\mathbf{u}$ were shown to be:

$$\mathbf{b}_{BLUE} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}$$

and

$$\mathbf{u}_{BLUP} = \mathrm{D}\mathrm{Z}^t\mathrm{V}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{BLUE}\right).$$

The mixed model solutions, $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{u}}$ of mixed model equations, however, are given by:

$$\tilde{\mathbf{b}} = \left[\mathbf{X}^t \left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{X}\right]^{-1}\mathbf{X}^t \left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\mathbf{y}$$

and

$$\tilde{\mathbf{u}} = \mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right).$$

Comparing the above equations, we notice that equivalence between the mixed model solution $\tilde{\mathbf{b}}$ and $\mathbf{b}_{BLUE}$ is obtained if the identity:

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \tag{3.35}$$

is true, where $\mathbf{T} = \left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}$. [The proof of this result is due to *Henderson (1963)*, but a similar proof was also given by *Woodbury M.A.* in 1950, in an unpublished paper, though].

To prove that identity (3.35) is true, that is the inverse of variance-covariance matrix $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^t + \mathbf{R}$ is $\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}$, it suffices to show that $\mathbf{V}\mathbf{V}^{-1} = \mathbf{I}$, with $\mathbf{I}$ being the identity matrix. Indeed,

$$\begin{aligned}
\mathbf{V}\mathbf{V}^{-1} &= \left(\mathbf{Z}\mathbf{D}\mathbf{Z}^t + \mathbf{R}\right)\left(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right)\\
&= \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} + \mathbf{R}\mathbf{R}^{-1} - \mathbf{R}\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \underset{\mathbf{RR}^{-1}=\mathbf{I}}{=}\\
&= \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} + \mathbf{I} - \mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\\
&= \mathbf{I} + \left(\mathbf{Z}\mathbf{D}\mathbf{T}^{-1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}\right)\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \underset{replace\ \mathbf{T}^{-1}}{=}\\
&= \mathbf{I} + \left[\mathbf{Z}\mathbf{D}\left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1}\right) - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}\right]\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\\
&= \mathbf{I} + \left(\mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{Z}\mathbf{D}\mathbf{D}^{-1} - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}\right)\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\\
&= \mathbf{I} + \left(\mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{Z} - \mathbf{Z}\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}\right)\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\\
&= \mathbf{I} + 0\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} = \mathbf{I}.
\end{aligned}$$

Now, as concerns the equivalence of mixed model solution $\tilde{\mathbf{u}} = \mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right)$ and the BLUP formula $\mathbf{u}_{BLUP} = \mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}_{BLUE}\right) = \mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\mathbf{b}}\right)$, since

$\mathbf{y}-\mathbf{X}\tilde{\mathbf{b}}$ is common to both formulas, we only have to show that:

$$\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}=\mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1}, \tag{3.36}$$

where $\mathbf{T}=\left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}$. The above equality is true, since:

$$\begin{aligned}
\mathbf{D}\mathbf{Z}^t\mathbf{V}^{-1} \quad & \underset{use\ (3.35)}{=} \mathbf{D}\mathbf{Z}^t\left(\mathbf{R}^{-1}-\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}\right) \\
&= \mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}-\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \\
&= \left(\mathbf{D}\mathbf{T}^{-1}-\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\right)\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \underset{replace\ \mathbf{T}^{-1}}{=} \\
&= \left[\mathbf{D}\left(\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1}\right)-\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\right]\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \\
&= \left(\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}\mathbf{D}^{-1}-\mathbf{D}\mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z}\right)\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1} \\
&= \mathbf{I}\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}=\mathbf{T}\mathbf{Z}^t\mathbf{R}^{-1}.
\end{aligned}$$

3.3.4 Variance Component Estimation

Up to this point, both fixed effects $\mathbf{b}$ and random effects $\mathbf{u}$ estimators of the GLMM (3.17) were obtained upon the assumption that the variance-covariance matrices of the latter model, $\mathbf{D}$, $\mathbf{R}$ and accordingly $\mathbf{V}=\mathbf{Z}\mathbf{D}\mathbf{Z}^t+\mathbf{R}$ are known, (positive definite) matrices. Equivalently, this means that the elements of $\mathbf{D}$ and $\mathbf{R}$ (usually called variance components or variance parameters), are known. In practice however, the assumption of known variance components almost never holds. If by $\boldsymbol{\theta}=(\theta_1,\theta_2,...,\theta_q)^t$ we denote the parameter vector that comprises all the unknown variance components, say q in total, included in $\mathbf{V}=\mathbf{Z}\mathbf{D}\mathbf{Z}^t+\mathbf{R}$, then clearly mixed model analysis involves two, complementary to each other, estimation issues:

(a) The estimation of the vectors of fixed and random effects, $\mathbf{b}$ and $\mathbf{u}$

and

(b) The estimation of the unknown variance components $\boldsymbol{\theta}$, included in $\mathbf{D}$, $\mathbf{R}$.

Thus in consequence, prior to the estimation of fixed and random vectors, $\mathbf{b}$ and $\mathbf{u}$ respectively (e.g. by using ML or BLUE/BLUP theory), the variance components $\boldsymbol{\theta}$ need to be estimated. The general procedure is to obtain estimates of the variance components by one of the usual variance component estimation methods, and then replace the variance parameters of $\mathbf{D}$ and $\mathbf{R}$ with these estimates, in order to proceed with the estimation of fixed and random effects. Consequently, once the variance components are estimated, [let $\hat{\mathbf{D}}$, $\hat{\mathbf{R}}$ and $\hat{\mathbf{V}} = \mathbf{Z}\hat{\mathbf{D}}\mathbf{Z}^t + \hat{\mathbf{R}}$ denote the estimated variance-covariance matrices obtained by replacing their elements (the variance components $\boldsymbol{\theta}$) with its estimations], then the ML (or BLUE, or GLS) estimate of the fixed effects vector $\mathbf{b}$ would be simply given by replacing the estimated $\hat{\mathbf{V}}$ in place of $\mathbf{V}$ in equation (3.23), hence:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathrm{X}^t\hat{\mathrm{V}}^{-1}\mathrm{y}. \tag{3.37}$$

The above described two-stage procedure for the estimation of variance components is one possible course of action. Nowadays however, much of the attention has been focused into methods that provide us with estimations of fixed effects and variance components simultaneously, using one unified procedure.

Estimation of variance components is a very extensive topic, and various estimation methods have developed for the specific subject [notice the difference with the estimation of variance components of models concerning balanced data (ANOVA type models, Chapter 2), which rests almost entirely upon one method, the so-called analysis of variance method, consisting of equating mean squares to their expected values]. A vast literature on the issue of variance component estimation exists. For a comprehensive review on this extensive subject we refer to *Searle et al. (1992)*.

The main methods for estimating the variance components of the GLMM in a single stage, found in the literature are: maximum likelihood (ML) estimation, restricted maximum likelihood (REML) estimation, and minimum norm quadratic unbiased estimation (MINQUE) (*Rao, 1971*), with the first two (close related to each other) methods being the most often met. The basic feature that distinguishes ML/REML and MINQUE method is that the minimum norm quadratic unbiased estimation, unlike ML and REML, has been developed to estimate variance components without relying on any distributional assumptions. On the other hand though, MINQUE imposes some restrictions on the variance-covariance structure of vector $\mathbf{y}$. In what follows we attempt a description of ML and REML methods, not in full extension though, since that we will return on the specific subject of ML and REML estimation of variance components in Chapter 5, where we discuss the implementation of the GLMM on longitudinal data.

3.3.4.1 Maximum Likelihood Estimation of Variance Components

The maximum likelihood method for variance-component estimation was formally introduced by *Hartley* and *Rao (1967)*. [A first, informal suggestion though, of the method has already been discussed in *Crump (1951)*]. In their article, *Hartley* and *Rao* discuss the ML (variance-components) estimation of the generalized GLMM:

$$\mathbf{y} = \mathbf{Xb} + \sum_{i=1}^{k} \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}, \tag{3.38}$$

which contains, as one can observe, more than one random effects term, compared to the GLMM $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, and under this notion is considered to be a generalization of the latter model. In spite of presenting thoroughly the work of *Hartley* and *Rao,* it would be more constructive to sketch the basic steps of variance components estimation by maximum likelihood of the GLMM: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$. To begin with, let us consider the previous GLMM, and additionally, as regard the distributional behavior of the random terms of the model, let us take $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ to be multivariate normal random variables, such that:

$$\begin{aligned} \mathbf{u} &\sim N_q(\mathbf{0}, \mathbf{D}) \\ &\quad \textit{and} \\ \boldsymbol{\varepsilon} &\sim N_n(\mathbf{0}, \mathbf{R}). \end{aligned}$$

Obviously, the variance components that we seek to estimate are embedded within the variance-covariance matrices $\mathbf{D}$, $\mathbf{R}$. Now, to avoid very complicated computations, as well as for presentational purposes it is convenient to assume that $\mathbf{D} =\sigma^2\mathbf{A}$ and $\mathbf{R} =\sigma_e^2\mathbf{I}$, where $\mathbf{A}$ known matrix and $\mathbf{I}$ the identity matrix [in this way we assume that the model contains only two variance components,namely $\boldsymbol{\theta} = (\sigma^2, \sigma_e^2)^t$].

The estimation procedure in general, is very similar to the ML estimation of the fixed-effects vector $\mathbf{b}$. In order to derive the ML estimator $\hat{\mathbf{b}}$ of $\mathbf{b}$, one has to equate to zero the partial derivative with respect to $\mathbf{b}$ of the $(n \times 1)$ response vector $\mathbf{y}$ log-likelihood $[\partial (\ln L) \diagup \partial \mathbf{b} = \mathbf{0}]$. Similarly, to obtain the ML estimates of the variances σ^2, σ_e^2 (say $\hat{\sigma}^2$, $\hat{\sigma}_e^2$), the standard procedure is to solve:

$$\left\{ \begin{array}{c} \frac{\partial(\ln L)}{\partial\sigma^2} = 0 \\ \\ \frac{\partial(\ln L)}{\partial\sigma_e^2} = 0 \end{array} \right\}. \tag{3.39}$$

The likelihood function of the (normaly, multivariate distributed) response vector $\mathbf{y} \sim N_n(\mathbf{Xb}, \mathbf{V})$, is given by:

$$L(\mathbf{Xb}, \mathbf{V}; \mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{V} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{Xb})^t \mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})\right\}, \tag{3.40}$$

and accordingly the log-likelihood λ is:

$$\begin{aligned} \lambda(\mathbf{Xb}, \mathbf{V}; \mathbf{y}) &= \ln L \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln \mid \mathbf{V} \mid -\frac{1}{2}(\mathbf{y} - \mathbf{Xb})^t \mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) \\ &= \text{constant} - \frac{1}{2}\ln \mid \mathbf{V} \mid -\frac{1}{2}(\mathbf{y} - \mathbf{Xb})^t \mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}). \end{aligned} \tag{3.41}$$

In addition, due to the fact that $\mathbf{D} =\sigma^2\mathbf{A}$ and $\mathbf{R} =\sigma_e^2\mathbf{I}$, the variance-covariance matrix $\mathbf{V}$ is written as:

$$\mathbf{V} = \mathrm{ZDZ}^t + \mathrm{R} =\sigma^2\mathrm{ZAZ}^t + \sigma_e^2\mathbf{I}.$$

To start with the $\hat{\sigma}^2$ estimator, we have to calculate $\partial\lambda/\partial\sigma^2$. To this end:

$$\begin{aligned}\frac{\partial\,(\ln L)}{\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\left[-\frac{n}{2}\ln\,(2\pi) - \frac{1}{2}\ln \mid \mathbf{V} \mid -\frac{1}{2}\,(\mathbf{y}-\mathbf{X}\mathbf{b})^t\,\mathbf{V}^{-1}\,(\mathbf{y}-\mathbf{X}\mathbf{b})\right] \\ &= 0 - \frac{1}{2}\frac{\partial \ln \mid \mathbf{V} \mid}{\partial\sigma^2} - \frac{1}{2}\frac{\partial}{\partial\sigma^2}\left[(\mathbf{y}-\mathbf{X}\mathbf{b})^t\,\mathbf{V}^{-1}\,(\mathbf{y}-\mathbf{X}\mathbf{b})\right].\end{aligned}$$

At this point, we can take advantage of the following general results of matrix theory (see, e.g. *Harville, 1997*) concerning first-order partial derivatives of determinants and inverse matrices:

Let $\mathbf{M}$ be a square matrix whose elements are functions of a scalar variable x. Then, if $\mathbf{M}$ is nonsingular and continuously differentiable the following holds:

$$\frac{\partial \mid \mathbf{M} \mid}{\partial x} = \mid \mathbf{M} \mid tr\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\right), \tag{3.42}$$

$$\frac{\partial \mathbf{M}^{-1}}{\partial x} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\mathbf{M}^{-1}. \tag{3.43}$$

In addition to the above results, by applying the well-known chain rule to the function $\ln \mid \mathbf{M} \mid$, and using (3.42), (3.43) we find that if $\mathbf{M}$ is continuously differentiable (at all points in its domain), then $\ln \mid \mathbf{M} \mid$ is continuously differentiable and the following holds:

$$\frac{\partial \ln \mid \mathbf{M} \mid}{\partial x} = tr\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial x}\right). \tag{3.44}$$

Implementing the above results in the current context of ML estimation of σ^2 and σ_e^2, (since the elements of the variance-covariance matrix $\mathbf{V}$ are functions of the variances σ^2, σ_e^2), we have:

$$\frac{\partial \ln \mid \mathbf{V} \mid}{\partial\sigma^2} = tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\sigma^2}\right),$$

as well as

$$\frac{\partial \mathbf{V}^{-1}}{\partial\sigma^2} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial\sigma^2}\mathbf{V}^{-1},$$

and consequently, $\partial\lambda/\partial\sigma^2$ becomes:

$$\begin{aligned}\frac{\partial\lambda}{\partial\sigma^2} &= -\frac{1}{2}tr\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2}\right)-\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\frac{\partial\mathbf{V}^{-1}}{\partial\sigma^2}\left(\mathbf{y}-\mathbf{Xb}\right)\\ &= -\frac{1}{2}tr\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2}\right)+\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\sigma^2}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{Xb}\right),\end{aligned}$$

and by considering that $\partial\mathbf{V}/\partial\sigma^2=\partial\left(\sigma^2\mathbf{ZAZ}^t+\sigma_e^2\mathbf{I}\right)/\partial\sigma^2=\mathbf{ZAZ}^t$, we get:

$$\frac{\partial\lambda}{\partial\sigma^2}=-\frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{ZAZ}^t\right)+\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\mathbf{V}^{-1}\mathbf{ZAZ}^t\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{Xb}\right).$$

As regards now the estimation of the remaining variance component σ_e^2, calculation of $\partial\lambda/\partial\sigma_e^2$ is similar to the above, using though that $\partial\mathbf{V}/\partial\sigma_e^2=\partial\left(\sigma^2\mathbf{ZAZ}^t+\sigma_e^2\mathbf{I}\right)/\partial\sigma_e^2=\mathbf{I}$. Thus, it is:

$$\frac{\partial\lambda}{\partial\sigma_e^2}=-\frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{I}\right)+\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\mathbf{V}^{-1}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{Xb}\right).$$

Having obtained the above formulas for $\partial\left(\ln L\right)/\partial\sigma^2$ and $\partial\left(\ln L\right)/\partial\sigma_e^2$, what remains to derive the ML estimators $\hat{\sigma}^2$, $\hat{\sigma}_e^2$ is to solve system (3.39), which now has become:

$$\left\{\begin{array}{c}-\frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{ZAZ}^t\right)+\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\mathbf{V}^{-1}\mathbf{ZAZ}^t\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{Xb}\right)=0\\ -\frac{1}{2}tr\left(\mathbf{V}^{-1}\mathbf{I}\right)+\frac{1}{2}\left(\mathbf{y}-\mathbf{Xb}\right)^t\mathbf{V}^{-1}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{Xb}\right)=0\end{array}\right\}.$$

But there are certain difficulties we are confronted with, in order to solve the above system; first of all, as one may observe, the two equations contain the (unknown) fixed parameter $\mathbf{b}$. Hence, it is evident that in order to proceed with the estimation of the variance components, it is necessary for $\mathbf{b}$ to be replaced by one of its estimators, e.g. the ML estimator, (which in fact also corresponds to the generalized least squares and best linear unbiased estimator), $\hat{\mathbf{b}}=\left(\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\hat{\mathbf{V}}^{-1}\mathbf{y}$, where $\hat{\mathbf{V}}=\hat{\sigma}^2\mathbf{ZAZ}^t+\hat{\sigma}_e^2\mathbf{I}$.

Trying to summarize all the above, we can state that to acquire estimates for the variance components $(\sigma^2,\ \sigma_e^2)$, one has to proceed solving the system of three equations (known and as ML equations):

$$\left\{ \begin{array}{c} tr\left(\hat{\mathbf{V}}^{-1}\right) = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^{t} \hat{\mathbf{V}}^{-1}\hat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \\ tr\left(\hat{\mathbf{V}}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^{t}\right) = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^{t} \hat{\mathbf{V}}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^{t}\hat{\mathbf{V}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \\ \hat{\mathbf{b}} = \left(\mathbf{X}^{t}\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{t}\hat{\mathbf{V}}^{-1}\mathbf{y} \end{array} \right\}.$$

The motivation behind this system is simple; while the first two equations provide the estimates of the variance components (contained within the variance-covariance matrix $\mathbf{V}$), unfortunately they also include the (unknown) fixed parameter $\mathbf{b}$. Hence, naturally this leads to introducing the third equation into the system, which provides us with the ML estimator of $\mathbf{b}$, $\hat{\mathbf{b}}$. Substituting this estimate in the first two equations enables us to obtain $\hat{\sigma}^2$and $\hat{\sigma}_e^2$. The inherent difficulty that arises is that the ML estimator $\hat{\mathbf{b}}$ is a function of $\hat{\mathbf{V}}$, and thus previously to obtaining $\hat{\mathbf{b}}$, estimators of the variance components are required. Due to this, it is evident that to solve the set of these equations, there is no closed form solution for any of $\mathbf{b}$, σ^2, σ_e^2. As a consequence, there is no simple one-step solution for the above system.

In general, to handle situations like this, where no theoretical solution of the ML equations can be obtained, we have no other alternative but to resort to numerical optimization techniques. By this we refer to algorithms, iterative in nature, developed for solving theoretical problems through numerical analysis. Among a wide variety of iterative numerical procedures for solving ML equations and obtaining maximum likelihood estimates (MLEs) of the variance components, two are the most popular; the *Newton-Raphson* (NR) algorithm and the *Expectation-Maximization* (EM) algorithm [see *Dempster et al. (1977)*]. These procedures, based on some starting values for the parameters, iteratively update the estimates until sufficient convergence has been obtained. Since a detailed description of both algorithms, as well as their implementation on the General Linear Mixed Model (GLMM) for longitudinal data are a basic subject of Chapter 5, we will avoid to pursue here a further discussion on these iterative algorithms.

3.3.4.2 Restricted Maximum Likelihood Estimation of Variance Components

Despite the fact that the method of maximum likelihood (ML) has become one of the most popular techniques for the estimation of variance components, unfortunately there exists a serious drawback that produces a negative effect on the latter estimates. More specific, the ML estimators of variance components fail to take into account the degrees of freedom lost in estimating the fixed effects **b**, and are thus biased (generally, they tend to be downwardly biased).

In simpler words, to perform variance-component ML estimation, all fixed effects are assumed to be known (constants), without error. However, in practice this is rarely true, and consequently estimations of the fixed effects are used instead. This results in losing one degree of freedom each time a fixed effect parameter is estimated. Exactly that loss in the degrees of freedom ML method fails to take into account, producing downward bias in the maximum likelihood estimators of variance components. This results in variance estimates that are generally too small, suggesting more precision than we actually have.

Having recognized the bias problem, much research attention focused in finding ways to tackle it sufficiently. In order to overcome it, *Patterson* and *Thompson (1971)* proposed the restricted (also called residual) maximum likelihood (REML) approach, which is essentially based upon a modification of the well-known maximum likelihood method. It is worth noting though, that the first suggestions of REML go back in early sixties, where initially *Thompson (1962)* introduced the idea of REML for the purpose of obtaining unbiased estimates of variances and avoiding negative estimates of variances, and later *Patterson (1964)* utilized the same ideas in a components-of-variance problem arising in the analysis of rotation experiments. These two papers however did not pursue thoroughly the detailed discussion of REML since it was not of primary concern. The model *Patterson* and *Thompson* considered for the demonstration of their "modified"

ML method was the general linear model (GLM) of the form[5]:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon},$$

where, as usual, $\mathbf{y}$ represents a $(n \times 1)$ vector of responses, $\mathbf{b}$ is a $(p \times 1)$ vector of treatment parameters (the fixed-effects), and finally $\boldsymbol{\varepsilon}$ corresponds to the random vector of order $(n \times 1)$, which is normally distributed with zero mean and variance-covariance matrix $\mathbf{V}$, i.e. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{V})$. Further, they assumed $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{H}$, with $\mathbf{H} = \mathbf{Z}\mathbf{D}\mathbf{Z}^t + \mathbf{I}$. As a result, the response vector $\mathbf{y}$ is also assumed to be normally distributed with $\mathbf{y} \sim N_n(\mathbf{X}\mathbf{b}, \mathbf{V})$. The main task was the estimation of the fixed vector $\mathbf{b}$, as well as the estimation of σ^2 and $\mathbf{D}$. Estimation of variance parameters σ^2 and $\mathbf{D}$ via ML makes use of the likelihood of $\mathbf{y}$, which unfortunately includes $\mathbf{b}$ since $\mathbf{y} \sim N_n(\mathbf{X}\mathbf{b}, \mathbf{V})$, resulting to (downward) biassed variance estimates. What *Patterson* and *Thompson* thought in order to resolve the bias problem was that instead of maximizing the "full" data log-likelihood (as *Hartley* and *Rao* did), to use a modified maximum likelihood procedure, where only the portion of the likelihood that does not depend on the fixed effects vector $\mathbf{b}$ is maximized (this is why Patterson and Thompson found suitable using the terminology "restricted" to name their method). They accomplished that, by partitioning the (full) data, contained in $\mathbf{y}$, into two separate parts, so that one of them to be free of the fixed effects. Maximizing this part yields the variance component estimators, which are called restricted maximum likelihood (REML) estimators.

Hence, seeking for a suitable set of data that would not depend on the fixed effects, Patterson and Thompson considered a linear transformed[6] set of data, say $\mathbf{K}\mathbf{y}$, on the presumption that the latter does not depend on $\mathbf{b}$ any more, as the response vector $\mathbf{y}$ did. [Here, $\mathbf{K}$ is taken to be a square matrix of $(n \times n)$ order, for the product $\mathbf{K}\mathbf{y}$ to be well-defined].

[5] Hartley and Rao (1967) have also considered the same model, in order to demonstrate the method of maximum likelihood for variance component estimation.

[6] Alternatively, is known as error contrast

By demanding the linear combination $\mathbf{Ky}$ to be independent of the fixed-effects, in essence we require the distribution of the (random vector) $\mathbf{Ky}$ to be free of $\mathbf{b}$. To this end, matrix $\mathbf{K}$ should be chosen such that the mean of $\mathbf{Ky}$ will not include $\mathbf{b}$, as was the case with vector $\mathbf{y}$. Namely:

$$E\left(\mathbf{Ky}\right) = \mathbf{0}. \tag{3.45}$$

Equivalently, this yields:

$$E\left(\mathbf{Ky}\right) = \mathbf{0} \Leftrightarrow \mathbf{K}E\left(\mathbf{y}\right) = \mathbf{0} \Leftrightarrow \mathbf{KXb} = \mathbf{0} \Leftrightarrow \mathbf{KX} = \mathbf{0}, \tag{3.46}$$

since by definition, $\mathbf{b}$ is always different from zero. The question now, is to determine a suitable matrix $\mathbf{K}$, that shares property (3.45). Patterson and Thompson's proposal was to take $\mathbf{K}$ be of the following form:

$$\mathbf{K} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t. \tag{3.47}$$

Indeed, it is easy to show that the specific choice of $\mathbf{K}$ verifies (3.45), since:

$$\mathbf{Ky} = \left\{\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right\}\mathbf{y} = \mathbf{y} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y},$$

and thus,

$$\begin{aligned} E\left(\mathbf{Ky}\right) &= E\left(\mathbf{y}\right) - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t E\left(\mathbf{y}\right) \\ &= \mathbf{Xb} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Xb} \\ &= \mathbf{Xb} - \mathbf{XIb} = \mathbf{Xb} - \mathbf{Xb} = \mathbf{0}. \end{aligned}$$

Harville (1977) utilized the ideas of *Patterson* and *Thompson (1971)* to extend REML approach for estimation of variance components in the more general mixed-effects model (GLMM): $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$. He also described justifications for the REML estimation from both Bayesian and Frequentist theory points. Furthermore, Harville claimed, through a sufficient argument, that REML approach to variance components estimation loses no information and thus REML estimates are efficient in the same sense as are ML estimators. Papers by *Harville (1977)* and *Corbeil* and *Searle (1976)* discuss in detail REML estimation as well as desirable properties of REML estimates.

Let us see now, how the REML likelihood function of the GLMM: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon}$, $[\mathbf{u} \sim N_q(\mathbf{0}, \mathbf{D}), \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \mathbf{R})]$ is derived. Essentially, the latter likelihood is the likelihood function of the already defined error contrast $\mathbf{Ky}$ (which distribution does not depend on the fixed effects anymore). We denote this likelihood function of $\mathbf{Ky}$ by L_{REML} to avoid confusion with the standard ML function, L. Optimization of this (restricted) maximum likelihood function yields the REML estimates of the variance components. We have already seen that the mean of vector $\mathbf{Ky}$ is zero, independent of $\mathbf{b}$. Moreover, its variance-covariance matrix is:

$$Var(\mathbf{Ky}) = \mathbf{K}Var(\mathbf{y})\,\mathbf{K}^t = \mathbf{KVK}^t,$$

Thus, by imposing once again a $n-$variate Normal distribution for the random vector $\mathbf{Ky}$, we may write:

$$\mathbf{Ky} \sim N_n\left(\mathbf{0}, \mathbf{KVK}^t\right), \tag{3.48}$$

and consequently, the corresponding likelihood function of $\mathbf{Ky}$ will be:

$$L_{REML} = (2\pi)^{-\frac{n-r(\mathbf{X})}{2}} \mid \mathbf{KVK}^t \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Ky})^t \left(\mathbf{KVK}^t\right)^{-1}(\mathbf{Ky})\right\}, \tag{3.49}$$

where the matrix $\mathbf{K}$ has rank equal to $n - r(\mathbf{X})$ $[r(\mathbf{X})$ is the rank of $\mathbf{X}]$. The log-likelihood of L_{REML}, denoted by λ_{REML}, is then given by:

$$\begin{aligned}\lambda_{REML} &= \ln L_{REML} \\ &= -\frac{1}{2}[n - r(\mathbf{X})]\ln(2\pi) - \frac{1}{2}\ln \mid \mathbf{KVK}^t \mid -\frac{1}{2}\mathbf{y}^t\mathbf{K}^t\left(\mathbf{KVK}^t\right)^{-1}\mathbf{Ky}\end{aligned}$$

and since the term $-1/2\,[n - r\,(\mathbf{X})]\ln(2\pi)$ is just a constant, we have:

$$\lambda_{REML} = const. - \frac{1}{2}\ln \mid \mathbf{KVK}^t \mid - \frac{1}{2}\mathbf{y}^t\mathbf{K}^t\left(\mathbf{KVK}^t\right)^{-1}\mathbf{Ky}. \tag{3.50}$$

Using now the two important results (*Searle, 1979*):

$$\ln \mid \mathbf{KVK}^t \mid = \ln \mid \mathbf{V} \mid + \ln \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid, \tag{3.51}$$

$$\mathbf{y}^t\mathbf{K}^t\left(\mathbf{KVK}^t\right)^{-1}\mathbf{Ky} = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right), \tag{3.52}$$

that hold for any $\mathbf{K}$ as long as $\mathbf{KX} = \mathbf{0}$ and $r\,(\mathbf{K}) = n - r\,(\mathbf{X})$, log-likelihood function λ_{REML} can be rewritten as:

$$\begin{aligned}
\lambda_{REML} &= \\
&= const. - \frac{1}{2}\left(\ln \mid \mathbf{V} \mid + \ln \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid\right) - \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \\
&= const. - \frac{1}{2}\ln \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid - \frac{1}{2}\ln \mid \mathbf{V} \mid - \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \underset{(3.41)}{=} \\
&= const. - \frac{1}{2}\ln \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid + \lambda\left(\mathbf{X}\hat{\mathbf{b}}, \mathbf{V}; \mathbf{y}\right), \qquad (3.53)
\end{aligned}$$

where $\hat{\mathbf{b}}$ denotes the (ML) estimator of $\mathbf{b}$, $\hat{\mathbf{b}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y}$. Observing (3.53), it is evident that the difference between the REML log-likelihood λ_{REML} and the ordinary ML log-likelihood λ (given in 3.41) is caused by the extra term:

$$-\frac{1}{2}\ln \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid.$$

Another interesting point arising from (3.53), is that λ_{REML} does not contain $\mathbf{b}$ but instead uses its estimate $\hat{\mathbf{b}}$. In this way, the REML approach takes into account that $\mathbf{b}$ is a parameter and not a constant (as was the case with maximum likelihood estimation), and thus the resulting variance component estimates are unbiased.

The appealing feature of REML estimation, is that formulation of REML log-likelihood

λ_{REML} does not really require matrix $\mathbf{K}$ to be of the specific form (3.47) in order to produce, via its optimization, the REML estimates of variance components. In fact, it can be shown that any set of linearly independent error contrasts may be used in place of the particular error contrast $\mathbf{Ky}$, where $\mathbf{K} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$, as long as their expectations are zero. Thus, the REML estimators of variance components are invariant to the choice of the specific error contrast. To prove the above claim, (see, e.g. *Diggle et al., 1994*), let us consider $\mathbf{B}$ to be a $n \times (n-p)$ matrix satisfying the following restrictions:

$$\mathbf{B}^t\mathbf{B} = \mathbf{I}, \tag{3.54}$$

$$\mathbf{B}\mathbf{B}^t = \mathbf{K}. \tag{3.55}$$

The maximum likelihood estimator of the fixed effects, $\hat{\mathbf{b}}$ is given as we have already showed (assuming the variance components to be known) by:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y} = \mathbf{Gy},$$

where $\mathbf{G} = \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}$. The probability density function (p.d.f.) of this estimator, since $\hat{\mathbf{b}} \sim N_p\left[\mathbf{b}, \left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\right]$, is expressed as:

$$f\left(\hat{\mathrm{b}}\right) = (2\pi)^{-\frac{p}{2}} \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\hat{\mathrm{b}} - \mathrm{b}\right)^t \left(\mathrm{X}^t\mathrm{V}^{-1}\mathrm{X}\right)\left(\hat{\mathrm{b}} - \mathrm{b}\right)\right\}. \tag{3.56}$$

Further, as concern the p.d.f. of the response vector $\mathbf{y} \sim N_n\left(\mathbf{Xb}, \mathbf{V}\right)$, is (as we have already showed) given by:

$$f\left(\mathrm{y}\right) = (2\pi)^{-\frac{n}{2}} \mid \mathrm{V} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \mathbf{Xb}\right)^t \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{Xb}\right)\right\}. \tag{3.57}$$

If we consider now the linear combination of $\mathbf{y}$, formed this time by multiplying the (rather arbitrary) $[(n-p) \times n]$ matrix $\mathbf{B}^t$ with $\mathbf{y}$, we are able to show that the following result holds:

$$E\left(\mathbf{B}^t\mathbf{y}\right) = \mathbf{0}.$$

Indeed, by observing that:

$$E\left(\mathbf{B}^t\mathbf{y}\right) = \mathbf{B}^t E\left(\mathbf{y}\right) = \mathbf{B}^t\mathbf{X}\mathbf{b} \underset{\mathbf{B}^t\mathbf{B}=\mathbf{I}}{=} \mathbf{B}^t\mathbf{B}\mathbf{B}^t\mathbf{X}\mathbf{b} \underset{\mathbf{B}\mathbf{B}^t=\mathbf{K}}{=} \mathbf{B}^t\mathbf{K}\mathbf{X}\mathbf{b},$$

and since:

$$\mathbf{K}\mathbf{X} = \left\{\mathbf{I} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right\}\mathbf{X} = \mathbf{X} - \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I} = \mathbf{X} - \mathbf{X} = \mathbf{0},$$

it becomes evident that the mean of the error contrast $\mathbf{B}^t\mathbf{y}$ is zero, without doubt. Moreover, it can be shown that $\mathbf{B}^t\mathbf{y}$ and the ML estimator $\hat{\mathbf{b}}$ are independent, since:

$$\begin{aligned} Cov\left(\mathbf{B}^t\mathbf{y}, \hat{\mathbf{b}}\right) &= E\left[\left(\mathbf{B}^t\mathbf{y} - \mathbf{0}\right)\left(\hat{\mathbf{b}} - \mathbf{b}\right)^t\right] = E\left[\mathbf{B}^t\mathbf{y}\left(\hat{\mathbf{b}}^t - \mathbf{b}^t\right)\right] \underset{\hat{\mathbf{b}}=\mathbf{G}\mathbf{y}}{=} \\ &= E\left[\mathbf{B}^t\mathbf{y}\left(\mathbf{y}^t\mathbf{G}^t - \mathbf{b}^t\right)\right] = E\left(\mathbf{B}^t\mathbf{y}\mathbf{y}^t\mathbf{G}^t - \mathbf{B}^t\mathbf{y}\mathbf{b}^t\right) \\ &= \mathbf{B}^t E\left(\mathbf{y}\mathbf{y}^t\right)\mathbf{G}^t - \mathbf{B}^t E\left(\mathbf{y}\right)\mathbf{b}^t, \end{aligned}$$

and using the well-known result of Multivariate Analysis:

$$Var\left(\mathbf{y}\right) = E\left(\mathbf{y}\mathbf{y}^t\right) - E\left(\mathbf{y}\right)E\left(\mathbf{y}\right)^t,$$

we have:

$$\begin{aligned} Cov\left(\mathbf{B}^t\mathbf{y}, \hat{\mathbf{b}}\right) &= \mathbf{B}^t\left[Var\left(\mathbf{y}\right) + E\left(\mathbf{y}\right)E\left(\mathbf{y}\right)^t\right]\mathbf{G}^t - \mathbf{B}^t E\left(\mathbf{y}\right)\mathbf{b}^t \\ &= \mathbf{B}^t\left(\mathbf{V} + \mathbf{X}\mathbf{b}\mathbf{b}^t\mathbf{X}^t\right)\mathbf{G}^t - \mathbf{B}^t\mathbf{X}\mathbf{b}\mathbf{b}^t \\ &= \mathbf{B}^t\mathbf{V}\mathbf{G}^t + \mathbf{B}^t\mathbf{X}\mathbf{b}\mathbf{b}^t\mathbf{X}^t\mathbf{G}^t - \mathbf{B}^t\mathbf{X}\mathbf{b}\mathbf{b}^t. \end{aligned} \tag{3.58}$$

The prove that the latter is equal to zero, we only have to show that each term equals to zero. Indeed, it is:

$$\mathbf{X}^t\mathbf{G}^t = \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} = \mathbf{I},$$

and

$$\begin{aligned}\mathbf{B}^t\mathbf{V}\mathbf{G}^t &= \mathbf{B}^t\mathbf{V}\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \\ &= \mathbf{B}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \underset{\mathbf{B}^t\mathbf{B}=\mathbf{I}}{=} \mathbf{B}^t\mathbf{B}\mathbf{B}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \underset{\mathbf{B}\mathbf{B}^t=\mathbf{K}}{=} \\ &= \mathbf{B}^t\mathbf{K}\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \underset{\mathbf{K}\mathbf{X}=0}{=} 0.\end{aligned}$$

Considering the above results, (3.58) becomes:

$$Cov\left(\mathbf{B}^t\mathbf{y}, \hat{\mathbf{b}}\right) = 0 + \mathbf{B}^t\mathbf{X}\mathbf{b}\mathbf{b}^t\mathbf{I} - \mathbf{B}^t\mathbf{X}\mathbf{b}\mathbf{b}^t = \mathbf{0}. \tag{3.59}$$

Now, due to the independence [since $Cov\left(\mathbf{B}^t\mathbf{y}, \hat{\mathbf{b}}\right) = 0$] of $\mathbf{B}^t\mathbf{y}$ and $\hat{\mathbf{b}}$, the p.d.f. of response vector $\mathbf{y}$ that jointly accounts for fixed effects $\mathbf{b}$ and variance components $\boldsymbol{\theta}$ can be expressed as the product of the (independent to each other) probability density functions of $\hat{\mathbf{b}}$ and $\mathbf{B}^t\mathbf{y}$, i.e.:

$$f\left(\mathbf{y}; \mathbf{b}, \boldsymbol{\theta}\right) \propto f\left(\hat{\mathbf{b}}; \mathbf{b}\right) \times f\left(\mathbf{B}^t\mathbf{y}; \boldsymbol{\theta}\right). \tag{3.60}$$

Therefore, p.d.f. of $\mathbf{B}^t\mathbf{y}$ is proportional to the following ratio:

$$f\left(\mathbf{B}^t\mathbf{y}\right) \propto \frac{f\left(\mathbf{y}\right)}{f\left(\hat{\mathbf{b}}\right)}, \tag{3.61}$$

and considering the expressions for $f\left(\hat{\mathbf{b}}\right)$, $f\left(\mathbf{y}\right)$ given in (3.56) and (3.57), respectively, the above becomes:

$$f\left(\mathbf{B}^t\mathbf{y}\right) \propto \frac{(2\pi)^{-\frac{n}{2}} \mid \mathbf{V} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)^t \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)\right\}}{(2\pi)^{-\frac{p}{2}} \mid \mathbf{X}^t\mathbf{V}^{-1}\mathbf{X} \mid^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\hat{\mathbf{b}} - \mathbf{b}\right)^t\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)\left(\hat{\mathbf{b}} - \mathbf{b}\right)\right\}}.$$

We can make use now of the following standard result of the general linear model (GLM):

$$(\mathbf{y}-\mathbf{X}\mathbf{b})^t\,\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})=\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)+\left(\hat{\mathbf{b}}-\mathbf{b}\right)^t\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)\left(\hat{\mathbf{b}}-\mathbf{b}\right),$$

that enables us to re-express $f\left(\mathbf{B}^t\mathbf{y}\right)$ as:

$$f\left(\mathbf{B}^t\mathbf{y}\right)\propto(2\pi)^{-\frac{n-p}{2}}\mid\mathbf{V}\mid^{-\frac{1}{2}}\mid\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\mid^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)\right\}. \tag{3.62}$$

The above density function of linear transformation of vector $\mathbf{y}$, $\mathbf{B}^t\mathbf{y}$ also corresponds to the likelihood function of $\mathbf{B}^t\mathbf{y}$, thus consequently the log-likelihood function of $\mathbf{B}^t\mathbf{y}$ simply results by taking the logarithm of $f\left(\mathbf{B}^t\mathbf{y}\right)$, which yields:

$$\ln f\left(\mathbf{B}^t\mathbf{y}\right)\propto const.-\frac{1}{2}\ln\mid\mathbf{V}\mid-\frac{1}{2}\ln\mid\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\mid-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right)^t\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}\right).$$

Hence, the log-likelihood of $\mathbf{B}^t\mathbf{y}$ is essentially identical to the log-likelihood λ_{REML} in (3.53). Therefore, practically any error contrast $\mathbf{B}^t\mathbf{y}$ (where $\mathbf{B}^t$ a rather arbitrary matrix, compared to the particular form of matrix $\mathbf{K}$) can be adequately used in place of error contrast $\mathbf{Ky}$ to derive REML estimators of variance components.

As in the case of (full) Maximum Likelihood estimation, the REML estimates for the fixed effects and the variance components are obtained by maximizing λ_{REML} given in equation (3.53) with respect to $\mathbf{b}$ and the variance components simultaneously, due to the no-closed form problem already discussed. Once again, numerical iterative techniques must be employed to determine the REML estimates for the variance parameters as well as the ML estimates of the fixed effects. One therefore resorts to general type algorithms, such as the Expectation-Maximization (EM) and Newton-Raphson (NR) algorithms, which have proved to be the most popular among various iterative algorithms.

Chapter 4

Exploratory Longitudinal Data Analysis

4.1 Graphical Illustration of Longitudinal Data

In longitudinal data analysis, as in every other statistical method, we distinguish in general two basic, mutually connected to each other, components: *exploratory* and *confirmatory* analysis. Exploratory analysis is essentially a graphical display analysis, serving the purpose of visualizing patterns in the data. Confirmatory analysis is a model-based analysis, drawing inferences on the data by testing statistical hypotheses, thus is considered the formal component of the analysis. In this Chapter, we will present graphical plots and methods which have been developed specifically for displaying longitudinal data, in an effort to make the preliminary graphical investigation (exploratory analysis) as informative as possible.

To begin with, we have to note that illustrating longitudinal data is a much more complicated procedure compared to the plotting of classical univariate data. Hence, the usage of univariate displays used to examine separate variables, such as stem-and-leaf plots, quantile plots, histograms, box plots etc., is very limited in exploring longitudinal data since they give no (or very little) information about the relationships between measurements of different subjects.

Furthermore, longitudinal data differ from the classical multivariate data, as it was previously discussed, since the former consist of repeated measurements on a single variable, while the latter compares the relation of a series of random variables. This departure between longitudinal and other multivariate data establishes not only dissimilar statistical methods for the analysis of these data, but also the graphical presentation of these data has apparent differences. In fact, we will see that longitudinal data share the advantage of providing interesting graphical ways of data representation, compared to the general multivariate data; they are a special form of multivariate data that make multivariate graphics both practical and useful. They can be displayed in graphical plots which are easily interpretable and most important, indicative of basic features of the data structure.

This last property of longitudinal data plots, assigns them a more significant role than just as a tool for visualizing the data. The major features of longitudinal data are the mean of y_i over time, and also the covariance structure of y_i. Graphical plots, such as the parallel axis plot and the Draftman's display plot that are presented below, can provide great assistance at understanding these features of the data.

Consider for example the assumption of correlation between the observations within the ith unit. This assumption requires the definition of a specific covariance structure for the vector $\mathbf{y}_i = (y_{i1,}y_{i2}, \ldots.., y_{in_i})^t$. This feature, along with other features of longitudinal data can be explored at a preliminary stage by using the previously mentioned plots, and this type of preliminary analysis where we use information based on graphical output rather than formal statistical modeling analysis is usually known as *exploratory (longitudinal) data analysis.*

The convenience in using and interpretate such plots is of great importance and lies in the fact that these plots are constructed without fitting any model, and consequently provide important assistance in specifying features of the best-suited model for the data, in an a priori way, without having to fit a large variety of models each of them with a different structure, and decide a posteriori through the statistical analysis which is the best among them. In this way, we are avoiding a large amount of numerical calculations and time, that could occur when fitting a possibly misspecified or overparameterized model.

The question that naturally arises, is in what way exactly these graphical techniques assist in determining either a suitable covariance structure of the within-subject observations or other useful feature of the data, without using any formal statistical analysis based on statistical modeling;

The general idea of the whole procedure used to incorporate in the data analysis the information found in the graphical output is simple and it finds implementation not just in the modelling of longitudinal data, but also in various fields of data analysis and modelling. In words, following *Weiss (1997)*, the idea is as follows; initially, we plot the data e.g. by using the parallel plot. From the plot we identify any feasible structure existing in the data. Afterwards, we fit the data using a model that incorporates the structure we identified in the plots. Now, that we have fitted a specific model, the advantage of using again graphics, this time not of the 'raw' data, but plots based on the residuals obtained from the fitted model is important. At this stage two alternatives may occur: either an additional structure can be identified and can be used in building another model, or no other structure is identifiable. We will demonstrate later how this general procedure works for each of the presented plots, separately.

In the next sections we present two of the most common plots used for visualizing and exploring longitudinal data, along with the basic tools that each one of them provides in the graphical analysis of longitudinal data.

4.2 The 'Parallel Axis' Plot

One of the basic and most common ways to visualize longitudinal data is through the 'parallel axis' plot or 'profile' plot. A parallel plot can be constructed by plotting each repeated measurement y_{ij} (the jth measurement on the ith unit), against t_{ij}, the corresponding time that measurement was taken, and then connecting the points at times

$t_{i(j-1)}$ and t_{ij}.

Specifically, suppose we have n_i repeated measurements on subject i. These measurements, in vector notation, are: $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots.., y_{in_i})^t$. Plotting this n_i-dimensional vector in a parallel axis graph, against the corresponding time points consists of drawing lines from the values y_{ij} to the values $y_{i(j+1)}$. This means that we connect only two consecutive points (measurements) in time, and not between other times. So, instead of just drawing a simple scatterplot of each subjects' points, we connect them by a solid line, gaining the advantage of identifying the subjects' pattern much easier, through the inspection of a continuous shape that reveals the whole pattern of changes of individual observations, rather than separate points.

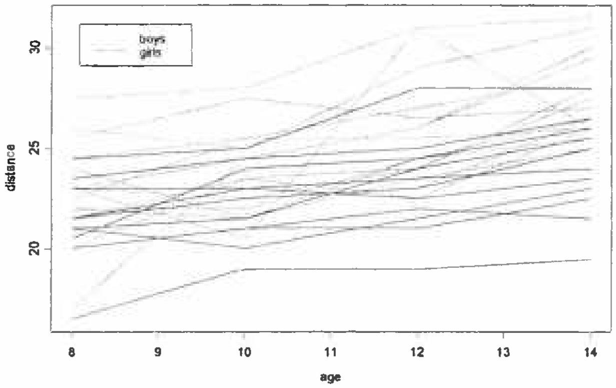

Figure 4.1 : *Parallel plot of orthodontic distance growth for boys and girls*

In Figure 4.1, we have a typical example of data plotted in a parallel plot. The data shown on this plot are a classic longitudinal data set, first presented by *Potthoff* and *Roy* (1964), and come from an orthodontic study of 16 boys and 11 girls between the ages of 8 and 14 years. The response variable is the distance (in millimeters) between the pituitary and the pterygomaxillary fissure, which was measured at 8,10,12 and 14 years for each boy and girl. The 4 measurements on each subject i (that is the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})^t$), can be identified as solid (broken) lines connecting only the consecutive measurements of each individual.

The numerus lines shown in a parallel plot, as the one in Figure 4.1, have the effect of causing various problems. Most common problem that appears in such graphical plots is that in cases of large number of individuals, the plot tends to become dense, with numerus lines covering one another and consequently difficult to interpretate. This undesirable feature is also known as '*overplotting*'(exactly this inconvenience of parallel plots has given them the alternative name of 'spaghetti' plots).

The parallel plot is a basic plotting technique for longitudinal data, mainly because it is a plot that takes advantage of the special structure of such data. By speaking about special structure, we refer to the difference of longitudinal data in comparison to multivariate data, since for the former it is feasible to compare two different measurements $y_{ij}, y_{ij'}$ of vector y_i, a comparison that in the general multivariate data case is between measurements of different variables, hence becomes unreasonable and inappropriate.

Later in this section, we are going to concentrate our attention in demonstrating how effective the parallel plot is, not just as a convenient way of viewing longitudinal data, but most importantly, we will show the usefulness of these plots in determining the structure of the data and furthermore how well a particular model fits the data. Before attempting on this though, it would be useful to give some insight into the particular interesting issue of the origin of the parallel axis plot and its introduction in statistics as an exploratory plotting technique.

The first references to the parallel axis plot, as a tool for visualizing statistical multivariate data in general, can be found in the articles of *Inselberg* and *Wegman,* referring to it as the 'parallel coordinate' plot. Parallel coordinates were originally proposed by *Inselberg (1985)* as a new way to represent multidimensional information. While Inselberg viewed parallel coordinate plot mainly as a device for computational geometry, *Wegman (1990)* focuses the attention on statistical basis, explaining the similarities of the parallel coordinate plot with the conventional scatterplot, and furthermore demonstrating how simple mathematical properties of parallel plot can be utilized in order to derive useful statistical interpretations of multivariate data plotted with the aid of such plots.

The necessity that urged statisticians to employ the parallel coordinate plot as a (multidimensional) multivariate visualization technique was in fact caused by the insufficiency of the classic scatterplot (or scatter diagram) beyond the three dimensions. Parallel (coordinate) plot can be considered as a plotting device for displaying points in high-dimensional spaces, in particular for dimensions above three. As such it is a graphical alternative to the scatterplot, in those situations where the latter does not work well. Visualization of data through a scatterplot has become one of the best and most common ways to look for relationships and patterns among variables. It is simple to understand, yet it conveys much information about the data. Unfortunately, its usefulness proves significant only in two-dimensional planes, where the values of two variables (*e.g.* X, Y) are plotted in an orthogonal Cartesian coordinate system, with the two axes corresponding to the two variables. Essentially, this is done by plotting each two-dimensional point (x,y) of pairs of measurements on the two variables X, Y. The generalization of this plotting technique, to d-dimensions (even for just d=3, the three axes Cartesian coordinate system in space), makes it problematic as an exploratory tool for detecting existing patterns by visual inspection, hence the need for alternative methods for displaying multidimensional data was apparent. Among these alternative methods one can find and the parallel coordinate plot. Parallel plots allow us to visualize points of three or higher dimension better than Cartesian scatterplots.

The main difference, is that instead of the scatterplot which tries to preserve *orthogonality* of the d-dimensional coordinate axes, now the axes are drawn *parallel* to each other. For example, a d-dimensional vector point $\mathbf{y} = (y_1, y_2, \ldots.., y_d)^t$ is plotted by plotting y_1 on axis 1, y_2 on axis 2 and so on through y_d on axis d. The points plotted in this manner are joined by a broken line. Thus, in simple words, what is actually done is a geometrical coordinate transformation from the standard Cartesian system to the introduced parallel system. By this way, the advantage of being able to view d-dimensional data ($d \geqslant 3$) using a two-dimensional system is gained.

Such a transformation from Cartesian to parallel coordinates is not a simple task, and cannot be easily defined in the framework of the standard Euclidean geometry. An alternative way to establish mathematically this transformation, is provided by a non-euclidean geometry, *projective geometry*[1]. Thus the transformation from a Cartesian coordinate system to a parallel coordinate system is accomplished by using basic theory of projective geometry, and consequently is a projective transformation. In the following section we explain, not in full detail since this goes us far beyond the statistical content, how such a projective transformation is possible.

4.2.1 Transforming Cartesian to Parallel Coordinates

In standard euclidean geometry it can be proved that two lines in a 2-dimensional plane determine a point if the lines intersect, but do not determine a point if the lines are parallel. This does not hold for projective geometry, where we define that even parallel lines intersect, with the intersection point being at infinity. As a consequence, a new point (intersection point of parallel lines) is added to the euclidean plane for each set of parallel lines. These new points are called *ideal points* while the original euclidean points are called *ordinary points*.

Moving a little further, we are going to present the basic definitions and axioms of projective geometry that assist in the definition and comprehension of the projective transformation from Cartesian coordinates to parallel coordinates. Among them is the fundamental definition of the *projective plane* that follows immediately.

Definition 4.1: *Consider a set* $\mathbb{P}\neq\emptyset$*, where its elements are called points, a set* $L\neq\emptyset$*, where its elements are called lines and finally consider* I *to be a relation between* $\mathbb{P}$ *and* L*,* $(I \subset \mathbb{P}\times L)$*, which we call incidence.*

A projective plane is a triplet $(\mathbb{P},L,I)$ *satisfying the following axioms:*

- *for every points* $P,Q\in\mathbb{P}$ *with* $P\neq Q$*, there exists exactly one line* $\ell\in L$ *such that:* $(P,\ell)\in I$*,* $(Q,\ell)\in I$*.*

[1] Projective geometry is a non-euclidean geometry since that, in contrast to euclidean geometry, parallel lines are not defined by its theory.

- *for every lines $k,\ell \in L$ with $k \neq \ell$, there exists exactly one point $P \in \mathcal{P}$ such that $(P,k) \in I$, $(P,\ell) \in I$.*

- *there exist at least 4 distinct points with every 3 of them to be non-colinear.*

Remark 4.1: *From the second axiom of the previous definition it follows that two different lines have always a common point (intersection point). Thus, in the projective plane, parallelism between lines is not defined.*

The following are derived directly from the definition of the projective plane, and constitute the basic axioms of projective geometry:

1) There exists at least one line

2) On each line there are at least three points

3) Not all points lie on the same line

4) Two distinct points lie on one and only line

5) Two distinct lines meet in one and only one point

6) Through each point there exist at least three lines

7) There is at least one-to-one correspondence between the real numbers and all but one point on a line

Observing 2 and 6 axioms, we see that one derives from the other if we interchange the notions point and line. In fact it can been proved that this is a general situation in projective geometry, since any statement about points and lines is true with the words points and lines interchanged. This general situation characterizing projective geometry is known as the *duality principle,* and is a factor playing a significant role in defining the transformation from Cartesian to parallel coordinates, as we are going to see briefly.

The idea is to consider both Cartesian and parallel planes to be projective planes. As a consequence, the transformation from the Cartesian coordinate projective plane into the parallel coordinate projective plane is thus a transformation from one projective plane to another (projective transformation). For convenience, we describe the transformation for $d = 2$, that is for the two-dimensional plane. Consider the two-dimensional point

(x_1, x_2) which determines a point in an orthogonal Cartesian coordinate system. In a parallel axis system that same point is represented by drawing a straight line between the value x_1 on x_1-axis and x_2 on x_2-axis. This suggests the interesting duality that points in the Cartesian plane map into lines in the parallel plane.

Such a point-line duality between (projective) planes is made possible only in the framework of projective geometry, due to the previously mentioned duality principle. In projective geometry, according to the duality principle, every conclusion concerning a projective plane is true also for its *dual plane*, that is the plane that arises by interchanging points and lines. More formally, a dual projective plane is defined via the following proposition as follows.

Proposition 4.1: *Consider (Ꝑ,L,I) to be a projective plane, and also consider the triplet (Ꝑ*,L*,I*) where Ꝑ* = L, L* = Ꝑ and I* is defined such that for (P*,ℓ*)∈Ꝑ*×L* will be (P*,ℓ*)∈ I* if-f (P,ℓ)∈ I. Then (Ꝑ*,L*,I*) is also a projective plane, and is called the dual projective plane of (Ꝑ,L,I).*

To demonstrate the framework of this transformation, we describe the procedure for the two-dimensional case. The task thus is to show how the transformation from an orthogonal Cartesian xy coordinate system (scatterplot) to a parallel coordinate system is possible, and furthermore to ensure that this transformation preserves the structures seen in a scatterplot, hence allowing us to use it in the same manner to the former, for visual detection and exploration of patterns in (multivariate) data.

In ordinary Euclidean space we have the cartesian coordinates, where each two-dimensional point is presented by (x, y). Since the transformation will be defined within the framework of projective geometry, it is necessary to present the (analog to the cartesian) projective coordinates called *natural homogeneous coordinates*. For this purpose, consider the parallel lines (ε_1), (ε_2) presented by the equations:

$$(\varepsilon_1): \ Ax + By + C = 0$$

$$(\varepsilon_2): Ax + By + C^{'} = 0$$

Trying to solve these equations simultaneously no solution is obtained. However, in the projective plane the solution of the two equations is the ideal point (intersection point of two parallel lines). Indeed, observe that we can rewrite the previous equations as:

$$(\varepsilon_1) \quad : \quad Ax + By + Cz = 0$$

$$(\varepsilon_2) \quad : \quad Ax + By + C^{'}z = 0$$

The representation of points in the projective plane is done, using the natural homogeneous coordinates, by triples (x, y, z). Now, solving both equations for the common part $Ax + By$ we have: $-Cz+ C^{'}z = 0 \Rightarrow \left(C - C^{'}\right) z = 0$ and since $C \neq C^{'}$ (ε_1, ε_2: parallel lines)$\Rightarrow$ $z = 0$. Therefore the point $(x, y, 0)$ represents an ideal point. Also, for the presentation of ordinary points we want $z = 1$ so that the ordinary equation $Ax+By+C = 0$ holds. Thus the cartesian point (x, y) which corresponds to an ordinary point in natural homogeneous coordinates is represented[2] as $(x, y, 1)$.

As a final assistance in defining the (projective) transformation from cartesian to parallel coordinates, the following proposition is necessary.

Proposition 4.2: *A projective transformation has an analytic representation as* $\mathbf{X} = \mathbf{AX}^{'}$, *where* $\mathbf{A}$ *is a nonsingular matrix and* $\mathbf{X}$, $\mathbf{X}^{'}$ *points on two projective planes on which the specific transformation is applied.*

At last, we are in a position now to define the projective transformation from cartesian coordinate planes to parallel coordinate planes (*Wegman, 1990).* Consider a line L in the cartesian 2-dimensional plane, given by $L : y = mx + b$. Then consider two points lying on that line, say $(a, ma + b)$ and $(c, mc + b)$. For simplicity consider the *tu*-orthogonal

[2] Notice that if $Ax + By + C = 0$, then also $Apx + Bpy + Cp = 0$ hence (px, py, p) is also a valid representation of point (x, y).

cartesian axes mapped into the xy-parallel axes (see Figure 4.2).

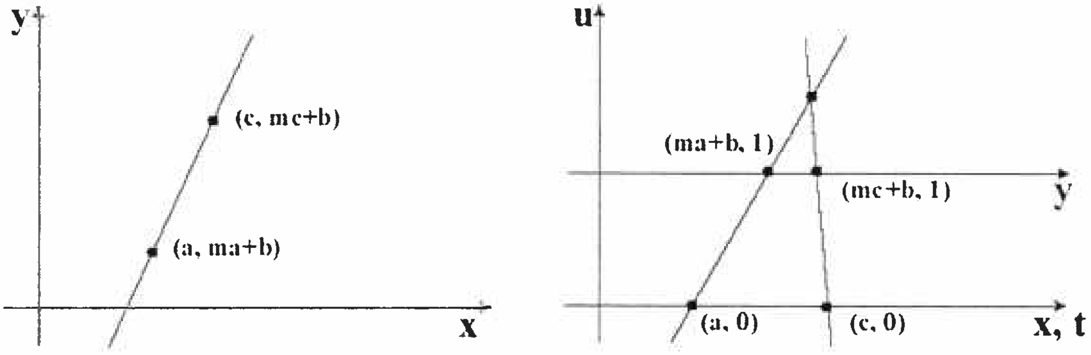

Figure 4.2 : *Cartesian and Parallel coordinate plots of two points* $(a, ma+b)$, *and* $(c, mc+b)$

The point $(a, ma+b)$ in the cartesian system maps into the line joining $(a, 0)$ to $(ma+b, 1)$ in the parallel system. Similarly, point $(c, mc+b)$ maps into the line joining $(c, 0)$ to $(mc+b, 1)$. Conversely, it is straightforward to show that the above lines of the parallel system intersect at the point $L^* : \left(b\,(1-m)^{-1}, (1-m)^{-1}\right)$, which depends only on m and b, the parameters of the original line in the cartesian plot. Thus, in accordance to Proposition 4.1, L^* is the dual of L and we have the interesting duality result (a characteristic of projective planes already discussed), that points in cartesian coordinates map into lines in parallel coordinates and conversely lines in cartesian coordinates map into points in parallel coordinates.

The cartesian to parallel system transformation can be theoretically defined now, and Proposition 8 will serve as the basic tool in order to achieve this. Indeed, consider first that in natural homogeneous coordinates the line $L : y = mx + b$ is represented by the triplet $(m, -1, b)$. Additionally, the point $L^* : \left(b\,(1-m)^{-1}, (1-m)^{-1}\right)$ as an ordinary point can be represented as $\left(b\,(1-m)^{-1}, (1-m)^{-1}, 1\right)$ or equivalently if we multiply by $1-m$ (recall notation 2), as $(b, 1, 1-m)$. Given the above, if we consider the nonsingular matrix **A**:

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & -1 & -1 \\ 1 & 0 & 0 \end{pmatrix}$$

it is straightforward to show that

$$(b, 1, 1 - m) = (m, -1, b)\,\mathbf{A}$$

and consequently, in accordance to Proposition 4.2, the transformation from lines in cartesian orthogonal coordinates represented by $(m, -1, b)$, to points in parallel coordinates represented by $(b, 1, 1 - m)$ is a linear projective transformation, depending on the particularly simple matrix **A**. To complete the formal specification of the cartesian/parallel transformation, we have to define the transformation once again, this time by transforming points in cartesian orthogonal coordinates to lines in parallel coordinates. To achieve this, once again we have to express our coordinates in projective geometry notation using the natural homogeneous coordinates; to start with, assume a point in the cartesian coordinate system given by (x_1, x_2). It is evident that in natural homogeneous coordinates (x_1, x_2) is written as $(x_1, x_2, 1)$. Now, the point (x_1, x_2) in parallel axes is represented by the line joining point $(x_1, 0)$ to $(x_2, 1)$. What remains is to express this line in natural homogeneous coordinates. In do this, we first find the equation of the line, which is given by

$$y = (x_2 - x_1)^{-1}\, x - x_1\,(x_2 - x_1)^{-1}\,. \tag{4.1}$$

Recall that a line $L : y = mx + b$ is represented in natural homogeneous coordinates by the triplet $(m, -1, b)$. Accordingly, the line in Figure 4.2 is represented by the triplet $\left((x_2 - x_1)^{-1}, -1, -x_1\,(x_2 - x_1)^{-1}\right)$ or equivalently (multiplying with $x_2 - x_1$), by $(1, x_1 - x_2, -x_1)$. It is straightforward now to verify that the linear transformation from point $(x_1, x_2, 1)$ of cartesian coordinates to line $(1, x_1 - x_2, -x_1)$ of parallel coordinates is given by

$$(1, x_1 - x_2, -x_1) = (x_1, x_2, 1)\,\mathbf{B}$$

where

$$\mathrm{B} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

4.2.2 Identifying Statistical Information from Parallel Plots

As already stated, the most interesting feature of Inselbergs' parallel coordinate plot is its feasibility to be considered, from a statistical point of view, an exploratory data analysis tool capable of diagnosing two-dimensional features such as correlations and nonlinear structures, as well as multi-dimensional features such as clustering, hyperplanes and the modes. Of the above, the most interesting, at least in the case of presenting longitudinal/repeated measures data, is the visual detection of correlation between data.

In *Wegman (1990)* we find an adequate explanation on exactly how the parallel coordinate plot can become useful in detecting the existence of correlation as well as the degree of this correlation of the data plotted in this way. We once again need to consider the points $(a, ma + b)$ and $(c, mc + b)$ of line $L : y = mx + b$ (Figure 4.2). Moving from Wegman's general multivariate setting towards a longitudinal perspective, these 2-dimensional points could be taken to represent in a longitudinal study the two dimensional vectors $\mathbf{y}_i = \left(y_{ij}, y_{i(j+1)}\right)^t$ and $\mathbf{y}_{i'} = \left(y_{i'j}, y_{i'(j+1)}\right)^t$ each one consisting of two repeated measurements on subjects i and i' $\left(i \neq i'\right)$ respectively, taken at the consecutive times t_j and t_{j+1}. As already shown, point $(a, ma + b)$ in a parallel plot is represented by the line connecting points $(a, 0)$ and $(ma + b, 1)$ and point $(c, mc + b)$ is represented by the line connecting points $(c, 0)$ and $(mc + b, 1)$. We are going to show that the relative position of these two lines determines in essence the degree of correlation between the pairs $(a, ma + b)$, $(c, mc + b)$ or equivalently (from a longitudinal point of view) the degree of correlation between the neighboring times t_j and t_{j+1}. First, consider that the intersection point of the two lines is given by $\left(b\,(1-m)^{-1}, (1-m)^{-1}\right)$. Then, if $0 < (1-m)^{-1} < 1$ which means that the intersection of the two lines occur between the parallel axes, apparently $m < 0$ ($\frac{1}{1-m} < 1 \Rightarrow m < 0$). On the other hand, in the case of $(1-m)^{-1} > 1$ or $(1-m)^{-1} < 0$ so that the intersection occurs external to the region between the parallel lines, m is positive. If there was a way to relate the sign of m and the correlation of the two points $(a, ma+b)$, $(c, mc+b)$ then we would have obtained an interesting criterion, stating that for highly negatively correlated points their dual line representations in a parallel plot tend to cross near a single point between the two parallel coordinate axes while for highly positively correlated data lines tend not to intersect between the parallel coordinate axes.

Indeed, as regard this relation, it is a relatively simple task to show that if $m < 0$ the points are highly negative correlated, and if $m > 0$ we have highly positive correlated points. By denoting as X the random variable comprising the abscissas of points on line $L: y = mx + b$, then $Y = mX + b$ is a random variable that includes the corresponding ordinates of points on the same line:

X	Y
a	$ma+b$
c	$mc+b$
$\vdots$	$\vdots$

Corollary 4.1, deriving from the following theorem will assist in defining the above mentioned relation.

Theorem 4.1: *Suppose X, Y random variables of which their correlation coefficient $\rho(X,Y)$ exists. If a,b,c and d are constant real numbers with $ac \neq 0$, then:*

$$\rho\,(aX+b, cY+d) = \left\{ \begin{array}{c} \rho\,(X,Y)\ \ if\ ac > 0 \\ -\rho\,(X,Y)\ \ if\ ac < 0 \end{array} \right\}.$$

From the above theorem, by taking $X = Y$, $a = 1$ and $b = 0$, evidently we have the following corollary:

Corollary 4.1: *Suppose X, Y random variables linearly dependent, that is $Y = aX + b$ with a, b constant real numbers and $a \neq 0$. Then:*

$$\rho(X,Y) = \rho(X, aX+b) = \left\{ \begin{array}{c} 1 \; if \; a > 0 \\ -1 \; if \; a < 0 \end{array} \right\}.$$

Implementation of the above corollary, in the case of X, $Y = mX + b$ gives:

$$\rho(X, mX+b) = \left\{ \begin{array}{c} 1 \; if \; m > 0 \\ -1 \; if \; m < 0 \end{array} \right\}.$$

which in words states that if $m > 0$, that is the parallel lines intersect outside the parallel axes, we have a strong positive correlation between random variables X and $mX + b$ described above, corresponding to a strong positive correlation between the data y_{ij}, $y_{i'j}$ and $y_{i(j+1)}$, $y_{i'(j+1)}$ obtained at the neighboring times t_j and t_{j+1}. Similarly, when $m < 0$ (hence intersection occurs somewhere between the parallel lines), this is a strong indication of negative correlated data taken at consecutive times t_j, t_{j+1}.

Besides the usefulness of parallel axis plots in detecting the existence of correlation (either positive or negative) between within-subject measurements taken at neighboring (consecutive) times, it should be noted that the underlying plot exhibits another important utility; the basic structure of longitudinal data can be revealed from a parallel plot. As stated before, the basic objective in the statistical analysis of longitudinal data is to come up with a suitable statistical model that fits the data adequately. The General Linear Mixed Model (GLMM in abbreviation) for longitudinal data presented in Chapter 5, for instance, one of the most recently developed approaches in longitudinal data modeling, has become quite popular.

In the context of the latter model, the parallel plot may be a useful first exploratory tool for understanding the structure of the data (mainly the between-subjects structure). Moreover, all this interesting information obtained by the graphical representation of the data via the parallel plot can be conveyed into the longitudinal data modeling in order to assist in the construction of the most suitable model that will fit the data in the best possible way. Thus, observing the parallel plot of the 'raw' data allows us to detect basic features associated with the fixed as well as the random parameters (e.g. fixed/random intercepts and slopes) of the mixed-effects model for longitudinal data.

Some basic interpretation of parallel plots in this setting is given by *Weiss* and *Lazaro (1992).* In their article, the authors describe how to read information by using either the parallel plot of the 'raw' data or the parallel plot of the residuals. As demonstrated by Weiss and Lazaro, the presence of specific patterns in a parallel axis plot of the data can be clearly suggestive about the most possible model formulation. In the sequel, we describe how to read information in a parallel plot of the 'raw' data. To this end, we note that individual parallel flat lines in a parallel plot usually indicate the need for a random intercept, individual parallel sloped lines indicate the need for a fixed slope as well as for a random intercept, and differing slopes indicate the need for random slopes. The easiest way to illustrate all the preceding is by means of some examples of real, repeated measures, data sets. For instance, Figure 4.3 shows the rat body weights data taken from *Box (1950).*

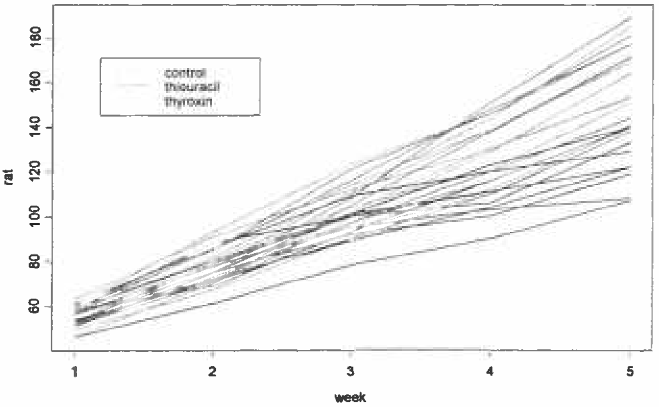

Figure 4.3 : *Rat body weights plotted in a paralel axis plot.*

The data consist of rat body weights in grams, from a toxicology study. In this experiment, 27 rats were randomly allocated to each of three treatment groups; control, treatment with the thyroxin additive, and treatment with the thiouracil additive. The rats were weekly weighted for a period of five weeks. Thus, a total number of five repeated measurements were obtained from each rat, and therefore we can classify the specific data as longitudinal data.

By viewing the plot of Figure 4.3, one can immediately see that the individual body weights generally increase with time and that the marginal variance increases with time too. (The marginal variance at the last time point, that is week 5, is significantly larger compared to the marginal variance of the preceding time points). In general, widening-shaped patterns such as the above are indicative of increasing variability and usually a model that involves a fixed intercept but a random slope seems to be the most appropriate model to fit the specific data. Thus in short, in a parallel plot where the individual lines at time point 1 begin close together but spread with increasing time, a fixed intercept and a random slope are a typical choice. Moreover, it is also possible from Figure 4.3 to detect the presence of a strong positive correlation between the neighboring within-subject measurements, since as one can easily observe the individual line segments do not cross in general. (With negative correlated data, the line segments would criss-cross between two consecutive time points). Similarly, a model with both a random slope and an intercept might be indicated by a parallel plot where individual lines begin separated and spread further. Finally, for longitudinal data with a parallel plot where individual cases appear more or less as parallel lines (as, for example, in the dental study data of Potthoff and Roy depicted by Figure 4.1), there is a strong indication that a random intercept should be included in the initial model.

4.2.3 Casement Plots

Casement plots (*Chambers et al., 1983*) provide a useful way of reducing clutter that can usually shown on a single parallel axis plot, by stratifying the overall observations using more than one parallel plots, according to a covariate. For instance, consider the Potthoff and Roy data as presented in a single parallel plot (Figure 4.1). Plots of this form, due to the fact that comprise the entire set of observations on all subjects may become very complicated and consequently any preliminary attempts to recognize possible structures from these plots can be rather problematic. By presenting the observations, not in a single plot, but via plots that each of them includes observations associated with the different levels of some covariate enables us to detect more easily any differences between the levels of the covariate.

Figure 4.4 presents a typical casement plot of the Potthoff and Roy data. In this case the data have been grouped in such a way so that they can be distinguished as concerns the two different levels (i.e. boys, girls) of the covariate 'sex'.

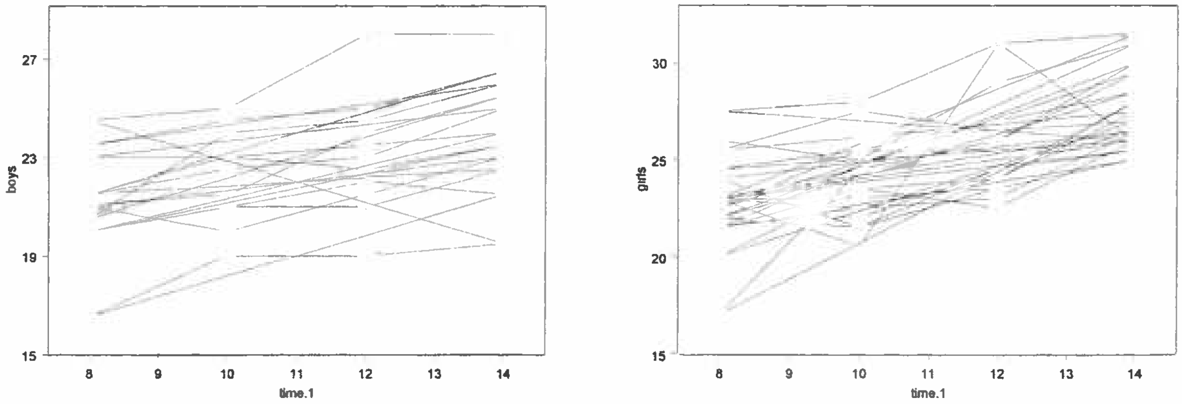

Figure 4.4: Casement plots of the Potthoff & Roy data. The orthodontic growth patterns of the 16 boys and the 11 girls are presented with different parallel plots per gender.

It is evident that by observing these two separate plots is much more easier to see differences (e.g. differences associated with the variation between boys and girls) among the two groups, than by using the single parallel axis plot of Figure 4.1. Finally, as concern the appearance of casement plots in the literature, it should be mentioned that the casement plots have been used by *Crowder* and *Hand (1990, Chapter 2)* although not by name.

4.3 The Draftman's Display

As an exploratory tool, the parallel axis plot satisfactorily conveys important information about the data's behavior, such as changes of overall (between subjects) variance across time or changes in the shape of each subject's curve. However, it has not proven to be useful for the inspection of the within-subjects variance-covariance structure of longitudinal and repeated measures data.

The main plot for checking the data's, within-subject, covariance structure (without first having to fit a model) is the so called Draftman's display[3] as described by *Dawson et al. (1997)*. Consider once again a typical, balanced, longitudinal data set, where i subjects ($i = 1, 2, ..., m$) are repeatedly measured over time. Suppose that measurements are equally collected for all subjects at the n occasions: $\mathbf{t} = (t_1, t_2, ..., t_n)^t$. A Draftman's display of data such as the above is a matrix of scatterplots (also known as scatterplot matrix) of observations from the same subject, at times t_1 and t_2, times t_1 and t_3,..., times t_n and t_{n-1}. Basically, it shows how observations on the same subject, but at different time points, are related. Inspecting this array of scatterplots, it is possible to gain an indication of the correlation (and hence the covariance structure) between the within-subjects measurements, at various time lags.

For example, if the correlations at all lags are of about the same magnitude, then a compound symmetric structure seems reasonable to describe the within-subjects covariance. If the correlations are shown to decay exponentially with the time lag, then a autoregressive covariance structure can be considered as the most appropriate to describe the dependence of observations within a subject. To demonstrate application of

[3] The specific graph was originated by *Chambers et al. (1983)* as a technique for detecting clustering and outliers.

Draftman's display in practice, we now apply it on the data described in *Box (1950)*, consisting of rats body weights in grams, from a toxicology study.

To get an impression of the within-subject dependency structure, the Draftman's display of the above data has been drawn and is shown in Figure 4.5.

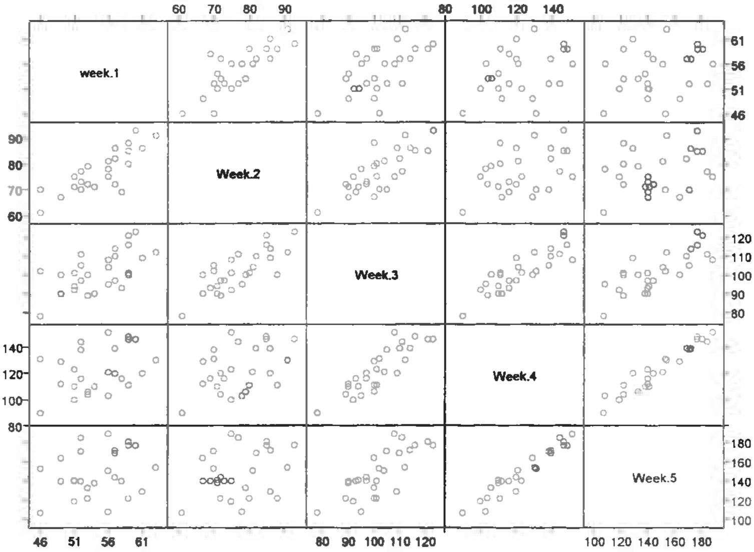

Figure 4.5 : *Draftman's display of the rat body weight data.*

The trends of decreasing correlations with increasing interval between measurement times is apparent in the above scatterplots. That is, the scatterplots for the measures close in time show stronger correlations than the scatterplots for the measures apart in time. For example, scatterplots of sequential observations (i.e. week 1 against week 2 measurements, week 2 against week 3 measurements, etcetera) are shown to exhibit a significant linear trend, while the scatterplot that plots observations collected at week 1 against observations collected at week 5 shows no trace of correlations (all points on this plot appear to be randomly scattered). The basic conclusion of practical usage is that measurements taken close together in time are generally more strongly correlated that those taken further apart in time.

Typically, within-subject variation of this nature is usually represented by assuming an autoregressive type covariance structure (see Section 5.4). Of course, various other covariance structures may be indicated by the inspection of a Draftman's display, depending each time on the specific data's within-subject interrelations (e.g. suppose a Draftman's display that exhibits total absence of a linear trend in any of the pairwise scatterplots. Clearly, a plot of this form suggests that the within-subject observations are essentially independent and thus an analogous, independent type, covariance structure should be assumed).

Finally, it should be mentioned at this point, that despite its practical usefulness Draftman's display has not shown to perform adequately well under all situations. *Weiss (1997)* reports that while in the cases where the between-subjects variance is constant across time the Dtaftman's display gives de-facto a picture of the (within-subjects) correlation structure, on the other hand when the variance is not constant this changing variance may easily overwhelm the correlation information. As a solution to overcome this problem, both *Weiss (1997)* and *Dawson (1997)* recommend in the non-constant variance situations to apply Draftman's display not to the observed data, but to transformed data obtained by the removal of variability associated with differences in the means and variances over time [one usually uses the standardized data y_{ij}/s_j, obtained by dividing the repeated measurements y_{ij} collected at time t_j $(j = 1, 2, ..., n)$ with the sample standard deviation s_j of the response variable at time t_j].

Moreover, it is worth noting that Draftman's display is most effective for studying equally spaced longitudinal data (balanced designs). In the case of unbalanced longitudinal data, comparisons across the various scatterplots of the Draftman's display may become difficult. Other plots are useful for checking the covariance structure under this setting. For instance, an alternative way of visualizing the association among within-subjects repeated measurements with irregular observation times is the semi-variogram (*Diggle, 1988*).

Chapter 5

General Linear Mixed Model for Longitudinal Data

5.1 Introduction

In Chapter 2, we have discussed in detail that the classical methods (ANOVA, MANOVA), for analyzing longitudinal and repeated measures data unfortunately find application only under special circumstances, mostly due to the restrictions they impose. To begin with, a basic requirement of both classical methods is the one of keeping the design balanced. This, although does not causes serious problems in many types of designs (e.g. agricultural or industrial experiments, where the researchers usually have full control over experimental conditions), often there are situations where it becomes a major drawback. Such situations are those where the experimental units are humans. Take for example a medical study, that is designed to measure a specific characteristic of patients participating in the study, at predetermined regular time intervals. Many years of practice showed that due to various reasons (patients' drop-out of the study, or failing to return at the specific designed times), the balance brakes down. Thus, the bottom line is that in many practical situations the requirement for balanced design is an unachivable ideal.

In addition to the above, we have shown that the analysis of variance F-tests for the (fixed or random) effects of the ANOVA model depend on the sphericity assumption for their validity (see *Huynh* and *Feldt, 1970*). However, longitudinal data obtained in applied settings, (particularly in the medical and behavioral sciences), will rarely conform to restrictive assumptions for the covariance structure of the model such as the ones of sphericity or compound symmetry. In contrast, the multivariate analysis of variance method presumes a complete arbitrary covariance structure. As a consequence, the method does not attempt to take in account the two sources of random variation arising in longitudinal data, the within-unit variation (random error) and the between-unit variation (random effects). This corresponds in having to estimate a large number of variance components. For the above reasons, other more recently developed strategies have been advocated for the analysis of longitudinal and repeated measurements–strategies that may be generally more valid for data obtained in applied settings.

One of the newer approaches is based on a mixed model methodology, that results in a very general model for handling longitudinal data that not only allows for various parameterizations of the covariance structure, but in addition can handle unequal numbers of observations for each subject, as well as unequal time-spacing of the observations. This approach enables practicing statisticians to choose from various covariance structures for the model formulation, rather than having to presume a certain type of structure (ANOVA case), or a complete unspecified structure (MANOVA case). A consequence to this, is more efficient estimates of the parameters of the model and more powerful tests of the models effects (fixed/random).

The General Linear Mixed Model for Longitudinal Data has been proposed by *Laird* and *Ware (1982)* in their significative, computationally oriented article under the title "random-effects models for longitudinal data". Although their model is partly based upon the General Linear Mixed Model presented in Chapter 3, as well as in the significant work of *Harville (1977)*, they were the first to introduce the GLMM theory to longitudinal studies, making their work seminar in the specific field of analysis of longitudinal data. Thus, it is not surprising to find many authors referring to the general linear mixed model for longitudinal data as the "Laird-Ware model". In fact, a wide variety of names are also used in the statistical literature to describe the latter model and its versions, reflecting the diversity of its use in many fields. These names include: general linear mixed (effects) model; Laird-Ware model; two-stage random effects model; multilevel linear model; two-level hierarchical model (because there are two levels of random variation: $\mathbf{u}_i$ and ε_i); 'empirical Bayes' model, and random regression coefficients.

5.2 The 'Laird-Ware' Model

5.2.1 Model and Notation

Following *Laird* and *Ware (1982)*, the general linear mixed model for longitudinal data (or the two-stage random effects model, as it was originally called in the specific article) can be written as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i, \qquad (i = 1, ..., m)\,, \tag{5.1}$$

where,

- $\mathbf{y}_i$ is the $(n_i \times 1)$ vector of responses for the ith subject $\left[\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t\right]$, taken at times $\mathbf{t}_i = (t_{i1}, t_{i2}, ..., t_{in_i})^t$,
- $\mathbf{X}_i$ is a $(n_i \times p)$ design matrix that characterizes the systematic part of the response, e.g. depending on covariates and time,
- $\mathbf{b}$ is a $(p \times 1)$ vector of (population-specific) fixed parameters, namely the fixed effects,
- $\mathbf{Z}_i$ is a $(n_i \times q)$ design matrix that characterizes random variation in the response due to among-unit sources,
- $\mathbf{u}_i$ is a $(q \times 1)$ vector of (subject-specific) random effects, and finally
- ε_i is a $(n_i \times 1)$ vector of within-unit errors, usually called random error.

Since each subject's response vector $\mathbf{y}_i$ consists of n_i measurements, it is evident that the total number of observations included in the longitudinal model will be $N = \sum_{i=1}^{m} n_i$, in total.

Furthermore, as concerns the distributional behavior of the random terms of model (5.1), it is customary to specify a fully parametric form for both the subject-specific random effects $\mathbf{u}_i$ and the random errors $\boldsymbol{\varepsilon}_i$. Normality is the most common parametric assumption for the distribution of $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ (*Laird* and *Ware, 1982*), that is we assume:

$$\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathrm{D}\right) \quad and \quad \varepsilon_i \sim N_{n_i}\left(0, \mathrm{R}_i\right), \tag{5.2}$$

where $\mathbf{D}$ is a $(q \times q)$ variance-covariance matrix that characterizes variation due to between-subjects sources, and $\mathbf{R}_i$ a $(n_i \times n_i)$ variance-covariance matrix that characterizes variance and correlation due to within-subjects sources (i.e. the variation that occurs due to measurement error or due to biological within-unit fluctuations). Also, it is assumed that $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ are distributed independently for $i = 1, ..., m$. Further, notice that the above model assumes homogeneity of variance only for $\mathbf{u}_i$ (constant variance-covariance matrix $\mathbf{D}$ for all subjects $i = 1, ..., m$).

In the above GLMM of Laird and Ware, the primary constraint upon the variance-covariance matrices $\mathbf{D}$, $\mathbf{R}_i$ is that matrix $\mathbf{D}$ has to be positive-semidefinite[1], whereas no specific assumption is made on $\mathbf{R}_i$. From (5.2) it follows that each response vector $\mathbf{y}_i$ follows a n_i-variate normal distribution, with mean vector $E\left(\mathbf{y}_i\right) = \mathbf{X}_i\mathbf{b}$ and with variance-covariance matrix $Var\left(\mathbf{y}_i\right) = \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$. Hence, we may write:

$$\mathbf{y}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i\right). \tag{5.3}$$

[1] A square matrix $\mathbf{A}$ is positive-semidefinite (or non-negative definite) if for any column vector of constants $\mathbf{x}$, it is $\mathbf{x}^t\mathbf{A}\mathbf{x} \geqslant 0$.

5.2.2 Variations in Presenting the "Laird-Ware" Model

Due to its general applicability, model (5.1) has enjoyed great popularity, and applications of it have been and are continuing to be made in many fields of statistical analysis. In this Section we display some of the most common alternative representations and variations of the 'Laird-Ware' model, found in the literature.

The "Laird-Ware" Model as a GLMM

Alternatively, we can write the combined model for all data in a (single) matrix form by letting:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{pmatrix} = \left(\mathbf{y}_1^t, \mathbf{y}_2^t, ..., \mathbf{y}_m^t\right)^t,$$

the $(N \times 1)$ vector that comprises the repeated measurements of all m subjects, and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix},$$

where $\mathbf{X}$ is a $(N \times p)$ matrix, $\mathbf{u}$ is the $(mq \times 1)$ vector containing each subject's random effects and $\boldsymbol{\varepsilon}$ the $(N \times 1)$ vector of random errors. Further, if we define[2]

$$\begin{aligned} \mathbf{V} &= diag\left(\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_m\right), \\ \mathbf{D} &= diag\left(\mathbf{D}, \mathbf{D}, ..., \mathbf{D}\right), \\ \mathbf{Z} &= diag\left(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_m\right) \text{ and} \end{aligned}$$

[2] The notation $diag\,(\alpha_1, ..., \alpha_n)$ implies a matrix with diagonal elements $\alpha_1, ..., \alpha_n$ and all off-diagonal elements zero.

$$\mathbf{R} \;=\; diag\left(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_m\right),$$

then the linear model for vector $\mathbf{y}$ has exactly the same form as the General Linear Mixed Model (GLMM), discussed in Chapter 3:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \boldsymbol{\varepsilon},$$

with the only difference being that the implied variance-covariance structure for vector $\mathbf{y}$ is block-diagonal with the m $\mathbf{V}_i$ matrices making up the diagonal of the variance-covariance matrix $\mathbf{V}$. Thus, under this notion, the "*Laird-Ware*" model can be considered as a special case of the GLMM.

The "Laird-Ware" Model as a Two-Stage Model

As noted previously, the "Laird-Ware model" through a barrage of literature following the article of *Laird* and *Ware (1982),* has been presented under various names, including the appellation of two-stage model. The specific name is attributed to the model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ due to the fact that the latter can be formulated in two stages, as we demonstrate in detail now.

One of the most interesting features of "Laird-Ware" model, is its duality concerning the source of random variation. The model incorporates two sources of random variation; the random vector $\mathbf{u}_i$, aiming to describe the variation between the i individuals, as well as the random error $\boldsymbol{\varepsilon}_i$ corresponding to the within individuals variation. The two stages serve the purpose of introducing into the model these two sources of randomness, one at each stage, separately. Specifically, the first of the two stages, describes the distribution of each response $\mathbf{y}_i$ within individuals. For this reason, the random term $\mathbf{u}_i$ becomes a non-random parameter by conditioning the response $\mathbf{y}_i$ on $\mathbf{u}_i$. Considering that the mean and variance-covariance matrix of the conditional variable $\mathbf{y}_i \mid \mathbf{u}_i$ is given by $E\left(\mathbf{y}_i \mid \mathbf{u}_i\right)=\mathbf{X}_i\mathbf{b}+\mathbf{Z}_i\mathbf{u}_i$ and $Var\left(\mathbf{y}_i \mid \mathbf{u}_i\right)=\mathbf{R}_i$, at stage 1 we have:

$$\textit{Stage } 1: \qquad \mathbf{y}_i \mid \mathbf{u}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b}+\mathbf{Z}_i\mathbf{u}_i, \mathbf{R}_i\right).$$

(Notice that up to this point, the variance-covariance of $\mathbf{y}_i$ is $\mathbf{R}_i = Var\left(\boldsymbol{\varepsilon}_i\right)$, hence stage 1 does not allow other source of random variation, except from the $\boldsymbol{\varepsilon}_i$).

The second stage describes the variability between individuals by specifying a distribution for the random effects $\mathbf{u}_i$. Usually, the assumption is that $\mathbf{u}_i$ follows a multivariate normal distribution with zero mean and variance-covariance matrix $\mathbf{D}$. Hence, at stage 2 we have:

$$\textit{Stage } 2: \qquad \mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right).$$

It is important to realize that the above two-stage model is just an alternative presentation of the "Laird-Ware model", $\mathbf{y}_i = \mathbf{X}_i\mathbf{b}+\mathbf{Z}_i\mathbf{u}_i+\boldsymbol{\varepsilon}_i$. Indeed, based on the information provided by stages 1 and 2, it is an easy task to derive the model equation (5.1). To show this, we only use standard results of Probability Theory, and especially the definitions of the marginal density function and the conditional density function based on a two-dimensional random variable. More specific, if we consider $\mathbf{y}_i$ and $\mathbf{u}_i$ to formulate the two-dimensional random variable $\left(\mathbf{y}_i, \mathbf{u}_i\right)^t$, then it is well-known that the marginal density function (p.d.f) of $\mathbf{y}_i$ is given by:

$$f\left(\mathbf{y}_i\right)=\int f\left(\mathbf{y}_i, \mathbf{u}_i\right) d\mathbf{u}_i=\int f\left(\mathbf{u}_i\right) f\left(\mathbf{y}_i \mid \mathbf{u}_i\right) d\mathbf{u}_i, \tag{5.4}$$

where by $f\left(\mathbf{u}_i\right)$ and $f\left(\mathbf{y}_i \mid \mathbf{u}_i\right)$ we denote the density function of $\mathbf{u}_i$ and $\mathbf{y}_i \mid \mathbf{u}_i$, respectively. Now, from stages 1 and 2, $f\left(\mathbf{u}_i\right)$ and $f\left(\mathbf{y}_i \mid \mathbf{u}_i\right)$ can be easily constructed as:

$$f\left(\mathbf{u}_i\right)=\frac{1}{(2\pi)^{\frac{q}{2}} \mid \mathbf{D} \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i\right\}, \tag{5.5}$$

and

$$f\left(\mathbf{y}_i \mid \mathbf{u}_i\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \mathbf{R}_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right\}, \quad (5.6)$$

By substituting (5.5) and (5.6) in (5.4) and performing the calculations, it can be shown that the probability density function of $\mathbf{y}_i$, $f\left(\mathbf{y}_i\right)$, corresponds to a n_i–dimensional normal distribution with mean vector $\mathbf{X}_i\mathbf{b}$ and variance-covariance matrix $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$, that is:

$$\mathbf{y}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i\right),$$

which is just an alternative way of stating that $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$, with $\mathbf{y}_i$ following a multivariate, n_i–dimensional normal distribution.

The "Laird-Ware" Model as a Growth-Curve Model

We are closing the discussion on the most often presented variations in the literature of the GLMM for longitudinal data, with a different approach as concern the model's formulation. We are going to see how the latter model can be expressed, this time in the context of growth curve analysis[3]. As is standard when working under a growth curve perspective, for the formulation of model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$, the primary concern is on specifying the part of the model that characterizes the growth curve of each individual. Then, based on this specification, we proceed with the modeling of the parameters of the individual growth curves as linear functions of individual characteristics [*Laird et al. (1987)*].

Namely, this type of formulation consists of two separate stages, as described above; in the first stage we assume:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \varepsilon_i,$$

[3] Growth curve analysis applies to data consisting of repeated measurements over time, in which a single characteristic has been measured at n different occasions on each individual. The interest is mainly on forecasting the future growth of individuals. [For more details on the subject, we refer to *Potthoff* and *Roy (1964), Khatri (1966)* and *Rao (1965)].*

where the only introduced term is $\boldsymbol{\beta}_i$, which is the random vector that defines the ith individual's growth curve ($i = 1, ..., m$). Additionally, as concern the distributional behavior of $\boldsymbol{\beta}_i$, we have:

$$\boldsymbol{\beta}_i \sim N\left(\mathbf{A}_i\mathbf{b}, \mathbf{D}\right),$$

where $\mathbf{A}_i$ a $(q \times p)$ design matrix and $\mathbf{b}$, $\mathbf{D}$ (as usual) a $(p \times 1)$ vector of fixed parameters and the $(q \times q)$ variance-covariance matrix of random vector $\mathbf{u}_i$, respectively. Now, we are able to proceed with the second stage, which completes the specification of the model. To do so, we calculate the mean and variance-covariance of the response vector $\mathbf{y}_i$ as:

$$E\left(\mathbf{y}_i\right) = \mathbf{Z}_i E\left(\boldsymbol{\beta}_i\right) + E\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{Z}_i\mathbf{A}_i\mathbf{b} = \mathbf{X}_i\mathbf{b},$$

and

$$\begin{aligned} Var\left(\mathbf{y}_i\right) &= \mathbf{Z}_i Var\left(\boldsymbol{\beta}_i\right)\mathbf{Z}_i^t + Var\left(\boldsymbol{\varepsilon}_i\right) \\ &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i, \end{aligned}$$

where we have set $\mathbf{Z}_i\mathbf{A}_i = \mathbf{X}_i$. From the above, it follows that $\boldsymbol{\beta}_i = \mathbf{A}_i\mathbf{b} + \mathbf{u}_i$, and the equivalence of growth curve model: $\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$ to the 'Laird-Ware' model: $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ is apparent, since:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i = \mathbf{Z}_i\left(\mathbf{A}_i\mathbf{b} + \mathbf{u}_i\right) + \boldsymbol{\varepsilon}_i \\ &= \mathbf{Z}_i\mathbf{A}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i. \end{aligned}$$

The "Laird-Ware" Model in a Bayesian framework

The Bayesian approach to the formulation and representation of the 'Laird-Ware' random effects model consists of three distinct stages. As is well-known, the distinguishing feature of a typical Bayesian model is the specification of prior distributions for all parameters in the model. These ideas are accommodated in the GLMM for longitudinal data by the following 3-stage Bayesian model:

$$\underline{stage\ 1}: \quad (describes\ the\ within - subject\ variation)$$
$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad (i = 1, 2, ..., m)$$
$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$$

$$\underline{stage\ 2}: \quad (describes\ the\ between - subject\ variation)$$
$$\mathbf{u}_i \sim N_q(0, \mathbf{D})$$

$$\underline{stage\ 3}: \quad (specifies\ the\ prior\ distributions)$$
$$\mathbf{b} \sim N_p(\mathbf{b}^*, \mathbf{H})$$
$$\mathbf{D}^{-1} \sim Wishart\ and\ \mathbf{R}_i^{-1} \sim Wishart$$

In the above setting, the parameters $\mathbf{b}^*$, $\mathbf{H}$ and those characterizing the independent Wishart distributions for $\mathbf{D}^{-1}$ and $\mathbf{R}_i^{-1}$ are assumed to be known. The choice of Normal and Wishart priors, is an example of the usual Bayesian strategy of using conjugate[4] priors.

At this point it should be noted that although the primary emphasis of the present thesis has been on describing the modeling framework as well as on the estimation methods for the 'Laird-Ware' model from a classical frequentist perspective, Bayesian approaches to the analysis and estimation of the 'Laird-Ware' model that have been considered in the literature are also presented. In general, there is a strong similarity between the inferential procedures arrived at from both the frequentist or the Bayesian point of view as concerns the 'Laird-Ware' model.

Lindley and *Smith's (1972)* work is of key importance to Bayesian type approaches to the linear mixed model analysis. Their idea was to add a third stage to the two-stage model to incorporate prior distributions for the parameters. *Fearn (1975)* developed the ideas of Lindley and Smith for specific application to growth curve models. Other related

[4] A conjugate prior is one for which the resulting posterior distributions of interest come from the same distributional family.

references include *Butler* and *Louis (1992), Strenio et al. (1983), Louis (1991), Geisser (1970)* and *Searle et al. (1992).*

Further, we should also remark that the iterative techniques recently developed in the Bayesian theory for estimating posterior distributions, namely Markov chain Monte Carlo (MCMC) methods, and in particular the Gibbs sampling algorithm (see, e.g., *Gelfand* and *Smith, 1990*) have been adopted by many authors for the analysis and parameter estimation of the general linear mixed-effects model for longitudinal data (*Laird* and *Ware, 1982*). Within this framework, *Gelfand* and *Smith (1990)* consider a full Bayesian analysis of the (Gaussian) linear mixed model of Laird and Ware by proposing a Gibbs sampling scheme. A wide variety of alternative Gibbs samplers have been implemented by several authors in longitudinal modeling applications, for instance see *Lange et al. (1992); Carlin (1996); Carlin* and *Louis (1996)* and *Liu* and *Rubin (1995).* In the following years, further refinements and improvements (e.g. convergence improvements) on the work of Gelfand and Smith appeared in the literature, including *Chib* and *Carlin (1999); Vines et al. (1996)* and *Gelfand et al. (1995, 1996)* among others.

Alternative Parametric Specifications

Most of the existing work on the 'Laird-Ware' model is based on the parametric specification (5.2) for the random terms $\mathbf{u}_i$ and ε_i. That is, whenever a fully parametric distributional assumption is made for the random effects $\mathbf{u}_i$ and the error terms ε_i it is almost always taken to be the normal model (5.2).

However, this approach, despite its major advantages, has a few disadvantages, too. Specifically, a typical linear mixed-effects model for longitudinal data specified by (5.1) and (5.2) (i.e. with normal distributions to characterize the distributional behavior of $\mathbf{u}_i$ and ε_i) suffers at some point from lack of robustness against outlying observations in the same manner as other statistical regression models based on the normal distribution. To this end, other parametric assumptions have been proposed in the literature for the 'Laird-Ware' model, assumptions that replace the well-established Gaussian distribution with distributions that handle more adequately any possible occurring outliers. *Pinheiro et al. (2001)* for instance, following a robust statistical approach originally considered by *Lange et al. (1989),* replace the multivariate normal distributions of (5.2) for the $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ by multivariate t-distributions with known or unknown degrees of freedom which are allowed to vary with the subject. The motivation for this was the fact that the t-distribution, with its heavier tails, appears as a suitable robust alternative to handle outlying individuals, compared to the Gaussian distribution.

Further contributions, associated with the GLMM and its parametric specification via the t-distribution, may be found in *Wakefield et al. (1994)* and *Racine-Poon (1992),* restricted only to the distribution of $\mathbf{u}_i$ though. Ideas similar to *Pinheiro et al.'s (2001)* are those by *Pendergast* and *Broffitt (1986),* who have also considered the multivariate t-distribution for the more restrictive field of growth curve models.

Pinheiro et al. (2001), proceeding as in *Lange et al. (1989),* replace the normal distributions used for the distributional specification of $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ with the t-distribution [i.e. $\mathbf{u}_i \sim t_q(\mathbf{0}, \mathbf{D}, v_i)$ and $\boldsymbol{\varepsilon}_i \sim t_{n_i}(\mathbf{0}, \mathbf{R}_i, v_i)$, where v_i denotes the multivariate t-distribution degrees-of-freedom (d.f.) for the ith subject]. Thus, the robust variant of the Laird-Ware model (multivariate t model), considered by *Pinheiro et al. (2001)* is written as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad (i = 1, ..., m)\,, \tag{5.7}$$

where random terms $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$ are assumed to be mutually independent, with:

$$\mathbf{u}_i \overset{ind.}{\sim} t_q(\mathbf{0}, \mathbf{D}, v_i) \quad \textit{and} \quad \varepsilon_i \overset{ind.}{\sim} t_{n_i}(\mathbf{0}, \mathbf{R}_i, v_i)\,. \tag{5.8}$$

Besides the formal specification of the multivariate t model, the authors also consider maximum likelihood (ML) estimation of the parameters in the latter model (namely, the fixed effects and the variance components), by describing three EM-type algorithms (*Dempster et al., 1977*) for the ML estimation with known and unknown degrees of freedom. Furthermore, by conducting simulation studies, *Pinheiro et al. (2001)* deduce that the multivariate t model substantially outperforms the typical Laird-Ware model with Gaussian specification when outliers are present in the data, even in moderate amounts, making thus the former as a robust alternative to the latter.

Non-Gaussian linear mixed-effects models for longitudinal data have been also taken under investigation by *Verbeke* and *Lesaffre (1996),* who assume the random effects $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ to be a sample from a mixture of g normal distributions instead of the usual normality assumption (5.2). They refer to their model as the 'heterogeneity model' (as it is specifically designed to take into account the presence of subgroups/clusters among the $\mathbf{u}_i$'s), and propose estimation and inferential techniques similar to those for the Laird-Ware model (e.g. EM algorithm, likelihood ratio tests). Mixture densities have been extensively used in biology and medicine for the purposes of modeling unobserved population heterogeneity [for a review on mixture models we refer the interested reader to *Böhning* and *Seidel (2003); Everitt* and *Hand (1981); McLachlan* and *Basford (1988)* and *McLachlan* and *Peel (2000)* among others]. As remarked by *McLachlan, Peel* and *Bean (2003),* especially for multivariate data (and consequently for longitudinal data) of a continuous nature, attention has focussed on the use of multivariate Gaussian components of the mixture distribution because of their computational convenience. In this setting, *Verbeke* and *Lesaffre (1996)* proceed by considering the random effects vectors $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ to be distributed according to the mixture: $pN(\boldsymbol{\mu}_1, \mathbf{D}) + (1 - p) N(\boldsymbol{\mu}_2, \mathbf{D})$, i.e. the random effects $\mathbf{u}_i$ following a mixture of two (multivariate) normals with proportions p and $(1 - p)$.

Finally, the interested reader may also be referred to *Magder* and *Zeger (1996)* for another parametric model for the random effects $\mathbf{u}_i$.

Nonparametric/Semiparametric Approaches

Parametric specifications for the random terms $\mathbf{u}_i$ and ε_i, such as the ones in (5.2), are by far the best-accepted and most widely applied approaches for analyzing longitudinal data using the basic general linear mixed model (5.1). In particular, most of the existing work and methods concerning model (5.1) has been based on assuming Gaussian distributions for both random effects vector $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ and random error term $\boldsymbol{\varepsilon}_i$ $(i = 1, 2, ..., m)$.

However, it is also possible either not to make any assumptions about the distribution of the random parameters included in the model at all (nonparametric approach), or to resort to some kind of compromise between a parametric and a nonparametric specification (semiparametric approach). Adaptation of distribution-free approaches, such as the above, has gained some attention in the recent years mainly due to the fact that the latter approaches offer flexibility and are less restrictive compared to the, more formal, parametric approaches. The lack of full distributional assumptions, however, often makes such methods difficult to interpret as well as computationally intensive.

Standard nonparametric smoothing methods, such as smoothing splines and kernel methods are mostly considered for the proposed nonparametric/semiparametric models. The literature on nonparametric smoothers for independent data is extensive. We refer the interested reader to *Hastie* and *Tibshirani (1990)* and *Wahba (1990)* for a comprehensive overview.

In the current context of longitudinal data, a comparatively limited literature exists, mainly concerning special types of mixed models for specific data sets. For instance, *Anderson* and *Jones (1995)* use smoothing spline structure to model the random effects, while *Wypij et al. (1993)* and *Wang* and *Taylor (1995)* use spline smoothers to model the non-random terms of their model, namely the fixed effects. Similar ideas have been used in *Shi et al. (1996)* who devised splines to model both the fixed effects and the random effects. *Zeger* and *Diggle (1994)* use a kernel smoother to model the mean CD4 cell numbers in HIV seroconverters and *Diggle* and *Verbyla (1997)* consider modeling the covariance structure using local linear smoothing with kernel weights. Also, *Verbyla et al. (1999)* assume a spline-based structure for the random effects. A more general family of nonparametric mixed-effects models has been proposed by *Wang (1996)*, who uses general smoothing spline models for the modeling of the fixed effects in the GLMM, while the random effects are modeled parametrically.

5.3 Estimation/Prediction of Fixed and Random Effects

5.3.1 Estimation of Fixed Effects b

As we have already done in Chapter 3 with the General Linear Mixed Model (GLMM), we once again begin by the (point) estimation of the fixed effects vector $\mathbf{b}$ of "Laird-Ware" model (5.1). Although several methods for estimating fixed-effects $\mathbf{b}$ are available (e.g. maximum likelihood, restricted maximum likelihood, generalized least squares), we will consider maximum likelihood estimation here. Thus, assuming that each subjects' i $(i = 1, 2, ..., m)$, response vector $\mathbf{y}_i$ is multivariate normal with **known** variance-covariance matrix $\mathbf{V}_i$ [i.e. $\mathbf{y}_i \sim N_{n_i}(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i)$], it can be shown [see *Laird* and *Ware (1982)*], that the maximum likelihood estimate of $\mathbf{b}$, say $\hat{\mathbf{b}}$, is given by:

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{y}_i, \tag{5.9}$$

where $\mathbf{V}_i = Var\left(\mathbf{y}_i\right) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$. Indeed, on assuming that the vector of measurements of ith subject, $\mathbf{y}_i$ follows a multivariate normal probability with mean $\mathbf{X}_i\mathbf{b}$ and variance-covariance matrix $\mathbf{V}_i$, then the (marginal) probability density function of $\mathbf{y}_i$ is given by:

$$f\left(\mathbf{y}_i; \mathbf{X}_i\mathbf{b}, \mathbf{V}_i\right) = \frac{1}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right)\right\}. \tag{5.10}$$

But there are m in total, independent to each other, vectors $\mathbf{y}_i$ $(i = 1, ..., m)$ and thus the likelihood function of all measurements $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)^t$ of the model (5.1) will be:

$$L\left(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i; \mathbf{y}\right) =$$

$$
\begin{aligned}
&= \prod_{i=1}^{m} f\left(\mathbf{y}_i;\mathbf{X}_i\mathbf{b},\mathbf{V}_i\right) = \prod_{i=1}^{m} (2\pi)^{-\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)\right\} \\
&= (2\pi)^{-\sum\limits_{i=1}^{m}\frac{n_i}{2}} \left(\prod_{i=1}^{m} \mid \mathbf{V}_i \mid\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)\right\},
\end{aligned}
$$

hence the corresponding log-likelihood λ is calculated as

$$
\begin{aligned}
\lambda &= \ln L = -\sum_{i=1}^{m}\frac{n_i}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\prod_{i=1}^{m} \mid \mathbf{V}_i \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right) \\
&= -\sum_{i=1}^{m}\frac{n_i}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{m}\left(\ln \mid \mathbf{V}_i \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right).
\end{aligned}
$$

To obtain the maximum likelihood estimator of fixed effects vector $\mathbf{b}$, $\hat{\mathbf{b}}$, it suffices to maximize log-likelihood λ. For this, we must calculate the partial derivative of λ with respect to $\mathbf{b}$ and equate the resulting derivative to zero. Indeed, doing so we have:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \mathbf{b}} &= 0 - 0 - \frac{1}{2}\frac{\partial}{\partial \mathbf{b}}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right) \\
&= -\frac{1}{2}\sum_{i=1}^{m}\left[\frac{\partial}{\partial \mathbf{b}}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)\right] = -\frac{1}{2}\sum_{i=1}^{m}\left[-2\mathbf{X}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)\right] \\
&= \sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right),
\end{aligned}
$$

and equating to zero, we have:

$$
\begin{aligned}
&\frac{\partial \ln L}{\partial \mathbf{b}} = 0 \Leftrightarrow \sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right) = 0 \\
\Leftrightarrow\ &\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i = \sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\mathbf{b} \\
\Leftrightarrow\ &\hat{\mathbf{b}} = \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i,
\end{aligned}
$$

assuming that the inverse of $\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i$ exists, of course.

Maximum Likelihood estimator $\hat{\mathbf{b}}$ has proven to share good optimality properties that even hold without the assumption of normality for the error terms $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$ and consequently for the response vector $\mathbf{y}_i$. Among these properties, it is worthwhile noticing that $\hat{\mathbf{b}}$ is consistent, asymptotically normal and fully efficient, provided though that the variance-covariance matrix $\mathbf{V}_i$ correctly specifies the $Var\left(\mathbf{y}_i\right)$. If $Var\left(\mathbf{y}_i\right) \neq \mathbf{V}_i$, $\hat{\mathbf{b}}$ is still consistent and asymptotically normal, but not efficient any longer.

Mean and Variance-covariance Matrix of the ML Estimator $\hat{\mathbf{b}}$

It is straightforward to conclude that the ML[5] estimator $\hat{\mathbf{b}}$, is also an unbiased estimator of the fixed-effects vector $\mathbf{b}$, since:

$$\begin{aligned} E\left(\hat{\mathbf{b}}\right) &= E\left[\left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{y}_i\right] \\ &= \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} E\left(\mathbf{y}_i\right), \end{aligned} \tag{5.11}$$

and by regarding that $E\left(\mathbf{y}_i\right) = \mathbf{X}_i \mathbf{b}$, we have

$$\begin{aligned} E\left(\hat{\mathbf{b}}\right) &= \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mathbf{b} \\ &= \mathbf{I}\mathbf{b} = \mathbf{b}, \end{aligned}$$

where $\mathbf{I}$ is the identity matrix.

Further, as concerns the calculation of a suitable formula for the variance-covariance

[5] This ML estimator is also the (Aitken's) generalized least squares (GLS) estimator of $\mathbf{b}$, as it can be easily shown.

matrix of $\hat{\mathbf{b}}$, we have:

$$
\begin{aligned}
Var\left(\hat{\mathbf{b}}\right) &= Var\left[\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right] \\
&= \left[\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right] Var\left(\mathbf{y}_i\right)\left[\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right]^t \\
&= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{V}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right)^t\left[\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\right]^t \underset{\mathbf{V}_i^{-1}\mathbf{V}_i=\mathbf{I}}{=} \\
&= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right)^t\left[\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^t\right]^{-1} \\
&= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \\
&= \mathbf{I}\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} = \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}. \qquad (5.12)
\end{aligned}
$$

[for the above calculations, we have used that if $\mathbf{A}$ a square matrix, then $\left(\mathbf{A}^{-1}\right)^t = \left(\mathbf{A}^t\right)^{-1}$, as well as that $\mathbf{V}_i^t = \mathbf{V}_i$, since $\mathbf{V}_i$ is a variance-covariance matrix].

It is worth noting that in the case where variance-covariance matrix $\mathbf{V}_i$ has been mispecified (i.e. $Var\left(\mathbf{y}_i\right) \neq \mathbf{V}_i$), equation (5.10) does no longer provide a valid estimate for $Var\left(\hat{\mathbf{b}}\right)$. In an attempt to handle situations of this kind, *Liang* and *Zeger (1986)* suggest using an alternative formula for $Var\left(\hat{\mathbf{b}}\right)$, given by:

$$
\begin{aligned}
&Var\left(\hat{\mathbf{b}}\right) = \\
&= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\hat{\mathbf{b}}\right)\left(\mathbf{y}_i-\mathbf{X}_i\hat{\mathbf{b}}\right)^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}.
\end{aligned}
$$

5.3.2 Prediction of Random Effects $\mathbf{u}_i$

We have already discussed in the context of the GLMM distinctions between fixed and random effects as well as the subject that concerns their estimation. What has been proposed for the estimation of the fixed and the random terms of the GLMM also applies to its extension, the Laird-Ware model. Hence, although in most practical longitudinal studies it is of primarily interest to estimate the fixed effects vector $\mathbf{b}$ of the model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$, $(i = 1, ..., m)$, and the variance components (the elements of the variance-covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$), on the other hand there are situations where it is often useful to calculate estimates (usually called predictors) for the random effects $\mathbf{u}_i$ as well.

Harville (1976) obtained estimates of the, m in total, subject-specific random effects $\mathbf{u}_i$, $(i = 1, ..., m)$ using an extended version of the Gauss-Markov theorem for random effects that produces the best linear unbiased predictor (BLUP) of $\mathbf{u}_i$ (denoted by $\mathbf{u}_{i,BLUP}$). The procedure is similar to that presented in Chapter 3, where we have shown derivation of the best linear unbiased predictor for the random vector $\mathbf{u}$ of mixed-effects model $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$, therefore we give the resulting formula for $\mathbf{u}_{i,BLUP}$, omitting the unnecessary calculations. To this end, if $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$, and both $\mathbf{R}_i$ and $\mathbf{D}$ are assumed to be known, then the BLUP of random effects $\mathbf{u}_i$, $(i = 1, ..., m)$ has been shown to be (*Laird* and *Ware, 1982*):

$$\mathbf{u}_{i,BLUP} = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{BLUE}\right), \tag{5.13}$$

where $\mathbf{b}_{BLUE}$ as usual denotes the best linear unbiased estimator (BLUE) of $\mathbf{b}$, $\mathbf{b}_{BLUE} = \left(\sum\limits_{i=1}^{m} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \sum\limits_{i=1}^{m} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i$. (Note that $\mathbf{b}_{BLUE}$ and the ML estimator $\hat{\mathbf{b}}$ of 5.2 are identical).

Another possible way to come up with a predictor for $\mathbf{u}_i$, is to use an extended likelihood approach. This likelihood-based procedure produces a ML solution for $\mathbf{u}_i$ by optimizing the log-likelihood that results from all two-dimensional random vectors

$(\mathbf{y}_i, \mathbf{u}_i)^t$, $i = 1, \ldots, m$. To illustrate the method, let us consider again GLMM (5.1). Moreover, consider the conditional random variable of $\mathbf{y}_i$ given $\mathbf{u}_i$, namely $\mathbf{y}_i \mid \mathbf{u}_i$. Under the usual normality assumption imposed for model (5.1), it is clear that $\mathbf{y}_i \mid \mathbf{u}_i$ is also normally distributed, with:

$$\mathbf{y}_i \mid \mathbf{u}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i, \mathbf{R}_i\right). \tag{5.14}$$

This gives the following expression for the probability density function of $\mathbf{y}_i \mid \mathbf{u}_i$:

$$f\left(\mathbf{y}_i \mid \mathbf{u}_i\right) = (2\pi)^{-\frac{n_i}{2}} \mid \mathbf{R}_i \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right\}$$

Now, the distribution of random effects term $\mathbf{u}_i$, $i = 1, \ldots, m$ is $N_q\left(\mathbf{0}, \mathbf{D}\right)$ with corresponding density given as:

$$f\left(\mathbf{u}_i\right) = (2\pi)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i\right\}.$$

By definition, the conditional density function of $\mathbf{y}_i$ given $\mathbf{u}_i$ is:

$$f\left(\mathbf{y}_i \mid \mathbf{u}_i\right) = \frac{f\left(\mathbf{y}_i, \mathbf{u}_i\right)}{f\left(\mathbf{u}_i\right)},$$

thus, consequently the joint density function of $(\mathbf{y}_i, \mathbf{u}_i)^t$ which also corresponds to the likelihood of $(\mathbf{y}_i, \mathbf{u}_i)^t$ will be:

$$f\left(\mathbf{y}_i, \mathbf{u}_i\right) = f\left(\mathbf{u}_i\right) \times f\left(\mathbf{y}_i \mid \mathbf{u}_i\right). \tag{5.15}$$

Moreover, the (joint) likelihood of all pairs $(\mathbf{y}_i, \mathbf{u}_i)^t$, $i = 1, \ldots, m$ is calculated as:

$$L_m = \prod_{i=1}^{m} f\left(\mathbf{y}_i, \mathbf{u}_i\right) =$$

$$= \prod_{i=1}^{m} (2\pi)^{-\frac{n_i+q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \mid \mathbf{R}_i \mid^{-\frac{1}{2}} \times$$

$$\times \exp\left\{-\frac{1}{2}\left[\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i + \left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right]\right\},$$

which becomes, without the constant $(2\pi)^{-\sum\limits_{i=1}^{m}(n_i+q/2)}$:

$$L_m = \mid \mathbf{D} \mid^{-\frac{m}{2}} \left(\prod_{i=1}^{m} \mid \mathbf{R}_i \mid\right)^{-\frac{1}{2}} \times$$
$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i + \left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right]\right\}.$$

The corresponding log-likelihood $\lambda_m = \ln L_m$ of L_m is:

$$\begin{aligned}
\lambda_m = \ln L_m = \ln \prod_{i=1}^{m} f\left(\mathbf{y}_i, \mathbf{u}_i\right) = \\
= -\frac{m}{2}\ln \mid \mathbf{D} \mid - \frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{R}_i \mid - \\
-\frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i + \left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right].
\end{aligned}$$

ML solution for $\mathbf{u}_i$ is obtained by differentiating this log-likelihood with respect to $\mathbf{u}_i$ and setting the result equal to zero. We start with the calculation of $\partial \ln L_m / \partial \mathbf{u}_i$. It is:

$$\begin{aligned}
\frac{\partial \ln L_m}{\partial \mathbf{u}_i} &= \\
&= -\frac{1}{2}\frac{\partial\left(\sum\limits_{i=1}^{m}\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i\right)}{\partial \mathbf{u}_i} - \frac{1}{2}\frac{\partial\left[\sum\limits_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)\right]}{\partial \mathbf{u}_i} \\
&= -\frac{1}{2}\frac{\partial \mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i}{\partial \mathbf{u}_i} - \frac{1}{2}\frac{\partial\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)^t \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right)}{\partial \mathbf{u}_i},
\end{aligned}$$

and using (3.32) result of matrix derivation, we have:

$$\begin{aligned}
\frac{\partial \ln L_m}{\partial \mathbf{u}_i} &= -\frac{1}{2}2\mathbf{D}^{-1}\mathbf{u}_i - \frac{1}{2}2\left(-\mathbf{Z}_i^t\right)\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right) \\
&= \mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i\right) - \mathbf{D}^{-1}\mathbf{u}_i
\end{aligned}$$

$$\begin{aligned} &= \mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{y}_i - \mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{X}_i\mathbf{b} - \mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i\mathbf{u}_i - \mathbf{D}^{-1}\mathbf{u}_i \\ &= \mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right) - \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{-1}\right)\mathbf{u}_i. \end{aligned} \tag{5.16}$$

Setting the above partial derivative equal to zero yields the following ML estimator of $\mathbf{u}_i$:

$$\hat{\mathbf{u}}_i = \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right),$$

with the presumption that the inverse matrix of $\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{-1}$ exists. Since the above formula contains the unknown fixed-effects vector $\mathbf{b}$, the latter can be replaced by one of its estimators (i.e. either its ML estimator $\hat{\mathbf{b}}$, or its BLUE estimator $\mathbf{b}_{BLUE}$, given by the same formula). Hence, we may write:

$$\hat{\mathbf{u}}_i = \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right). \tag{5.17}$$

In fact, as it has already stated, it can be shown that the latter (ML) estimator of $\mathbf{u}_i$ is identical to the best linear unbiased predictor (BLUP) of $\mathbf{u}_i$, given in (5.13). Indeed, to prove this, we only have to show that:

$$\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1} \equiv \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}. \tag{5.18}$$

Recalling the important general result (3.36) (developed by Henderson for application to the GLMM of Chapter 3), and using it in the case of the Laird-Ware model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$, (i.e. replacing $\mathbf{Z}$, $\mathbf{R}$ and $\mathbf{V}$ of the GLMM by $\mathbf{Z}_i$, $\mathbf{R}_i$ and $\mathbf{V}_i$ of the Laird-Ware model, respectively) one may see that (5.18) holds. Thus, estimator $\hat{\mathbf{u}}_i$ can be equivalently written as:

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right), \tag{5.19}$$

or

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_{BLUE}\right), \tag{5.20}$$

since $\hat{\mathbf{b}} = \mathbf{b}_{BLUE} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{y}_i$.

Alternatively, and due to the fact that the subject-specific parameters $\mathbf{u}_i$, $(i = 1, ..., m)$ are assumed random, it is most natural and appealing as well to implement Bayesian techniques for their estimation (prediction). We have already shown that conditional on $\mathbf{u}_i$, the response vector $\mathbf{y}_i$ follows a n_i-variate normal distribution with mean vector $\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i$ and with variance-covariance matrix $\mathbf{R}_i$ (equation 5.14). In combination with the distributional assumptions for $\mathbf{u}_i$ and by using typical Bayesian methodology one can easily deduce (see, e.g. *Smith, 1973; Lindley* and *Smith, 1972*) that, conditionally on $\mathbf{y}_i$, $\mathbf{u}_i$ follows a $q-$variate normal distribution with mean:

$$\hat{\mathbf{u}}_i = \mathbf{D}\mathbf{Z}_i^t \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b} \right), \tag{5.21}$$

which may be used in practice as an estimator of random effects vector $\mathbf{u}_i$. Naturally, in practice, since the fixed-effects vector $\mathbf{b}$ is unknown, it has to be replaced by its ML or Best Linear Unbiased estimator, $\hat{\mathbf{b}}$ and $\mathbf{b}_{BLUE}$ respectively, thus yielding prediction formulas for the $\mathbf{u}_i$ identical to the BLUP and ML presented earlier. The resulting predictions for the random effects, using a Bayesian methodology, are called Empirical Bayes (EB) estimates.

Determining the Variance-covariance Matrix of $\hat{\mathrm{u}}_i$

As originally proposed by *Laird* and *Ware (1982)*, since $\hat{\mathbf{u}}_i$ is a linear function of $\mathbf{y}_i$ (like the fixed-effects estimate $\hat{\mathbf{b}}$), an analytic expression for its standard error may be easily derived as:

$$Var\left(\hat{\mathbf{u}}_i\right) = \mathbf{D}\mathbf{Z}_i^t \left\{ \mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \mathbf{X}_i^t \mathbf{V}_i^{-1} \right\} \mathbf{Z}_i \mathbf{D}.$$

To prove the above expression for the variation of $\hat{\mathbf{u}}_i$, observe that from (5.19) we obtain:

$$Var\left(\hat{\mathbf{u}}_i\right) = Var\left(\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i - \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\hat{\mathbf{b}}\right).$$

Then, by replacing $\hat{\mathbf{b}} = \left(\sum\limits_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum\limits_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i$, the above becomes:

$$\begin{aligned}
Var\left(\hat{\mathbf{u}}_i\right) = Var&\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i - \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\} \\
= \; & Var\left(\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right) + Var\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\} \\
& -Cov\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i, \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\} \\
& -Cov\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i, \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\}. \qquad (5.22)
\end{aligned}$$

We start to calculate now each part of (5.22), to come up with the desired expression for $Var\left(\hat{\mathbf{u}}_i\right)$. As concerns the first part, it is:

$$\begin{aligned}
& Var\left(\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right) = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}Var\left(\mathbf{y}_i\right)\left(\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\right)^t \\
= \; & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{V}_i\left(\mathbf{V}_i^{-1}\right)^t\mathbf{Z}_i\mathbf{D} = \mathbf{D}\mathbf{Z}_i^t\mathbf{I}\left(\mathbf{V}_i^t\right)^{-1}\mathbf{Z}_i\mathbf{D} = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}, \qquad (5.23)
\end{aligned}$$

since $\mathbf{V}_i$, $\mathbf{D}$ are variance-covariance matrices and thus $\mathbf{V}_i^t = \mathbf{V}_i$, $\mathbf{D}^t = \mathbf{D}$. For the second part, we have:

$$\begin{aligned}
& Var\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\} \\
= \; & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}Var\left(\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right)\cdot \\
& \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}Var\left(\mathbf{y}_i\right)\mathbf{V}_i^{-1}\mathbf{X}_i\cdot\\
&\quad\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{V}_i\mathbf{V}_i^{-1}\mathbf{X}_i\right)\cdot\\
&\quad\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)\cdot\\
&\quad\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}. \qquad (5.24)
\end{aligned}
$$

Finally, by performing simple calculus, the first of the last two terms of formula (5.22) becomes:

$$
\begin{aligned}
&Cov\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i,\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}Cov\left(\mathbf{y}_i,\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right)\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}Cov\left(\mathbf{y}_i,\mathbf{X}_1^t\mathbf{V}_1^{-1}\mathbf{y}_1+\ldots+\mathbf{X}_m^t\mathbf{V}_m^{-1}\mathbf{y}_m\right)\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}Cov\left(\mathbf{y}_i,\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right)\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}Var\left(\mathbf{y}_i\right)\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{V}_i\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}
\end{aligned}
$$

$$= \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}, \tag{5.25}$$

Similarly, the second of the last two terms of (5.22) becomes:

$$\begin{aligned}
&Cov\left\{\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i, \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i\right\}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} Cov\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i, \mathbf{y}_i\right)\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} Cov\left(\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{y}_i, \mathbf{y}_i\right)\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \mathbf{X}_i^t\mathbf{V}_i^{-1}Var\left(\mathbf{y}_i\right)\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{V}_i\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}, \qquad (5.26)
\end{aligned}$$

i.e., it is identical to (5.25). Thus, by considering (5.23), (5.24), (5.25) and (5.26), equation (5.22) can be rewritten as:

$$\begin{aligned}
&Var\left(\hat{\mathbf{u}}_i\right) = \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} + \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}-\\
&\mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} - \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}\\
= {} & \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D} - \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}.
\end{aligned}$$

Noticing that $\mathbf{D}\mathbf{Z}_i^t$ and $\mathbf{Z}_i\mathbf{D}$ are common factors in both parts of the above, we conclude that:

$$Var\left(\hat{\mathbf{u}}_i\right)=\mathbf{D}\mathbf{Z}_i^t\left\{\mathbf{V}_i^{-1}-\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right\}\mathbf{Z}_i\mathbf{D}. \qquad (5.27)$$

A drawback of the latter expression (5.27) is that it fails to take into account the variation in $\hat{\mathbf{u}}_i-\mathbf{u}_i$. Specifically, it underestimates the variability of $\hat{\mathbf{u}}_i-\mathbf{u}_i$ since it ignores the variation of random effects $\mathbf{u}_i$. Thus, in order to assess the variation in $\hat{\mathbf{u}}_i-\mathbf{u}_i$ it is preferable to use (*Laird* and *Ware, 1982*):

$$\begin{aligned}Var\left(\hat{\mathbf{u}}_i-\mathbf{u}_i\right)&=\mathbf{D}-Var\left(\hat{\mathbf{u}}_i\right)\underset{(5.27)}{=}\\ &=\mathbf{D}-\mathbf{D}\mathbf{Z}_i^t\left\{\mathbf{V}_i^{-1}-\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}\right\}\mathbf{Z}_i\mathbf{D}.\end{aligned}$$

5.4 The Covariance Structure of the "Laird-Ware" Model

In longitudinal studies, a very interesting and challenging problem is that of determining an adequate way to model the heterogeneity of longitudinal data, and not in a few situations applied statisticians may have primary interest in the covariance structure[6] of the proposed model. Since observations on different subjects are assumed independent, the covariance structure refers to the covariance pattern of measurements on the same subject. Characterizing this within-subject covariance structure essentially consists of specifying $Var\left(\mathbf{y}_i\right)=\mathbf{V}_i$ as a function of a (relatively small) number of parameters, in order to obtain a parsimonious parameterization of $\mathbf{V}_i$. Of course, this should be done at a stage prior to the inferential stage of the analysis. That is, we must conclude with

[6] This opportunity, in being able to choose among various covariance structures is in fact one of the most important advantages of the GLMM for longitudinal data, in comparison to classical methods (such as ANOVA or MANOVA), where we are forced to presume either a too restrictive variance-covariance matrix (i.e. ANOVA) or a too vague variance-covariance matrix (i.e. MANOVA).

a suitable form for $\mathbf{V}_i$ before fitting the model, estimating the model's parameters and conducting any tests of significance for the model's effects. It is important to model $\mathbf{V}_i$ carefully, since it affects both the efficiency of the estimates of the fixed effects, say $\hat{\mathbf{b}}$, and the validity of the estimate of $Var\left(\hat{\mathbf{b}}\right)$. To this end, various approaches have been presented, and this section intends to give some insight into the subject. Specifically, we present the most common choices of covariance structures, for the formulation of $\mathbf{V}_i$ matrix. Moreover, we will discuss standard methodology found in the literature, developed for the selection of the covariance structure of Laird-Ware model.

As already stated, variation of model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ is attributed to the model's two random sources, the random effects $\mathbf{u}_i$ and the random error $\boldsymbol{\varepsilon}_i$ chosen so that:

$$\mathbf{u}_i \sim N_q\left(0, \mathbf{D}\right) \qquad \boldsymbol{\varepsilon}_i \sim N_{n_i}\left(0, \mathbf{R}_i\right).$$

(In particular, $\mathbf{u}_i$ is used to describe the between-subjects variability, while $\boldsymbol{\varepsilon}_i$ describes the within-subjects variability). This is so, due to the fact that $Var\left(\mathbf{y}_i\right) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$, hence $Var\left(\mathbf{y}_i\right)$ is based on $\mathbf{D}$ and $\mathbf{R}_i$. Consequently, the specification the covariance structure for (5.1) model must be done through $\mathbf{D}$ and $\mathbf{R}_i$. In the following, we see how the variance-covariance matrices of $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ (that is $\mathbf{D}$ and $\mathbf{R}_i$), are parameterized in order to model the covariance structure of the response vector $\mathbf{y}_i$.

5.4.1 The Variance-Covariance Matrix of $\mathbf{u}_i$

Most work related to the covariance specification of Laird-Ware model, has been focused on specifying a suitable covariance pattern for variance-covariance matrix $\mathbf{D}$. Especially the recent advances in the statistical software have led to an impressively wide variety of possible candidate forms for parameterization of $\mathbf{D}$. A large selection of covariance structures are available, varying from the most simple to extremely complex, indicating the intensive work conducted in the last years on the specific area. The majority of these patterns is intended to deal with the problem of incorporating the (possible) serial correlation of within-subject measurements into the model. More specifically, according to *Littell et al. (2000)*, for most of these models the covariance between any two observations on the same subject depends only on the length of the time interval between those measurements (most often known as the 'lag'). Such correlation structures, where the correlations between measurements on the same subject depend only on the time difference of those measurements are called **stationary**. Also, usually the variance for each measurement is assumed to be constant over time.

In the sequel, we describe five of the most representative covariance structures, covering a wide range of the available choices, from a completely unstructured covariance pattern to more complex covariance patterns borrowed from Time Series analysis.

The Unstructured Structure: One example of a very simple covariance structure very often used for the specification of variance-covariance matrix **D** is the "unstructured" structure. The latter, specifies a variance-covariance matrix with no particular pattern, essentially leaving **D** completely unspecified, since that variance parameters within **D** are all different to each other, and are given by:

$$(UN): \qquad Cov\left(y_{ij}, y_{ik}\right) = \sigma_{jk}, \;\; for \; all \; i, j, k. \tag{5.28}$$

There is a major potential problem however with using the unstructure covariance structure; this generality in the parameterization brings the disadvantage of having to estimate a very large number of variance and covariance parameters. Indeed, for the $(q \times q)$ variance-covariance matrix **D**, an unstructured approach requires a total number of $q\left(q+1\right)/2$ variance parameters to specify **D**, raising the cost in computational efforts and thus leading to severe computational problems when fitting models of such structure.

The Compound Symmetric Structure: The compound symmetry (CS) structure specifies the variances and covariances included in each subject's variance-covariance matrix $Var\left(\mathbf{y}_i\right)$ to be:

$$(CS): \quad Cov\left(y_{ij}, y_{ik}\right) = \sigma_1^2 \; for \; j \neq k, \quad Var\left(y_{ij}\right) = \sigma^2 + \sigma_1^2 \; for \; j = k, \tag{5.29}$$

that is, observations on the same subject have homogeneous covariance σ_1^2 as well as homogeneous variance $\sigma^2 + \sigma_1^2$. As one notices, CS corresponds to the (unique) structure assumed by the Analysis of Variance (ANOVA) method for longitudinal data, presented in Chapter 2. Although the specific structure is not too realistic for modeling longitudinal data (since CS assumes equal covariances across time whereas for repeated observations on the same subject it is more possible covariances close in time to be greater than covariances distal in time), it has been established as a familiar choice for modeling longitudinal data via the Laird-Ware model.

There are various ways to specify a CS structure for $Var\left(\mathbf{y}_i\right)$; a typical procedure is to replace the $(n_i \times q)$ matrix $\mathbf{Z}_i$ in $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ by a $(n_i \times 1)$ vector of ones. In doing so, the random effects vector $\mathbf{u}_i$ consequently reduces to a univariate random variable, say u_i with $u_i \sim N\left(0, \sigma_1^2\right)$. By further assuming that $Var\left(\boldsymbol{\varepsilon}_i\right) = \sigma^2 \mathbf{I}_{n_i}$, it is:

$$\begin{aligned} Var\left(\mathbf{y}_i\right) &= \mathbf{Z}_i Var\left(\mathbf{u}_i\right) \mathbf{Z}_i^t + Var\left(\boldsymbol{\varepsilon}_i\right) \\ &= \mathbf{1}_{n_i}\sigma_1^2\mathbf{1}_{n_i}^t + \sigma^2\mathbf{I}_{n_i} = \sigma_1^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}^t + \sigma^2\mathbf{I}_{n_i} \\ &= \sigma_1^2\mathbf{J}_{n_i} + \sigma^2\mathbf{I}_{n_i}, \end{aligned} \tag{5.30}$$

where, as usual, $\mathbf{J}_{n_i}$ denotes a $(n_i \times n_i)$ matrix of ones and $\mathbf{I}_{n_i}$ is a $(n_i \times n_i)$ identity matrix.

Another commonly used way is to define $\mathbf{D}$ or $\mathbf{R}_i$, to be zero and define the other matrix to be compound symmetric, e.g.: $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i} + \sigma_1^2\mathbf{J}_{n_i}$.

The First-Order Autoregressive Structure: For data collected over time on the same subject (e.g. longitudinal data), often within-subject serial correlation is present. As *Jones (1993)* states, when data are serially correlated, observations that are closer together in time tend to have higher correlations than observations that are farther apart. Covariance structures, such as compound symmetry or the unstructured fail to take into account this possibility (unstructured structure makes no assumptions at all about covariance and correlation, while CS assumes equal correlations among measurements, no matter how far apart in time the measurements are placed).

The need aroused to model adequately this type of correlation, led to the introduction of familiar Time Series models into the analysis of longitudinal data via the GLMM. Especially, for equally spaced repeated measurements[7] (i.e. when time periods between observations are evenly spaced), usually the first-order autoregressive [AR(1)] covariance structure is employed.

By definition (e.g. see *Brockwell* and *Davis, 1996*), an AR(1) time series process with zero mean [i.e. $E\left(y_t\right) = 0$], is given by the equation:

$$y_t = \rho y_{t-1} + Z_t, \qquad t = 0, \pm 1, \ldots \tag{5.31}$$

where $\{Z_t\}$ is an i.i.d. random variable with $Z_t \sim N\left(0, \sigma^2\right)$. Further, Z_t is uncorrelated with y_s, for every $s < t$. ρ is called the autocorrelation coefficient and the restriction imposed on ρ is that $\mid \rho \mid < 1$. We call this model autoregressive due to the specific nature of equation (5.31), where the present value y_t is 'regressed' on the previous value y_{t-1}. On assuming that each within-subject's time sequence of repeated measurements specifies an AR(1) model of the form (5.31) it is easy to show that variances and covariances of the ith subject's repeated measurements $\mathbf{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{in_i}\right)^t$ are given by:

$$AR\left(1\right): \qquad Cov\left(y_{ij}, y_{ik}\right) = {\sigma^*}^2 \rho^{|j-k|}, \; for\ all\ j,k = 1,2,\ldots,n_i, \tag{5.32}$$

where ${\sigma^*}^2 \equiv \sigma^2/1 - \rho^2$. Indeed, under an AR(1) formulation we have that the following

[7] For unequally spaced observations usually *continuous* time processes [e.g. a continuous AR(1) process, denoted as CAR(1)], are employed to describe within-subject serial correlation (for more details we refer to *Jones, 1993*).

equation holds for the $y_{i1}, y_{i2}, ..., y_{in_i}$ repeated observations:

$$y_{ij} = \rho y_{i,j-1} + Z_j, \qquad j = 1, 2, ..., n_i. \tag{5.33}$$

The variance of each y_{ij} is calculated, considering that $E\left(y_{ij}\right) = 0$, as:

$$Var\left(y_{ij}\right) = E\left(y_{ij}^2\right).$$

Hence, for determining $Var\left(y_{ij}\right)$ we equivalently have to find $E\left(y_{ij}^2\right)$. By multiplying both sides of equation (5.33) with y_{ij} and take expectations, we get:

$$\begin{aligned}
& y_{ij}^2 = y_{ij}\left(\rho y_{i,j-1} + Z_j\right) \\
& \Rightarrow E\left(y_{ij}^2\right) = E\left[y_{ij}\left(\rho y_{i,j-1} + Z_j\right)\right] \\
& \Rightarrow E\left(y_{ij}^2\right) = E\left[\left(\rho y_{i,j-1} + Z_j\right)\left(\rho y_{i,j-1} + Z_j\right)\right],
\end{aligned}$$

and since y_{ij} uncorrelated with Z_j, we have:

$$E\left(y_{ij}^2\right) = \rho^2 E\left(y_{ij}^2\right) + \sigma^2 \Rightarrow E\left(y_{ij}^2\right)\left(1-\rho^2\right) = \sigma^2 \Rightarrow E\left(y_{ij}^2\right) = \frac{\sigma^2}{1-\rho^2},$$

hence $Var\left(y_{ij}\right) = \sigma^2/\ 1 - \rho^2$. As concern covariances between two repeated measurements, say y_{ij} and $y_{i,j-k}$, observe that:

$$\begin{aligned}
Cov\left(y_{ij}, y_{i,j-k}\right) &= Cov\left(\rho y_{i,j-1} + Z_j, y_{i,j-k}\right) \\
&= Cov\left(\rho y_{i,j-1}, y_{i,j-k}\right) + Cov\left(Z_j, y_{i,j-k}\right) \\
&= \rho Cov\left(y_{i,j-1}, y_{i,j-k}\right),
\end{aligned}$$

since $Cov\left(Z_j, y_{i,j-k}\right) = 0$. For example, for $k = 1$, the covariance between two consecutive measurements y_{ij}, $y_{i,j-1}$ is $Cov\left(y_{ij}, y_{i,j-1}\right) = \rho Cov\left(y_{i,j-1}, y_{i,j-1}\right) = \rho Var\left(y_{i,j-1}\right) = \rho\sigma^2/1 - \rho^2 = \sigma^{*^2}\rho$. Hence, in any case $Cov\left(y_{ij}, y_{ik}\right) = \sigma^{*^2}\rho^{|j-k|}$ is true. For the AR(1) structure, observe that variances are equal $\left(\sigma^2/\ 1 - \rho^2\right)$, and the covariances decrease exponentially depending on the lag (separation in time of measurements): $| j - k |$.

Finally, closing the discussion on the AR(1) structure, we have to notice that the specific structure in contrast to others (i.e. unstructured) is an example of a very parsimonious structure since its parameterization requires only two parameters, ρ and σ^2, and thus adaptation of AR(1) type structures for modeling the within-subjects covariance in a GLMM for longitudinal data has become quite popular; *Potthoff* and *Roy (1964)* were the first to consider a first-order autoregressive AR(1) structure for equally spaced observations in a balanced growth curve situation. More recent, *Jones (1985, 1991), Mansour et al. (1985)* and *Pantula* and *Pollock (1985)* discuss longitudinal data analyses with a random subject effect and AR(1) error structure. A thorough review on analyzing longitudinal data via linear models as well as a general discussion on AR(1) models can be found in *Ware (1985)*.

The Toeplitz Structure: The Toeplitz structure (sometimes called the general autoregressive structure), resembles that of an order-one autoregressive structure. In particular, one may regard the Toeplitz structure as a generalization of the AR(1) structure. The Toeplitz structure [similarly to the AR(1) structure], assumes covariances that depend only on lags, but the difference now is that the covariances do not all depend on the parameter ρ, as is the case with the autoregressive structure. In the case of the AR(1) structure, all covariances partially based on ρ or powers of ρ. Here, covariances are more arbitrarily defined, the only restriction being that covariances along every diagonal be equal. Hence, while for the parameterization of a (4×4) matrix $\mathbf{D}$ as an AR(1) structure we would have used:

$$\mathbf{D} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

the same matrix, being parameterized by the Toeplitz structure would be expressed as:

$$\mathbf{D} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

Notice that the above, by taking σ^2 inside the matrix, can be rewritten as:

$$\sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & \sigma^2\rho_1 & \sigma^2\rho_2 & \sigma^2\rho_3 \\ \sigma^2\rho_1 & \sigma^2 & \sigma^2\rho_1 & \sigma^2\rho_2 \\ \sigma^2\rho_2 & \sigma^2\rho_1 & \sigma^2 & \sigma^2\rho_1 \\ \sigma^2\rho_3 & \sigma^2\rho_2 & \sigma^2\rho_1 & \sigma^2 \end{pmatrix},$$

and setting $\sigma^2\rho_1 = \sigma_1$, $\sigma^2\rho_2 = \sigma_2$, $\sigma^2\rho_3 = \sigma_3$ results in the following form for variance-covariance matrix $\mathbf{D}$:

$$\mathrm{D} = \begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix},$$

which is just another, alternative way to express a variance-covariance matrix of Toeplitz form [for more on Toeplitz matrices we refer to *Grenander (1958)* and *Littell et al. (2000)*]. Thus, variances and covariances of a matrix of Toeplitz covariance structure are namely:

$$(TOEP): \quad Cov\,(y_{ij}, y_{ik}) = \sigma_{|j-k|} \; for \; j \neq k, \qquad Var\,(y_{ij}) = \sigma^2 \; for \; j = k. \qquad (5.34)$$

The Banded Type Structures: Structures as the AR(1) or Toeplitz, allow for the presence of correlation among observations made on the same subject. But there is the possibility the correlation between two widely separated observations to be negligible. For situations like this, it may be appropriate to 'band' the **D** matrix by setting correlations between the observations that are widely separated in time to zero. For example, by choosing to parameterize **D** as [where say **D** is of order (4 × 4)]:

$$\mathrm{D} = \begin{pmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{pmatrix},$$

we essentially say that correlation is present only between consecutive observations, whereas correlation between observations separated by two or more time intervals is practically zero. This specific banded model is known as the one-dependent model, which in its general form expresses that:

$$Cov\left(y_{ij}, y_{i,j+1}\right) = \sigma_{|j-(j+1)|} = \sigma_1, \qquad Cov\left(y_{ij}, y_{i,j+k}\right) = 0 \; for \; k > 1, \qquad Var\left(y_{ij}\right) = \sigma^2.$$

Of course, depending on how far in time we want correlation to exist, the one-dependent model can be extended to a two-dependent or higher dependent model, producing each time different banded covariance structures (e.g. the two-dependent covariance structure states that observations that are one and two steps apart in time are correlated). Table 5.1 summarizes the diferent covariance structure specifications that have been discussed.

Table 5.1

Covariance structure specifications considered in the literature for the Laird-Ware model

Structure	*Formula*
Unstructured	$\begin{pmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}$
Compound Symmetry	$\begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{pmatrix}$
First – Order Autoregressive	$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$
Toeplitz	$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$
Banded Main Diagonal	$\begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{pmatrix}$

(continued)

Table 5.1 (continued)

Structure	*Formula*
Toeplitz with Two Bands	$\begin{pmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{pmatrix}$
Heterogeneous CS	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$
Heterogeneous AR(1)	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$
First − Order Factor Analytic	$\begin{pmatrix} \lambda_1^2 + d_1 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2 + d_2 & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + d_3 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2 + d_4 \end{pmatrix}$
Huynh − Feldt	$\begin{pmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2} - \lambda & \frac{\sigma_1^2+\sigma_3^2}{2} - \lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2} - \lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2} - \lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2} - \lambda & \frac{\sigma_3^2+\sigma_2^2}{2} - \lambda & \sigma_3^2 \end{pmatrix}$

5.4.2 The Variance-Covariance Matrix of ε_i

In contrast to the various options available when parameterizing the between-subjects variance-covariance matrix $\mathbf{D}$, the assumptions on (within-subjects) variance-covariance matrices $\mathbf{R}_i$ are rather simplified. Although the various covariance structures presented in section 5.4.1 for the modeling of $\mathbf{D}$ can be similarly used for describing the covariance structure of random error ε_i, through $\mathbf{R}_i$, often we resort to less complex solutions. In fact, the most common choice for $\mathbf{R}_i$ is the simple covariance structure:

$$\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i},$$

where $\mathbf{I}_{n_i}$ the $(n_i \times n_i)$ identity matrix and σ^2 is a (unique) variance parameter used to describe the within-subjects variability in the data. The simple structure specifies that even measurements on the same subject are independent, and all measurements have homogeneous variance, i.e.:

$$Simple\ (SIM): \quad Cov\,(y_{ij}, y_{ik}) = 0\ for\ j \neq k, \quad Var\,(y_{ij}) = \sigma^2\ for\ j \neq k.$$

In other words, this parameterization suggests that the variance is the same across each ith $(i = 1, ..., m)$ individual's separate measurements $y_{i1}, y_{i2}, ..., y_{in_i}$ and furthermore, these measurements were taken sufficiently far apart in time so that the possibility of correlation among them is practically considered negligible. Hence, this model choice essentially assumes that all the variability in the data which is not taken into account by the random effects $\mathbf{u}_i$ (that models the between-subjects variability), is purely measurement error.

The Laird-Ware model with this specific additional restriction of $\varepsilon_i \sim N_{n_i}\left(\mathbf{0},\sigma^2 \mathbf{I}_{n_i}\right)$ is called the **conditional-independence model**, due to that conditional on the random effects $\mathbf{u}_i$, it is $Var\,(\mathbf{y}_i) = \sigma^2 \mathbf{I}_{n_i}$ which implies that the n_i responses on individual i are independent. Note that this commonly used variant of Laird-Ware model assumes homogeneity of both variance terms $\mathbf{u}_i$, ε_i (constant variance-covariance matrices $\mathbf{D}$, $\sigma^2 \mathbf{I}$

for all subjects i), while model (5.1) assumes homogeneity of variance only for $\mathbf{u}_i$.

Remark 5.1: *The above methodology for the parameterization of* $\mathbf{D}$ *and* $\mathbf{R}_i$ *illustrate what has become standard in the modeling of longitudinal data via the Laird-Ware model. Of course, it should be mentioned that there are other approaches for introducing serial correlation into the model, except of the one already described which consists of specifying the simple, conditional-independence structure for the covariance of* $\boldsymbol{\varepsilon}_i$, *using the random effects* $\mathbf{u}_i$ *to give any additional structure to* $Var(\mathbf{y}_i)$. *Jones (1993), for example, introduces serial correlation into the within-subject errors* $\boldsymbol{\varepsilon}_i$ *by letting* $\mathbf{R}_i$ *to have a first-order autoregressive* $[AR(1)]$ *structure, rather than just assuming that the errors and consequently observations on the same unit are uncorrelated. For alternative parameterizations of the within-subject variation see also Verbeke et al. (1998) and Lesaffre et al. (2000).*

In general, several authors have considered various alternative ways for the parameterization of $\mathbf{D}$ *and* $\mathbf{R}_i$, *including the study of rather extreme situations, such as the one of setting one of* $\mathbf{D}$, $\mathbf{R}_i$ *equal to zero (e.g. setting* $\mathbf{D} = \mathbf{0}$ *essentially corresponds to elimination of the* $\mathbf{u}_i$ *term) and specifying a covariance structure for the other remaining term.*

5.4.3 Selecting the "Best" Covariance Structure

The feasibility to choose among a broad variety of covariance structures for Laird-Ware model is one of the greatest advantages this model accommodates into the analysis of longitudinal data. By selecting a structure that best fits the true covariance of the data results in obtaining the best possible efficient estimates of fixed effects $\mathbf{b}$, and consequently in performing more powerful and valid rests upon the latter effects.

On the other hand however, this availability has an obvious impact on the implementation of Laird-Ware model since modeling can be considerably more complicated under the additional burden of having to select between more than one covariance structures. Choosing the most appropriate structure is not an easy task, and various methods are now utilized, including indices of goodness-of-fit, comparisons of covariance estimates and graphical techniques. In particular, for practical applications usually likelihood-based methods are used. Note though, that a straightforward application of the model's likelihood function L by forming a likelihood statistic based entirely on L cannot be seen as a suitable choice for model comparison, since its value will always increase as more variance components are added. Hence, other alternative approaches for model choice have become standard in use. Models with the same fixed effects, but with different covariance structures can be compared using again statistics based on the likelihood function, this time though adjusting for the number of variance parameters. The other approach is to use likelihood ratio tests (LRTs). In the following sections we present both methods, describing the formulas of statistics and specifying the situations each of those methods apply.

5.4.3.1 Likelihood Ratio Test (LRT)

The likelihood ratio test (LRT) for selecting among two models, the one with the 'best' covariance structure applies only to certain circumstances. More specific, it provides valid inferences in the special situation where one model is a 'constrained' version of the other. By this, we mean that the two models are nested, i.e. the simpler model can be obtained by restricting some of the variance parameters in the more complex model. Under this notion, the LRT can be used to test the null hypothesis that the model with more variance parameters is not a significantly better fit than the simpler model with the fewer parameters.

As concern now the formula of the statistic, if by ℓ_1 we denote the value of -2 times the logarithm of the likelihood from the first model, that is $\ell_1 = -2\log L_1$ and accordingly

ℓ_2 is the value from the second model (i.e. $\ell_2 = -2 \log L_2$), then the LRT is given by[8]:

$$\ell_1 \left(\hat{\theta}\right) - \ell_2 \left(\hat{\theta}\right) \sim x_d^2, \tag{5.35}$$

where $\hat{\theta}$ is the ML (or REML) estimate of the unknown variance components θ that maximizes L_1 and L_2 respectively, and d is the difference in the number of variance components fitted between the two models. Finally, x_d^2 denotes the chi-square distribution with d degrees of freedom. [To clear things a little, what is actually done here is first fitting the two nested model, then obtain estimations of each model's variance components θ and finally (using these estimated $\hat{\theta}$) calculate statistic 5.35]. A large value of the difference $\ell_1 \left(\hat{\theta}\right) - \ell_2 \left(\hat{\theta}\right)$ leads to the rejection of the null hypothesis that the two models are the same (i.e. the extra variance parameters do not improve the fit) and thus conclude that the best model is the second, the one with the extra parameters.

5.4.3.2 Akaike's Information Criterion (AIC)

So far we have discussed only comparisons of nested covariance structures, for which the likelihood ratio test can be validly used. However, for comparisons between non-nested models the LRT is not suitable for making inferences. Alternatively, covariance structures can be objectively compared using other selection procedures, such as Akaike's Information Criterion (AIC) and a modified AIC, the Schwarz's Information Criterion (SIC).

AIC, which initially developed for decision theory (see *Akaike, 1974*), is a statistic based on the (maximized) likelihood, with the advantage of penalizing though this likelihood for the number of parameters fit to the data to avoid overfitting. Under this perspective, AIC statistic is given by:

$$AIC = \log L \left(\hat{\theta}\right) - q, \tag{5.36}$$

[8] It has proven to be more convenient when constructing likelihood ratio tests, to work with -2 loglikelihood than with the likelihood itself.

where once again $\hat{\theta}$ denotes the ML/REML estimates of variance components θ, and q is the number of the (estimated) variance components. The model which has the largest value of AIC is selected as the 'best' model (i.e. the model with a covariance structure that fits best to the data).

Some authors find it more convenient, instead of using (5.36) to work with another formula for AIC that incorporates $\ell = -2\log L$ into the statistic. Thus, in many texts, the following formula for AIC can be met:

$$AIC = -2\left[\log L\left(\hat{\theta}\right) - q\right] = -2\log L\left(\hat{\theta}\right) + 2q = \ell\left(\hat{\theta}\right) + 2q. \tag{5.37}$$

In this case, due to the modification, best model is the model with the lowest AIC value.

5.4.3.3 Schwarz's Information Criterion (SIC)

Schwarz (see *Schwarz, 1978*), suggested a modification of Akaike's Information Criterion that, as he has proven, increases penalty for overfitting compared to AIC. Schwarz Information Criterion (or Schwarz Bayesian Information Criterion) in addition to the number of estimated variance parameters q, is formed in such a way so that to take into account and the (total) number of repeated observations $N = \sum_{i=1}^{m} n_i$. More precisely, SIC (or SBC) is expressed as:

$$SIC = \log L\left(\hat{\theta}\right) - \frac{q}{2}\log N, \tag{5.38}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, q the number of the variance parameters and N the total number of observations in the model. For the case where the estimations of θ, $\hat{\theta}$, are obtained by the restricted maximum likelihood method, the SIC statistic is slightly modified to:

$$SIC = \log L\left(\hat{\theta}\right) - \frac{q}{2}\log\left(N - p\right), \tag{5.39}$$

where the extra term p is the number of the fixed effects in the model. As was the case with AIC, once again the model with the largest SIC value is the most preferable.

5.4.3.4 Other Information Criteria

More complicated information criteria for selecting an appropriate covariance structure have been the subject of ongoing research in the recent years. These criteria though, are of questionable applicability, since in most practical studies standard choice criteria (i.e. AIC, SIC) are used instead. Among them, worthwhile mentioning is the information criterion proposed by *Bozdogan (1987)*, who suggests increasing the penalty term slightly more, and called the resulting statistic the consistent Akaike Information Criterion (CAIC):

$$CAIC = \log L\left(\hat{\theta}\right) - \frac{q}{2}\left(1 + \log N\right). \tag{5.40}$$

Another, less frequently used criterion, given by the SAS (*Littell et al., 1996*) Procedure PROC MIXED, is the Hannan & Quinn information criterion (HQIC):

$$HQIC = \log L\left(\hat{\theta}\right) - q \log\left(\log N\right). \tag{5.41}$$

(For more information on the above criteria, the interested reader is referred to the SAS manuals).

5.4.3.5 Graphical Techniques

Complementary to the above criteria, informal graphical techniques that have developed for deciding among various candidate covariance structures are available. In particular, graphs such as the Draftman's display (see Section 4.3) or the empirical semivariogram (see sub-Section 5.7.2) can help with the specification of the Laird-Ware model's covariance structure. As is known, repeated measurements on the same subject are usually (positively) correlated. Moreover, measurements taken close together in time are potentially more highly correlated than those taken far apart in time. Inspection of a Draftman's display and/or a semivariogram allows us to determine the appropriate (within-subject) covariance structure by visualizing the relationships and (possible) correlations between the repeated measurements within each subject. Once a correlation pattern has been detected from these plots, a suitable covariance structure may be used to allow for this pattern of correlation to be incorporated in the model.

5.4.4 Concluding Remarks

Although the ability for various parameterizations makes clear an advantage of the 'Laird-Ware' model over other longitudinal data modeling approaches, however it is not usually practical to test a large number of covariance structures in a single application. Especially, covariance structures of the most complex form are rarely used in practice. Most common strategies suggest to start with the fit of simple structures, such as the compound symmetric or the first-order autoregressive. More complex structures can be tested and should be accepted only if they prove to be significantly better, compared to the simpler structures. As *Brown* and *Prescott (1999)* point out, numerical evaluations have shown that for many real datasets, especially those with a few repeated measurements on each subject, the estimates of the fixed effects $\mathbf{b}$ differ little between models using different covariance structures. In any case, one usually utilizes either a general unstructured variance-covariance matrix $\mathbf{D}$ (i.e. a symmetric positive definite $(q \times q)$ matrix which does not assume the random effects matrix $\mathbf{D}$ to be of any specific form), or a compound symmetric variance-covariance matrix $\mathbf{D}$.

As far as concerns the selection of a covariance structure among various models, the usual procedure is to compare the values of AIC and SIC on all models, that is choose the model that exceeds the largest value on both information criteria. This is quite possible to occur, since in many cases the two criteria are likely to come up with equivalent results. In any other case, where one model has the largest AIC value and another model has the largest SIC value, what is practically done is to trust one of the two information criteria. Closing the discussion on information criteria, it is worthwhile mentioning a recent criticism to the two widely applied criteria, AIC and SIC. In *Keselman (1998)* for example, a study was conducted in order to see how effective the latter information criteria are in detecting the correct covariance structure. The authors simulated longitudinal data arising from specific population covariance structures, with a total number of eleven covariance structures being fit with the SAS procedure, PROC MIXED. AIC and SIC were both utilized to detect each time the correct covariance structure. Unfortunately, the results indicated that neither criterion was much effective in finding the correct structure. Specifically, the Akaike criterion only resulted in the correct structure being selected 47 percent of the time, whereas the Schwarz criterion resulted in the correct structure being selected just 35 percent of the time. Authors present as a possible explanation for the poor performance of both criteria, the fact that the (wrong) structures chosen by the criteria might be very close approximations to the true covariance structures.

Nonetheless, despite criticism, information criteria and especially AIC and SIC still remain the most frequently used approaches on testing for the best covariance structure between various Laird-Ware models, and are included as a standard tool in most statistical software for modeling longitudinal data through mixed effects analysis.

5.5 Unknown Variance Components

So far we have considered estimation of fixed effects vector $\mathbf{b}$ and prediction (estimation) of random effects $\mathbf{u}_i$, $(i = 1, ..., m)$ of the Laird-Ware model (5.1), on the presumption that the covariance structure of the latter model is known. That is, when the variance components $\boldsymbol{\theta}$ (i.e. the elements of the variance-covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$) are known, and thus $\mathbf{D}$, $\mathbf{R}_i$ and $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \mathbf{R}_i$ are considered to be known, closed-form expressions for the estimates of fixed and random effects $\mathbf{b}$ and $\mathbf{u}_i$ can be derived without particular difficulties (see section 5.3).

However, the specific assumption rarely holds in practice, as it has been already mentioned in Chapter 3. In most longitudinal studies conducted via the Laird-Ware model, variance-covariance matrices $\mathbf{D}$, $\mathbf{R}_i$ are unknown. Hence, since equations for the estimators of fixed and random effects involve these unknown matrices, estimates of $\mathbf{D}$, $\mathbf{R}_i$ and $\mathbf{V}_i$ are necessary in order to derive closed form solutions for the estimators of $\mathbf{b}$

and $\mathbf{u}_i$. In this situation, where all the variance components $\boldsymbol{\theta}$ of the Laird-Ware model are not known, but an estimate, say $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is available, the common strategy to proceed with fixed and random effects estimation is to set $\hat{\mathbf{V}}_i = \mathbf{V}_i\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{Z}_i\mathbf{D}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{Z}_i^t + \mathbf{R}_i\left(\hat{\boldsymbol{\theta}}\right)$ and estimate $\mathbf{b}$ and $\mathbf{u}_i$ using again the already derived equations of section 5.3, this time replacing $\mathbf{V}_i$, $\mathbf{D}$ and $\mathbf{R}_i$ with their estimates $\hat{\mathbf{V}}_i = \mathbf{V}_i\left(\hat{\boldsymbol{\theta}}\right)$, $\hat{\mathbf{D}} = \mathbf{D}\left(\hat{\boldsymbol{\theta}}\right)$ and $\hat{\mathbf{R}}_i = \mathbf{R}_i\left(\hat{\boldsymbol{\theta}}\right)$. Thus, we may write:

$$\hat{\mathbf{b}}\left(\hat{\boldsymbol{\theta}}\right) = \left(\sum_{i=1}^{m} \mathbf{X}_i^t\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t\hat{\mathbf{V}}_i^{-1}\mathbf{y}_i, \tag{5.42}$$

to be the ML estimator of fixed-effects vector $\mathbf{b}$, and accordingly:

$$\hat{\mathbf{u}}_i\left(\hat{\boldsymbol{\theta}}\right) = \hat{\mathbf{D}}\mathbf{Z}_i^t\hat{\mathbf{V}}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right), \tag{5.43}$$

as the ML estimator of the random effects $\mathbf{u}_i$, $(i = 1, \ldots, m)$, under the assumption that model's variance components (and thus model's variance-covariance matrices) are unknown.

The literature that is concerned with the topic of estimation of variance components in mixed models is quite extensive. Although several methods are available, much of the attention has been focussed onto two competitive methods; one is the standard maximum likelihood (ML) method, introduced to variance components estimation by *Hartley* and *Rao (1967)* and the other is the modified ML technique known as the restricted maximum likelihood (REML) method (*Patterson* and *Thompson, 1971*). Both methods as well as their adaptation on the GLMM for the estimation of the model's variance components have been reviewed in sections 3.3.4.1 and 3.3.4.2, therefore we will avoid any additional comments on the two methods. The common approach to perform estimation of variance components $\boldsymbol{\theta}$ of the Laird-Ware model is in combination with the estimation of the fixed effects $\mathbf{b}$, such that both estimation procedures are conducted in a simultaneous way. As stated before (Chapter 3), the troublesome complication arising when trying to obtain simultaneously estimates of $\mathbf{b}$ and $\boldsymbol{\theta}$, is that in the end we conclude with a system of equations of no closed form, since equation that gives the estimator of $\mathbf{b}$ is a function of the unknown variance components $\boldsymbol{\theta}$, and similarly equation that provides variance components estimator $\hat{\boldsymbol{\theta}}$ contains the unknown $\mathbf{b}$. Consequently, no simple one-step solution can be obtained, as was the case with estimation of fixed and random effects when variance components are known. The only possible way to proceed is to employ numerical procedures of iterative nature to derive the desired estimates. These iterative schemes may be satisfactory applied to perform both ML and REML estimation of the variance components. Most notable among these iterative algorithms are the widely applied Newton-Raphson (N-R) algorithm, and the more recently developed Expectation-Maximization (EM) algorithm. [Others however, (see e.g. *Vonesh* and *Carter, 1987*), have proposed non-iterative estimation procedures for the Laird-Ware model parameters].

In the sequel, we attempt to describe the current status and recent developments associated with the ML/REML estimation of covariance structure (i.e. the variance components) of the Laird-Ware model for longitudinal data via EM and N-R algorithms. To this end, the remainder of the current section is organized as follows; in subsection 5.5.1 we present the basic theory of the EM algorithm, while in the next subsection 5.5.2 application of the latter numerical iterative algorithm either for ML estimation (subsections 5.5.2.1 and 5.5.2.2) or REML estimation (subsection 5.5.2.3) is discussed. In subsections 5.5.3 and 5.5.4 the basics of the other (restricted) likelihood maximization method, the N-R algorithm are reviewed. Further, we consider all necessary formulas for the implementation of the iterative N-R to derive estimates for both ML estimates (subsection 5.5.5.1) and REML estimates (subsection 5.5.5.2) for the variance components in mixed effects model for longitudinal data (5.1).

5.5.1 The EM Algorithm

A major difficulty of maximum likelihood estimation is that in many situations, no theoretical solution on the likelihood equations is available, leaving no other choice than to resort to numerical optimization techniques. A widely applied general approach to numerical computation of ML estimates is the so called: *Expectation-Maximization algorithm* (EM algorithm). The EM algorithm has become in recent years one of the most well-known and popular techniques in applied statistics. Since its definition by *Dempster et al. (1977)* in their fundamental article, it has been used successfully in a wide variety of applications, from mixture models density estimation to maximum likelihood estimation of variance components.

EM is a general-purpose algorithm for both ML and REML estimation in a wide variety of situations, best described as 'incomplete-data' problems. In simple words, the EM algorithm is essentially an iterative, numerical technique that is based on the fact that if certain data values were not missing, ML/REML estimation would computationally be much easier. Under this notion, one could say that the generality of EM algorithm is in doubt since it only applies to missing data problems. But this is not true, however; EM applies not only to evidently incomplete-data situations, but also in a whole variety of situations where the incompleteness of the data is not natural or evident. For example, in many occasions, even if the estimation problem is not one of incomplete-data, it is often preferable to express it as an incomplete-data problem and apply the EM algorithm. In this way, what is actually done is to associate with a given (assumed to be) incomplete-data problem for which ML estimation is extremely difficult to perform, a complete-data problem for which maximization of the likelihood is much easier. (It is worth noting that this is the approach followed in applying EM algorithm to Laird-Ware model, as we shall see in the sequel). Trying to make a general concluding remark on EM algorithm, we can say that it is a simple and easy to implement, iterative procedure for ML/REML maximization in incomplete (or at least assumed as being incomplete) data problems. As regards its implementation, we should note that despite its good properties (i.e. simplicity, numerical stability), it also suffers from several drawbacks, the main one being its very slow convergence in most situations.

In the following, we briefly describe the general framework of the EM algorithm, presenting the basic features of the algorithm in a simple way as possible. In do this, we first start by establishing some necessary additional notation. First of all, the 'incomplete data' term in its general form implies the existence of two sample spaces, $\mathcal{Y}_{obs}$ (observed sample space) and $\mathcal{Y}_{com}$ (complete sample space), as well as a many-to-one transformation (or mapping) from $\mathcal{Y}_{com}$ to $\mathcal{Y}_{obs}$. If we let H denote this transformation, then a random variable $\mathbf{Y}_{obs}$ defined in $\mathcal{Y}_{obs}$ is related to random variable $\mathbf{Y}_{com}$ defined in $\mathcal{Y}_{com}$ through this transformation, i.e.:

$$\mathbf{Y}_{obs} = H\left(\mathbf{Y}_{com}\right). \tag{5.44}$$

In addition, let us consider the 'observed' (incomplete) data $\mathbf{y}_{obs}$, which is a realization of the random variable $\mathbf{Y}_{obs}$, and accordingly the 'complete' data $\mathbf{y}_{com}$, realized from random variable $\mathbf{Y}_{com}$. Essentially, $\mathbf{y}_{com}$ consists of the observed data $\mathbf{y}_{obs}$, plus the missing data $\mathbf{y}_{mis}$. Thus, we may write $\mathbf{y}_{com} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$. In other words, we have just hypothesized a problem, where instead of observing some 'complete data', $\mathbf{y}_{com}$, we observe only a portion of these data, namely the 'incomplete data' $\mathbf{y}_{obs}$. The ultimate purpose is to estimate (usually via maximum likelihood estimation) a vector of some parameters θ, taking its values in the convex set Ω ($\theta \in \Omega$). If the density function of r.v. $\mathbf{Y}_{com}$ is $f\left(\mathbf{y}_{com}; \theta\right)$ and the density function of $\mathbf{Y}_{obs}$ is $f\left(\mathbf{y}_{obs}; \theta\right)$, then it is:

$$f\left(\mathbf{y}_{obs}; \theta\right) = \int_{H(\mathbf{y}_{com}) = y_{obs}} f\left(\mathbf{y}_{com}; \theta\right) d\mathbf{y}_{com}. \tag{5.45}$$

Now, let us denote by $L\left(\theta; \mathbf{y}_{com}\right) = f\left(\mathbf{y}_{com}; \theta\right)$ the complete-data likelihood function of parameter $\theta \in \Omega$, and by $L\left(\theta; \mathbf{y}_{obs}\right) = f\left(\mathbf{y}_{obs}; \theta\right)$ the observed-data likelihood function. As already mentioned, EM finds application in those statistical problems, where ML (or REML) estimation of θ is much simpler through maximization of the 'complete-data' log-likelihood $\ln L\left(\theta; \mathbf{y}_{com}\right)$ than maximization of the 'observed-data' log-likelihood $\ln L\left(\theta; \mathbf{y}_{obs}\right)$. Hence, the EM algorithm requires maximizing $\ln L\left(\theta; \mathbf{y}_{com}\right)$. But there is an evident complication in performing the latter maximization. The complete data $\mathbf{y}_{com}$ are not available (only the incomplete data are available in practice), therefore it is not possible to perform directly the optimization of the complete data log-likelihood

$\ln L(\theta; \mathbf{y}_{com})$. To overcome this, we replace $\ln L(\theta; \mathbf{y}_{com})$ by its conditional expectation, given the observed data $\mathbf{y}_{obs}$. This is the so-called expectation step of the algorithm.

In general, expectation-maximization algorithm is an iterative procedure that consists of two steps; the expectation step (E-step) and the maximization step (M-step). Since the E- and M-steps involve parameter θ which is unknown, it becomes necessary to use an iterative procedure which starts by providing some initial value for θ and then iterate between the two steps until convergence is reached. More specific, if $\theta^{(k)}$ denotes the estimate of θ at the kth iteration ($k = 0, 1, 2, ...$), then at the $(k+1)$st iteration the E-step calculates the expected log-likelihood of the complete data, say $Q\left(\theta, \theta^{(k)}\right)$, given the observed data $\mathbf{y}_{obs}$ and the current estimate $\theta^{(k)}$. The M-step then, simply finds a new estimate of θ, $\theta^{(k+1)}$ by maximizing $Q\left(\theta, \theta^{(k)}\right)$. Hence, the E- and M-steps of EM algorithm (for the kth iteration), can be presented as:

$$E-step: \quad \text{Calculate} \quad Q\left(\theta, \theta^{(k)}\right) = E\left[\ln L(\theta; \mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \theta^{(k)}\right] \;,$$

and

$$M-step: \quad \text{Choose } \theta^{(k+1)} \text{ that maximizes } Q\left(\theta, \theta^{(k)}\right).$$

After providing the initial estimates for θ, $\theta^{(0)}$, the above iterative scheme is repeated until the produced sequence $\left\{\theta^{(k)}\right\}$ reaches convergence. As usual, we say that we have reached convergence when the difference:

$$L_{obs}\left(\theta^{(k+1)}\right) - L_{obs}\left(\theta^{(k)}\right)$$

changes only by an arbitrarily small amount. In the above, L_{obs} denotes the likelihood of the observed data $\mathbf{y}_{obs}$ that we seek to maximize with respect to θ.

Closing this brief description of EM algorithm, we note that *Dempster et al. (1977)* among other general properties of EM, show that each iteration of the algorithm increases $L_{obs}(\theta) = L(\theta; \mathbf{y}_{obs})$, which in words verifies the monotonous increase of the likelihood.

5.5.2 The Implementation of the EM to the "Laird-Ware" Model

The EM algorithm developed by *Dempster et al. (1977)*, is a general purpose algorithm for obtaining maximum likelihood (ML) or restricted maximum likelihood (REML) estimates for some unknown parameters in the case of 'incomplete data' problems. In general, the algorithm consists of two distinct steps, the E- and M-steps. The E-step calculates the conditional expectation of the complete-data log-likelihood given the observed (incomplete) data and the current estimates of the unknown parameters, while the M-step computes the new estimates of the parameters by maximizing the conditional expectation obtained at the E-step. The process iterates between the E-step and the M-step, until the estimates reach convergence. Actually, as pointed out by *Dempster et al. (1977)*, EM algorithm can be applied even in estimation problems where there are no missing data, in the actual sense. A typical example of such type of implementation for the EM algorithm is the GLMM for longitudinal data, proposed by *Laird* and *Ware (1982)*.

Let us recall once again the (Gaussian) general linear mixed model (*Laird* and *Ware, 1982*):

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad (i = 1, ..., m)\,,$$

where as usual, $\mathbf{y}_i$ are vectors of length n_i containing the repeated measurements on the ith subject and $\boldsymbol{\varepsilon}_i$ are error vectors of the same length, independently distributed as $N_{n_i}(\mathbf{0}, \mathbf{R}_i)$. Further, $\mathbf{X}_i$ is a $(n_i \times p)$ design matrix of covariates and $\mathbf{b}$ is a corresponding $(p \times 1)$ vector of fixed effects while $\mathbf{Z}_i$ is a $(n_i \times q)$ design matrix which corresponds to the $(q \times 1)$ vector of subject-specific random effects $\mathbf{u}_i$ $[\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{D})]$. In the context of the specific model, to implement the EM algorithm, the observed (incomplete) data are considered to be of course the measurements $\mathbf{y}_i$, $(i = 1, ..., m)$ actually collected on each subject. However, the complete data are taken to be the observed data plus the unobservable random effects and random error terms, namely $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$. Thus, the missing data cannot be viewed as data in the traditional statistical sense. In this way, that is by treating the latent variables (such as $\mathbf{u}_i$, ε_i) as missing values, we are able to apply EM even in estimation problems where there are no missing data in actuality.

The first to describe implementation of EM algorithm for both ML and REML estimation of the variance components (in $\mathbf{D}$ and $\mathbf{R}_i$) of model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$ was *Laird* and *Ware (1982)*. Since then, many authors have discussed computational details of the above implementation. For example, *Laird et al. (1987)* reviewed and continued to examine the application of EM to the latter model. In addition, *Lindstrom* and *Bates (1988)* examined methods to improve computational efficiency of the EM algorithm, initially proposed by Laird and Ware. Their improvements involved a reparameterization of the covariance structure using a Cholesky decomposition to avoid problems with the parameter space, as well as computational improvements indented to speed the algorithm's convergence rate. Other work on the specific field includes *Jennrich* and *Schluchter (1986)*, *Meng* and *van Dyk (1998)*, *Jones (1993, Section* 2.6*)* and *McLachlan* and *Krishnan (1997, Section* 5.9*)*.

5.5.2.1 Maximum Likelihood Estimation via the EM Algorithm

We first discuss the use of Expectation-Maximization algorithm for maximum likelihood estimation, following ideas from *Laird* and *Ware (1982)*, who use the EM algorithm to estimate the fixed-effects vector $\mathbf{b}$ and the (unknown) variance components associated with the, introduced by the authors, general linear mixed model:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i, \qquad (i = 1, ..., m),$$

where

$$\mathbf{u}_i \sim N_q(0, \mathbf{D}) \quad \textit{and} \quad \varepsilon_i \sim N_{n_i}(0, \mathbf{R}_i).$$

Note that Laird and Ware came up with all necessary formulas that define the EM algorithm for the particular choice for the variance-covariance matrix of random errors

$\boldsymbol{\varepsilon}_i$ (conditional-independence model):

$$\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i},$$

where $\mathbf{I}_{n_i}$, as usual, denotes an identity matrix of order $(n_i \times n_i)$. They also illustrated their developed methodology on two datasets arising in the study of effects of atmospheric pollutants on pulmonary function.

Under this conditional-independence setting (i.e. $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$), *Laird* and *Ware (1982)* notice that if random terms $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ were observable then we could easily find simple closed-form ML estimates of the model's variance components (which in this case is the variance σ^2 and the elements of variance-covariance matrix $\mathbf{D}$), based on quadratic forms in $\mathbf{u}_i$ and ε_i $(i = 1, ..., m)$, given by the 'sufficient' statistics:

$$\hat{\sigma}^2 = \frac{\sum\limits_{i=1}^{m} \boldsymbol{\varepsilon}_i^t \boldsymbol{\varepsilon}_i}{N}, \tag{5.46}$$

and

$$\hat{\mathbf{D}} = \frac{\sum\limits_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^t}{m}, \tag{5.47}$$

where $N = \sum\limits_{i=1}^{m} n_i$ denotes the total number of measurements, and m is the number of subjects. Evidently, as it should be, equation (5.46) produces a scalar while from equation (5.47) we obtain a matrix of order $(q \times q)$.

However, random vector $\mathbf{u}_i$ and random error term ε_i cannot be observed and thus Laird and Ware by treating them as missing data, gave an interesting variation[9] of the EM algorithm to estimate $\mathbf{b}$, σ^2 and $\mathbf{D}$. In this case the E-step of the iterative algorithm,

[9] As we shall see later, the algorithm described by Laird and Ware corresponds to a particular version of EM algorithm and not to the EM originated by Dempster et al. (1977).

say at the $(k+1)$st iteration, is given by the following equations:

$$
\begin{aligned}
&\underline{E-step}:\\
&E\left\{\sum_{i=1}^{m}\boldsymbol{\varepsilon}_i^t\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\} = \sum_{i=1}^{m}\left\{trVar\left[\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right] + \hat{\boldsymbol{\varepsilon}}_i^t\hat{\boldsymbol{\varepsilon}}_i\right\},\\
&and\\
&E\left\{\sum_{i=1}^{m}\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\} = \sum_{i=1}^{m}\left\{Var\left[\mathbf{u}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right] + \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t\right\},
\end{aligned}
$$

while for the M-step, at the $(k+1)$st iteration, we have:

$$
\begin{aligned}
&\underline{M-step}:\ calculate\\
&\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{(k+1)^{-1}}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{(k+1)^{-1}}\mathbf{y}_i,\\
&\sigma^{2^{(k+1)}} = \sum_{i=1}^{m}\left\{trVar\left[\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right] + \hat{\boldsymbol{\varepsilon}}_i^t\hat{\boldsymbol{\varepsilon}}_i\right\}/N,\\
&and\\
&\mathbf{D}^{(k+1)} = \sum_{i=1}^{m}\left\{Var\left[\mathbf{u}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right] + \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t\right\}/m,
\end{aligned}
$$

where $\mathbf{V}_i^{(k+1)} = \mathbf{Z}_i\mathbf{D}^{(k+1)}\mathbf{Z}_i^t + \sigma^{2^{(k+1)}}\mathbf{I}_{n_i}$, $\boldsymbol{\theta}^{(k)} = \left(\mathbf{b}^{(k)},\sigma^{2^{(k)}},\mathbf{D}^{(k)}\right)$ is the estimate of $\boldsymbol{\theta} = (\mathbf{b},\sigma^2,\mathbf{D})$ obtained at the kth iteration, $\hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i\left(\boldsymbol{\theta}^{(k)}\right) = E\left(\mathbf{u}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right)$, and $\hat{\boldsymbol{\varepsilon}}_i = E\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right) = \mathbf{y}_i - \mathbf{X}_i\mathbf{b}^{(k)} - \mathbf{Z}_i\hat{\mathbf{u}}_i\left(\boldsymbol{\theta}^{(k)}\right)$. As one can observe, the expectation step of the above algorithm involves the determination of the conditional expected values of the numerators of the sufficient statistics (equations 5.46, 5.47), given the observed data $\mathbf{y}_i$ $(i = 1,...,m)$ and the current estimates $\boldsymbol{\theta}^{(k)} = \left(\mathbf{b}^{(k)},\sigma^{2^{(k)}},\mathbf{D}^{(k)}\right)$. Then, at the maximization step, we use the expectations calculated at the E-step to obtain the updated values $\sigma^{2^{(k+1)}}$, $\mathbf{D}^{(k+1)}$ and $\mathbf{b}^{(k+1)}$ of σ^2, $\mathbf{D}$ and $\mathbf{b}$. The iterative process continues until convergence is reached (i.e. the change in new versus old estimates is insignificant). These values, obtained at convergence, are essentially the ML estimates of σ^2, $\mathbf{D}$ and $\mathbf{b}$.

In the following, we attempt to give some insight on the exact methodology that led to the formulation of the specific equations and we try to describe in as much detail as possible the steps of the EM algorithm generated by Laird and Ware. However, for the purpose of generalization, instead of studying the implementation of EM on the conditional-independence model, we study a more broad model where a slightly different covariance structure for the random errors $\boldsymbol{\varepsilon}_i$ is assumed. To this end, we assume $\boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0},\sigma^2\mathbf{R}_i\right)$, where $\mathbf{R}_i$ this time is a known $(n_i \times n_i)$ positive definite matrix, in contrast to the article of Laird and Ware, where an identity square matrix $\mathbf{I}_{n_i}$ is considered. More specifically, to derive the equations that define the E- and M-steps of the iterative EM, we consider the mixed effects model of the following form:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad (i = 1, ..., m)\,, \tag{5.48}$$

where

$$\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right) \quad and \quad \boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0},\sigma^2\mathbf{R}_i\right) \,. \tag{5.49}$$

In the above, $\mathbf{R}_i$ is a known, positive definite matrix of order $(n_i \times n_i)$, and $\mathbf{D}$ is an unknown $(q \times q)$ positive-semidefinite symmetric matrix. It follows from (5.48), (5.49) that the response vectors $\mathbf{y}_i$ are marginally distributed as:

$$\mathbf{y}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i\right),$$

where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i$. The goal is to estimate the variance components of the above model, that is the parameters of variance-covariance matrix $\mathbf{D}$, along with the unknown positive scalar σ^2. The intuitive idea behind EM algorithm, is to think of the response vectors $\mathbf{y}_i$ $(i = 1, ..., m)$ as being the incomplete (observed) data, while as missing data we regard the (unobserved) random effects $\mathbf{u}_i$ $(i = 1, ..., m)$. Under this perspective, it is obvious that the 'complete data' would be the vector:

$$\mathbf{x}_{com} = \left(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m\right)^t, \tag{5.50}$$

where each vector $\mathbf{x}_i$ $(i = 1, ..., m)$ collects together the observed data $\mathbf{y}_i$ plus the 'unobserved data' $\mathbf{u}_i$ for the ith subject. That is $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{y}_i)^t = \begin{pmatrix} \mathbf{u}_i \\ \mathbf{y}_i \end{pmatrix}$. Normaly, to obtain ML estimates of the variance components (σ^2 and $\mathbf{D}$ in this particular problem), one should maximize the log-likelihood of the full observed data vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)^t$:

$$\begin{aligned} \lambda &= \ln L\left(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i; \mathbf{y}\right) \\ &= const. - \frac{1}{2}\sum_{i=1}^{m}\left(\ln \mid \mathbf{V}_i \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right). \qquad (5.51) \end{aligned}$$

However, as already stated, the EM-type approach bypasses the need of directly maximizing this incomplete data log-likelihood function, by iteratively maximizing an expected complete-data log-likelihood function. More precisely, due to that $\mathbf{y}_i \sim N_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i\right)$ and $\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right)$, it is reasonable to assume that the complete data vector $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{y}_i)^t$ for subject i follows a multivariate normal distribution with some mean vector $\boldsymbol{\mu}_i$ and some variance-covariance matrix $\boldsymbol{\Sigma}_i$, i.e.:

$$\mathbf{x}_i \sim N_{q+n_i}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right). \qquad (5.52)$$

Thus, evidently the corresponding log-likelihood function of the complete data vector $\mathbf{x}_{com}$ can be expressed as:

$$\begin{aligned} \lambda_{com} &= \ln L_{com}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i; \mathbf{x}_{com}\right) \\ &= const. - \frac{1}{2}\sum_{i=1}^{m}\left(\ln \mid \boldsymbol{\Sigma}_i \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)^t \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}_i\right). \qquad (5.53) \end{aligned}$$

Formally, the EM algorithm maximizes the observed data log-likelihood function (5.51) by iteratively maximizing the complete data log-likelihood function (5.53). Each iteration consists of the two distinct steps; the E- and M-steps. The E-step computes the conditional expectation of the above complete data log-likelihood given the observed data $\mathbf{y}$ and the current parameter estimates, while the M-step of the algorithm simply maximizes the ensuing conditional expected complete data log-likelihood obtained at the

E-step with respect to the variance components and fixed-effects $\mathbf{b}$. The iterative scheme is repeated until convergence is reached. The two steps are presented right away.

Formulation of the Expectation Step: In order to compute the conditional expectation of complete data log-likelihood (5.53), given the observed data $\mathbf{y}$ we essentially require the following conditional moments of the unobservable random effects vector $\mathbf{u}_i$, namely:

$$E\left(\mathbf{u}_i \mid \mathbf{y}_i\right) \tag{5.54}$$

and

$$E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i\right). \tag{5.55}$$

To proceed with calculating the above expectations, for start we need to define the exact form of the mean vector $\boldsymbol{\mu}_i$ and variance-covariance matrix Σ_i of $\mathbf{x}_i$. Based on the multivariate normal distribution theory, it can be shown that the complete data vector $\mathbf{x}_i = \left(\mathbf{u}_i, \mathbf{y}_i\right)^t$ has a multivariate normal distribution with mean vector:

$$\boldsymbol{\mu}_i = \begin{pmatrix} E\left(\mathbf{u}_i\right) \\ E\left(\mathbf{y}_i\right) \end{pmatrix},$$

and variance-covariance matrix:

$$\Sigma_i = \begin{pmatrix} Var\left(\mathbf{u}_i\right) & Cov\left(\mathbf{u}_i, \mathbf{y}_i\right) \\ Cov\left(\mathbf{y}_i, \mathbf{u}_i\right) & Var\left(\mathbf{y}_i\right) \end{pmatrix}.$$

Thus, we may state that the joint distribution vector $\mathbf{x}_i = \left(\mathbf{u}_i, \mathbf{y}_i\right)^t$ can be specified as:

$$\begin{pmatrix} \mathbf{u}_i \\ \mathbf{y}_i \end{pmatrix} \sim N_{q+n_i}\left[\begin{pmatrix} E\left(\mathbf{u}_i\right) \\ E\left(\mathbf{y}_i\right) \end{pmatrix}, \begin{pmatrix} Var\left(\mathbf{u}_i\right) & Cov\left(\mathbf{u}_i, \mathbf{y}_i\right) \\ Cov\left(\mathbf{y}_i, \mathbf{u}_i\right) & Var\left(\mathbf{y}_i\right) \end{pmatrix}\right]. \tag{5.56}$$

We already know that $E\left(\mathbf{u}_i\right) = \mathbf{0}$, $E\left(\mathbf{y}_i\right) = \mathbf{X}_i\mathbf{b}$, $Var\left(\mathbf{u}_i\right) = \mathbf{D}$ and

$Var\left(\mathbf{y}_i\right) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i$. In addition, it is:

$$\begin{aligned} Cov\left(\mathbf{u}_i, \mathbf{y}_i\right) &= Cov\left(\mathbf{u}_i, \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i\right) \\ &= Cov\left(\mathbf{u}_i, \mathbf{Z}_i\mathbf{u}_i\right) = Cov\left(\mathbf{u}_i, \mathbf{u}_i\right)\mathbf{Z}_i^t \\ &= Var\left(\mathbf{u}_i\right)\mathbf{Z}_i^t = \mathbf{D}\mathbf{Z}_i^t, \end{aligned}$$

and similarly,

$$\begin{aligned} Cov\left(\mathbf{y}_i, \mathbf{u}_i\right) &= Cov\left(\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \mathbf{u}_i\right) \\ &= Cov\left(\mathbf{Z}_i\mathbf{u}_i, \mathbf{u}_i\right) = \mathbf{Z}_i Cov\left(\mathbf{u}_i, \mathbf{u}_i\right) \\ &= \mathbf{Z}_i Var\left(\mathbf{u}_i\right) = \mathbf{Z}_i\mathbf{D}. \end{aligned}$$

Consequently, the distributional form of $\left(\mathbf{u}_i, \mathbf{y}_i\right)^t$ can be re-expressed as:

$$\begin{pmatrix} \mathbf{u}_i \\ \mathbf{y}_i \end{pmatrix} \sim N_{q+n_i}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{X}_i\mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i^t \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i \end{pmatrix}\right]. \tag{5.57}$$

To obtain the required $E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ and $E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i\right)$, what remains is to consider the following well-known result from multivariate normal distribution theory:

Proposition 5.1: *Let $\mathbf{y}_1$, $\mathbf{y}_2$ be random vectors of order $(p \times 1)$ and $(q \times 1)$ respectively. If $\mathbf{y}_1$ and $\mathbf{y}_2$ are partitioned into a single vector $\mathbf{y}$ such that*

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \sim N_{p+q}\left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}\right],$$

then the conditional distribution of $\mathbf{y}_1$ given $\mathbf{y}_2$ follows a p-variate normal distribution with mean

$$E\left(\mathbf{y}_1 \mid \mathbf{y}_2\right) = \boldsymbol{\mu}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right),$$

and variance-covariance matrix

$$Var\left(\mathbf{y}_1 \mid \mathbf{y}_2\right) = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21},$$

hence

$$\mathbf{y}_1 \mid \mathbf{y}_2 \sim N_p\left[\boldsymbol{\mu}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right), \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right].$$

Now, both $E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ and $E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i\right)$ are directly obtainable by applying the above proposition to equation (5.57). First, for $E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ it is:

$$\begin{aligned} E\left(\mathbf{u}_i \mid \mathbf{y}_i\right) &= \mathbf{0} + \mathbf{D}\mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i\right)^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right) \\ &= \mathbf{D}\mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i\right)^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right), \qquad (5.58) \end{aligned}$$

which, after some simple matrix manipulations, becomes:

$$\begin{aligned} E\left(\mathbf{u}_i \mid \mathbf{y}_i\right) &= \mathbf{D}\mathbf{Z}_i^t\left[\mathbf{R}_i\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t\mathbf{R}_i^{-1} + \sigma^2\mathbf{I}_{n_i}\right)\right]^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right) \\ &= \mathbf{D}\mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t\mathbf{R}_i^{-1} + \sigma^2\mathbf{I}_{n_i}\right)^{-1}\mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right). \qquad (5.59) \end{aligned}$$

By the same proposition, the conditional variance-covariance $Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ is given by:

$$\begin{aligned} Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right) &= \mathbf{D} - \mathbf{D}\mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{R}_i\right)^{-1}\mathbf{Z}_i\mathbf{D} \\ &= \mathbf{D} - \mathbf{D}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\mathbf{D}. \qquad (5.60) \end{aligned}$$

The above results for $E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ and $Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ can be updated by making use of the following trivial matrix identity:

$$\left(\mathbf{Z}_i^t\mathbf{Z}_i + \sigma^2\mathbf{D}^{-1}\right)\mathbf{D}\mathbf{Z}_i^t = \mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \sigma^2\mathbf{I}_{n_i}\right). \qquad (5.61)$$

where $\mathbf{D}$ is assumed nonsingular, and from which it follows that:

$$\mathbf{D}\mathbf{Z}_i^t\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t+\sigma^2\mathbf{I}_{n_i}\right)^{-1}=\left(\mathbf{Z}_i^t\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t. \tag{5.62}$$

Using the latter identities, $E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ can be re-written as:

$$E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)=\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right), \tag{5.63}$$

and similarly $Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right)$ becomes:

$$Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right)=\mathbf{D}-\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i\mathbf{D}, \tag{5.64}$$

which once again after some simple manipulations [common factor the $\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}$], results with the following more compact form:

$$\begin{aligned}
Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right) &= \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\left[\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)\mathrm{D}-\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i\mathbf{D}\right] \\
&= \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i\mathbf{D}+\sigma^2\mathbf{D}^{-1}\mathbf{D}-\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i\mathbf{D}\right) \\
&= \left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\sigma^2\mathbf{I}_{n_i}=\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)^{-1}\left(\sigma^{-2}\mathbf{I}_{n_i}\right)^{-1} \\
&= \left[\sigma^{-2}\mathbf{I}_{n_i}\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^2\mathbf{D}^{-1}\right)\right]^{-1}=\left(\sigma^{-2}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\mathbf{D}^{-1}\right)^{-1}.
\end{aligned} \tag{5.65}$$

[For the above, we have used the well-known matrix property $\left(\mathbf{AB}\right)^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$]. Let us now move on to the calculation of $E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i\right)$. To achieve this, we will use a familiar result of multivariate analysis; if $\mathbf{y}$ denotes a random vector with mean vector $E\left(\mathbf{y}\right)$ and variance-covariance matrix $Var\left(\mathbf{y}\right)$, then the following result holds:

$$Var\left(\mathbf{y}\right)=E\left(\mathbf{y}\mathbf{y}^t\right)-E\left(\mathbf{y}\right)E\left(\mathbf{y}\right)^t \Rightarrow E\left(\mathbf{y}\mathbf{y}^t\right)=Var\left(\mathbf{y}\right)+E\left(\mathbf{y}\right)E\left(\mathbf{y}\right)^t. \tag{5.66}$$

The above result (5.66), in conjunction with (5.65), gives:

$$E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i\right) = Var\left(\mathbf{u}_i \mid \mathbf{y}_i\right)+E\left(\mathbf{u}_i \mid \mathbf{y}_i\right)E\left(\mathbf{u}_i^t \mid \mathbf{y}_i\right)$$

$$
\begin{aligned}
&= \left(\sigma^{-2}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\mathbf{D}^{-1}\right)^{-1}+E\left(\mathbf{u}_i\mid\mathbf{y}_i\right)E\left(\mathbf{u}_i^t\mid\mathbf{y}_i\right)\\
&= \left(\sigma^{-2}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\mathbf{D}^{-1}\right)^{-1}+\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t, \qquad (5.67)
\end{aligned}
$$

where for notational convenience we have denoted $\hat{\mathbf{u}}_i=E\left(\mathbf{u}_i\mid\mathbf{y}_i\right)$ and $\hat{\mathbf{u}}_i^t=E\left(\mathbf{u}_i^t\mid\mathbf{y}_i\right)$.

All the above, lead naturally to the following formulation for the E-step of the EM algorithm, say at some $(k+1)$st iteration:

$$\underline{E-step}:\ u\sin g\ the\ current\ estimates\ \boldsymbol{\theta}^{(k)}=\left(\mathbf{b}^{(k)},\sigma^{2^{(k)}},\mathbf{D}^{(k)}\right),\ calculate$$

$$\hat{\mathbf{u}}_i=E\left(\mathbf{u}_i\mid\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right)=\left(\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\sigma^{2(k)}\mathbf{D}^{(k)^{-1}}\right)^{-1}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}^{(k)}\right)$$

and

$$E\left(\mathbf{u}_i\mathbf{u}_i^t\mid\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right)=\left(\sigma^{-2(k)}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i+\mathbf{D}^{(k)^{-1}}\right)^{-1}+E\left(\mathbf{u}_i\mid\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right)E\left(\mathbf{u}_i^t\mid\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right).$$

Formulation of the Maximization Step

We consider now the formulation of the second distinct step of the EM algorithm, namely the M-step. Recall that the M-step is associated with finding the values of $\mathbf{b}$, σ^2 and $\mathbf{D}$ that maximize the expected complete data log-likelihood, given the observed data and the parameter values obtained at the previous iteration. Thus, at the $(k+1)$st iteration of EM, we have to find the values $\mathbf{b}^{(k+1)}$, $\sigma^{2^{(k+1)}}$ and $\mathbf{D}^{(k+1)}$ that maximize the latter log-likelihood function:

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)=E\left[\ln L_{com}\left(\boldsymbol{\theta};\mathbf{x}_{com}\right)\mid\mathbf{y}_{obs},\boldsymbol{\theta}^{(k)}\right], \qquad (5.68)$$

where in this case $\boldsymbol{\theta}\equiv(\mathbf{b},\sigma^2,\mathbf{D})$. By the direct maximization of $Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(k)}\right)$ we obtain:

$$
\begin{aligned}
\mathbf{b}^{(k+1)} &= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\left[\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{y}_i-\mathbf{Z}_iE\left(\mathbf{u}_i\mid\mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right)\right]\\
&= \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\left(\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{y}_i-\mathbf{Z}_i\hat{\mathbf{u}}_i\right). \qquad (5.69)
\end{aligned}
$$

$$\mathbf{D}^{(k+1)} = \frac{1}{m}\sum_{i=1}^{m} E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right), \tag{5.70}$$

and

$$\sigma^{2^{(k+1)}} = \frac{1}{n}\sum_{i=1}^{m} E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right), \tag{5.71}$$

to be the updated values for $\mathbf{b}$, $\mathbf{D}$ and σ^2 respectively. Though the acquired equation (5.69) for fixed effects vector $\mathbf{b}$ is adequately analytical, expressions for $\mathbf{D}$ and σ^2 require further manipulations in order to become suitable for the algorithm's implementation. Below, we present all the necessary calculations that lead to an analytical M-step. Consider first $\mathbf{D}^{(k+1)}$. From (5.67) we obtain:

$$\begin{aligned}
\mathbf{D}^{(k+1)} &= \frac{1}{m}\sum_{i=1}^{m} E\left(\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[Var\left(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) + E\left(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) E\left(\mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\left(\sigma^{-2^{(k)}}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1} + E\left(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) E\left(\mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[\left(\sigma^{-2^{(k)}}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1} + \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t\right],
\end{aligned} \tag{5.72}$$

where $\hat{\mathbf{u}}_i = E\left(\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)$ =[from equation (5.58)]= $\mathbf{D}^{(k)}\mathbf{Z}_i^t\mathbf{V}_i^{(k)^{-1}}\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}^{(k)}\right)$ and $\hat{\mathbf{u}}_i^t = E\left(\mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)$. Now, as concerns $\sigma^{2^{(k+1)}}$, observe that the conditional expectation $E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)$ contained in equation (5.71), can be rewritten as:

$$E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) = E\left[tr\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] = E\left[tr\left(\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right].$$

[The first equality derives form the fact that quantity $\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i$ is a quadratic form[10] and due to that every quadratic form is a scalar, thus $\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i$ is equal to its trace. For the second equality we have used the matrix property $tr\left(\mathbf{AB}\right) = tr\left(\mathbf{BA}\right)$, for $\mathbf{A} \equiv \boldsymbol{\varepsilon}_i^t$

[10] We define as a quadratic form of a random vector $\mathbf{y}$ every function of the form $\mathbf{y}^t\mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is every known matrix whose dimension complies with the dimension of $\mathbf{y}$.

and $\mathbf{B} \equiv \mathbf{R}_i^{-1}\varepsilon_i]$. Further manipulations give:

$$
\begin{aligned}
& E\left[tr\left(\mathbf{R}_i^{-1}\varepsilon_i\varepsilon_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] = trE\left(\mathbf{R}_i^{-1}\varepsilon_i\varepsilon_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) \\
& \underset{(5.62)}{=} tr\left[\mathbf{R}_i^{-1}E\left(\varepsilon_i\boldsymbol{\varepsilon}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] \\
& = tr\left\{\mathbf{R}_i^{-1}\left[Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) + E\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right)E\left(\boldsymbol{\varepsilon}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right]\right\} \\
& = tr\left[\mathbf{R}_i^{-1}Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] + tr\left[\mathbf{R}_i^{-1}E\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right)E\left(\boldsymbol{\varepsilon}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right].
\end{aligned}
$$

Thus, if we denote $\hat{\boldsymbol{\varepsilon}}_i = E\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right)$and $\hat{\boldsymbol{\varepsilon}}_i^t = E\left(\boldsymbol{\varepsilon}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)$, $E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right)$ becomes:

$$
\begin{aligned}
E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) &= tr\left[\mathbf{R}_i^{-1}Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] + tr\left(\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i^t\right) \\
&= tr\left[\mathbf{R}_i^{-1}Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right)\right] + tr\left(\hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i\right) \\
&= tr\left[\mathbf{R}_i^{-1}Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i\right)\right] + \hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i. \qquad (5.73)
\end{aligned}
$$

(In the above we have replaced the trace of $\hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i$ with $\hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i$ itself, since $\hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i$ is a quadratic form).

Now since:

$$
\begin{aligned}
Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right) &= Var\left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) \\
&= Var\left(\mathbf{Z}_i\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) = \mathbf{Z}_iVar\left(\mathbf{u}_i \mid \mathbf{y}_i, \theta^{(k)}\right)\mathbf{Z}_i^t \\
&= \mathbf{Z}_i\left(\sigma^{-2(k)}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1}\mathbf{Z}_i^t, \qquad (5.74)
\end{aligned}
$$

combining (5.73) and (5.74) yields the following expression for $E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \theta^{(k)}\right)$:

$$
E\left(\boldsymbol{\varepsilon}_i^t\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right) = tr\left[\mathbf{R}_i^{-1}\mathbf{Z}_i\left(\sigma^{-2(k)}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1}\mathbf{Z}_i^t\right] + \hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i.
$$

Using the above, we may hence re-express $\sigma^{2^{(k+1)}}$ as:

$$\sigma^{2^{(k+1)}} = \frac{1}{n}\sum_{i=1}^{m}\left\{tr\left[\mathbf{R}_i^{-1}\mathbf{Z}_i\left(\sigma^{-2^{(k)}}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1}\mathbf{Z}_i^t\right] + \hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i\right\}. \qquad (5.75)$$

In summary, the maximization step, for the $(k+1)$st iterative step, is notably illustrated by the following:

$\underline{M-step}$: *update* $\boldsymbol{\theta} = (\mathbf{b}, \sigma^2, \mathbf{D})$ *with* $\boldsymbol{\theta}^{(k+1)} = \left(\mathbf{b}^{(k+1)}, \sigma^{2^{(k+1)}}, \mathbf{D}^{(k+1)}\right)$ *by*

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\left(\mathbf{X}_i^t\mathbf{R}_i^{-1}\mathbf{y}_i - \mathbf{Z}_i\hat{\mathbf{u}}_i\right),$$

$$\sigma^{2^{(k+1)}} = \frac{1}{n}\sum_{i=1}^{m}\left\{tr\left[\mathbf{R}_i^{-1}\mathbf{Z}_i\left(\sigma^{-2^{(k)}}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1}\mathbf{Z}_i^t\right] + \hat{\boldsymbol{\varepsilon}}_i^t\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i\right\},$$

and

$$\mathbf{D}^{(k+1)} = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\sigma^{-2^{(k)}}\mathbf{Z}_i^t\mathbf{R}_i^{-1}\mathbf{Z}_i + \mathbf{D}^{(k)^{-1}}\right)^{-1} + \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t\right].$$

It is important to note that the above presented maximization step specifically concerns, as noted earlier, a Laird-Ware model that specifies $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \sigma^2\mathbf{R}_i)$. Of course, in the slightly varying situation of 'conditional independence' model $[\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})]$ considered in the seminal article of *Laird* and *Ware (1982)*, we simply have to set $\mathbf{I}_{n_i}$ in place of $\mathbf{R}_i$ to derive the corresponding M-step.

5.5.2.2 An Important Remark

By comparing the above formulation of the maximization step and the maximization step proposed by *Laird* and *Ware (1982)* (see page 170), clearly an important difference is indicated. As one may observe, the two formulations are generally consistent, except for the formulas providing the ML estimator of fixed effects vector $\mathbf{b}$. Thus, the difference between the standard EM algorithm, previously described, and the EM scheme of Laird and Ware lies in the way the fixed-effects vector $\mathbf{b}$ is calculated.

In fact, the algorithm described by Laird and Ware does not correspond to the standard EM algorithm proposed by *Dempster et al. (1977)*. As *Liu* and *Rubin (1994)* point out, the formulas given by *Laird* and *Ware (1982)* (which were mistakenly called an EM algorithm) are clearly justified by the theory underlying a variation of the EM algorithm and not EM itself, namely the **"expectation-conditional maximization either"** (ECME) algorithm. The specific algorithm was called a hybrid EM algorithm by *Jennrich* and *Schluchter (1986)* who also realized it is not an EM.

In short, the ECME algorithm, originated by *Liu* and *Rubin (1994)*, can be considered as being an extension to the standard EM algorithm. The "conditional maximization step" refers to the fact that instead of the EM algorithm's usual maximization step (M-step), the maximization step of ECME algorithm is undertaken conditional on some of the parameters. As regards the "either", it refers to the fact that with this extension some or all of these conditional maximization steps (CM-steps) can be replaced by steps that maximize the incomplete (observed) data log-likelihood function $\ln L\left(\theta; \mathbf{y}_{obs}\right)$ conditional on some of the parameters, and not the complete data log-likelihood $Q\left(\theta, \theta^{(k)}\right) = E\left[\ln L\left(\theta; \mathbf{y}_{com}\right) \mid \mathbf{y}_{obs}, \theta^{(k)}\right]$ as is the case with standard EM algorithm. For a detailed review on the ECME algorithm we refer the interested reader to *Liu* and *Rubin (1994)* and *McLachlan* and *Krishnan (1997)*.

As concerns the particular algorithm of Laird and Ware, observe that the value of $\mathbf{b}$, obtained by the maximization step at some $(k+1)$st iteration is, as already shown, given by:

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{y}_i.$$

The above formula gives in fact the ML estimator of $\mathbf{b}$, obtained by maximizing the observed data (and not the complete-data) log-likelihood (see Section 5.3.1). Further, it is of interest to note that the updated value $\mathbf{b}^{(k+1)}$ of $\mathbf{b}$ does not depend on the variance parameter values obtained at the previous iteration step (i.e. $\sigma^{2^{(k)}}$, $\mathbf{D}^{(k)}$) as it should be the case for the usual EM algorithm, but instead requires the updated values $\sigma^{2^{(k+1)}}$ and

$\mathbf{D}^{(k+1)}$. Thus, what Laird and Ware consider as a unified M-step, in fact consists of two separate CM-steps of the ECME algorithm. While the one CM-step proceeds identically to the standard EM algorithm calculating the variance components σ^2 and $\mathbf{D}$ as on the M-step of the EM algorithm, the second CM-step calculates $\mathbf{b}^{(k+1)}$ by maximizing the observed data log-likelihood given the updated values of the other two parameters, namely $\sigma^{2^{(k+1)}}$ and $\mathbf{D}^{(k+1)}$. Taking all the above under consideration, we can now reformulate the variant of EM described by Laird and Ware as follows:

$$\underline{E-step}: \ This\ is\ the\ same\ as\ the\ E-step\ of\ EM$$

$$\underline{CM-step\ 1}: \ calculate$$

$$\sigma^{2^{(k+1)}} = \sum_{i=1}^{m} \left\{ trVar\left[\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\boldsymbol{\varepsilon}}_i^t \hat{\boldsymbol{\varepsilon}}_i \right\} /N,$$

and

$$\mathbf{D}^{(k+1)} = \sum_{i=1}^{m} \left\{ Var\left[\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^t \right\} /m.$$

$$\underline{CM-step\ 2}: \ u\sin g\ \sigma^{2^{(k+1)}},\ \mathbf{D}^{(k+1)}\ calculate$$

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{y}_i.$$

The examination of convergence behavior of the Laird-Ware ECME algorithm in terms of number of iterations and computational time required for convergence merits special attention. As Liu and Rubin claimed through the use of numerical examples, this extension of the standard EM algorithm is nearly always faster than EM in terms of required iterations and moreover can be faster in total computer time by orders of magnitude. Thus, it is generally advisable to resort to the Laird-Ware ECME algorithm which appears to converge more quickly compared to the standard EM algorithm, often criticized for its slow convergence.

5.5.2.3 Restricted Maximum Likelihood Estimation via the E-M Algorithm

So far, we discussed maximum likelihood (ML) estimation of fixed effects and variance components in the Laird-Ware model. As is well known, performing ML estimation for the Laird-Ware model requires maximization of the following "full-data" log-likelihood:

$$\begin{aligned}
&\lambda_{ML}\left(\mathbf{X}_i\mathbf{b},\mathbf{V}_i;\mathbf{y}\right) = \ln L\left(\mathbf{X}_i\mathbf{b},\mathbf{V}_i;\mathbf{y}\right)\\
= &-\sum_{i=1}^{m}\frac{n_i}{2}\ln\left(2\pi\right)-\frac{1}{2}\sum_{i=1}^{m}\left(\ln\mid\mathbf{V}_i\mid\right)-\frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)\\
= &\; const.-\frac{1}{2}\sum_{i=1}^{m}\left(\ln\mid\mathbf{V}_i\mid\right)-\frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right)^t\mathbf{V}_i^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}\right). \qquad (5.76)
\end{aligned}$$

An important problem with ML estimation of variance parameters is that this method produces biased estimators of those parameters due to the fact that the ML estimates of the variance components fail to take into account the degrees of freedom lost in estimating fixed effects. Instead, the restricted maximum likelihood (REML) method, which amounts to maximizing the part of the likelihood that is invariant to fixed effects, corrects for this bias. Specifically, the crucial aspect of the REML approach is that a linear combination (also known as the "error contrast") of the i^{th} subject's vector of measurements, namely $\mathbf{Ky}_i$ is used instead of the 'raw' data vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t$. As already noted in Chapter 3, the matrix $\mathbf{K}$ must be chosen so that $\mathbf{Ky}_i$ will be invariant to $\mathbf{X}_i\mathbf{b}$. Thus, essentially we choose a $\mathbf{K}$ matrix of order $(n_i \times n_i)$ so that:

$$E\left(\mathbf{Ky}_i\right) = \mathbf{0}, \;\; (i.e.\; \mathbf{KX}_i = \mathbf{0})\,.$$

Also, as concerns the variance-covariance matrix of error contrast $\mathbf{Ky}_i$, we have:

$$Var\left(\mathbf{Ky}_i\right) = \mathbf{K}Var\left(\mathbf{y}_i\right)\mathbf{K}^t = \mathbf{KV}_i\mathbf{K}^t.$$

It follows from the above (imposing once again the necessary normality assumption for each error contrast $\mathbf{Ky}_i$), that the 'restricted' data vector $\mathbf{Ky}_i$ $(i = 1, 2, ..., m)$ follows a n_i−variate normal distribution with $\mathbf{Ky}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{KV}_i\mathbf{K}^t\right)$ and corresponding p.d.f. given by:

$$f\left(\mathbf{Ky}_i\right) = (2\pi)^{-\frac{n_i - r(\mathbf{X}_i)}{2}} \mid \mathbf{KV}_i\mathbf{K}^t \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{Ky}_i\right)^t \left(\mathbf{KV}_i\mathbf{K}^t\right)^{-1} \mathbf{Ky}_i\right\}.$$

Having that in mind, the next step is to calculate the restricted maximum likelihood as follows:

$$\begin{aligned} L_{REML} &= \prod_{i=1}^{m} f\left(\mathbf{Ky}_i\right) \\ &= const. \times \left(\prod_{i=1}^{m} \mid \mathbf{KV}_i\mathbf{K}^t \mid\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{y}_i^t\mathbf{K}^t\left(\mathbf{KV}_i\mathbf{K}^t\right)^{-1}\mathbf{Ky}_i\right]\right\}. \end{aligned}$$

To calculate the corresponding restricted log-likelihood function (or log-restricted-likelihood function) it suffices to take the natural logarithm of L_{REML}. Indeed, we have:

$$\begin{aligned} \lambda_{REML} &= \ln L_{REML} \\ &= const. - \frac{1}{2}\ln\left(\prod_{i=1}^{m} \mid \mathbf{KV}_i\mathbf{K}^t \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{y}_i^t\mathbf{K}^t\left(\mathbf{KV}_i\mathbf{K}^t\right)^{-1}\mathbf{Ky}_i\right] \\ &= const. - \frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{KV}_i\mathbf{K}^t \mid - \frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{y}_i^t\mathbf{K}^t\left(\mathbf{KV}_i\mathbf{K}^t\right)^{-1}\mathbf{Ky}_i\right] \qquad (5.77) \end{aligned}$$

Now we only need to utilize once more the two important results due to Searle [equations (3.51) and (3.52)] that can be re-expressed as:

$$\ln \mid \mathbf{KV}_i\mathbf{K}^t \mid = \ln \mid \mathbf{V}_i \mid + \ln \mid \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid,$$

$$\mathbf{y}_i^t\mathbf{K}^t\left(\mathbf{KV}_i\mathbf{K}^t\right)^{-1}\mathbf{Ky}_i = \left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right),$$

By substituting the above results to (5.77) we derive the following expression for

λ_{REML}:

$$\begin{aligned}
&\lambda_{REML} = \\
= \;& const. - \frac{1}{2}\sum_{i=1}^{m}\left(\ln \mid \mathbf{V}_i \mid + \ln \mid \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right) \\
= \;& const. - \frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid - \frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right)^t \mathbf{V}_i^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}\right),
\end{aligned}$$

which, without the additive constant, and from (5.76) becomes:

$$\lambda_{REML} = -\frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid + \lambda_{ML}\left(\mathbf{X}_i\hat{\mathbf{b}}, \mathbf{V}_i; \mathbf{y}\right). \tag{5.78}$$

Observe that the restricted log-likelihood function λ_{REML} differs from standard log-likelihood function λ_{ML} only in the extra term $-1/2\sum_{i=1}^{m}\ln \mid \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid$. Also, λ_{REML} does not depend any longer on the fixed-effects vector $\mathbf{b}$ which now has been replaced by its estimator $\hat{\mathbf{b}}$.

Having defined the restricted log-likelihood for REML estimation, we now move on to the description of the computing formulas for implementing the EM algorithm to calculate restricted maximum likelihood (REML) estimates of the fixed-effects vector $\mathbf{b}$ and the variance components of the Laird-Ware model (5.48). In particular, we focus our attention on describing the variant of EM algorithm (namely the ECME algorithm) introduced by *Laird* and *Ware (1982)* for REML estimation instead of the standard EM algorithm as the former seems to have substantial computational advantages over the latter (faster convergence rate) and is the most commonly considered between the two algorithms. For the sake of a clear presentation, let us start the discussion with considering once more the iterative scheme of the EM variant of Laird and Ware for ML estimation, described in the previous section. Specifically, after providing the starting values for the (unknown) parameters $\boldsymbol{\theta} = (\mathbf{b}, \sigma^2, \mathbf{D})$, the E- and the two CM-steps of ECME algorithm at some iteration $k+1$ as already shown may be expressed via the following iterative computational scheme:

$\underline{E-step}: \ u\sin g \ the\ current\ estimates\ \boldsymbol{\theta}^{(k)} = \left(\mathbf{b}^{(k)}, \sigma^{2^{(k)}}, \mathbf{D}^{(k)}\right), \ calculate$

$$E\left\{\sum_{i=1}^{m} \varepsilon_i^t \varepsilon_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right\} = \sum_{i=1}^{m}\left\{trVar\left[\varepsilon_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\varepsilon}_i^t \hat{\varepsilon}_i\right\},$$

and

$$E\left\{\sum_{i=1}^{m} \mathbf{u}_i \mathbf{u}_i^t \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right\} = \sum_{i=1}^{m}\left\{Var\left[\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^t\right\}.$$

$\underline{CM-step\ 1}: \ calculate$

$$\sigma^{2^{(k+1)}} = \sum_{i=1}^{m}\left\{trVar\left[\varepsilon_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\varepsilon}_i^t \hat{\varepsilon}_i\right\}/N,$$

and

$$\mathbf{D}^{(k+1)} = \sum_{i=1}^{m}\left\{Var\left[\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right] + \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^t\right\}/m.$$

$\underline{CM-step\ 2}: \ u\sin g\ \sigma^{2^{(k+1)}}, \ \mathbf{D}^{(k+1)}\ calculate$

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{y}_i.$$

It is easy to realize that the above optimization algorithm can be used to define the ECME iterations for the calculation of the REML estimates for $\mathbf{b}$, σ^2 and $\mathbf{D}$, too. Only this time, the quantities $Var\left[\mathbf{u}_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right]$ and $Var\left[\varepsilon_i \mid \mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right]$ have to be replaced by $Var\left[\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right]$ and $Var\left[\varepsilon_i \mid \mathbf{K}\mathbf{y}_i, \boldsymbol{\theta}^{(k)}\right]$ respectively, since to perform REML estimation requires utilization of error contrasts $\mathbf{K}\mathbf{y}_i$ instead of the 'full' data vectors $\mathbf{y}_i$ $(i = 1, 2, ..., m)$. Hence, to describe the EM iterative equations for the Laird-Ware model in terms of REML estimation, what is additionally required is the calculation of the following variance-covariance matrices of the conditional random variables $\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i$

and $\boldsymbol{\varepsilon}_i \mid \mathbf{K}\mathbf{y}_i$, namely:

$$Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right) \quad and \quad Var\left(\varepsilon_i \mid \mathbf{K}\mathbf{y}_i\right).$$

Beginning with the calculation of $Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right)$, we may use Proposition 5.1 to write:

$$Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right) = Var\left(\mathbf{u}_i\right) - Cov\left(\mathbf{u}_i, \mathbf{K}\mathbf{y}_i\right)\left[Var\left(\mathbf{K}\mathbf{y}_i\right)\right]^{-1} Cov\left(\mathbf{K}\mathbf{y}_i, \mathbf{u}_i\right). \tag{5.79}$$

It is straightforward to calculate the above variances and covariances included in the expression for $Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right)$. Indeed, we have:

$$\begin{aligned}
Var\left(\mathbf{u}_i\right) &= \mathbf{D}, \\
Var\left(\mathbf{K}\mathbf{y}_i\right) &= \mathbf{K}Var\left(\mathbf{y}_i\right)\mathbf{K}^t = \mathbf{K}\mathbf{V}_i\mathbf{K}^t, \\
Cov\left(\mathbf{u}_i, \mathbf{K}\mathbf{y}_i\right) &= Cov\left[\mathbf{u}_i, \mathbf{K}\left(\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i\right)\right] \\
&= Cov\left(\mathbf{u}_i, \mathbf{K}\mathbf{Z}_i\mathbf{u}_i\right) = Var\left(\mathbf{u}_i\right)\mathbf{Z}_i^t\mathbf{K}^t \\
&= \mathbf{D}\mathbf{Z}_i^t\mathbf{K}^t
\end{aligned}$$

and

$$\begin{aligned}
Cov\left(\mathbf{K}\mathbf{y}_i, \mathbf{u}_i\right) &= Cov\left[\mathbf{K}\left(\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i\right), \mathbf{u}_i\right] \\
&= Cov\left(\mathbf{K}\mathbf{Z}_i\mathbf{u}_i, \mathbf{u}_i\right) = \mathbf{K}\mathbf{Z}_i Var\left(\mathbf{u}_i\right) \\
&= \mathbf{K}\mathbf{Z}_i\mathbf{D}.
\end{aligned}$$

By substitution of the above in (5.79), the updated $Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right)$ now becomes:

$$Var\left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i\right) = \mathbf{D} - \mathbf{D}\mathbf{Z}_i^t\mathbf{K}^t\left(\mathbf{K}\mathbf{V}_i\mathbf{K}^t\right)^{-1}\mathbf{K}\mathbf{Z}_i\mathbf{D}. \tag{5.80}$$

Considering a result from *Searle (1979)* on REML estimation for the General Linear Mixed Model (GLMM), (which modified applies to the Laird-Ware model, too), we may write that:

$$\mathbf{K}^t\left(\mathbf{K}\mathbf{V}_i\mathbf{K}^t\right)^{-1}\mathbf{K}=\mathbf{P}_i, \tag{5.81}$$

where

$$\mathbf{P}_i=\mathbf{V}_i^{-1}-\mathbf{V}_i^{-1}\mathbf{X}_i\left(\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{X}_i^t\mathbf{V}_i^{-1}. \tag{5.82}$$

Substituting (5.81) in (5.80) yields the following expression for the conditional variance of random effects vector $\mathbf{u}_i$ given $\mathbf{Ky}_i$, namely $Var\left(\mathbf{u}_i \mid \mathbf{Ky}_i\right)$:

$$Var\left(\mathbf{u}_i \mid \mathbf{Ky}_i\right)=\mathbf{D}-\mathbf{D}\mathbf{Z}_i^t\mathbf{P}_i\mathbf{Z}_i\mathbf{D}=\mathbf{D}\left(\mathbf{I}-\mathbf{Z}_i^t\mathbf{P}_i\mathbf{Z}_i\mathbf{D}\right). \tag{5.83}$$

Furthermore, by similar routine algebraic operations it may be possible to derive the following representation for the second required conditional variance, namely $Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{Ky}_i\right)$:

$$Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{Ky}_i\right)=\sigma^2 tr\left(\mathbf{I}-\sigma^2\mathbf{P}_i\right), \tag{5.84}$$

with $\mathbf{P}_i$ representing again the projection matrix already defined in (5.82).

We are thus in position now, introducing the obtained expressions (5.83) and (5.84) for $Var\left(\mathbf{u}_i \mid \mathbf{Ky}_i\right)$ and $Var\left(\boldsymbol{\varepsilon}_i \mid \mathbf{Ky}_i\right)$ respectively into the iterative scheme presented in page 180, to conclude with the following 'Laird-Ware' ECME numerical optimization algorithm for REML estimation: The E-step and the two CM-steps, at the $(k+1)$st iteration, for the simultaneous estimation of $\mathbf{b}$, σ^2 and $\mathbf{D}$ are (after providing the necessary starting values $\mathbf{b}^{(0)}$, $\sigma^{2^{(0)}}$, $\mathbf{D}^{(0)}$):

$$\underline{E-step}:\ u\sin g\ the\ current\ estimates\ \boldsymbol{\theta}^{(k)}=\left(\mathbf{b}^{(k)},\sigma^{2^{(k)}},\mathbf{D}^{(k)}\right),\ calculate$$

$$E\left\{\sum_{i=1}^{m}\boldsymbol{\varepsilon}_i^t\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\}=\sum_{i=1}^{m}\left\{\sigma^{2^{(k)}}tr\left(\mathbf{I}-\sigma^{2^{(k)}}\mathbf{P}_i^{(k)}\right)+\hat{\boldsymbol{\varepsilon}}_i^t\hat{\boldsymbol{\varepsilon}}_i\right\},$$
$$E\left\{\sum_{i=1}^{m}\mathbf{u}_i\mathbf{u}_i^t \mid \mathbf{y}_i,\boldsymbol{\theta}^{(k)}\right\}=\sum_{i=1}^{m}\left\{\mathbf{D}^{(k)}\left(\mathbf{I}-\mathbf{Z}_i^t\mathbf{P}_i^{(k)}\mathbf{Z}_i\mathbf{D}^{(k)}\right)+\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^t\right\}.$$

$$\underline{CM - step\ 1}: \ calculate$$

$$\sigma^{2(k+1)} = \sum_{i=1}^{m} \left\{ \sigma^{2(k)} tr \left(\mathbf{I} - \sigma^{2(k)} \mathbf{P}_i^{(k)} \right) + \hat{\varepsilon}_i^t \hat{\varepsilon}_i \right\} / N,$$
$$\mathbf{D}^{(k+1)} = \sum_{i=1}^{m} \left\{ \mathbf{D}^{(k)} \left(\mathbf{I} - \mathbf{Z}_i^t \mathbf{P}_i^{(k)} \mathbf{Z}_i \mathbf{D}^{(k)} \right) + \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^t \right\} / m.$$

$$\underline{CM - step\ 2}: \ u\sin g\ \sigma^{2(k+1)},\ \mathbf{D}^{(k+1)}\ calculate$$

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k+1)^{-1}} \mathbf{y}_i,$$

where $\mathbf{V}_i^{(k+1)^{-1}} = \mathbf{Z}_i \mathbf{D}^{(k+1)} \mathbf{Z}_i^t + \sigma^{2(k+1)} \mathbf{I}_{n_i}$, $\hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i \left(\boldsymbol{\theta}^{(k)} \right) = E \left(\mathbf{u}_i \mid \mathbf{K}\mathbf{y}_i, \boldsymbol{\theta}^{(k)} \right) =$
$= \mathbf{D}^{(k)} \mathbf{Z}_i^t \mathbf{V}_i^{(k)^{-1}} \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b}^{(k)} \right)$, $\hat{\varepsilon}_i = E \left(\boldsymbol{\varepsilon}_i \mid \mathbf{K}\mathbf{y}_i, \boldsymbol{\theta}^{(k)} \right) = \mathbf{y}_i - \mathbf{X}_i \mathbf{b}^{(k)} - \mathbf{Z}_i \hat{\mathbf{u}}_i \left(\boldsymbol{\theta}^{(k)} \right)$
and $\mathbf{P}_i^{(k)} = \mathbf{V}_i^{(k)^{-1}} - \mathbf{V}_i^{(k)^{-1}} \mathbf{X}_i \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{(k)^{-1}} \mathbf{X}_i \right)^{-1} \mathbf{X}_i^t \mathbf{V}_i^{(k)^{-1}}$. The above scheme is iterated until convergence of the parameters is reached. (For an early discussion on REML estimation of the 'Laird-Ware' model via the EM algorithm compare also *Laird et al., 1987*; and *Lindstrom* and *Bates, 1988*).

5.5.3 The Newton-Raphson Algorithm

In numerical analysis, the most effective way of finding the roots of nonlinear equations is to devise iterative algorithms, which start with an initial estimate of the root and converge to the exact value of the root in a finite number of steps. One of the oldest and at the same time most powerful methods for solving such one-dimensional or multidimensional equations when an algebraically solution to this problem cannot be achieved, is the so-called *Newton-Raphson method* (N-R in abbreviation). As early as 1685, in the book of Wallis: Algebra, it is mentioned that the idea of this method was due to Newton. Several years later (1690) the method, slightly modified was published by Raphson and since then the method is known as the Newton-Raphson method.

To understand the general framework of the method, let us consider some (one-dimensional) function $f(x)$, of which we seek to find its zeros, i.e. to solve equation

$f(x) = 0$. To achieve this, the N-R method depends on the following iterative relation (which derives from the well-known Taylor series expansion of a function at a neighborhood of a point):

$$x_{k+1} = x_k - \frac{1}{f'(x_k)} f(x_k), \qquad k = 0, 1, 2, ... \tag{5.85}$$

where x_k is the current value, $f(x_k)$ represents the value of the function evaluated at x_k, and $f'(x_k)$ is the derivative at x_k. x_{k+1} represents the next value obtained by the iterative scheme (5.85). To find a root of $f(x) = 0$ we only need to start with an initial value x_0 and then iterate the above relation until convergence is reached.

5.5.4 Maximum Likelihood Estimation via the N-R Algorithm

As noted already, an appealing feature of the Newton-Raphson algorithm is that it is not restricted only to one dimension, but it can easily generalize to multiple dimensions. Consequently, N-R can be applied to situations that require numerical evaluation of the roots of n-dimensional functions, where $n > 1$. In the statistical framework for example, the N-R method (and variations) has been extensively applied to solving equations of multidimensional functions. N-R has become a necessary tool for the maximization of, very often multidimensional, likelihood functions when no theoretical solution could be obtained. Its usefulness is more apparent in problems that require the ML estimation of unknown variance parameters, i.e. ML estimation of variance components of the General Linear Mixed Model. Maximizing the log-likelihood function of the GLMM, evidently requires a generalization of N-R algorithm since that the likelihood function is a multidimensional function depending on the variance parameters. In particular, if $\boldsymbol{\theta}$ denotes the vector that contains the variance parameters that we seek to estimate, the N-R algorithm obtains the ML estimates $\overset{\wedge}{\boldsymbol{\theta}}$ by starting with some initial value $\boldsymbol{\theta}^{(0)}$ and then iterating to converge to a final solution by using:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \left(\mathrm{H}^{(k)}\right)^{-1} \frac{\partial \lambda}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}^{(k)}}, \tag{5.86}$$

where $\lambda = \ln L(\boldsymbol{\theta})$ denotes the log-likelihood, $\partial\lambda/\partial\boldsymbol{\theta}$ is a column vector consisting of the partial derivatives of log-likelihood λ with respect to each parameter of $\boldsymbol{\theta}$ evaluated at the estimate $\boldsymbol{\theta}^{(k)}$, and $\mathbf{H}$ is the Hessian[11] matrix of all second-order partial derivatives of the log-likelihood with respect to the variance components. In fact, the inverse of the Hessian matrix, $\mathbf{H}^{-1}$ is a measure of the curvature of the likelihood surface given the current estimates, whereas $\partial\lambda/\partial\boldsymbol{\theta}$ measures the slope (directionality) of the likelihood. Therefore, their product in (5.86) gives a projected degree of movement of vector $\boldsymbol{\theta}$ towards an improved set of values to be used in the next iteration.

A very interesting variation of the N-R algorithm, often met in statistics, is the modified *Scoring algorithm of Fisher* (or the Fisher's scoring algorithm). This algorithm results from N-R by simply replacing the inverse of the Hessian matrix in (5.86) with its expected value, which after allowing for a change in sign turns out to be defined by the inverse of Fisher's information matrix, namely $-\mathbf{I}^{-1}$. Thus, the Fisher's scoring iterative equation is given by:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left(\mathrm{I}^{(k)}\right)^{-1} \frac{\partial\lambda}{\partial\boldsymbol{\theta}} \,|_{\theta^{(k)}}, \tag{5.87}$$

for $k = 1, 2, \ldots.$ The significance of Fisher's scoring algorithm mainly lies on the fact that in most times the information matrix is easier to compute, compared to the Hessian matrix, because some of the correlation terms of I are zero. Further, Fisher's scoring has been shown to be more robust to 'poor' starting values than the strict N-R algorithm.

5.5.5 Implementation of N-R to the "Laird-Ware" Model

We mentioned already that Gradient-type algorithms such as Newton-Raphson and its variant, Fisher scoring, have been extensively used in the Statistical field for likelihood maximization. More specific, N-R and variations can be used to obtain (in a simultaneous way) both ML estimates of the fixed parameters of the model and ML/REML estimates of the variance components of the model.

[11] A Hessian matrix, is a matrix which contains all second-order partial derivatives of a real-valued function.

In the literature, especially in the recent years, many articles have been occupied with the issue of estimating variance components in the Laird-Ware model by the application of Newton-Raphson algorithm. As regards a thorough treatment on the latter issue, full implementations of the N-R and Fisher scoring methods for maximum (or restricted maximum likelihood) estimation in the Laird-Ware model were discussed by *Jennrich* and *Schluchter (1986)* and by *Lindstrom* and *Bates (1988)*. Jennrich and Schluchter applies the N-R algorithm to a more general model, that includes the mixed-effects model of Laird and Ware as a special case. Their presentation involves illustration of general formulas for ML/REML estimation of fixed effects vector $\mathbf{b}$ and variance components of the unbalanced repeated measures GLM: $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \boldsymbol{\varepsilon}_i$ $(i = 1, 2, ..., m)$ via the N-R method, as well as formulas for the modified N-R, the Fisher's scoring algorithm.

Lindstrom and Bates, in their more computationally-oriented article, present in great detail derivative formulas for implementing the N-R algorithm to obtain ML/REML estimates of the GLMM $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ $(i = 1, 2, ..., m)$. Additionally, they consider improvements to the general N-R method proposed by Jennrich and Schluchter, improvements that have the purpose of reducing the number of iterations required for the convergence of the algorithm and consequently improving the overall convergence behavior of the algorithm.

The following sections have as a primary aim to illustrate and adequately describe all the necessary computations needed for the implementing N-R algorithm to the standard Laird-Ware model. Moreover, we try to present these, computationally expensive, formulas of *Jennrich* and *Schluchter (1986)* and *Lindstrom* and *Bates (1988)* in a simple way as possible, without at the same time moving far away from the spirit of both sets of authors though.

5.5.5.1 ML Estimation via the Newton-Raphson Algorithm

Following *Jennrich* and *Schluchter (1986)*, we may express the form of the iterative N-R algorithm that, when reaches convergence, produces ML estimates of the fixed effects vector $\mathbf{b}$ and the variance components $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_q)^t$simultaneously, as:

$$\begin{pmatrix} \mathrm{b}^{(k+1)} \\ \theta^{(k+1)} \end{pmatrix} = \begin{pmatrix} \mathrm{b}^{(k)} \\ \theta^{(k)} \end{pmatrix} - \left(\mathrm{H}^{(k)}\right)^{-1} \begin{pmatrix} \partial\lambda/\partial\mathrm{b}\,|_{\mathrm{b}^{(k)}} \\ \partial\lambda/\partial\boldsymbol{\theta}\,|_{\boldsymbol{\theta}^{(k)}} \end{pmatrix}, \tag{5.88}$$

where $\partial\lambda/\partial\mathbf{b}$, $\partial\lambda/\partial\boldsymbol{\theta}$ are the first partial derivatives of the log-likelihood function λ with respect to $\mathbf{b}$ and $\boldsymbol{\theta}$, respectively and $\mathbf{H}$ denotes the Hessian matrix, containing the second partial derivatives of λ:

$$\mathbf{H} = \begin{pmatrix} \partial^2\lambda/\partial\mathbf{b}\partial\mathbf{b} & \partial^2\lambda/\partial\mathbf{b}\partial\boldsymbol{\theta} \\ \partial^2\lambda/\partial\boldsymbol{\theta}\partial\mathbf{b} & \partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta} \end{pmatrix}. \tag{5.89}$$

The above iterative scheme (5.88) indicates, in simple words, that in order to obtain at each iteration step of the algorithm the 'new' estimates $\mathbf{b}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$ what is required additionally to the 'present' estimates $\mathbf{b}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$, is first and second partial derivatives of log-likelihood λ with respect to $\mathbf{b}$ and variance components, evaluated at the current estimate values $\mathbf{b}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$. The remainder of the current section thus, will be devoted to the presentation of all the above necessary derivatives for the implementation of the Newton-Raphson algorithm. Before proceed with the calculations though, it should be mentioned that for the derivation of the following formulas we are taking under consideration that the variance-covariance matrix $\mathbf{V}_i$ is symmetric (i.e. $\mathbf{V}_i = \mathbf{V}_i^t$). This is in compliance with the style adopted by Jennrich and Schluchter. Lindstrom and Bates, in their article, follow an slightly alternate approach that maintains the distinction between $\mathbf{V}_i^t$ and $\mathbf{V}_i$ (e.g. for Lindstrom and Bates $\partial\mathbf{V}_i^t/\partial\boldsymbol{\theta}$ does not coincide with $\partial\mathbf{V}_i/\partial\boldsymbol{\theta}$, while for Jennrich and Schluchter does). Considering the above, we are ready now to continue with the calculation of the first and second partial derivatives.

First of all, recall that the log-likelihood function λ that we seek to maximize is given by:

$$\lambda = const. - \frac{1}{2}\sum_{i=1}^{m} \ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum_{i=1}^{m} (\mathbf{y}_i - \mathrm{X}_i\mathrm{b})^t\, \mathrm{V}_i^{-1}\, (\mathbf{y}_i - \mathrm{X}_i\mathrm{b})\,. \tag{5.90}$$

Hence, evidently, obtaining $\partial\lambda/\partial\mathbf{b}$ requires calculation of $\partial \ln \mid \mathbf{V}_i \mid /\partial\mathbf{b}$ as well as calculation of $\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\, \mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]/\partial\mathbf{b}$, while to obtain a formula for the first partial derivative with respect to the other parameter $\boldsymbol{\theta}$, namely $\partial\lambda/\partial\boldsymbol{\theta}$ what is needed is the calculation of all partial derivatives $\partial \ln \mid \mathbf{V}_i \mid /\partial\theta_j$, $\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\, \mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]/\partial\theta_j$, where $j = 1, 2, ..., q$ index the variance components (all elements of the parameter vector $\boldsymbol{\theta}$).

We start with the calculation of $\partial\lambda/\partial\mathbf{b}$; clearly, it is $\partial \ln \mid \mathbf{V}_i \mid /\partial\mathbf{b} = \mathbf{0}$, thus what is only needed is:

$$\frac{\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\, \mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}} = -2\mathbf{X}_i^t\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\,. \tag{5.91}$$

To derive the above, we have used the standard result from matrix derivatives, already presented in (3.32). From (5.90) and (5.91) we get:

$$\begin{aligned}
\frac{\partial\lambda}{\partial\mathbf{b}} &= \frac{\partial\left[const. - \frac{1}{2}\sum\limits_{i=1}^{m} \ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum\limits_{i=1}^{m} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\, \mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}} \\
&= \mathbf{0} - \mathbf{0} - \frac{1}{2}\sum_{i=1}^{m} \frac{\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\, \mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}} \\
&= -\frac{1}{2}\sum_{i=1}^{m} \left[-2\mathbf{X}_i^t\mathbf{V}_i^{-1}\,(\mathrm{y}_i - \mathrm{X}_i\mathrm{b})\right] \\
&= \sum_{i=1}^{m} \mathbf{X}_i^t\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathrm{X}_i\mathrm{b})\,.
\end{aligned} \tag{5.92}$$

As concern now $\partial\lambda/\partial\theta_j$, we once again make use of the useful results (3.43) and (3.44) of Chapter 3. From (3.44) the derivative $\partial \ln \mid \mathbf{V}_i \mid /\partial\theta_j$ can be expressed as:

$$\frac{\partial \ln \mid \mathbf{V}_i \mid}{\partial\theta_j} = tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right), \tag{5.93}$$

while using (3.43) we get:

$$\begin{aligned} &\frac{\partial\left[(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j} = (\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\frac{\partial\mathbf{V}_i^{-1}}{\partial\theta_j}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}) \\ = \;& (\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\left(-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\right)(\mathbf{y}_i-\mathbf{X}_i\mathbf{b}) \\ = \;& -(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\,. \end{aligned} \tag{5.94}$$

Formulas (5.93), (5.94) are sufficient to allow us to come up with an expression for the first-order partial derivative of the log-likelihood with respect to variance parameter θ_j. Indeed:

$$\begin{aligned} &\frac{\partial\lambda}{\partial\theta_j} = \frac{\partial\left[const. - \frac{1}{2}\sum\limits_{i=1}^{m}\ln\mid\mathbf{V}_i\mid - \frac{1}{2}\sum\limits_{i=1}^{m}(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j} \\ = \;& \mathbf{0} - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial\ln\mid\mathbf{V}_i\mid}{\partial\theta_j} - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial\left[(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j} \\ = \;& -\frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right) - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial\left[(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j}, \end{aligned} \tag{5.95}$$

and by noticing that quantity $(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})$ is a quadratic form (therefore a scalar), thus it is equal to its trace, the above can be reformulated as:

$$\begin{aligned} &\frac{\partial\lambda}{\partial\theta_j} = -\frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right) - \frac{1}{2}\sum_{i=1}^{m}tr\frac{\partial\left[\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\right]}{\partial\theta_j} \\ = \;& -\frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right) - \frac{1}{2}\sum_{i=1}^{m}tr\left[\frac{\partial\mathbf{V}_i^{-1}}{\partial\theta_j}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\right] \\ = \;& -\frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right) + \frac{1}{2}\sum_{i=1}^{m}tr\left[\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t\right] \\ = \;& \frac{1}{2}\sum_{i=1}^{m}tr\left[\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})\,(\mathbf{y}_i-\mathbf{X}_i\mathbf{b})^t - \mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right] \\ = \;& \frac{1}{2}\sum_{i=1}^{m}tr\left[\mathbf{V}_i^{-1}\left(\mathbf{e}_i\mathbf{e}_i^t-\mathbf{V}_i\right)\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right], \end{aligned} \tag{5.96}$$

where, for the sake of notational convenience mainly, we have set in the above $\mathbf{e}_i = (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})$, $\mathbf{e}_i^t = (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t$. To summarize, thus far we have calculated the first partial derivatives of log-likelihood λ with respect to $\mathbf{b}$ and θ_j (shown in (5.92) and (5.96) respectively), required for the implementation of N-R algorithm for deriving ML estimates of both fixed-effects vector $\mathbf{b}$ and variance components θ_j $(j = 1, 2, ..., q)$ in the Laird-Ware model.

In addition, what is needed to complete the set of crucial formulas for N-R implementation is the elements of the Hessian matrix $\mathbf{H}$, namely the second partial derivatives: $\partial^2\lambda/\partial\mathbf{b}\partial\mathbf{b}$, $\partial^2\lambda/\partial\mathbf{b}\partial\boldsymbol{\theta}$, $\partial^2\lambda/\partial\boldsymbol{\theta}\partial\mathbf{b}$ and $\partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}$. To start with, as regards $\partial^2\lambda/\partial\mathbf{b}\partial\mathbf{b}$, observe that (using 5.91):

$$\begin{aligned}
\frac{\partial^2\left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}\partial\mathbf{b}} &= \frac{\partial}{\partial\mathbf{b}}\left\{\frac{\partial\left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}}\right\} \overset{}{\underset{(5.91)}{=}} \\
&= \frac{\partial}{\partial\mathbf{b}}\left[-2\mathbf{X}_i^t\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right] = -2\mathbf{X}_i^t\mathbf{V}_i^{-1}\frac{\partial\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})}{\partial\mathbf{b}} \\
&= -2\mathbf{X}_i^t\mathbf{V}_i^{-1}\,(-\mathbf{X}_i) = 2\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i. \qquad (5.97)
\end{aligned}$$

This, combined with the obvious result:

$$\frac{\partial^2 \ln \mid \mathbf{V}_i \mid}{\partial\mathbf{b}\partial\mathbf{b}} = \mathbf{0}, \qquad (5.98)$$

yields the following expression for the partial derivative $\partial^2\lambda/\partial\mathbf{b}\partial\mathbf{b}$:

$$\begin{aligned}
\frac{\partial^2\lambda}{\partial\mathbf{b}\partial\mathbf{b}} &= \mathbf{0} - \mathbf{0} - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial^2\left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t\,\mathbf{V}_i^{-1}\,(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}\partial\mathbf{b}} \\
&= -\frac{1}{2}\sum_{i=1}^{m}2\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i = -\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i. \qquad (5.99)
\end{aligned}$$

We now turn to the calculation of the second partial derivative of λ with respect to one variance component, say θ_j, and vector $\mathbf{b}$. First, the second partial derivative of the term $(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})$ with respect to θ_j and $\mathbf{b}$ is given by:

$$\begin{aligned}
\frac{\partial^2 \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j \partial\mathbf{b}} &= \frac{\partial}{\partial\theta_j}\left\{\frac{\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}}\right\} \\
&= \frac{\partial}{\partial\theta_j}\left[-2\mathbf{X}_i^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right] = -2\mathbf{X}_i^t \frac{\partial \mathbf{V}_i^{-1}}{\partial\theta_j} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \\
&= -2\mathbf{X}_i^t \left(-\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial\theta_j} \mathbf{V}_i^{-1}\right) (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \\
&= \mathbf{X}_i^t \left(2\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial\theta_j} \mathbf{V}_i^{-1}\right) (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) .
\end{aligned} \tag{5.100}$$

Moreover, it is:

$$\frac{\partial^2 \ln \mid \mathbf{V}_i \mid}{\partial\theta_j \partial\mathbf{b}} = \frac{\partial}{\partial\theta_j}\left(\frac{\partial \ln \mid \mathbf{V}_i \mid}{\partial\mathbf{b}}\right) = \mathbf{0}. \tag{5.101}$$

Using (5.100) and (5.101), we may easily calculate $\partial^2\lambda/\partial\theta_j\partial\mathbf{b}$ as follows:

$$\begin{aligned}
\frac{\partial^2\lambda}{\partial\theta_j\partial\mathbf{b}} &= \frac{\partial}{\partial\theta_j}\left(\frac{\partial\lambda}{\partial\mathbf{b}}\right) = \frac{\partial}{\partial\theta_j}\left\{\mathbf{0} - \mathbf{0} - \frac{1}{2}\sum_{i=1}^{m} \frac{\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\mathbf{b}}\right\} \\
&= -\frac{1}{2}\sum_{i=1}^{m} \frac{\partial^2 \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j\partial\mathbf{b}} = -\frac{1}{2}\sum_{i=1}^{m} \mathbf{X}_i^t \left(2\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\right)(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \\
&= -\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) .
\end{aligned} \tag{5.102}$$

Finally, as concerns the derivation of an expression for $\partial^2\lambda/\partial\theta_j\partial\theta_r$, straightforward matrix algebra is used again to give the following result for the second partial derivative of the log-likelihood term $(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})$, with respect to θ_j and θ_r, $(j, r = 1, 2, ..., q)$:

$$\begin{aligned}
\frac{\partial^2 \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\theta_j \partial\theta_r} &= \frac{\partial}{\partial\theta_j}\left\{\frac{\partial \left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]}{\partial\theta_r}\right\} \\
&= \frac{\partial}{\partial\theta_j}\left[(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \frac{\partial \mathbf{V}_i^{-1}}{\partial\theta_r} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right] \\
&= \frac{\partial}{\partial\theta_j}\left[-(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial\theta_r}\mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]
\end{aligned}$$

$$= -(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \frac{\partial}{\partial\theta_j}\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\mathbf{V}_i^{-1}\right)(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\,. \tag{5.103}$$

Additionally we have:

$$\frac{\partial^2 \ln\mid\mathbf{V}_i\mid}{\partial\theta_j\partial\theta_r} = \frac{\partial}{\partial\theta_j}\left(\frac{\partial\ln\mid\mathbf{V}_i\mid}{\partial\theta_r}\right) \underset{(5.93)}{=} \frac{\partial}{\partial\theta_j}\left[tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right)\right],$$

and since $\mathbf{V}_i^{-1}\partial\mathbf{V}_i/\partial\theta_r$ is continuously differentiable, then $tr\left(\mathbf{V}_i^{-1}\partial\mathbf{V}_i/\partial\theta_r\right)$ is also continuously differentiable, thus:

$$\frac{\partial}{\partial\theta_j}\left[tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right)\right] = tr\left[\frac{\partial}{\partial\theta_j}\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right)\right],$$

and consequently it is:

$$\frac{\partial^2 \ln\mid\mathbf{V}_i\mid}{\partial\theta_j\partial\theta_r} = tr\left[\frac{\partial}{\partial\theta_j}\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right)\right].$$

Now, by applying the following rule that concerns the partial derivative of the product of two functions f, $g : \mathcal{R}^n \mapsto \mathcal{R}$, and states that $\frac{\partial(fg)}{\partial x} = \frac{\partial f}{\partial x}g + f\frac{\partial g}{\partial x}$, we get:

$$\begin{aligned}\frac{\partial^2 \ln\mid\mathbf{V}_i\mid}{\partial\theta_j\partial\theta_r} &= tr\left[\frac{\partial\mathbf{V}_i^{-1}}{\partial\theta_j}\frac{\partial\mathbf{V}_i}{\partial\theta_r} + \mathbf{V}_i^{-1}\frac{\partial}{\partial\theta_j}\left(\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right)\right] \\ &= tr\left(-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r} + \mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\theta_j\partial\theta_r}\right). \end{aligned} \tag{5.104}$$

By applying (5.103), (5.104) it can be shown (*Jennrich* and *Schluchter, 1986*) that $\partial^2\lambda/\partial\theta_j\partial\theta_r$ is written as:

$$\frac{\partial^2\lambda}{\partial\theta_j\partial\theta_r} = -\frac{1}{2}\sum_{i=1}^{m} tr\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\left[2(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t - \mathbf{V}_i\right]\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r} +$$
$$+\frac{1}{2}\sum_{i=1}^{m} tr\mathbf{V}_i^{-1}\left[(\mathbf{y}_i - \mathbf{X}_i\mathrm{b})(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t - \mathbf{V}_i\right]\mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\theta_j\partial\theta_r}.$$

5.5.5.2 REML Estimation via the Newton-Raphson Algorithm

In the current section, we consider implementation of Newton-Raphson algorithm for calculating restricted maximum likelihood (REML) estimates of variance components in the Laird-Ware model: $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$, $i = 1, 2, ..., m$. REML estimation differs from standard ML estimation in that while the latter obtains estimates by maximizing the 'full' data $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)^t$ log-likelihood function:

$$\lambda_{ML} = const. - \frac{1}{2}\sum_{i=1}^{m} \ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum_{i=1}^{m} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}),$$

REML method maximizes the log-likelihood function of a set of error contrasts $\mathbf{Ky}$, where $\mathbf{K}$ is a $(N \times N)$ full-rank matrix. As regards the log-likelihood function of this 'restricted' data $\mathbf{Ky}$, most often denoted by λ_{REML}, we have already seen that it is expressed as:

$$\begin{aligned} &\lambda_{REML} = \\ &= const. - \frac{1}{2}\ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid - \frac{1}{2}\sum_{i=1}^{m} \ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum_{i=1}^{m} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})^t \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \\ &= -\frac{1}{2}\ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i \mid + \lambda_{ML}, \end{aligned} \tag{5.105}$$

where λ_{ML} corresponds to the log-likelihood function of the standard maximum likelihood method.

The N-R algorithm for the REML shares the following, iterative scheme with the N-R version for ML:

$$\begin{pmatrix} \mathbf{b}^{(k+1)} \\ \theta^{(k+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{b}^{(k)} \\ \theta^{(k)} \end{pmatrix} - \left(\mathbf{H}_{REML}^{(k)}\right)^{-1} \begin{pmatrix} \partial\lambda_{REML}/\partial\mathbf{b} \mid_{\mathbf{b}^{(k)}} \\ \partial\lambda_{REML}/\partial\boldsymbol{\theta} \mid_{\boldsymbol{\theta}^{(k)}} \end{pmatrix}, \tag{5.106}$$

where

$$\mathbf{H}_{REML} = \begin{pmatrix} \partial^2\lambda_{REML}/\partial\mathbf{b}\partial\mathbf{b} & \partial^2\lambda_{REML}/\partial\mathbf{b}\partial\boldsymbol{\theta} \\ \partial^2\lambda_{REML}/\partial\boldsymbol{\theta}\partial\mathbf{b} & \partial^2\lambda_{REML}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta} \end{pmatrix} \tag{5.107}$$

Thus, the only difference with ML estimation via N-R is that the usual log-likelihood λ_{ML} is replaced by λ_{REML}. Observing now equation (5.105), one easily sees that the only difference between λ_{REML} and the standard log-likelihood λ_{ML} is just caused by the extra term:

$$-\frac{1}{2}\ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid .$$

For the derivation thus of first and second order partial derivatives of λ_{REML} with respect to $\mathbf{b}$ and the components of $\boldsymbol{\theta}$, that are required by the iterative N-R algorithm of (5.106), additionally to the derivatives already found in the previous section, we only have to find expressions for the following six partial derivatives of term $\ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid$, namely $\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\mathbf{b}$, $\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\theta_j$ (required for calculation of $\partial\lambda_{REML}/\partial\mathbf{b}$ and $\partial\lambda_{REML}/\partial\boldsymbol{\theta}$, respectively), and $\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\mathbf{b}\partial\mathbf{b}$, $\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\mathbf{b}\partial\theta_j$, $\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\theta_j\partial\mathbf{b}$, $\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid /\partial\theta_j\partial\theta_r$ (required for the calculation of the elements of $\mathbf{H}_{REML}$).

Since, for obvious reasons,

$$\frac{\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial\mathbf{b}} = \mathbf{0},$$

$$\frac{\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial\mathbf{b}\partial\mathbf{b}} = \mathbf{0},$$

and

$$\frac{\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial\theta_j\partial\mathbf{b}} = \mathbf{0} = \frac{\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial\mathbf{b}\partial\theta_j},$$

what remains in suspense, is calculation of the following partial derivatives:

$$\frac{\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial \theta_j}, \text{ and } \frac{\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial \theta_j \partial \theta_r},$$

for $j, r = 1, 2, ..., q$. Thus, by using again result (3.44), the first-order partial derivative of $\ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid$ with respect to the variance component θ_j $(j = 1, 2, ..., q)$ is:

$$\begin{aligned}\frac{\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial \theta_j} &= tr \left[\left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \frac{\partial \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i}{\partial \theta_j} \right] \\ &= tr \left[\left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \frac{\partial \mathbf{V}_i^{-1}}{\partial \theta_j} \mathbf{X}_i \right],\end{aligned}$$

and by setting $\mathbf{A} = \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i$ for notational convenience, the above becomes:

$$\begin{aligned}\frac{\partial \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial \theta_j} &= tr \left(\mathbf{A}^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \frac{\partial \mathbf{V}_i^{-1}}{\partial \theta_j} \mathbf{X}_i \right) \\ &= \sum_{i=1}^{m} tr \left[\mathbf{A}^{-1} \mathbf{X}_i^t \left(-\mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_j} \mathbf{V}_i^{-1} \right) \mathbf{X}_i \right] \\ &= -\sum_{i=1}^{m} tr \left(\mathbf{A}^{-1} \mathbf{X}_i^t \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \theta_j} \mathbf{V}_i^{-1} \mathbf{X}_i \right). \qquad (5.108)\end{aligned}$$

Finally, based upon the previous result it can be shown (*Lindstrom & Bates, 1988*) that derivative $\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid / \partial \theta_j \partial \theta_r$, necessary for the N-R algorithm, is formed as:

$$\frac{\partial^2 \ln \mid \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i \mid}{\partial \theta_j \partial \theta_r} =$$

$$= -tr\left[\mathbf{A}^{-1}\sum_{i=1}^{m}\left(\mathbf{X}_i^t\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\mathbf{X}_i\right)\times\mathbf{A}^{-1}\sum_{i=1}^{m}\left(\mathbf{X}_i^t\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\mathbf{V}_i^{-1}\mathbf{X}_i\right)\right]$$
$$+tr\left\{\mathbf{A}^{-1}\sum_{i=1}^{m}\left[\mathbf{X}_i^t\mathbf{V}_i^{-1}\left(\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}-\frac{\partial^2\mathbf{V}_i}{\partial\theta_j\partial\theta_r}+\frac{\partial\mathbf{V}_i}{\partial\theta_r}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\right)\mathbf{V}_i^{-1}\mathbf{X}_i\right]\right\}.$$

5.5.6 Implementing the Fisher Scoring Algorithm

In this section, we consider in brief one of the most significant variants of the iterative Newton-Raphson, namely the Fisher scoring algorithm, and its implementation to the Laird-Ware model for computing estimates of the variance components of the latter model. Fisher scoring differs from standard N-R algorithm in that all second derivatives of the log-likelihood function are replaced by their expectations. Hence, in other words, Fisher scoring simply replaces the Hessian matrix **H** by its expectation. *Jennrich* and *Schluchter (1986)*, considers Fisher scoring, introducing all necessary formulas for the computation of variance components in the Laird-Ware model through this algorithm. Specifically, as pointed out by Jennrich, the expectations of the second partial derivatives of the Hessian matrix are given by:

$$E\left(\frac{\partial^2\lambda}{\partial\mathbf{b}\partial\mathbf{b}}\right)=-\sum_{i=1}^{m}\mathbf{X}_i^t\mathbf{V}_i^{-1}\mathbf{X}_i, \tag{5.109}$$

$$E\left(\frac{\partial^2\lambda}{\partial\mathbf{b}\partial\boldsymbol{\theta}}\right)=E\left(\frac{\partial^2\lambda}{\partial\boldsymbol{\theta}\partial\mathbf{b}}\right)=0, \tag{5.110}$$

and

$$E\left(\frac{\partial^2\lambda}{\partial\theta_j\partial\theta_r}\right)=-\frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_j}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\theta_r}\right), \tag{5.111}$$

for every $j, r = 1, 2, ..., q$. As in every Fisher scoring algorithm in general, the expected Hessian matrix:

$$E(\mathbf{H})=\begin{pmatrix} E\left(\partial^2\lambda/\partial\mathbf{b}\partial\mathbf{b}\right) & E\left(\partial^2\lambda/\partial\mathbf{b}\partial\boldsymbol{\theta}\right)\\ E\left(\partial^2\lambda/\partial\boldsymbol{\theta}\partial\mathbf{b}\right) & E\left(\partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}\right)\end{pmatrix}$$

is easier to compute, compared to the Hessian matrix $\mathbf{H}$ of (5.86), due to that its off-diagonal elements $E\left(\partial^2\lambda/\partial\mathbf{b}\partial\boldsymbol{\theta}\right)$, $E\left(\partial^2\lambda/\partial\boldsymbol{\theta}\partial\mathbf{b}\right)$ are zero. As a consequence to this, the estimates of $\mathbf{b}$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_q)^t$ can be obtained by solving separate equations, differing in this way from the iterative single equation (5.85) that simultaneously estimates both $\mathbf{b}$ and $\boldsymbol{\theta}$ via N-R algorithm. Thus, for Fisher scoring, the 'new' estimates $\mathbf{b}^{(k+1)}$ are obtained by the maximum likelihood (or the generalized least squares) step (see equation 5.9):

$$\mathbf{b}^{(k+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^t \mathbf{V}_i^{-1} \mathbf{y}_i, \tag{5.112}$$

whereas 'new' estimates for the vector of variance components $\boldsymbol{\theta}$ are computed by the iterative scheme:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mathbf{I}_{\boldsymbol{\theta\theta}}^{-1} \frac{\partial\lambda}{\partial\boldsymbol{\theta}}, \tag{5.113}$$

where $\mathbf{I}_{\boldsymbol{\theta\theta}}$ is the negative of the matrix $E\left(\partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}\right)$.

There are several motivations for employing Fisher scoring instead of N-R algorithm. For start, Fisher scoring behaves better far from the solution (i.e. exhibits better handling of poor starting values), whereas it has reasonable convergence near the solution. A further advantage is that calculation of $E\left(\mathbf{H}\right)$ is easier than calculating $\mathbf{H}$ itself, reducing in this way the computational efforts. These reasons made implementation of the former algorithm a possible choice on variance component estimation, especially for models such as the GLMM for longitudinal data. Even though Fisher scoring is still not considered to be the standard method for estimating the variance components $\boldsymbol{\theta}$ in the Laird-Ware model, some of its virtues are quite appreciable in practice. For example, SAS procedure PROC MIXED, uses Fisher scoring for the first iteration and then N-R for the remaining iterations as the default fitting method.

5.5.7 Comparison of the EM and Newton-Raphson Algorithms

Thus far, two general purpose algorithms for iterative (ML/REML) estimation of the variance components in the Laird-Ware model were considered; the Expectation-Maximization and the Newton-Raphson algorithm. The main question that arises at this point is obvious; does any of the two algorithms performs better compared to the other, or their differences as concern their performance just tend to be indistinguishable in general.

The latter question of determining the 'best', if any, among these two algorithms is not an easy question to answer. Comparison of EM algorithm and derivative based algorithms (such as N-R and Fisher scoring) has been the focus of many authors since the development of the expectation-maximization algorithm in the late seventies (*Dempster et al., 1977*). Attempting to come up with a basic conclusion, we can say that both algorithms share their own advantages, suffering at the same time from several drawbacks, and thus the decision to use either the EM or the N-R algorithm should depend on the specific statistical problem, weighing each time the trade-offs between the two procedures.

Specifically, as concern the convergence rate of the two algorithms, N-R exhibits a quadratic rate of convergence, while EM has typically (very) slow linear convergence. In fact, this slow convergence rate is the main disadvantage of EM algorithm. The more variance components there are to estimate, the longer will be the number of iterates to reach convergence. Although various acceleration schemes (i.e. variants of the EM algorithm) have been proposed to improve convergence rate, they generally require significant analytical work, increasing thus the complexity of the algorithm. Another cause of slowness of the EM algorithm, besides its convergence rate, usually arises when the E- or the M-step does not admit an analytical solution. In situations like this it is necessary to resort to iterative methods for the computation of the expectation or the maximization. For example, a common case is where Monte-Carlo approximations of the E-step are used.

On the other hand, however, EM algorithm has an additional positive feature not shared by derivative type methods such as N-R and Fisher scoring; its convergence is global, which means that the algorithm converges to the solution from any starting point. This fact is guaranteed by the monotonous increase of the likelihood at each iteration step (see *Dempster et al., 1977* for the proof). This property of EM establishes the latter algorithm as a very stable algorithm, in contrast to N-R type algorithms that do not guarantee convergence and may diverge from starting points that are not appropriately chosen. In addition, Newton-Raphson is a computationally intensive algorithm that requires heavy analytical preparatory work compared to other likelihood maximization methods, since the necessary calculations of the gradient and the Hessian matrix can be very complex in general. Moreover, implementation of the latter method may present numerical difficulties, particularly when the number of parameters to be estimated is large. Instead, EM is considered to be quite simple and easy to implement, reducing thus substantially the computational efforts needed for its implementation.

Another interesting point is the insufficiency of EM algorithm to compute an estimate of the parameters' variance-covariance matrix (standard error of the parameter). This disadvantage is not shared by N-R which provides consistent standard errors for the parameter estimates, automatically. Extensions of the EM algorithm intended to fix this problem have been proposed (e.g. *Meng* and *Rubin, 1991*), but they result in increasing the complexity of the implementation, canceling out the most important advantage of EM, its simplicity.

In the context of "Laird-Ware" model now, *Lindstrom* and *Bates (1988)* compare in detail the EM and N-R algorithms, as well as a variant of EM, the EM with Aitken's acceleration, as methods for obtaining estimates of the variance components. For this purpose they use two different data sets to compare the behavior of the three algorithms. Their analysis verified once again the slow convergence of the EM algorithm. For both data sets, the number of iterations required for the Newton-Raphson algorithm was quite small compared with the corresponding number for the EM. Although Aitken's acceleration improved the convergence rate of EM substantially in some situations, this does not generalize to all cases.

The authors concluded that although N-R is not guaranteed to converge (EM will always converge to a local maximum of the likelihood surface), the prohibitively large number of iterations required by the EM algorithm is still a serious drawback in spite of the acceleration schemes used for speeding convergence. Thus, they propose N-R estimation as the most preferable method, especially in cases where the number of random effects q is very small. This is because when q is relatively small it has been observed that there is no significant computational penalty in using the N-R algorithm, in comparison to EM algorithm. *Jennrich* and *Schluchter (1986)* also recommend using N-R type methods when q is small, while for large q they prefer using their generalized EM (GEM) algorithm (see, e.g. *Dempster et al., 1977* for more on GEM algorithm).

As a final remark, we can say that both EM and derivative based (e.g. Newton-Raphson, Fisher scoring) iterative methods can be well-applied to the model $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$, ($i = 1, 2, ..., m$) for likelihood maximization. However, while in the past the EM algorithm has been preferred over the N-R algorithm (mainly due to that each iteration could be computed more quickly), nowadays the advances in statistical software has forced practicing statisticians to usually use N-R based procedures for the estimation of all parameters in the latter model.

5.6 Testing for Fixed Effects

While estimation of effects (both fixed and random) in the GLMM for longitudinal data is usually of prime importance, tests of hypotheses associated with the fixed effects vector $\mathbf{b}$ will inevitably be required to assess the significance of the latter effects. Thus, for practical use one may be interested in testing the significance of the entire fixed-effects parameter vector $\mathbf{b}$ that summarizes all fixed effects parameters included in the Laird-Ware model: $\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i$, ($i = 1, 2, ..., m$). Furthermore, often questions of interest may be phrased in terms of a linear combination of the elements of vector $\mathbf{b}$.

Generally, the most convenient way to draw inferences on the above questions is to set up a hypothesis associated with the question of interest and then use a statistical test to check the specific hypothesis. In fact, each hypothesis associated with the $(p \times 1)$ vector $\mathbf{b}$ can be expressed by specifying appropriate full rank matrices $\mathbf{L}$ of dimension

$(r \times p)$ to form suitable contrasts (i.e. linear combinations) of $\mathbf{b}$, given by $\mathbf{Lb}$. Interest is focussed in the test of the following hypotheses:

$$\begin{gathered} H_0 : \mathbf{Lb} = \mathbf{h} \\ vs \\ H_1 : \mathbf{Lb} \neq \mathbf{h} \end{gathered} \tag{5.114}$$

where $\mathbf{h}$ is a specified $(r \times 1)$ vector. Most often, $\mathbf{h}$ will be equal to $\mathbf{0}$. Basically, three approaches are used to test the above null hypothesis H_0. Likelihood ratio tests can be used with large samples, providing one uses maximum likelihood (rather than REML) for model fitting. Also standard Wald tests are widely available. Finally, approximate F-tests can be carried out by dividing the Wald statistic by the numerator degrees-of-freedom and approximating the denominator degrees-of-freedom. Below we give a brief account of the preceding tests.

5.6.1 The Wald Test Statistic

The classic Wald test, originally proposed by *Wald (1943)* for testing hypotheses concerning the regression coefficients of linear regression models, has been successfully conveyed in longitudinal data analysis as a means of testing hypotheses concerning contrasts of the fixed effects vector $\mathbf{b}$ of the Laird-Ware model. Specifically, the Wald type procedure for testing the important class of hypotheses (5.114) exploits the distributional form of the ML estimator of $\mathbf{b}$, $\hat{\mathbf{b}}$. As already demonstrated, the ML estimate $\hat{\mathbf{b}}$ follows, approximately, a multivariate normal distribution with mean $E\left(\hat{\mathbf{b}}\right) = \mathbf{b}$ and variance-covariance matrix $Var\left(\hat{\mathbf{b}}\right) = \left(\sum\limits_{i=1}^{m} \mathbf{X}_i^t \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right)^{-1}$, i.e.:

$$\hat{\mathbf{b}} \sim N_p \left[\mathbf{b}, \left(\sum_{i=1}^{m} \mathbf{X}_i^t \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i\right)^{-1}\right]. \tag{5.115}$$

Since the fixed-effects vector $\mathbf{b}$ is unknown, it seems reasonable that an estimate of the quantity **Lb** could be obtained by substituting **b** with an estimator of it. For instance, we could use its ML estimator $\hat{\mathbf{b}}$. Obviously, in the light of (5.115), we obtain:

$$E\left(\mathbf{L}\hat{\mathbf{b}}\right) = \mathbf{L}E\left(\hat{\mathbf{b}}\right) = \mathbf{L}\mathbf{b},$$

and

$$Var\left(\mathbf{L}\hat{\mathbf{b}}\right) = \mathbf{L}Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^t = \mathbf{L}\left(\sum_{i=1}^{m}\mathbf{X}_i^t\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{L}^t.$$

Thus, the sampling distribution of the linear function $\mathbf{L}\hat{\mathbf{b}}$ can be approximated by:

$$\mathbf{L}\hat{\mathbf{b}} \sim N_r\left[\mathbf{L}\mathbf{b}, \mathbf{L}\left(\sum_{i=1}^{m}\mathbf{X}_i^t\hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1}\mathbf{L}^t\right]. \tag{5.116}$$

Now, to test null hypothesis H_0 of (5.114), we form the following statistic (**Wald statistic**):

$$T_W = \left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right)^t\left(\mathbf{L}Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^t\right)^{-1}\left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right). \tag{5.117}$$

Since $\mathbf{L}\hat{\mathbf{b}}$ is approximately normally distributed, it may be argued that statistic T_L follows (approximately) a chi-square distribution with r degrees of freedom. The reason for this is that, as $m \to \infty$, $\mathbf{L}\hat{\mathbf{b}} \sim N_r\left(\mathbf{L}\mathbf{b}, \mathbf{L}Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^t\right)$ and consequently under H_0 we have $\mathbf{L}\hat{\mathbf{b}} - \mathbf{h} \sim N_r\left(0, \mathbf{L}Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^t\right) \Rightarrow \left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right)^t\left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right)/\mathbf{L}Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^t \sim x_r^2$.

Thus the Wald test for testing H_0 may be conducted by comparing statistic T_L to an approximate x_r^2 critical value. We would reject H_0 at a (predetermined) significance level α if $T_L > x_r^2(1-\alpha)$.

Unfortunately, Wald type approaches as the above may suffer from a serious drawback; when the number of subjects m is not too large, the resulting inferences may not be too reliable. This is because these approaches rely on a normal approximation to the sampling distribution that may be a lousy approximation unless m is relatively large.

5.6.2 The Likelihood Ratio Test

An alternative to Wald approximate methods is that of the likelihood ratio test (LRT). The LRT, already discussed for model selection between models of various covariance structures (see subsection 5.4.3), is applicable in the situation in which we wish to test what are often called 'reduced' versus 'full' model hypotheses. That is, the LRT can be implemented for testing hypotheses concerning 'nested' models and only under this perspective can be used for testing hypotheses of fixed effects such as hypothesis (5.114). For the specific hypothesis thus, the full model is taken to be the model under H_1 (that is the general model that does not make any particular assumptions about the contrast $\mathbf{Lb}$), while, for the reduced model, we take the model for which the restriction of H_0 (i.e. $\mathbf{Lb} = \mathbf{h}$) is imposed. If by $\hat{L}_{full}$ and $\hat{L}_{red}$ we denote the values of the maximized likelihoods for the full (under H_1) and the reduced (under H_0) model respectively, then the **likelihood ratio statistic** is given by:

$$T_{LRT} = -2\ln\left(\frac{\hat{L}_{red}}{\hat{L}_{full}}\right) = -2\left(\ln \hat{L}_{red} - \ln \hat{L}_{full}\right). \tag{5.118}$$

Technical arguments may be used to show that, as the sample size m tends to ∞, the statistic T_{LRT} follows approximately a chi square distribution with degrees of freedom equal to the difference in number of parameters in the two models (i.e. number of parameters of the full model minus number of parameters of the reduced model). Thus, if we denote this difference with r, then we reject null hypothesis H_0 at level of significance α if $T_{LRT} > x_r^2\left(1-\alpha\right)$.

Despite the fact that the LRT is (as the Wald test) an approximate method based on large sample theory (i.e. large m), it has proven to be more reliable for small sample sizes than the Wald test (see for instance *Agresti, 1996*). Thus, the LRT is usually more preferable to Wald approaches when sample size m is not too large.

5.6.3 The F-test

As an alternative to the Wald and likelihood ratio test approaches for testing hypotheses of the form (5.114), one can use F-statistics of the following form:

$$F = \frac{\left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right)^{t} \left(\mathbf{L} Var\left(\hat{\mathbf{b}}\right)\mathbf{L}^{t}\right)\left(\mathbf{L}\hat{\mathbf{b}} - \mathbf{h}\right)}{rank\left(\mathbf{L}\right)}. \tag{5.119}$$

It can be shown that the above statistic follows an F-distribution. The number of the numerator degrees of freedom is equal to $rank\left(\mathbf{L}\right)$. The denominator degrees of freedom have to be approximated. For more details on approximate F-tests we refer the interested reader to *Fai* and *Cornelius (1996).*

5.7 Diggle's Extension to the "Laird-Ware" Model

Thus far, it has been clear that the interest as concerns the covariance structure of the Laird-Ware model is mainly focused on the random effects $\mathbf{u}_i$ variance-covariance matrix, namely $\mathbf{D}$. On the contrary, for the variance-covariance matrix $\mathbf{R}_i$ of random errors $\boldsymbol{\varepsilon}_i$, usually a very simple covariance structure is adopted. Take for example the very commonly used conditional-independence model $[Var\left(\boldsymbol{\varepsilon}_i\right) = \sigma^2 \mathbf{I}_{n_i}]$. A serious drawback when following this approach is the failure to take under consideration the possible stochastic variation between pairs of measurements taken on the same subject (within-subject variation).

Diggle (1988, 1990 Chapter 5) was the first to develope a parametric model that suggests an alternative specification for the error term's variance-covariance matrix, and may be viewed as an extension of the GLMM for longitudinal data. The framework for the formulation of the author's model is once again the typical longitudinal study, where repeated observations are taken on, say m, units (or subjects) at different time intervals. More formally, let y_{ij} denote the jth measurement taken on the ith subject. The total number of repeated measurements on each subject i is usually represented by the response vector $\mathbf{y}_i = \left(y_{i_1}, y_{i_2}, ..., y_{i_{n_i}}\right)^t$. In *Diggle (1988),* it is suggested a different approach in order to obtain a parsimonious parameterization of the covariance structure of the data. More specifically, Diggle introduced a model for longitudinal data that incorporates variation between subjects, measurement error, as well as the serial correlation between the measurements of the same subject. Moreover, this within-subject correlation is formulated in such a way so that it depends on the measurements' separation in time. According to *Diggle (1988),* a model that incorporates all of the above characteristics is:

$$y_{ij} = \mu_{ij} + Z_{ij} + U_i + W_i\left(t_{ij}\right), \tag{5.120}$$

for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. The Z_{ij} are i.i.d. variates corresponding to measurement error, with $Z_{ij} \sim N\left(0, \tau^2\right)$. The U_i are i.i.d. $N\left(0, \nu^2\right)$ variates representing the between-subjects variation, and finally $W_i(t_{ij})$ are independent stationary Gaussian processes with $E[W_i(t_{ij})] = 0$ and $Cov\{W_i(t_{ij}), W_i(t_{ik})\} = \sigma^2 \rho(\mid t_{ij} - t_{ik} \mid)$. $\rho(u)$ is some correlation function that is suitably specified to describe serial correlation of W_i process, and is such that $\rho(0) = 1$ and $\rho(u) \to 0$ as $u \to \infty$ (t_{ij} denotes the time at which the measurement y_{ij} is taken). Since the development of the latter model, several choices for $\rho(u)$ have been proposed. Two of the most popular among them are the exponential and Gaussian serial correlation functions, defined as:

$$\rho(u) = \exp(-\phi u) \quad \text{and} \quad \rho(u) = \exp(-\phi u^2),$$

respectively, for some $\phi > 0$. The most important qualitative difference between these two correlation functions is their behavior near $u = 0$ [theoretically, any correlation function of the form $\rho(u) = \exp\left(-\phi u^k\right)$ for any fixed value of k may be used]. Under model (5.120), it is:

$$\begin{aligned} Var(y_{ij}) &= Var\left(\mu_{ij} + Z_{ij} + U_i + W_i\right) \\ &= Var(Z_{ij}) + Var(U_i) + Var(W_i) \\ &= \tau^2 + \nu^2 + \sigma^2, \end{aligned} \tag{5.121}$$

and also

$$\begin{aligned} Cov(y_{ij}, y_{ik}) &= E\left[\left(y_{ij} - \mu_{ij}\right)(y_{ik} - \mu_{ik})\right] \\ &= E\{[Z_{ij} + U_i + W_i(t_{ij})][Z_{ik} + U_i + W_i(t_{ij})]\} \\ &= \ldots = \nu^2 + \sigma^2 \rho(\mid t_{ij} - t_{ik} \mid). \end{aligned} \tag{5.122}$$

Now, considering the entire vector of responses $\mathbf{y}_i$ for subject i, as shown by Diggle, the mean vector $E(\mathbf{y}_i)$ and variance-covariance matrix $Var(\mathbf{y}_i) = \mathbf{V}_i$ of each response vector $\mathbf{y}_i$ becomes:

$$E(\mathbf{y}_i) = \boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i}\right)^t, \tag{5.123}$$

and

$$\mathbf{V}_i = \tau^2 \mathbf{I}_{n_i} + \nu^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t + \sigma^2 \mathbf{H}_i, \tag{5.124}$$

respectively, where $\mathbf{H}_i$ is the $(n_i \times n_i)$ matrix whose (j,k)th element is $\rho\left(\mid t_{ij} - t_{ik} \mid\right)$, $\mathbf{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix, and $\mathbf{1}_{n_i}$ is a $(n_i \times 1)$ vector of ones. To see the close relation of Diggle's model with the standard 'Laird-Ware' model, notice that the former may be alternatively represented using matrix notation as a general linear mixed model of the form:

$$\mathbf{y}_i = \mathbf{I}_{n_i} \boldsymbol{\mu}_i + \mathbf{1}_{n_i} U_i + \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i}. \tag{5.125}$$

In (5.125), $\mathbf{I}_{n_i}$, $\boldsymbol{\mu}_i$, $\mathbf{1}_{n_i}$ and U_i are as already defined and $\boldsymbol{\varepsilon}_{(1)_i}$, $\boldsymbol{\varepsilon}_{(2)_i}$ represent random error terms chosen such that $\boldsymbol{\varepsilon}_{(1)_i} \sim N_{n_i}\left(\mathbf{0}, \tau^2 \mathbf{I}_{n_i}\right)$ and $\boldsymbol{\varepsilon}_{(2)_i} \sim N_{n_i}\left(\mathbf{0}, \sigma^2 \mathbf{H}_i\right)$. The term $\mathbf{I}_{n_i} \boldsymbol{\mu}_i$ defines the fixed part $(\mathbf{X}_i \mathbf{b})$, while term $\mathbf{1}_{n_i} U_i$ can be considered to be the random part $(\mathbf{Z}_i \mathbf{u}_i)$. Observe that by determining $\mathbf{1}_{n_i} U_i$ to be the random effects term, essentially we include only a random intercept in the model.

Model (5.125) is another way to express model (5.120), since both models yield the same mean and variance-covariance matrix for the multivariate Gaussian vector $\mathbf{y}_i$ [indeed, observe that $E\left(\mathbf{y}_i\right) = E\left(\mathbf{I}_{n_i} \boldsymbol{\mu}_i + \mathbf{1}_{n_i} U_i + \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i}\right) = \boldsymbol{\mu}_i$ and in addition that $Var\left(\mathbf{y}_i\right) = Var\left(\mathbf{I}_{n_i} \boldsymbol{\mu}_i + \mathbf{1}_{n_i} U_i + \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i}\right) = \mathbf{1}_{n_i} Var\left(U_i\right) \mathbf{1}_{n_i}^t + \tau^2 \mathbf{I}_{n_i} + \sigma^2 \mathbf{H}_i = \nu^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^t + \tau^2 \mathbf{I}_{n_i} + \sigma^2 \mathbf{H}_i$]. The main difference of Diggle's model (5.120), compared to a standard Laird-Ware model of the form $\mathbf{y}_i = \mathbf{X}_i \mathbf{b} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i$, is that the former essentially decomposes the unique error term $\boldsymbol{\varepsilon}_i$ of the Laird-Ware model into two components as:

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i},$$

so that $\boldsymbol{\varepsilon}_{(1)_i}$ represents the within-subject variation due to measurement error, and $\boldsymbol{\varepsilon}_{(2)_i}$ accounts for the extra random component, introduced by Diggle, to incorporate the (possible) dependence of the within-individual's responses.

5.7.1 An Additional Extension

Diggle et al. (1994) expanded further the model proposed by Diggle. In particular, they introduce an extended Laird-Ware model which can be expressed formally as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i}, \tag{5.126}$$

with

$$\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right),$$
$$\boldsymbol{\varepsilon}_{(1)_i} \sim N_{n_i}\left(\mathbf{0}, \tau^2\mathbf{I}_{n_i}\right),$$
$$\boldsymbol{\varepsilon}_{(2)_i} \sim N_{n_i}\left(\mathbf{0}, \sigma^2\mathbf{H}_i\right).$$

The terms $\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{b}$, $\mathbf{u}_i$ and $\mathbf{D}$ are as defined for the Laird-Ware model. As was the case with Diggle's model, the error terms $\boldsymbol{\varepsilon}_{(1)_i}$ and $\boldsymbol{\varepsilon}_{(2)_i}$ express the decomposition of the overall random error, say $\boldsymbol{\varepsilon}_i$, to measurement error and to a component of serial correlation, respectively. Further, $\boldsymbol{\varepsilon}_{(1)_i}$ is assumed to be independent of $\boldsymbol{\varepsilon}_{(2)_i}$.

Model (5.126) is a generalization of Diggle's model, in the sense that it considers a more general structure for both fixed and random effects of the model. For instance, the former includes a general random effects vector, while the latter considers only a single random intercept. For comparison, recall the matrix representation of Diggle's model: $\mathbf{y}_i = \mathbf{I}_{n_i}\boldsymbol{\mu}_i + \mathbf{1}_{n_i}U_i + \boldsymbol{\varepsilon}_{(1)_i} + \boldsymbol{\varepsilon}_{(2)_i}$. Here, the fixed effects vector $\mathbf{b}$ was specifically chosen to be $\boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, ..., \mu_{in_i}\right)^t$, the random effects vector $\mathbf{u}_i$ was taken to be the univariate random variable $U_i \sim N\left(0, \nu^2\right)$ and accordingly the design matrices $\mathbf{X}_i$, $\mathbf{Z}_i$ could only be $\mathbf{I}_{n_i}$ and $\mathbf{1}_{n_i}$, respectively. Now, as regards the mean vector and variance-covariance matrix for the response vector $\mathbf{y}_i$, we have:

$$E\left(\mathbf{y}_i\right) = \mathbf{X}_i\mathbf{b},$$

which is identical to the mean structure of the standard Laird-Ware model, and

$$Var\left(\mathbf{y}_i\right) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^t + \tau^2\mathbf{I}_{n_i} + \sigma^2\mathbf{H}_i.$$

A more detailed review concerning the extended model (5.126) as well as simplifications of it (e.g. Diggle's model), is given in *Diggle et al. (1994)*.

5.7.2 The Semivariogram

Analysis of longitudinal data via the linear mixed-effects model of Laird and Ware, includes among others the specification of an appropriate covariance structure that best fits the true covariance of the data. For models of the same fixed effects but of different covariance structures, comparisons can be conducted via the Akaike information criterion (AIC) and/or the Schwarz Bayesian information criterion (SBC), while likelihood ratio tests can be used for nested structures to identify better fitting models. Furthermore, parallel axis plots and Draftman's displays may provide significant help in selecting the model with the best covariance structure.

Diggle (1988) outlined another practical approach to the choice and validation of the covariance structure of his model (5.120), by using the empirical semivariogram. The semivariogram (*Matheron, 1963*) is a function, of widespread usage in geostatistics, for the estimation of the covariance structure. For a formal definition, let $Y(t)$ be a real-valued, second-order stationary random process. Following *Matheron (1963)*, the functions:

$$\gamma(u) = Var\left[Y(t) - Y(t+u)\right] = E\left[Y(t) - Y(t+u)\right]^2, \tag{5.127}$$

and

$$g(u) = \frac{1}{2} Var\left[Y(t) - Y(t+u)\right] = E\left\{\frac{1}{2}\left[Y(t) - Y(t+u)\right]^2\right\}, \tag{5.128}$$

for every $u \geq 0$, are known as the **theoretical variogram**[12] and the **theoretical semivariogram functions**, respectively.

The above defined theoretical functions cannot of course be determined in practice. Thus, instead of the theoretical semivariogram and variogram, we usually use estimators

[12] **The term variogram is attributed to *Matheron (1963)*. The specific function though, has been also called a *structure function* in physics (*Kolmogorov, 1941*) and a *mean squared difference* in the early time series literature (*Jowett, 1952*).**

of them, that are calculated by some realized values of process $Y(t)$ (namely the observed data in our case). Such estimators of the variogram and semivariogram functions are the empirical variogram and the empirical semivariogram, respectively. By definition, the empirical semivariogram (also referred to as the sample variogram or the semivariogram cloud), of a time series collection of data $\{y(t_j) : \ j = 1, 2, ..., n\}$ is a scatterplot formulated by plotting together all squared differences:

$$\frac{1}{2}\left[y(t_j) - y(t_k)\right]^2, \tag{5.129}$$

against all possible corresponding pairwise lags: $u_{jk} = t_j - t_k$ $(j \neq k)$. However, it is often very difficult for one to identify any structure of the data from the specific plot due to the fact that most of the times there are so many pairwise lags. To improve the appearance of such structureless plot and to allow the underlying semivariogram structure to be seen more easily, a common procedure is to average all squared differences $1/2\left[y(t_j) - y(t_k)\right]^2$ that share the same lag value u_{jk}. Also, alternatively points with similar but not necessarily the same u_{jk} value may be pooled (essentially, averaging over the squared differences of similar distances u_{jk} corresponds to fitting a smoothing curve to these differences). Then, we connect the averaged points that occur to form a curve. This procedure usually yields curves that are smooth in appearance, to describe the empirical semivariogram.

In particular, for constructing the empirical semivariogram of model (5.120), Diggle suggests using $y_{ij}^{*} = y_{ij} - \mu_{ij}$, instead of the 'raw' data y_{ij}. Thus, to formulate the corresponding semivariogram of model $y_{ij} = \mu_{ij} + Z_{ij} + U_i + W_i(t_{ij})$, $(i = 1, 2, ..., m)$, $(j = 1, 2, ..., n_i)$ all squared differences:

$$E\left[\frac{1}{2}\left(y_{ij}^{*} - y_{ik}^{*}\right)\right]^2, \tag{5.130}$$

for all $j \neq k$, must be calculated and plotted against $t_j - t_k$. Now, in order to examine exactly how the shape of the empirical semivariogram can provide useful information about the covariance structure of the data and consequently to assist in selecting a suitable model that will fit the data in the best possible way, consider first that each smoothed (pooled) point of the scatterplot of the sample semivariogram (5.130) can be written as:

$$\begin{aligned}\frac{1}{2}E\left(y_{ij}^{*}-y_{ik}^{*}\right)^{2} &= \frac{1}{2}Var\left(y_{ij}^{*}-y_{ik}^{*}\right)\\ &= \frac{1}{2}\left[Var\left(y_{ij}^{*}\right)+Var\left(y_{ik}^{*}\right)-2Cov\left(y_{ij}^{*},y_{ik}^{*}\right)\right]. \qquad (5.131)\end{aligned}$$

For calculating the above, we need to determine $Var\left(y_{ij}^{*}\right)$, $Var\left(y_{ik}^{*}\right)$ and $Cov\left(y_{ij}^{*},y_{ik}^{*}\right)$. As concerns $Var\left(y_{ij}^{*}\right)$ we have:

$$\begin{aligned}Var\left(y_{ij}^{*}\right) &= Var\left(Z_{ij}+U_{i}+W_{i}\left(t_{ij}\right)\right)\\ &= Var\left(Z_{ij}\right)+Var\left(U_{i}\right)+Cov\left[W_{i}\left(t_{ij}\right),W_{i}\left(t_{ij}\right)\right]\\ &= \tau^{2}+\nu^{2}+\sigma^{2}\rho\left(\mid t_{ij}-t_{ij}\mid\right)=\tau^{2}+\nu^{2}+\sigma^{2}\rho\left(0\right)\\ &= \tau^{2}+\nu^{2}+\sigma^{2}. \qquad (5.132)\end{aligned}$$

Similarly, it is:

$$Var\left(y_{ik}^{*}\right)=\tau^{2}+\nu^{2}+\sigma^{2}, \qquad (5.133)$$

and moreover

$$\begin{aligned}Cov\left(y_{ij}^{*},y_{ik}^{*}\right) &= Cov\left(Z_{ij}+U_{i}+W_{i}\left(t_{ij}\right),Z_{ik}+U_{i}+W_{i}\left(t_{ik}\right)\right)\\ &= ...=Cov\left(U_{i},U_{i}\right)+Cov\left(W_{i}\left(t_{ij}\right),W_{i}\left(t_{ik}\right)\right)\\ &= Var\left(U_{i}\right)+Cov\left(W_{i}\left(t_{ij}\right),W_{i}\left(t_{ik}\right)\right)\\ &= \nu^{2}+\sigma^{2}\rho\left(\mid t_{ij}-t_{ik}\mid\right). \qquad (5.134)\end{aligned}$$

Thus, using (5.132), (5.133) and (5.134), equation (5.131) becomes:

$$\frac{1}{2}E\left(y_{ij}^{*}-y_{ik}^{*}\right)^{2} = \frac{1}{2}\left[2\tau^{2}+2\nu^{2}+2\sigma^{2}-2\nu^{2}-2\sigma^{2}\rho\left(\mid t_{ij}-t_{ik}\mid\right)\right]$$

$$
\begin{aligned}
&= \tau^2 + \sigma^2 - \sigma^2 \rho(\mid t_{ij} - t_{ik} \mid) \\
&= \tau^2 + \sigma^2 \left[1 - \rho(\mid t_{ij} - t_{ik} \mid)\right]. \qquad (5.135)
\end{aligned}
$$

The last equation reveals the important feature incorporated by the empirical semivariogram; that is the latter is a function of the correlation $\rho \equiv \rho(\mid t_{ij} - t_{ik} \mid)$. In particular, equation (5.135) specifies that as the correlation between the within-subjects measurements decreases, the corresponding semivariogram values increase and vice-versa. Consequently, by observing the semivariogram's shape we have a visual identification of whether correlation between the within-subjects measurements depends on their distance (lag) in time, or is constant for all lags $\mid t_{ij} - t_{ik} \mid$. Let us consider an example to illustrate the applicability of the empirical semivariogram on a real data set. The data are from *Box (1950)* and consist of the body weights of 27 rats, collected at a period of 5 weeks. The 27 rats were divided at random into 3 groups, each group associated with a specific treatment applied to the rats (rats 1-10 are on a control treatment, rats 11-17 have had a thyroxin treatment and rats 18-27 have had a treatment with thiouracil). The data have been already presented in Figure 4.3 via the standard parallel plot.

Although parallel plots are both practical and effective for viewing longitudinal data, unfortunately they cannot provide additional guidance in checking the data's covariance structure. As already stated, the main plot for the inspection of the covariance structure of longitudinal data is the Draftman's display (Section 4.3). Alternatively, use of empirical semivariogram may suggest a suitable correlation structure that can be incorporated into Diggle's model (5.120) to analyze the data. To calculate the empirical semivariogram in longitudinal studies, instead of the raw data y_{ij} or the scaled data y^*_{ij}, it is usual to use residuals obtained by subtracting from the observed data y_{ij} the fitted values in a plausible model (in this approach, one computes the residuals from a regression of all y_{ij}'s on all design variables and covariates that might have predictive value). The ordinary least squares (OLS) residuals are often preferable. Figure 5.1 shows the graphical representation of the empirical semivariogram[13] based on the OLS residuals obtained from fitting an ordinary least squares model to the rat body weights data. The separate points in the plot denote the semivariogram values, and the colored line (semivariogram curve) connects average semivariogram values within each time lag.

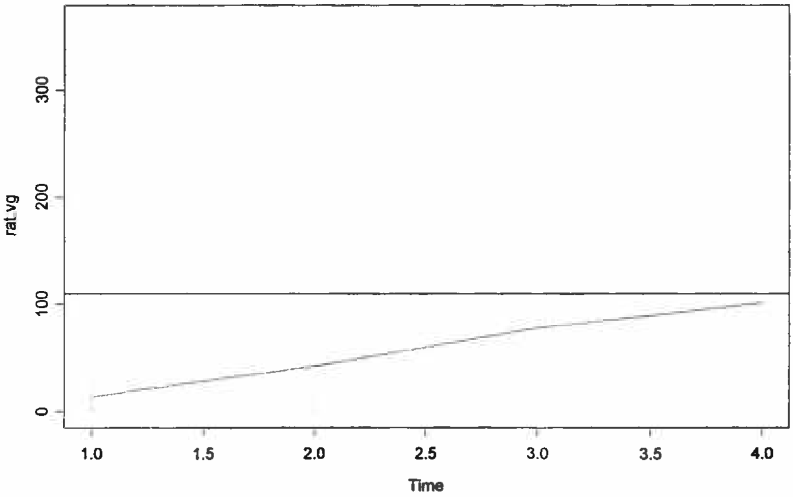

Figure 5.1 : *Empirical semivariogram of the residuals of the ordinary least squares fit to the rat body weights data.*

As one may observe, the graph's semivariogram curve clearly exhibits an increasing trend. As the lag between observations widens (from lag 1 to lag 4) the semivariogram increases, indicating a decrease in correlation between the within-subjects measurements (had this not been the case, the smoothed semivariogram curve would have resembled a horizontal line). This decrease in correlation while moving from data closer in time towards data further apart in time is indicative of a possible correlation structure. For instance, Diggle's model with a decreasing correlation function such as $\rho(u) = \exp(-\phi u)$ or $\rho(u) = \exp(-\phi u^2)$ appears to be a suitable choice for fitting the data.

[13] Calculation of the empirical semivariogram was performed by OSWALD (*Smith et al. 1994*) which is a package of functions for use with the S-Plus software, specifically designed for graphical display and analysis of longitudinal data.

5.8 Software for Linear Mixed-Model Analysis of Longitudinal Data

Although mixed models were initially developed by animal breeders to evaluate the genetic potential of specific animals, over the years application of mixed-model analysis has spread to other areas of research. An important application of the general linear mixed model (GLMM) is in the analysis of repeated measures and longitudinal data especially.

Recently, commercial computer software to analyze longitudinal data using mixed model methodology has become more widely available and more flexible. This was made possible, mainly due to the advances in computing, combined with the developments in the theory of linear mixed models. Several statistical software packages are designed to fit linear mixed-effects models with various covariance structures to longitudinal data. Commercial packages include SAS (*Littell et al., 1996*), S-PLUS (*Mathsoft Inc., 1997*), BMDP (*BMDP Statistical Software, 1990*), HLM (*Bryk et al. 1996*), STATA (*Stata Corporation, 1997*) and ML3 (*Prosser et al. 1991*).

Specifically, as concerns the SAS system, mixed-effects linear models can be implemented with either the `GLM` or the `PROC MIXED` procedures. However, the `GLM` procedure is actually a fixed effects procedure with accessory features such as the `RANDOM` statement, to make it useful for analyzing certain aspects of mixed model data. The `PROC MIXED` procedure on the other hand, was written from the start as a mixed model procedure and thus is considered more suitable for this type of longitudinal data analysis. It fits linear models for Gaussian response data and in addition to the standard Laird-Ware model it can fit the Diggle model. Generally, `PROC MIXED` allows for various parameterizations of the data's covariance structure (see Table 5.1) as well as an arbitrary number of random effects. The program provides maximum likelihood (ML) estimation for the fixed effects, ML/REML estimation for the variance components and empirical Bayes estimates for the random effects. The optimization algorithm required for the ML/REML estimation of both fixed effects and variance components is, by default, a combination of the two gradient algorithms of Newton-Raphson (N-R) and Fisher scoring (FS), which has shown to be more robust to poor starting values compared to the standard N-R algorithm.

A similar program is written in BMDP (BMDP-5V version), designed for the analysis of unbalanced repeated measures and longitudinal data. It also provides a large variety of options for the covariance structure. In contrast to the BMDP-5V, the BMDP-8V version does not share the ability to handle unbalanced longitudinal data.

The S-PLUS function `lme` fits a linear mixed-effects model (as described in *Laird* and *Ware*, *1982*), or a multilevel linear mixed-effects model (as described for example in *Longford, 1993*), using either ML or REML to estimate the variance components. The availability of choosing among various covariance patterns is currently offered by `lme`, too. To this end, different covariance structures can be used to represent the between-subjects variance-covariance matrix **D** of the Laird-Ware model, while the within-subject correlation structure can be flexibly modeled by specifying the appropriate pattern for matrix $\mathbf{R}_i$ from a large selection. Also notice that the (iterative) numerical estimation of fixed effects and variance components is carried out by means of the EM algorithm.

Similar procedures are also available in HLM, a package which is more popular though for fitting hierarchical linear models, hence is utilized mainly for educational and psychological research. Finally, ML3, which stands for Software for Three Level Analysis, was developed for applications within the fields of education and human growth. Last but not least, we must make a reference to another very frequently used alternative for mixed-model analysis of longitudinal data, the S-PLUS set of functions termed OSWALD (*Smith* and *Diggle, 1994*). OSWALD is a package of functions and data types for use with the `s` data analysis environment (e.g. S-Plus), specifically designed for the manipulation, graphical display and analysis of longitudinal data. A major motivation for utilizing OSWALD is that the latter has the ability to handle missing and incomplete longitudinal data (unbalanced designs). In fact, OSWALD has the added advantage of providing means to fit models under the assumption of informative dropout for the missing data mechanism.

Chapter 6

Nonlinear Mixed-effects Model for Longitudinal Data

6.1 Introduction

In the previous Chapter we have considered the analysis of continuous type longitudinal data where each subject's response is assumed to be linear in both the fixed effects and the random effects, by use of the General Linear Mixed Model (GLMM) for longitudinal data (*Laird* and *Ware, 1982*). In fact, the latter model has become extremely popular and the majority of work on methods for longitudinal data has been focused on data that can be modeled via the Laird-Ware model, that is data that can be modeled by an expectation function that is linear in its parameters. A large literature has grown on the particular subject and the model's use and implementation has been examined by many authors, e.g. *Laird* and *Ware (1982); Lindstrom* and *Bates (1988); Laird et al. (1987); Jennrich* and *Schluchter (1986); Lange* and *Laird (1989); Meng* and *van Dyk (1998)* and others.

However, there are often situations where longitudinal data are inherently nonlinear with respect to a given response function, say $f(\cdot)$. In particular, longitudinal data arising in the fields of population pharmacokinetics[1], biological growth and epidemiology

[1] The particular term *population pharmacokinetics* reflects the focused attention given to pharmacokinetic studies (i.e., studies carried out to characterize the kinetics of a drug) in populations different from those comprised solely of healthy volunteers (*Yuh et al., 1994*).

typically tend to be nonlinear in some unknown parameters of interest. For instance, consider the theophylline data considered by *Pinheiro* and *Bates (1995).* The specific data set is reported in Table A1 and refers to serum concentrations of the anti-asthmatic drug theophylline measured 11 times in twelve subjects (patients) over a 25-hour period.

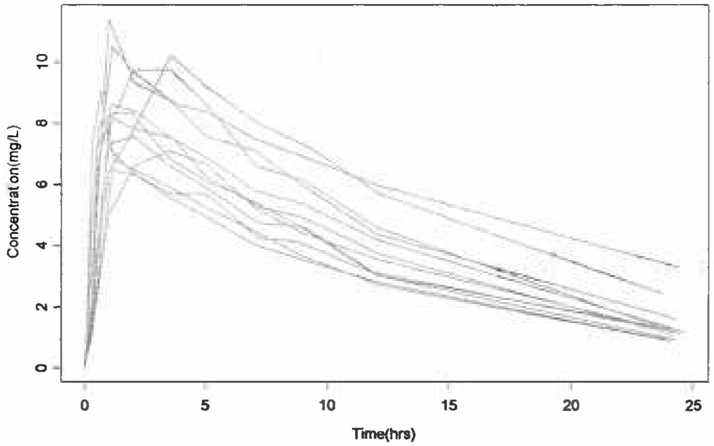

Figure 6.1 : *Theophilline concentrations* (*in* Mg/L) *of twelve patients.*

As easily seen by examining the parallel plot of the data (Figure 6.1), the general linear mixed model does not seem appropriate enough to describe the relationship between response (drug concentration) and covariate time, since the suggested from the plot relationship is clearly nonlinear. As a consequence, fitting a linear mixed-effects model to this type of data would no longer be the suitable choice and inferences drawn from such fits are usually invalid. Thus, the need to develop more general statistical models that allow for the expected responses to be nonlinear functions of the parameters has naturally arisen. The Nonlinear Mixed Effects Model (NLME in abbreviation), can be considered as a natural generalization of the general linear mixed model (GLMM) for longitudinal data introduced by *Laird* and *Ware (1982).* Many of the theoretical methods and analytical procedures used for the development and implementation of the GLMM for longitudinal data are extended in a straightforward manner to nonlinear mixed effects models. In contrast to the linear mixed-effects methodology, however, things are more complex with the nonlinear approach due to the fact that there exists a quite scattered literature on NLME models for longitudinal and repeated measures data, with most of the existing references being connected with specific inferential strategies. Several different methods for estimating the parameters in the nonlinear mixed-effects model have been proposed and in the remaining of the present chapter we are aiming at providing an adequate, unified presentation of up to date methods and issues for nonlinear longitudinal data in as much detail as possible.

The most important implication nonlinear mixed-model methodology has to confront is caused by the fact that the random effects enter the model in a nonlinear fashion. Consequently, there is no closed form expression for the marginal distribution of, say the ith subject's, response vector $\mathbf{y}_i$ $(i = 1, 2, ..., m)$. Thus, it is not as straightforward as in the linear case to write down the likelihood of data vector $\mathbf{y}_i$. In fact, the actual form of this likelihood will be quite complicated and will involve an integral with respect to the elements of random-effects vector $\mathbf{u}_i$, $(i = 1, 2, ..., m)$.

Several methods have been proposed for dealing with this challenging problem and various nonparametric, semiparametric and Bayesian methods have been developed. Their principal approach until now has been to try linearizing the model with respect to the random effects in a complete parametric context. The typical way to achieve this linearization is by approximating the marginal distribution of $\mathbf{y}_i$ $(i = 1, 2, ..., m)$. This is based on the approximation of the nonlinear function f by either linearizing the latter using the well-known Taylor series approximation or the (adaptive) Gaussian quadrature. Alternatively, the Laplacian approximation has been considered by many authors. Specifically, linearization of the f function using a Taylor expansion has been considered and examined in an extensive manner. In particular, *Sheiner* and *Beal (1980, 1985); Beal* and *Sheiner (1982, 1988, 1992); Beal (1984); Vonesh* and *Carter (1992)* and *Hirst et al. (1991)* employ a first-order Taylor series expansion to approximate f about the average random effect $E(\mathbf{u}_i) = \mathbf{0}$. *Solomon* and *Cox (1992)* also expand f about the mean of $\mathbf{u}_i$, using the four leading terms of the Taylor series expansion (fourth degree approximation) for a specific type of nonlinear model (an exponential growth model with random doubling time). An alternative approach that addresses linearization through Taylor series expansion is suggested by *Lindstrom* and *Bates (1990)*. They attempted to improve on the first-order approximation by expanding f not around $\mathbf{u}_i = \mathbf{0}$, as Sheiner and Beal did, but around estimates of the random effects $\mathbf{u}_i$. Moreover, they present a two-step iterative algorithm for maximum likelihood and restricted maximum likelihood estimation of the NLME variance components and fixed effects of the model. The closely related Laplacian approximation is discussed by *Beal* and *Sheiner (1992), Wolfinger (1993), Vonesh (1992, 1996)* and *Wolfinger* and *Lin (1997)* among others. In particular, *Wolfinger (1993)* shows that the Lindstrom & Bates algorithm for REML estimation can be derived using Laplace's approximation to the likelihood function. Also, *Pinheiro* and *Bates (1995)* and *Liu* and *Pierce (1994)* introduce the use of Gaussian quadrature rules (see, e.g. *Golub* and *Welsch, 1969*) for approximating the nonlinear function f.

A different approach, namely a nonparametric maximum likelihood procedure that makes no assumptions at all about the distributional form of the random effects, was first described by *Mallet (1986)*. In particular, the author proposes a nonparametric ML approach for estimating the distribution of the random effects of a NLME model. An alternative nonparametric approach is also presented by *Schumitzky (1991; 1993)*. In a semiparametric context, *Davidian* and *Gallant (1992; 1993)*, following ideas of *Gallant* and *Nychka (1987)*, proposed a particular inferential method referred to as the smooth nonparametric maximum likelihood (SNP) method.

There have been several articles addressing nonlinear mixed-effects models and estimation procedures for the latter in the analysis of longitudinal data from a Bayesian viewpoint. Bayesian parametric approaches involving development of a two-stage empirical Bayes estimators for the parameters of the NLME model have been considered by *Berkey (1982), Racine-Poon (1985), Berkey* and *Laird (1986), Pocock et al. (1981), Wakefield (1996), Wakefield* and *Racine-Poon (1985)* and *Wakefield et al. (1994)* among others. Also, in *Gelfand et al. (1990)* we find an iterative algorithm based on the Gibbs sampler (*Geman* and *Geman, 1984*) that generates samples of the random parameters based on a full hierarchical Bayesian specification.

Hence, as one can easily deduce, there has been a tremendous interest on NLME models for longitudinal and repeated measures data in the recent years and aim of the current Chapter is to present an analytical and critical overview of the preceding literature by covering the most important approaches proposed for modeling nonlinear, continuous response, longitudinal data via mixed-model methodology. We will focus on model formulation and inference. The rest of the Chapter is organized as follows; in the next section we briefly introduce the general form of the NLME model and some of its most important variations appeared in the literature. The following sections are devoted to the parameter estimation problem in the NLME model. The main focus is on the most salient and most widely applied of those methods. We discuss each of these methods in turn. In particular, section 6.3 presents the widely-used approximation method of Sheiner and Beal, based on a first-order Taylor series expansion. Ideas similar to the Beal and Sheiner approach, proposed by *Lindstrom* and *Bates (1990)* are discussed in section 6.4. Section 6.5 deals with a large-sample approximation, known as the Laplace approximation, which has proven to be close-connected to the Lindstrom and Bates approximate estimation method. Approximations to the log-likelihood function based on Gaussian quadrature rules are discussed in section 6.6. We continue our presentation on NLME modeling techniques for continuous longitudinal data with a brief overview on two alternative approaches found in the literature (that do not base inference on certain parametric specifications as is the case with the preceding methods), namely non- and semiparametric approaches (section 6.7). The considerable literature on Bayesian approaches is the topic of section 6.8. Finally, some remarks on the principal computer software programs and program procedures developed for the implementation of NLME model analysis in the context of longitudinal and repeated measures data are offered in section 6.9.

6.2 The Model and Some Notation

The standard model in situations where the repeated measurements on the given individuals are nonlinear in their parameters is the general nonlinear mixed-effects model. To establish notation for the NLME model, let us suppose once more a typical longitudinal (either balanced or unbalanced) design, where n_i repeated observations are taken on subject i, for $i = 1, 2, ..., m$ subjects in total. The total number of observations for all m subjects is denoted by $N = \sum_{i=1}^{m} n_i$. In addition, let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t$ be the response vector that comprises the multiple observations for the ith subject $(i = 1, 2, ..., m)$. The general nonlinear mixed-effects model may be written in the following form:

$$\mathbf{y}_i = f(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i) + \varepsilon_i \qquad (i = 1, 2, ..., m), \tag{6.1}$$

where:

- $\mathbf{b}$ is a $(p \times 1)$ fixed-effects parameter vector
- $\mathbf{u}_i$ is a $(q \times 1)$ vector of (between-subjects) random effects
- $\mathbf{T}_i$ is a $(n_i \times r)$ matrix of covariates and
- $\boldsymbol{\varepsilon}_i$ is the $(n_i \times 1)$ vector of (within-subjects) random errors

The function $f(\cdot)$ is assumed to be nonlinear and is used to model the relationship between the individual's responses and the covariates. Also, it is assumed that f is common for all individuals i $(i = 1, 2, ..., m)$.

The nonlinear model (6.1) can be considered as a straightforward generalization of the general linear mixed model (5.1). Indeed, in the case where function $f(\cdot)$ is linear in the parameters $\mathbf{b}$ and $\mathbf{u}_i$, that is of the form:

$$f(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i) = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i,$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are, respectively, $(n_i \times p)$ and $(n_i \times q)$ design matrices of known constants determined by $\mathbf{T}_i$, then, the nonlinear model (6.1) reduces to the already discussed GLMM for longitudinal and repeated measures data of *Laird* and *Ware (1982).* In other words, the nonlinear mixed model is similar in form to the GLMM (5.1), except that the expression $\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i$, which is linear in both the fixed effects $\mathbf{b}$ and the random effects $\mathbf{u}_i$ $(i = 1, 2, ..., m)$, is now replaced by the nonlinear expression $f\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)$.

To complete the description of the NLME model, what remains is to specify the distributional behavior of the between- and within-subjects random terms $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$, respectively. By analogy to the linear case, parametric specifications for $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$ is the most common approach for the nonlinear model. In particular, the Gaussian distribution is mostly used to characterize the distributional form of $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$. Thus, very often (see e.g., *Lindstrom* and *Bates (1990); Hirst et al. (1991); Wolfinger* and *Lin (1997)* among others), the random effects $\mathbf{u}_i$ are assumed to have a q-variate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $Var\left(\mathbf{u}_i\right) = \mathbf{D}$. The measurement errors $\boldsymbol{\varepsilon}_i$ are assumed to be distributed independently of the $\mathbf{u}_i$'s with a n_i-variate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $Var\left(\boldsymbol{\varepsilon}_i\right) = \mathbf{R}_i$, i.e.:

$$\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right) \quad and \quad \boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{R}_i\right). \tag{6.2}$$

The $(q \times q)$ matrix $\mathbf{D}$ is positive (semi)definite, and is assumed to be unknown. Also, as in the linear mixed model, most often $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$, the conditional-independence model ($\sigma^2 > 0$ and $\mathbf{I}_{n_i}$ is the identity matrix of order n_i). As an alternative to normality, a distribution such as the multivariate t (*Wakefield, 1996*) or a mixture of normal distributions (*Beal* and *Sheiner, 1992*) may be assumed to describe the shape of the population distribution of the between-subject random effects $\mathbf{u}_i$. However, in general, whenever a fully parametric distributional assumption is made for the random effects $\mathbf{u}_i$ it is almost always taken to be the normal model. The variance σ^2 and the elements of matrix $\mathbf{D}$, namely vector $\boldsymbol{\theta} = \left(\sigma^2, \mathbf{D}\right)$, are usually called the variance parameters (or the variance components). The parameters to be estimated in the NLME model defined by (6.1), (6.2) are the fixed effects vector $\mathbf{b}$, and the variance components. In most applications, however, the random effects parameters $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ are also of interest.

Alternatively, nonparametric (or semiparametric) approaches may be used, where no assumptions about the form of the population parameter distributions are made. Most of these nonparametric and semiparametric approaches are based on the methods proposed by *Mallet (1986)*. A typical nonparametric model specification allows the model's random parameters (in our case $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$) to arise from virtually any distribution, while in a semiparametric framework a flexible distributional form for the random effects $\mathbf{u}_i$ is assumed (e.g., *Davidian* and *Gallant 1992; 1993)*.

6.2.1 The NLME Model as a Two-Stage Model

Similarly to the linear mixed-effects model case, the standard (parametric) NLME model defined by (6.1), (6.2) can be undoubtedly re-expressed as a two-stage (hierarchical) model. In the first stage, we may summarize the data for the ith subject $(i = 1, 2, ..., m)$ as (*Vonesh, 1996; Davidian* and *Giltinan, 1995; Sheiner* and *Beal, 1985*):

$$\begin{gathered} \underline{stage\ 1}: \quad (describes\ the\ within - subject\ variation) \\ \mathbf{y}_i = f\left(\boldsymbol{\beta}_i\right) + \boldsymbol{\varepsilon}_i \quad (i = 1, 2, ..., m)\,, \\ \boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{R}_i\left(\boldsymbol{\beta}_i\right)\right), \end{gathered} \tag{6.3}$$

where $\mathbf{y}_i = \left(y_{i1}, y_{i2}, ..., y_{in_i}\right)^t$, $\boldsymbol{\beta}_i$ an unobservable $(t \times 1)$ vector of random parameters specific to the ith subject and associated to the covariates, and $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i}\right)^t$ the $(n_i \times 1)$ error vector. Stage one implicitly accounts for the within-individual variability (through the distributional specification of, within-subject, random errors $\boldsymbol{\varepsilon}_i$). The between-subject variation is incorporated into the model in the second stage:

$$\begin{gathered} \underline{stage\ 2}: \quad (describes\ the\ between - subject\ variation) \\ \boldsymbol{\beta}_i = \mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right), \\ \mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right), \end{gathered} \tag{6.4}$$

where $\mathbf{T}_i$ is a $(n_i \times r)$ matrix that comprises both the within- and between-subject covariates, $\mathbf{b}$ is a p-dimensional vector of fixed population parameters, $\mathbf{u}_i$ is a q-dimensional vector of between-subjects random effects and $\mathbf{d}(\cdot)$ is a t-dimensional vector of possibly nonlinear functions of $\mathbf{T}_i$, $\mathbf{b}$ and $\mathbf{u}_i$. Random effects $\mathbf{u}_i$ and common error terms $\boldsymbol{\varepsilon}_i$, as usual, are assumed independent to each other and between individuals. $\mathbf{D}$ and $\mathbf{R}_i(\boldsymbol{\beta}_i)$ are $(q \times q)$ and $(n_i \times n_i)$ variance-covariance matrices for the Gaussian distributed vectors $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$, respectively. As concerns the between-subject variance-covariance structure, the only restriction imposed on matrix $\mathbf{D}$ is that it is positive (semi)definite. The matrix $\mathbf{R}_i \equiv \mathbf{R}_i(\boldsymbol{\beta}_i) = \mathbf{R}_i\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}$ is analogous to the matrix $\mathbf{R}_i$ defined in (5.2) to represent the (within-subject) variance-covariance in the GLMM for longitudinal data. However, while in the linear case $\mathbf{R}_i$ was assumed to depend on i only through its dimension, here we choose a more flexible model, allowing $\mathbf{R}_i$ to depend on i through the subject-specific information (namely the $\mathbf{u}_i$) and mean response (namely vector $\mathbf{b}$), given by the vector of random parameters $\boldsymbol{\beta}_i$.

As one can observe, the second stage allows a nonlinear dependence of the $\boldsymbol{\beta}_i$ on the fixed and random effects and further allows the possibility that the dimensions of the random effects $\mathbf{u}_i$ and $\boldsymbol{\beta}_i$ may not coincide. A simpler variation of the two-stage NLME model (considered by many authors) hypothesizes a linear relationship between $\boldsymbol{\beta}_i$ and fixed and random effects, by specifying the second stage as:

$$\underline{stage\ 2}' : \quad (describes\ the\ between - subject\ variation)$$

$$\begin{aligned} \boldsymbol{\beta}_i &= \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i, \\ \mathbf{u}_i &\sim N_q(\mathbf{0}, \mathbf{D}), \end{aligned} \tag{6.5}$$

where $\mathbf{X}_i$, $\mathbf{Z}_i$ are $(t \times p)$ and $(t \times q)$ known design matrices of the fixed and the random parameters, respectively.

6.2.2 An Important Variation of the Standard NLME Model

While the main focus has been on the NLME model (6.1), various models considered in the literature as special cases of (6.1), have received analogous attention by many authors. For instance, *Vonesh* and *Carter (1992)*, *Gumpertz* and *Pantula (1992)* and *Hirst et al. (1991)* suggest models where $f(\cdot)$ is still nonlinear in the fixed effects $\mathbf{b}$, but linear in the random effects $\mathbf{u}_i$. The motivation behind this alternative specification is mainly simplicity, since that the distribution theory of their proposed models is significantly simpler than that of nonlinear model (6.1) because the response vectors $\mathbf{y}_i$ $(i = 1, 2, ..., m)$ are now linear in the random effects. Both sets of authors consider a model given by:

$$\mathbf{y}_i = f(\mathbf{b}, \mathbf{T}_i) + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i \qquad (i = 1, 2, ..., m), \tag{6.6}$$

where $\mathbf{y}_i$, $\mathbf{b}$, $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$, $\mathbf{T}_i$ and $\mathbf{Z}_i$ are as already given, and f a nonlinear function in the fixed-effects vector $\mathbf{b}$. Observe that now the random effects enter model (6.2) linearly (through the additive part $\mathbf{Z}_i \mathbf{u}_i$). Nevertheless, it is important to note that model (6.2) does not coincide with the original model (6.1) and, consequently, estimation and inferential methods for the latter does not remain true for the former model, and vice-versa.

6.3 The First-Order Method of Beal and Sheiner

Nonlinear mixed effects models for longitudinal data have received a great deal of attention in recent years, and many authors have proposed different estimation and inferential procedures for those models. Most of these approaches have appeared in the specific field of the population pharmacokinetics literature. Traditional approaches to fitting nonlinear mixed models for estimation of the model's parameters basically involve methods based on linearization using a Taylor series expansion. Two main linearization methods are popular, and both use Taylor series expansions in the random effects $\mathbf{u}_i$. One of them is the well-known first-order method proposed by *Beal (1984)*, *Beal* and *Sheiner*

(1982; 1988; 1992), and *Sheiner* and *Beal (1980; 1985)* in which a Taylor series is taken about $\mathbf{u}_i$ set to zero (the expected value of $\mathbf{u}_i$). The second is a first-order method, based on the method of Sheiner and Beal, proposed by *Lindstrom* and *Bates (1990).* By a first-order Taylor series expansion, a model is obtained that is linear in all random effects and approximates the nonlinear model.

Here, we consider first-order approximation of the nonlinear mixed effects model, developed by Sheiner and Beal. In doing this, let us assume having n_i repeated observations on subject i, $(i = 1, 2, ..., m)$; m is the number of subjects and $N = \sum_{i=1}^{m} n_i$ denotes the total number of observations. The data on each subject i can be summarized to the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^t$. To model longitudinal, continuous response, data of the above form, in the case where the hypothesized relationship between responses $\mathbf{y}_i$ and some specific covariates is nonlinear in unknown parameters of interest, one typically assumes a (two-stage) NLME model as given in (6.3), (6.4).

The usual objective in fitting nonlinear models such as the one described by (6.3) and (6.4), is to come up with estimates of the fixed effects parameter $\mathbf{b}$ and the variance components [namely the unique elements of matrices $\mathbf{D}$ and $\mathbf{R}_i$, usually denoted by the vector $\boldsymbol{\theta} = (\mathbf{D}, \mathbf{R}_i)^t$]. The additional difficulty in estimating the above parameters, compared to the estimation problem in the GLMM for longitudinal data, is that random effects $\mathbf{u}_i$ and random errors $\boldsymbol{\varepsilon}_i$ are no longer enter the model in an additive, linear fashion.

Similarly to the linear case (see Section 5.3), the most viable approach to inference on $\mathbf{b}$, $\mathbf{D}$ and $\mathbf{R}_i$ is probably through the use of maximum likelihood, which is based on the maximization of the likelihood function of all measurements $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)^t$. Let $L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y})$ denote this (full-data) likelihood. To obtain $L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y})$ we need the marginal likelihood of each ith subject's response vector:

$$\mathbf{y}_i = f(\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i, \tag{6.7}$$

namely $L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}_i)$, which coincides with the marginal probability density function of $\mathbf{y}_i$, given by[2]:

$$p\left(\mathbf{y}_i;\mathbf{b},\mathbf{D},\mathbf{R}_i\right)=\int p\left(\mathbf{y}_i \mid \mathbf{u}_i;\mathbf{b},\mathbf{R}_i\right)\cdot p\left(\mathbf{u}_i;\mathbf{D}\right)d\mathbf{u}_i, \tag{6.8}$$

where $p\left(\mathbf{y}_i \mid \mathbf{u}_i;\mathbf{b},\mathbf{R}_i\right)$ and $p\left(\mathbf{u}_i;\mathbf{D}\right)$ are the conditional density of response vector $\mathbf{y}_i$ given $\mathbf{u}_i$ and the probability density function of $\mathbf{u}_i$, respectively. (Notice that under the two-stage nonlinear model described by (6.3), (6.4), both density functions inside the integral are normal).

Then, using the independence assumption for the i subjects ($i=1,2,...,m$), it is:

$$\begin{aligned} L\left(\mathrm{b},\mathrm{D},\mathrm{R}_i;\mathrm{y}\right) &= \prod_{i=1}^{m} L\left(\mathrm{b},\mathrm{D},\mathrm{R}_i;\mathrm{y}_i\right)=\prod_{i=1}^{m} p\left(\mathrm{y}_i;\mathrm{b},\mathrm{D},\mathrm{R}_i\right) \\ &= \prod_{i=1}^{m}\int p\left(\mathbf{y}_i \mid \mathbf{u}_i;\mathbf{b},\mathbf{R}_i\right)\cdot p\left(\mathbf{u}_i;\mathbf{D}\right)d\mathbf{u}_i. \end{aligned} \tag{6.9}$$

For maximum likelihood (or restricted maximum likelihood) estimation of the parameters, the immediate objective is to maximize the above likelihood function for all data alone, over the parameters. [Equivalently, one may maximize the log-likelihood function $\lambda=\ln L\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}\right)$, or minimize $\ell=-2\ln L\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}\right)$ (known and as the objective function), to obtain ML or REML estimates of $\mathbf{b}$ and the variance components].

There is an obvious practical problem in treating (6.9) as a likelihood function for the parameters; evaluating (6.9) involves m integrations, and due to the fact that function f is nonlinear in the random effects $\mathbf{u}_i$, the required integrals are almost always intractable even when the random effects distribution is the Gaussian distribution. To circumvent this integration problem, Beal and Sheiner approximate the nonlinear model of (6.3), (6.4) by one that is linear in the random effects $\mathbf{u}_i$. They achieve this by expanding $\mathbf{y}_i =$

[2] For $\mathbf{X}$, $\mathbf{Y}$ continuous random variables with joint probability density function $f_{\mathbf{X},\mathbf{Y}}(x,y)$, the marginal probability density function of r.v. $\mathbf{X}$ is given by: $f_{\mathbf{X}}(x)=\int_{-\infty}^{+\infty} f_{\mathbf{X},\mathbf{Y}}(x,y)\,dy = \int_{-\infty}^{+\infty} f_{\mathbf{Y}}(y)\,f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}\mid\mathbf{Y})\,dy$, where $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}\mid\mathbf{Y})$ denotes the conditional density function of $\mathbf{X}$ given $\mathbf{Y}=y$.

$f(\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i \equiv f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} + \boldsymbol{\varepsilon}_i$ using a Taylor series expansion about the expectation vector of $\mathbf{u}_i$, $E(\mathbf{u}_i) = \mathbf{0}$. Their approach was initially advocated in the pharmacokinetics literature in the early 1980's, and since then is known as the 'first-order method'.

To outline the first-order approximation method, let us define (see, e.g. *Davidian, 2000*):

$$\mathbf{e}_i = \mathbf{R}_i^{-1/2}\boldsymbol{\varepsilon}_i \underset{from\ (6.7)}{=} \mathbf{R}_i^{-1/2}\left[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}\right], \tag{6.10}$$

where $\mathbf{R}_i^{1/2}$ is the square root matrix of variance-covariance matrix $\mathbf{R}_i$ [e.g. the Cholesky decomposition (see, e.g. *Gentle, 1998*) of matrix $\mathbf{R}_i$]. By performing straightforward calculations it may be shown that the previously defined vector $\mathbf{e}_i$ has zero mean vector and identity variance-covariance matrix, i.e. $E(\mathbf{e}_i) = \mathbf{0}$ and $Var(\mathbf{e}_i) = \mathbf{I}_{n_i}$.

Then, using (6.10), equation $\mathbf{y}_i = f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} + \boldsymbol{\varepsilon}_i$ can be rewritten as:

$$\mathbf{y}_i = f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} + \mathbf{R}_i^{1/2}\mathbf{e}_i. \tag{6.11}$$

The pioneering work by Sheiner and Beal was based on Taylor series expanding the above re-expressed first stage (6.11) of the NLME model, in $\mathbf{u}_i$ about the mean value $E(\mathbf{u}_i) = \mathbf{0}$. As is well-known, the general form of the Taylor series expansion of a function $f(x)$ about a point x_0 (that belongs to the domain of f), for approximating that function, is simply:

$$f(x) \simeq \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)(x - x_0)^n}{n!}, \tag{6.12}$$

where $f^{(n)}(x_0)$ represents the nth derivative of the function $f(\cdot)$ evaluated at the point x_0. (Naturally, in the case where f is a function of more than one variables, derivatives appearing in 6.12 are replaced by partial derivatives of f). The first-order approximation of $f(x)$ about x_0 essentially consists of using only the two leading terms of the Taylor series expansion (6.12), that is:

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) \tag{6.13}$$

This is generally known as the *linearization* of $f(x)$ around x_0. Thus, analogously to the above, in order to linearize (6.11), we consider the first-order approximation of $\mathbf{y}_i = f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} + \mathbf{R}_i^{1/2}\mathbf{e}_i$ with respect to $\mathbf{u}_i$, about the mean of the $\mathbf{u}_i$, $\mathbf{0}$. That is, we expand (by using only the first two terms of a Taylor series) with respect to $\mathbf{u}_i$, about $\mathbf{u}_i = \mathbf{0}$ the terms $f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}$ and $\mathbf{R}_i^{1/2}\mathbf{e}_i$, respectively. This yields:

$$\begin{aligned}
\mathbf{y}_i &\simeq f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\} + \frac{\partial}{\partial \mathbf{u}_i} f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}(\mathbf{u}_i - \mathbf{0}) + \mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{e}_i \\
&\quad + \frac{\partial}{\partial \mathbf{u}_i}\mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}(\mathbf{u}_i - \mathbf{0})\mathbf{e}_i \\
&= f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\} + \frac{\partial}{\partial \mathbf{u}_i} f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{u}_i + \mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{e}_i \\
&\quad + \frac{\partial}{\partial \mathbf{u}_i}\mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{u}_i\mathbf{e}_i. \qquad (6.14)
\end{aligned}$$

Now, if we omit from (6.14) the cross product term involving $\mathbf{u}_i\mathbf{e}_i$:

$$\frac{\partial}{\partial \mathbf{u}_i}\mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{u}_i\mathbf{e}_i,$$

as relatively 'small' compared to the leading three terms and, moreover, if we set:

$$\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\} \equiv \frac{\partial}{\partial \mathbf{u}_i} f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\},$$

and

$$\mathbf{e}_i^* \equiv \mathbf{R}_i^{1/2}\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{e}_i,$$

then the first-order approximation for $\mathbf{y}_i$ can be re-expressed as:

$$\mathbf{y}_i \simeq f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\} + \mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{0})\}\mathbf{u}_i + \mathbf{e}_i^*. \qquad (6.15)$$

The usefulness of the first-order linearization of response vector $\mathbf{y}_i = f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} + \mathbf{R}_i^{1/2}\mathbf{e}_i$, through (6.15) is obvious; comparing the approximate model (6.15) with the linear mixed effects model (5.1) of *Laird* and *Ware (1982),* shows that the two models have a similar form. The random effects $\mathbf{u}_i$ and the within-subject error term $\mathbf{e}_i^*$ enter model (6.15) in the same linear, additive fashion as is the case with the random terms of the Laird-Ware model.

From linear approximation (6.15) it evidently follows that the 'approximate' mean vector and variance-covariance matrix of $\mathbf{y}_i$ is given by:

$$E\left(\mathrm{y}_i\right) \simeq f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, 0\right)\right\}, \tag{6.16}$$

and

$$Var\left(\mathrm{y}_i\right) \simeq \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\} \mathbf{D}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\} + \mathbf{R}_i\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, 0\right)\right\}, \tag{6.17}$$

respectively. Essentially, the basis for the proposed methodology of Sheiner and Beal depends on assuming that approximation (6.15) is exact and consequently (6.16) and (6.17) are the actual mean and the actual variance-covariance matrix of $\mathbf{y}_i$ (and not some approximations). So, under the assumptions of normality of the random effects vector $\mathbf{u}_i$ and the error term $\mathbf{e}_i^*$ it follows that the (marginal) distribution of $\mathbf{y}_i$ is also the normal distribution, with parameters given by equations (6.16) and (6.17). Thus, we may write:

$$\mathbf{y}_i \sim N_{n_i}\left[f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\}, \mathbf{V}_i\left(\mathbf{b}, \mathbf{0}, \boldsymbol{\theta}\right)\right], \tag{6.18}$$

whereby $\mathbf{V}_i\left(\mathbf{b}, \mathbf{0}, \boldsymbol{\theta}\right) \equiv Var\left(\mathrm{y}_i\right) = \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\} \mathbf{D}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\} + \mathbf{R}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{0}\right)\right\}$, and $\boldsymbol{\theta} = \left(\mathbf{D}, \mathbf{R}_i\right)^t$ denotes the variance components. Accordingly, the probability density function of the (normally distributed) response vector $\mathbf{y}_i$ is readily specified as:

$$\begin{aligned} p\left(\mathrm{y}_i; \mathrm{b}, \mathrm{D}, \mathrm{R}_i\right) &= \left(2\pi\right)^{-\frac{n_i}{2}} \mid \mathbf{V}_i\left(\mathbf{b}, \mathbf{0}, \boldsymbol{\theta}\right) \mid^{-\frac{1}{2}} \times \\ &\times \exp\left\{-\frac{1}{2}\left[\mathrm{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, 0\right)\right\}\right]^t \mathbf{V}_i^{-1}\left(\mathbf{b}, \mathbf{0}, \boldsymbol{\theta}\right)\left[\mathbf{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, 0\right)\right\}\right]\right\}. \end{aligned} \tag{6.19}$$

The likelihood function of all (m in total) individuals is then calculated as (see 6.9):

$$
\begin{aligned}
L\left(\mathrm{b},\mathrm{D},\mathrm{R}_i;\mathrm{y}\right) &= \prod_{i=1}^{m} L\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}_i\right) = \prod_{i=1}^{m} p\left(\mathbf{y}_i;\mathbf{b},\mathbf{D},\mathbf{R}_i\right) \\
&= \prod_{i=1}^{m} (2\pi)^{-\frac{n_i}{2}} \mid \mathrm{V}_i \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\mathrm{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right]^t \mathbf{V}_i^{-1}\left[\mathrm{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i,\mathrm{b},0\right)\right\}\right]\right\} \\
&= (2\pi)^{-\sum_{i=1}^{m}\frac{n_i}{2}} \left(\prod_{i=1}^{m} \mid \mathrm{V}_i \mid\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right]^t \mathbf{V}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right]
\end{aligned}
$$

Calculation of the corresponding log-likelihood function, denoted by $\lambda\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}\right) = \ln L\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}\right)$ proceeds as follows:

$$
\begin{aligned}
\lambda &= \ln L = \\
&= -\sum_{i=1}^{m}\frac{n_i}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\prod_{i=1}^{m} \mid \mathrm{V}_i \mid\right) - \\
&\quad -\frac{1}{2}\sum_{i=1}^{m}\left[\mathrm{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i,\mathrm{b},0\right)\right\}\right]^t \mathrm{V}_i^{-1}\left[\mathrm{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right] \\
&= \text{const.} - \frac{1}{2}\sum_{i=1}^{m}\ln \mid \mathbf{V}_i \mid - \frac{1}{2}\sum_{i=1}^{m}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right]^t \mathbf{V}_i^{-1}\left[\mathrm{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i,\mathrm{b},0\right)\right\}\right].
\end{aligned}
$$

As emphasized earlier, the ML/REML zero-expansion estimation of Sheiner and Beal proceeds by (numerically) maximizing $\lambda\left(\mathbf{b},\mathbf{D},\mathbf{R}_\mathbf{i};\mathbf{y}\right)$ (instead of the intractable 6.9) over $\mathbf{b}$ and $\boldsymbol{\theta} = \left(\mathbf{D},\mathbf{R}_i\right)^t$ to derive ML/REML estimators of the above parameters. Equivalently, ML (or REML) estimates $\hat{\mathbf{b}}$, $\hat{\boldsymbol{\theta}}$ can be obtained by minimizing the objective function:

$$
\begin{aligned}
\ell &= -2\ln L\left(\mathbf{b},\mathbf{D},\mathbf{R}_i;\mathbf{y}\right) \\
&= \mathit{const.} + \sum_{i=1}^{m}\ln \mid \mathbf{V}_i \mid + \sum_{i=1}^{m}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right]^t \mathbf{V}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i,\mathbf{b},\mathbf{0}\right)\right\}\right],
\end{aligned}
$$

which is twice the negative marginal log-likelihood of $\mathbf{y}$.

Due to the non-closed nature of the above optimization problem, numerical iterative techniques (i.e. Newton-Raphson algorithm and its variants or quasi-Newton[3] algorithm) are inevitably implemented for achieving the required maximizations. For details on these computational techniques the interested reader is referred to *Beal* and *Sheiner (1992)* and *Boeckmann et al. (1992)*.

Along the same lines, *Hirst et al. (1991)* advocated the use of Taylor series expansion about the expectation vector of the $\mathbf{u}_i$, $E(\mathbf{u}_i) = \mathbf{0}$, for the approximation of the intermediate (compared to the nonlinear model 6.7) model $\mathbf{y}_i = f(\mathbf{b}, \mathbf{T}_i) + \mathrm{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$ where the response vectors $\mathbf{y}_i$ $(i = 1, 2, ..., m)$ are for simplicity taken to be linear in the random effects $\mathbf{u}_i$ (cf. subsection 6.2.2). Subsequently, the authors use the EM algorithm discussed by *Laird* and *Ware (1982)* to obtain ML/REML estimates of the fixed effects $\mathbf{b}$ and the variance parameters.

6.4 The 'Lindstrom-Bates' Linearization Method

As already demonstrated in the previous section, the basic idea of the Beal and Sheiner estimation method consists in approximating nonlinear model:

$$\mathbf{y}_i = f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, \mathrm{u}_i\right)\right\} + \mathbf{R}_i^{1/2}\mathbf{e}_i, \tag{6.20}$$

where $\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{D})$ and $\mathbf{e}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{I}_{n_i}\right)$, by taking a first-order Taylor series expansion (i.e. linearization) that expands (6.20) with respect to random effects $\mathbf{u}_i$, around its mean vector $E(\mathbf{u}_i) = \mathbf{0}$. However, under this setting, one may argue on the degree of accuracy of such approximation, mainly due to that individual aspect of the actual (nonlinear) model (6.20) is removed by replacing $\mathbf{u}_i$ with zero. In an attempt to improve on the 'first-order' method of Beal and Sheiner, *Lindstrom* and *Bates (1990)* suggested a first-order approximation of the nonlinear function $f(\cdot)$ of (6.20) not around $\mathbf{0}$ as Beal and

[3] Quasi-Newton procedures are methods that essentially use the same strategies with the N-R method, only that replace the Hessian matrix, namely $\mathbf{H}$, of N-R method by an approximation of $\mathbf{H}$. For more details on quasi-Newton algorithms see e.g. *Dennis & Morė (1977)*.

Sheiner did, but around $\hat{\mathbf{u}}_i$, where $\hat{\mathbf{u}}_i$ is an estimate of $\mathbf{u}_i$ [i.e., $\hat{\mathbf{u}}_i$ could be either the best linear unbiased predictor (BLUP) of $\mathbf{u}_i$, or the ML estimator of $\mathbf{u}_i$]. Moreover, they define estimators of the parameters of their proposed (approximated) model, by combining the least squares estimators for nonlinear fixed effects models and estimators for linear mixed effects models.

In the sequel, we illustrate the methodology of the approach described by Lindstrom and Bates, (which has come to be known as the 'Lindstrom-Bates' method in the literature), as well as the two-step iterative scheme developed by the aforementioned authors for the implementation of their estimation method in practice. As before, let us consider the modified model (6.20). We proceed in a similar fashion to the method of Beal and Sheiner, in the present case though we expand $\mathbf{y}_i = f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} + \mathbf{R}_i^{1/2}\mathbf{e}_i$ (by the use of a first-order Taylor series expansion) with respect to random effects $\mathbf{u}_i$ near the estimate $\hat{\mathbf{u}}_i$ of $\mathbf{u}_i$, rather than near $\mathbf{0}$. This can be done in the following way:

$$\begin{aligned}\mathbf{y}_i \;\simeq\;& f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} + \frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\left(\mathbf{u}_i - \hat{\mathbf{u}}_i\right) + \mathbf{R}_i^{1/2}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{e}_i \\ &+ \frac{\partial}{\partial \mathbf{u}_i}\mathbf{R}_i^{1/2}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\left(\mathbf{u}_i - \hat{\mathbf{u}}_i\right)\mathbf{e}_i. \qquad (6.21)\end{aligned}$$

As with the approximation methodology followed by Beal and Sheiner, we may omit the term involving cross-product $\left(\mathbf{u}_i - \hat{\mathbf{u}}_i\right)\mathbf{e}_i$ as relatively small compared to $\mathbf{u}_i$, $\mathbf{e}_i$. Moreover, if we define:

$$\mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} \equiv \frac{\partial}{\partial \mathbf{u}_i}\left[f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\right],$$

and treating $\hat{\mathbf{u}}_i$ as a fixed constant, it is straightforward to verify that first-order approximation (6.21) yields:

$$\begin{aligned}\mathbf{y}_i \;\simeq\;& f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} + \frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{u}_i - \frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i \\ &+ \mathbf{R}_i^{1/2}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{e}_i\end{aligned}$$

$$= \left[f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} - \mathbf{Z}_i\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i\right] + \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{u}_i$$
$$+\mathrm{R}_i^{1/2}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{e}_i. \qquad (6.22)$$

Finally, setting $\mathbf{e}_i^* \equiv \mathbf{R}_i^{1/2}\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, \hat{\mathrm{u}}_i\right)\right\}\mathrm{e}_i$ in the above yields the following approximate (linearized) model:

$$\mathbf{y}_i \simeq \left[f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} - \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i\right] + \mathbf{Z}_i\left\{\mathbf{d}\left(\mathrm{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{u}_i + \mathbf{e}_i^*. \qquad (6.23)$$

Hence, if the assumed distribution for $\mathbf{u}_i$ and $\mathbf{e}_i^*$ is normal, then from the above approximate model we easily deduce (since $\hat{\mathbf{u}}_i$ is considered to be a constant):

$$E\left(\mathrm{y}_i\right) \simeq f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, \hat{\mathrm{u}}_i\right)\right\} - \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i, \qquad (6.24)$$

and

$$Var\left(\mathbf{y}_i\right) \simeq \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathrm{D}\mathrm{Z}_i^t\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} + \mathrm{R}_i\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, \hat{\mathrm{u}}_i\right)\right\}. \qquad (6.25)$$

As *Lindstrom* and *Bates (1990)* point out, the specific method of approximation has been previously used in a similar setting by *Stiratelli et al. (1984).* Approximations (6.24), (6.25) for the mean vector $E\left(\mathbf{y}_i\right)$ and the variance-covariance matrix $Var\left(\mathbf{y}_i\right)$ of response vector $\mathbf{y}_i$ allow for estimation of the model's parameters via a two-step alternating algorithm.

Before proceeding with the description of the iterative algorithm considered by *Lindstrom* and *Bates (1990),* note that both approximations of the moments given by (6.24) and (6.25) require a suitable estimate $\hat{\mathbf{u}}_i$ for $\mathbf{u}_i$, in order to lead to estimates of $\mathbf{b}$ and $\boldsymbol{\theta} = \left(\mathbf{D}, \mathbf{R}_i\right)^t$. The authors suggest the following strategy for obtaining an estimate $\hat{\mathbf{u}}_i$ of $\mathbf{u}_i$: Let $p\left(\mathbf{u}_i \mid \mathbf{y}_i; \mathbf{D}\right)$ denote the posterior density of $\mathbf{u}_i$ given $\mathbf{y}_i$, $p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right)$ denote the conditional density of $\mathbf{y}_i$ given the random effects, and $p\left(\mathbf{u}_i; \mathbf{D}\right)$, $p\left(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i\right)$ be the (marginal) densities of $\mathbf{u}_i$ and $\mathbf{y}_i$, respectively. Typically, the conditional density

$p(\mathbf{u}_i \mid \mathbf{y}_i; \mathbf{D})$ is proportional to $p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i) \times p(\mathbf{u}_i; \mathbf{D})$. This may be easily seen, since the joint probability density function of $\mathbf{u}_i$ and $\mathbf{y}_i$ can be written as:

$$p(\mathbf{y}_i, \mathbf{u}_i) \equiv p(\mathbf{u}_i \mid \mathbf{y}_i; \mathbf{D}) \times p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i) = p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i) \times p(\mathbf{u}_i; \mathbf{D}),$$

and consequently it is:

$$p(\mathbf{u}_i \mid \mathbf{y}_i; \mathbf{D}) \propto p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i) \times p(\mathbf{u}_i; \mathbf{D}). \tag{6.26}$$

Now, as a result of the normality assumptions for both $\mathbf{u}_i$ and $\mathbf{y}_i \mid \mathbf{u}_i$, we obtain:

$$p(\mathbf{u}_i) = (2\pi)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i\right\}, \tag{6.27}$$

and

$$p(\mathbf{y}_i \mid \mathbf{u}_i) = (2\pi)^{-\frac{n_i}{2}} \mid \mathbf{R}_i \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right]\right\}. \tag{6.28}$$

Correspondingly, by combining (6.26), (6.27) and (6.28) and performing straightforward calculations yields:

$$\begin{aligned} & p(\mathbf{u}_i \mid \mathbf{y}_i) \propto p(\mathbf{y}_i \mid \mathbf{u}_i) \times p(\mathbf{u}_i) \\ \propto \; & (2\pi)^{-\frac{n_i+q}{2}} \mid \mathbf{R}_i \mid^{-\frac{1}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \\ & \times \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right] - \frac{1}{2}\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i\right\}, \end{aligned}$$

and taking logarithms in the above, ignoring the terms that are constants with respect to $\mathbf{u}_i$ (i.e. do not involve $\mathbf{u}_i$), we get:

$$\begin{aligned} \ln p(\mathbf{u}_i \mid \mathbf{y}_i) \propto & -\frac{1}{2}\ln \mid \mathbf{R}_i \mid - \frac{1}{2}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\right\}\right] \\ & -\frac{1}{2}\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i. \end{aligned} \tag{6.29}$$

Moreover, the term $-1/2 \ln \mid \mathbf{R}_i|$ involving the within-subject variance-covariance matrix $\mathbf{R}_i$ may be also omitted from the above (as a constant with respect to $\mathbf{u}_i$), due to the fact that in the specific setting assumed by Lindstrom and Bates, matrix $\mathbf{R}_i$ does not depend on $\boldsymbol{\beta}_i$ [i.e. $\mathbf{R}_i \neq \mathbf{R}_i(\boldsymbol{\beta}_i) = \mathbf{R}_i\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}$], hence $\mathbf{R}_i$ does not depend on $\mathbf{u}_i$. As a result, analogy (6.29) can be re-expressed (ignoring all constants) as:

$$\ln p(\mathbf{u}_i \mid \mathbf{y}_i) \propto -\frac{1}{2}\left[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}\right] - \frac{1}{2}\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i. \tag{6.30}$$

Lindstrom and *Bates (1990)* utilize $\ln p(\mathbf{u}_i \mid \mathbf{y}_i)$ and minimize it with respect to $\mathbf{u}_i$ for each $i = 1, 2, ..., m$ in order to deal with the problem of estimating random effects $\mathbf{u}_i$ $(i = 1, 2, ..., m)$.

Now, having cleared out the issue on how to estimate $\mathbf{u}_i$, we are in a position to illustrate the iterative two-step estimation scheme proposed by Lindstrom and Bates. Obviously, as with all numerical optimization algorithms, initial values (estimators) of the parameters to be estimated (in the current situation $\mathbf{b}$ and $\boldsymbol{\theta}$) are needed. Let us denote these initial values with $\mathbf{b}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$, respectively. (As usual, $\mathbf{b}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$ will denote the value of the parameters obtained at the kth iteration of the algorithm). For instance, $\mathbf{b}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ could be obtained from fitting the approximate model of Sheiner and Beal, i.e. the model derived by application of the standard first-order Taylor series expansion with respect to $\mathbf{u}_i$, about $E(\mathbf{u}_i) = \mathbf{0}$. Having found initial values $\mathbf{b}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ by the above procedure, as a next step, we substitute these in (6.30), and holding them fixed (treat them as constants) we maximize $\ln p(\mathbf{u}_i \mid \mathbf{y}_i)$ with respect to each $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ to acquire initial starting points for the $\mathbf{u}_i$'s, say $\mathbf{u}_i^{(0)}$ $(i = 1, 2, ..., m)$. Hence, with starting points $\mathbf{b}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and $\mathbf{u}_i^{(0)}$, $(i = 1, 2, ..., m)$ at hands, the 'Lindstrom-Bates' two-step iterative scheme, say at the $(k+1)$st iteration, may be formally expressed as follows:

1st step : *Substitute* $\mathbf{u}_i^{(k)}$ *(obtained at the kth iteration) for* $\hat{\mathbf{u}}_i$ *in the approximate* *linear expressions* (6.24), (6.25) *. Now, treating* $\mathbf{u}_i^{(k)}$ *as fixed, in a similar manner to the GLMM, update ML estimations of* $\mathbf{b}$ *and* $\boldsymbol{\theta}$ *using approximate linear model* (6.23) *by means of numerical iterative optimization algorithms, such as* $N-R$ *algorithm and EM algorithm. Call the updated estimators* $\mathbf{b}^{(k+1)}$ *and* $\boldsymbol{\theta}^{(k+1)}$.

<u>*2nd step*</u> : *Substituting* $\mathbf{b}^{(k+1)}$, $\boldsymbol{\theta}^{(k+1)}$ *obtained from the previous step, in equation* (6.30) *and holding these fixed, maximize* $\ln p(\mathbf{u}_i \mid \mathbf{y}_i)$ *with respect to* $\mathbf{u}_i$ *for each* i *in* m *separate maximizations to obtain "new" estimates* $\mathbf{u}_i^{(k+1)}$, $(i = 1, 2, ..., m)$ *. Set* $k = k + 1$ *and move back to the first step.*

The preceding algorithm alternates between these two steps until convergence is reached. The achieved (stabilized) values at convergence would evidently be the desired maximum likelihood (ML) estimates $\hat{\mathbf{b}}_{ML}$, $\hat{\boldsymbol{\theta}}_{ML}$ for fixed-effects vector $\mathbf{b}$ and variance components $\boldsymbol{\theta}$, respectively. *Wolfinger (1993)* points out that the two steps agree at convergence. He also shows that the above numerical procedure can be derived by the use of the Laplace approximation[4]. Though different from a theoretical point of view, the Laplacian approach has proven to be equivalent to the Lindstrom-Bates method, leading to the same approximate moments given in (6.24), (6.25), leading thus to the same approximate linear model (6.23). The specific results have been also confirmed in *Vonesh (1996),* and *Wolfinger* and *Lin (1997).* (The close-related to the Lindstrom-Bates approximate estimation method, Laplacian approximation of Wolfinger is the subject of the ongoing section 6.5).

Finally, it is interesting to note that *Lindstrom* and *Bates (1990)* refer to step 1 as the 'pseudo-data' step, because joint estimation of $\mathbf{b}$ and $\mathbf{u}_i$ by maximizing $\ln p(\mathbf{u}_i \mid \mathbf{y}_i)$ may be accomplished simultaneously by specifying an augmented-data nonlinear least squares problem (see *Lindstrom* and *Bates (1990)*; *Davidian* and *Giltinan (1995)* for

[4] The use of this type of approximation originates with Laplace, thus these approximations are usually called *Laplace* approximations. Laplace approximation has been alternatively called by Physicists the *saddle-point* approximation.

more details on the topic).

6.5 The Laplacian Method and Its Relation to the Lindstrom-Bates Method

The Laplace method for integrals (cf. *De Bruijn, 1961*) is a standard, widely applied, large-sample procedure for approximating integrals. Its popularity stems from the fact that the method can perform remarkably well in practice, even for modest amounts of data, despite that is based on large data limits, thus one should expect to perform very poorly for small data sets. A discomforting feature of the method, however, is its computational complexity [of $O\left(d^2 N\right)$, or greater, where d is the dimension of the model, and N the sample of the data]. As a result, the Laplace approximation can be a computational burden for large models.

In particular, Laplace's method approximates the integral of a function $f\left(\cdot\right)$, namely:

$$\int f\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta},$$

by fitting a Gaussian at the value $\hat{\boldsymbol{\theta}}$ that maximizes $f\left(\boldsymbol{\theta}\right)$, and computing the volume under that Gaussian (*Mackay, 1996*). This results in the following approximation:

$$\int f\left(\theta\right) d\theta \simeq f\left(\hat{\theta}\right) \left(2\pi\right)^{\frac{q}{2}} \mid -\nabla\nabla \ln f\left(\hat{\theta}\right) \mid^{-\frac{1}{2}}, \tag{6.31}$$

where q is the dimension of (vector) parameter $\boldsymbol{\theta}$ and $\nabla\nabla \ln f\left(\hat{\boldsymbol{\theta}}\right)$ denotes the matrix of second-order partial derivatives of $\ln f\left(\boldsymbol{\theta}\right)$ (with respect to $\boldsymbol{\theta}$) evaluated at the maximum $\hat{\boldsymbol{\theta}}$, i.e. $\nabla\nabla \ln f\left(\hat{\boldsymbol{\theta}}\right) = \partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t \left[\ln f\left(\boldsymbol{\theta}\right)\right] \mid_{\boldsymbol{\theta}=\overset{\wedge}{\boldsymbol{\theta}}}$.

In a statistical context, now, Laplacian approximations as described in (6.31) have been extensively used in Bayesian inference for the estimation of marginal posterior densities and predictive distributions. A vast literature on the specific field exists, including

Tierney and *Kadane (1986); Tierney et al. (1989); Leonard et al. (1989)* and *Kass* and *Raftery (1995)* among others.

Typically, the same approximation techniques applied in the above-mentioned Bayesian discussions can be adequately used for approximating NLME model (6.7) to derive estimates of the parameters of the nonlinear model. To demonstrate this, we need to consider again the marginal probability density function of $\mathbf{y}_i$, namely:

$$p\left(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i\right) = \int p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right) \cdot p\left(\mathbf{u}_i; \mathbf{D}\right) d\mathbf{u}_i. \tag{6.32}$$

Recall that to perform ML estimation of the parameters ($\mathbf{b}$ and $\boldsymbol{\theta}$) of the nonlinear model, one needs to calculate the above integral over the random effects $\mathbf{u}_i$. Unfortunately, a closed-form expression of (6.32) is typically not available, thus the need for numerical integration becomes evident. Approximate linearization methods described in sections 6.3 and 6.4 respectively, attempt to circumvent the problem by approximating the nonlinear response vector $\mathbf{y}_i$, $(i = 1, 2, ..., m)$ by a linear one, using standard first-order Taylor series expansions.

The Laplacian approach to estimation of $\mathbf{b}$, $\boldsymbol{\theta}$, proceeds in a more straightforward manner, approximating integral (6.32) via Laplace's approximation (6.31). Before developing the latter approximation we need some preliminaries. First, notice that for $f\left(\boldsymbol{\theta}\right) = e^{n\ell(\boldsymbol{\theta})}$, asymptotic approximation (6.31) becomes:

$$\begin{aligned} \int e^{n\ell(\boldsymbol{\theta})} d\boldsymbol{\theta} &\simeq e^{n\ell\left(\hat{\boldsymbol{\theta}}\right)} \left(\frac{2\pi}{n}\right)^{\frac{q}{2}} \left| -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t} \ln e^{n\ell\left(\hat{\boldsymbol{\theta}}\right)} \right|^{-\frac{1}{2}} \\ &= e^{n\ell\left(\hat{\boldsymbol{\theta}}\right)} \left(\frac{2\pi}{n}\right)^{\frac{q}{2}} \left| -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t} n\ell\left(\hat{\boldsymbol{\theta}}\right) \right|^{-\frac{1}{2}}, \end{aligned} \tag{6.33}$$

where q the dimension of $\boldsymbol{\theta}$, and $\ell\left(\boldsymbol{\theta}\right)$ a real-valued function of $\boldsymbol{\theta}$. Additionally, in the above and the following, we assume that the q-dimensional random effects vector $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ as well as the n_i-dimensional conditional response vector $\mathbf{y}_i$ given $\mathbf{u}_i$ are normally distributed, with corresponding p.d.f.'s given by (6.27) and (6.28), respectively.

Thus, substituting (6.27), (6.28) in (6.32) yields [for notational convenience, we write $p(\mathbf{u}_i)$, $p(\mathbf{y}_i \mid \mathbf{u}_i)$ and $p(\mathbf{y}_i)$ instead of $p(\mathbf{u}_i; \mathbf{D})$, $p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i)$ and $p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i)$] :

$$\begin{aligned} p(\mathbf{y}_i) &= \int p(\mathbf{y}_i \mid \mathbf{u}_i) \cdot p(\mathbf{u}_i)\, d\mathbf{u}_i \\ &= (2\pi)^{-\frac{n_i}{2}} (2\pi)^{-\frac{q}{2}} \mid \mathbf{R}_i \mid^{-\frac{1}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \times \qquad (6.34) \\ &\times \int \exp\left\{ -\frac{1}{2} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}]^t \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}] - \frac{1}{2} \mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i \right\} d\mathbf{u}_i. \end{aligned}$$

The crucial idea in developing a Laplacian approximation of the NLME model, is to substitute the integral contained in the above expression for the marginal p.d.f. of $\mathbf{y}_i$, namely $p(\mathbf{y}_i)$, using the Laplace formula (6.33). To this end, let us set:

$$\ell(\mathbf{u}_i) = -\frac{1}{2n_i} \left\{ [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}]^t \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}] + \mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i \right\}. \quad (6.35)$$

As a consequence, integral included in $p(\mathbf{y}_i)$ may be re-expressed as:

$$\int \exp\{n_i \ell(\mathbf{u}_i)\}\, d\mathbf{u}_i. \quad (6.36)$$

It is easily seen that the above integral is of similar form with the integral of (6.33), and we may thus utilize the latter formula to derive Laplacian approximation of the specific integral of interest. Clearly, as implied by the right-hand side of (6.33), we have to specify the second-order partial derivative $\partial^2 / \partial \mathbf{u}_i \partial \mathbf{u}_i^t \{\ell(\mathbf{u}_i)\}$ of function $\ell(\mathbf{u}_i)$ with respect to $\mathbf{u}_i$. The evaluation of derivative $\partial^2 / \partial \mathbf{u}_i \partial \mathbf{u}_i^t \{\ell(\mathbf{u}_i)\}$ requires the calculation of first-order partial derivative $\partial / \partial \mathbf{u}_i \{\ell(\mathbf{u}_i)\}$. The latter is calculated as:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{u}_i} \ell(\mathbf{u}_i) &= -\frac{1}{2n_i} \frac{\partial}{\partial \mathbf{u}_i} \left\{ [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}]^t \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}] \right\} \\ &\quad -\frac{1}{2n_i} \frac{\partial}{\partial \mathbf{u}_i} \left\{ \mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i \right\}, \end{aligned}$$

and using once more the matrix derivation result (3.32), we get:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{u}_i}\ell(\mathbf{u}_i) &= -\frac{1}{2n_i}\left\{-2\left[\frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\right\} \\
&\quad -\frac{1}{2n_i}\left\{2\mathbf{D}^{-1}\mathbf{u}_i\right\} \\
&= n_i^{-1}\left[\frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] \\
&\quad -n_i^{-1}\mathbf{D}^{-1}\mathbf{u}_i,
\end{aligned}$$

or equivalently, due to that we have already set $\mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \equiv \partial/\partial \mathbf{u}_i\left[f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]$,

$$\frac{\partial}{\partial \mathbf{u}_i}\ell(\mathbf{u}_i) = n_i^{-1}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] - n_i^{-1}\mathbf{D}^{-1}\mathbf{u}_i. \tag{6.37}$$

It is now possible to derive an analytic expression for $\partial^2/\partial \mathbf{u}_i \partial \mathbf{u}_i^t\left\{\ell(\mathbf{u}_i)\right\}$, using standard matrix algebra, as follows:

$$\begin{aligned}
&\frac{\partial^2}{\partial \mathbf{u}_i \partial \mathbf{u}_i^t}\ell(\mathbf{u}_i) = \frac{\partial}{\partial \mathbf{u}_i}\left\{\frac{\partial}{\partial \mathbf{u}_i}\ell(\mathbf{u}_i)\right\} \underset{(6.37)}{=} \\
&= n_i^{-1}\frac{\partial}{\partial \mathbf{u}_i}\left\{\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\right\} - n_i^{-1}\frac{\partial}{\partial \mathbf{u}_i}\left(\mathbf{D}^{-1}\mathbf{u}_i\right) \\
&= n_i^{-1}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\mathbf{R}_i^{-1}\frac{\partial}{\partial \mathbf{u}_i}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] \\
&\quad +n_i^{-1}\left[\frac{\partial}{\partial \mathbf{u}_i}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] - n_i^{-1}\mathbf{D}^{-1}\frac{\partial \mathbf{u}_i}{\partial \mathbf{u}_i} \\
&= -n_i^{-1}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\mathbf{R}_i^{-1}\frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \\
&\quad +n_i^{-1}\left[\frac{\partial}{\partial \mathbf{u}_i}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] - n_i^{-1}\mathbf{D}^{-1} \\
&= -n_i^{-1}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\mathbf{R}_i^{-1}\mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \\
&\quad +n_i^{-1}\left[\frac{\partial}{\partial \mathbf{u}_i}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\mathbf{R}_i^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] - n_i^{-1}\mathbf{D}^{-1}. \qquad (6.38)
\end{aligned}$$

Now, as *Vonesh (1996)* points out, it is more appealing in the specific case to use the (common) version of Laplace's approximation, where $\partial^2/\partial \mathbf{u}_i \partial \mathbf{u}_i^t\left\{\ell(\mathbf{u}_i)\right\}$ is replaced by its expectation $E\left[\partial^2/\partial\mathbf{u}_i\partial\mathbf{u}_i^t\{\ell(\mathbf{u}_i)\}\right]$. The latter is not difficult to compute. Indeed:

$$E\left\{\frac{\partial^2}{\partial\mathbf{u}_i\partial\mathbf{u}_i^t}\ell(\mathbf{u}_i)\right\} = -n_i^{-1}\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}$$
$$+n_i^{-1}\left[\frac{\partial}{\partial\mathbf{u}_i}\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}\right]\mathbf{R}_i^{-1}E\{[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}]\} - n_i^{-1}\mathbf{D}^{-1},$$

and since $E\{[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}]\} = \mathbf{0}$, it is readily available that:

$$E\left\{\frac{\partial^2}{\partial\mathbf{u}_i\partial\mathbf{u}_i^t}\ell(\mathbf{u}_i)\right\} = -n_i^{-1}\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\} - n_i^{-1}\mathbf{D}^{-1}. \quad (6.39)$$

Thus, recalling (6.33), the Laplacian-type approximation of integral (6.36), with $E\left[\partial^2/\partial\mathbf{u}_i\partial\mathbf{u}_i^t\{\ell(\mathbf{u}_i)\}\right]$ in place of $\partial^2/\partial\mathbf{u}_i\partial\mathbf{u}_i^t\{\ell(\mathbf{u}_i)\}$ though, is computed as follows:

$$\int \exp\{n_i\ell(\mathbf{u}_i)\}\,d\mathbf{u}_i$$
$$\equiv \int \exp\left\{-\frac{1}{2}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}]^t\mathbf{R}_i^{-1}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\mathbf{u}_i)\}] - \frac{1}{2}\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i\right\}d\mathbf{u}_i$$
$$\simeq \left(\frac{2\pi}{n_i}\right)^{\frac{q}{2}}\left|-\left(-n_i^{-1}\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\} - n_i^{-1}\mathbf{D}^{-1}\right)\right|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}]^t\mathbf{R}_i^{-1}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}] - \frac{1}{2}\hat{\mathbf{u}}_i^t\mathbf{D}^{-1}\hat{\mathbf{u}}_i\right\}$$
$$= (2\pi)^{\frac{q}{2}}n_i^{-\frac{q}{2}}\left|n_i^{-1}\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\} + n_i^{-1}\mathbf{D}^{-1}\right|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}]^t\mathbf{R}_i^{-1}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}] - \frac{1}{2}\hat{\mathbf{u}}_i^t\mathbf{D}^{-1}\hat{\mathbf{u}}_i\right\}$$
$$= (2\pi)^{\frac{q}{2}}n_i^{-\frac{q}{2}}n_i^{\frac{q}{2}}\left|\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\} + \mathbf{D}^{-1}\right|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}]^t\mathbf{R}_i^{-1}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}] - \frac{1}{2}\hat{\mathbf{u}}_i^t\mathbf{D}^{-1}\hat{\mathbf{u}}_i\right\}$$
$$= (2\pi)^{\frac{q}{2}}\left|\mathbf{Z}_i^t\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}\mathbf{R}_i^{-1}\mathbf{Z}_i\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\} + \mathbf{D}^{-1}\right|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}]^t\mathbf{R}_i^{-1}[\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i,\mathbf{b},\hat{\mathbf{u}}_i)\}] - \frac{1}{2}\hat{\mathbf{u}}_i^t\mathbf{D}^{-1}\hat{\mathbf{u}}_i\right\},$$

where $\hat{\mathbf{u}}_i$, likewise to the 'Lindstrom-Bates' method, is chosen to maximize function $\ell(\mathbf{u}_i)$. Hence, correspondingly, it is easy to verify (combining 6.34 and the above Laplace's formula) that marginal density $p(\mathbf{y}_i)$ is approximated via the Laplacian method as:

$$\begin{aligned} & p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i) \simeq \\ \simeq\ & (2\pi)^{-\frac{n_i}{2}} (2\pi)^{-\frac{q}{2}} \mid \mathbf{R}_i|^{-\frac{1}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} (2\pi)^{\frac{q}{2}} \left|\mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \mathbf{R}_i^{-1} \mathbf{Z}_i \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} + \mathbf{D}^{-1}\right|^{-\frac{1}{2}} \\ & \times \exp\left\{-\frac{1}{2} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\}]^t \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\}] - \frac{1}{2} \hat{\mathbf{u}}_i^t \mathbf{D}^{-1} \hat{\mathbf{u}}_i\right\} \\ =\ & (2\pi)^{-\frac{n_i}{2}} \mid \mathbf{R}_i|^{-\frac{1}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \left|\mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \mathbf{R}_i^{-1} \mathbf{Z}_i \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} + \mathbf{D}^{-1}\right|^{-\frac{1}{2}} \\ & \times \exp\left\{-\frac{1}{2} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\}]^t \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\}] - \frac{1}{2} \hat{\mathbf{u}}_i^t \mathbf{D}^{-1} \hat{\mathbf{u}}_i\right\}. \end{aligned} \qquad (6.40)$$

Now, as already noted, observe that estimator $\hat{\mathbf{u}}_i$ is defined as the value that maximizes $\ell(\mathbf{u}_i)$. Consequently, to obtain an exact expression for $\hat{\mathbf{u}}_i$ it suffices to solve:

$$\frac{\partial}{\partial \mathbf{u}_i} \ell(\mathbf{u}_i) = \mathbf{0},$$

or equivalently, from (6.37), to solve:

$$n_i^{-1} \mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\} \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i)\}] - n_i^{-1} \mathbf{D}^{-1} \mathbf{u}_i = \mathbf{0},$$

with respect to $\mathbf{u}_i$, which yields the following expression for the $\hat{\mathbf{u}}_i$ estimator:

$$\hat{\mathbf{u}}_i = \mathbf{D} \mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \mathbf{R}_i^{-1} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\}]. \qquad (6.41)$$

From (6.40), (6.41), and by applying appropriate general matrix results it is possible to derive (see, e.g. *Davidian, 2000*) the following modified form for $p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i)$:

$$\begin{aligned} & p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i) \simeq \\ \simeq\ & (2\pi)^{-\frac{n_i}{2}} \left|\mathbf{Z}_i \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \mathbf{D} \mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} + \mathbf{R}_i\right|^{-\frac{1}{2}} \\ & \times \exp\left\{-\frac{1}{2} [\mathbf{y}_i - f\{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} + \mathbf{Z}_i \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \hat{\mathbf{u}}_i]^t \cdot\right. \\ & \cdot \left[\mathbf{Z}_i \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} \mathbf{D} \mathbf{Z}_i^t \{\mathbf{d}(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i)\} + \mathbf{R}_i\right]^{-1} \cdot \end{aligned} \qquad (6.42)$$

$$\cdot \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} + \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i\right]\right\}.$$

Since each $\mathbf{y}_i$ $(i = 1, 2, ..., m)$ has (approximately) marginal density of the above form, which obviously is the density function of a n_i-variate, normally distributed random variable, we easily deduce from (6.42) that mean and variance-covariance of $\mathbf{y}_i$ are:

$$E\left(\mathbf{y}_i\right) \simeq f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} - \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\hat{\mathbf{u}}_i, \tag{6.43}$$

and

$$Var\left(\mathbf{y}_i\right) \simeq \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\}\mathbf{D}\mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i\right)\right\} + \mathbf{R}_i, \tag{6.44}$$

respectively. Notice that the above resulting moments are identical to the mean and variance-covariance approximations *Lindstrom* and *Bates (1990)* derived by their first-order linearization method (equations 6.24, 6.25), constituting thus the Laplacian approximation method as an appealing alternative to the 'Lindstrom-Bates' method. In terms of the above approximations for $E\left(\mathbf{y}_i\right)$ and $Var\left(\mathbf{y}_i\right)$, *Wolfinger (1993)* derived the same two-stage algorithm of *Lindstrom* and *Bates (1990)*, described in the previous section.

We conclude our brief discussion on the current approach with some additional contributions and references that exploit the Laplacian approximation method for application to the NLME model. These include *Vonesh (1996), Pinheiro* and *Bates (1995), Breslow* and *Clayton (1993), Wolfinger* and *Lin (1997), Davidian* and *Giltinan (1995)* and *Vonesh* and *Chinchilli (1997).*

6.6 The Method of Gaussian Quadrature

Among the principal methods for the numerical computation of a (definite) integral of a real-valued continuous function $f(\cdot)$, defined on a compact interval $[\alpha, \beta]$ which may be infinite in either or both directions, i.e. an integral of the form:

$$\int_{\alpha}^{\beta} f(x)\, dx, \tag{6.45}$$

is the so-called method of '*quadrature rule*'. The specific numerical integration technique approximates the integral of f from the ordinates of the function at particular absissae (the x values of the function) which are weighted and summed in order to give an approximation of the integral. More formally, a quadrature rule approximates a given integral, such as (6.45), by the weighted summation:

$$\sum_{i=1}^{n} w_i f(x_i), \tag{6.46}$$

where the x_i's $(i = 1, 2, ..., n)$ denote the nodes (or abscissas) and the w_i's are the weights (or coefficients) of the quadrature rule. Also, n denotes the number of evaluation points. As one can easily deduce, the basic problem in quadrature theory is to choose nodes and weights (independent of function f) so that:

$$\int_{\alpha}^{\beta} f(x)\, dx \simeq \sum_{i=1}^{n} w_i f(x_i). \tag{6.47}$$

Various quadrature rules have been proposed in the duration of time [see, e.g. *Conte* and *deBoor (1981), Chapter 7*]. A significant place among these rules is possessed by the Gaussian quadrature rule (*Gauss, 1816*). The specific quadrature method for the numerical (approximate) evaluation of integrals utilizes orthogonal polynomials, such as the Legendre, the Laguerre, the Chebyshev and the Hermite polynomials and their roots.

In short, the general form of a Gaussian quadrature rule may be written as:

$$\int_{\alpha}^{\beta} w(x)\, f(x)\, dx \simeq \sum_{i=1}^{n} w_i f(x_i), \tag{6.48}$$

where $w(x)$ is a weighting function (conveniently chosen to ensure the convergence of the integral of $w(x)\, f(x)$), and w_i, x_i $(i = 1, 2, .., n)$ are as previously defined. The basic idea of this approximation lies in the fact that we can find appropriate weights w_i and nodes x_i such that the rule (6.48) is exact if $f(x)$ is a polynomial of order less than $2n$. The nodes $\{x_i\}_{i=1}^{n}$ of this rule are then the zeros of the (orthogonal) polynomial. For a detailed discussion of the general theory of Gaussian quadrature methods, see *Abramowitz* and *Stegun (1964)*, *Davis* and *Rabinowitz (1984)* and *Golub* and *Welsch (1969)*.

With the latter background theory in mind, we can now proceed in describing how the Gaussian quadrature methodology finds application in the current context of nonlinear modeling for longitudinal data, by the (approximate) computation of the marginal likelihood function $L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y})$ of all repeated measurements $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m)^t$, given by (see Section 6.3):

$$L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}) = \prod_{i=1}^{m} p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i) = \prod_{i=1}^{m} \int p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i) \cdot p(\mathbf{u}_i; \mathbf{D})\, d\mathbf{u}_i. \tag{6.49}$$

In the above, $p(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i)$ represents the marginal probability density function of $\mathbf{y}_i$, and $p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i)$, $p(\mathbf{u}_i; \mathbf{D})$ denote the conditional density of response vector $\mathbf{y}_i$ given $\mathbf{u}_i$ and the probability density function of $\mathbf{u}_i$, respectively. Essentially, what is required is the calculation of integral $\int p(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i)\, p(\mathbf{u}_i; \mathbf{D})\, d\mathbf{u}_i$, which unfortunately does not have a closed-form expression in most situations. Assuming once again normality of the random effects $\mathbf{u}_i$, $(i = 1, 2, ..., m)$ and the within-subject errors $\boldsymbol{\varepsilon}_i$, $(i = 1, 2, ..., m)$, with a slightly modified structure for their variance-covariance matrices though, *Pinheiro* and *Bates (1995)* suggested a Gauss-Hermite quadrature procedure to calculate integral that appears in (6.49). The Gauss-Hermite quadrature rule is often used to approximate integrals in statistics, because of its relation to Normal densities.

Specifically, Gauss-Hermite integration evaluates infinite integrals (i.e. integrals having both their limits infinite) of a function that contains a term of the form $\exp\left(-x^2\right)$. Thus, Gauss-Hermite quadrature is strictly defined in terms of integrals of the form:

$$\int_{-\infty}^{+\infty} f\left(x\right) \exp\left(-x^2\right) dx. \tag{6.50}$$

The above general form of the Gauss-Hermite quadrature approximation, associated with integration in the range $(-\infty, +\infty)$, may, then, be written as:

$$\int_{-\infty}^{+\infty} f\left(x\right) \exp\left(-x^2\right) dx \simeq \sum_{i=1}^{n} w_i f\left(x_i\right), \tag{6.51}$$

where the nodes $\{x_i\}_{i=1}^{n}$ are now the zeros of the mth order Hermite polynomial and the weights $\{w_i\}_{i=1}^{n}$ are suitably corresponding weights (*Liu* and *Pierce, 1994*). The above formula is exact for any function of the form:

$$f\left(x\right) = \exp\left[-b\left(x-a\right)^2\right] \sum_{i=0}^{2n-1} c_i x^i.$$

Appropriate values for the nodes x_i and weights w_i $(n = 1, 2, ..., 10, 12, 16, 20)$ are tabulated in *Abramowitz* and *Stegun (1972).* Alternatively, one may use an algorithm developed by *Golub (1973).* The appropriateness of using formula (6.51) resolves from the fact that integral $\int p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right) p\left(\mathbf{u}_i; \mathbf{D}\right) d\mathbf{u}_i$, under the normality assumption for vectors $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$, should contain factors of the form $\exp\left(-x^2\right)$. To clarify things, let us consider nonlinear mixed effects model of *Pinheiro* and *Bates (1995).* In words, their model resembles (two-stage) NLME model as described by (6.3), (6.4), with the only exception being that the aforementioned authors adopt a different parameterization to express the variance-covariance matrices of $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$. That is, instead of assuming the Gaussian multivariate vectors $\mathbf{u}_i$, $\boldsymbol{\varepsilon}_i$ be distributed as $\mathbf{u}_i \sim N_q\left(\mathbf{0}, \mathbf{D}\right)$ and $\boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{R}_i\right)$

respectively, they propose the following model:

$$\mathbf{y}_i = f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathrm{b}, \mathrm{u}_i\right)\right\} + \varepsilon_i \qquad (i = 1, 2, ..., m)\,, \tag{6.52}$$

with

$$\mathbf{u}_i \sim N_q\left(0,\sigma^2\mathbf{D}\right) \quad \textit{and} \quad \boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0},\sigma^2\mathbf{I}_{n_i}\right), \tag{6.53}$$

whereas now $\sigma^2\mathbf{D}$ denotes the variance-covariance matrix of $\mathbf{u}_i$, and $\sigma^2\mathbf{I}_{n_i}$ denotes the variance-covariance matrix of $\boldsymbol{\varepsilon}_i$. As usual, $\mathbf{I}_{n_i}$ is the identity matrix of order n_i. Evidently, considering (6.53), probability density functions for vector $\mathbf{u}_i$ and conditional vector $\mathbf{y}_i \mid \mathbf{u}_i$ $(i = 1, 2, ..., m)$, namely $p\left(\mathbf{u}_i; \mathbf{D}\right)$ and $p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right)$ are given by:

$$p\left(\mathbf{u}_i\right) = \left(2\pi\right)^{-\frac{q}{2}} \mid \sigma^2\mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{u}_i^t\left(\sigma^2\mathbf{D}\right)^{-1}\mathbf{u}_i\right\},$$

and

$$\begin{aligned} & p\left(\mathbf{y}_i \mid \mathbf{u}_i\right) \\ = \;& \left(2\pi\right)^{-\frac{n_i}{2}} \mid \sigma^2\mathbf{I}_{n_i}\mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \left(\sigma^2\mathbf{I}_{n_i}\right)^{-1}\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]\right\}, \end{aligned}$$

respectively. It is not difficult to verify that the above pdf's, after some trivial manipulations become:

$$p\left(\mathbf{u}_i\right) = \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{\mathbf{u}_i^t\mathbf{D}^{-1}\mathbf{u}_i}{\sigma^2}\right\}, \tag{6.54}$$

and

$$p\left(\mathbf{y}_i \mid \mathbf{u}_i\right) = \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2}\frac{\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathrm{T}_i, \mathrm{b}, \mathrm{u}_i\right)\right\}\right]}{\sigma^2}\right\}. \tag{6.55}$$

Thus, from (6.54), (6.55), the integral that we want to calculate for the marginal distribution of $\mathbf{y}_i$ is easily seen that can be written as:

$$
\begin{aligned}
p\left(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i\right) &= \int p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right) p\left(\mathbf{u}_i; \mathbf{D}\right) d\mathbf{u}_i \\
&= \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]}{2\sigma^2}\right\} \\
&\cdot \exp\left\{-\frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} d\mathbf{u}_i,
\end{aligned}
$$

or, equivalently, as:

$$
\left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} d\mathbf{u}_i, \tag{6.56}
$$

since, by definition,

$$
\begin{aligned}
\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right\|^2 &= \sum_{i=1}^{m} \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^2 \\
&= \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]
\end{aligned}
$$

Pinheiro and *Bates (1995)* in their approach, use successive applications of one-dimensional Gauss-Hermite quadrature rules, to obtain an approximation of (6.56). Specifically, to illustrate the ideas, let (x_j, w_j; j = 1,2,...,p) denote the nodes and weights of the pth-order (one dimensional) Gauss-Hermite quadrature rule, based on the $N\left(0,1\right)$ kernel. Then, the Gauss-Hermite quadrature approximation (transforming first $\mathbf{u}_i$ as $\mathbf{u}_i \equiv \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}$), proceeds as follows:

$$
\begin{aligned}
&\left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} d\mathbf{u}_i \\
&= \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\right)^{-\frac{q}{2}} \exp\left\{-\frac{\left\|\mathbf{y}_i - f\left\{\mathrm{d}\left(\mathrm{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}\right)\right\}\right\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\left\|\mathbf{x}\right\|^2}{2}\right\} d\mathbf{x}
\end{aligned}
$$

$$\simeq \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \sum_{j_1}^{p} \cdots \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k}, \tag{6.57}$$

where q is the dimension of random effects vector $\mathbf{u}_i$, and $\mathbf{x}_{j_1,\dots,j_q}$ is an (abscissas) vector with elements $\left(x_{j_1}, x_{j_2}, \dots, x_{j_q}\right)$. Thus, for the resulting approximation to the overall likelihood function $L\left(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}\right)$ [i.e. the likelihood of the complete set of $N = \sum_{i=1}^{m} n_i$ measurements $\mathbf{y} = \mathbf{y} = \left(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m\right)^t$] required for parameter estimation, we have from (6.57) the following:

$$\begin{aligned} L\left(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}\right) &= \prod_{i=1}^{m} p\left(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i\right) \\ &\simeq \prod_{i=1}^{m} \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \sum_{j_1}^{p} \cdots \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k} \\ &\simeq \left(2\pi\sigma^2\right)^{-\frac{\sum n_i}{2}} \prod_{i=1}^{m} \left[\sum_{j_1}^{p} \cdots \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k} \right], \end{aligned}$$

and the corresponding approximation to the log-likelihood function of $L\left(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}\right)$, namely $\lambda = \ln L\left(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}\right)$ is:

$$\begin{aligned} \lambda &= \ln L\left(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}\right) \\ &= \ln \left\{ \left(2\pi\sigma^2\right)^{-\frac{\sum n_i}{2}} \prod_{i=1}^{m} \left[\sum_{j_1}^{p} \cdot\cdot \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k} \right] \right\} \\ &= -\frac{\sum n_i}{2} \ln\left(2\pi\sigma^2\right) + \ln \prod_{i=1}^{m} \left[\sum_{j_1}^{p} \cdot\cdot \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k} \right. \\ &= -\frac{N}{2} \ln\left(2\pi\sigma^2\right) + \sum_{i=1}^{m} \ln \left[\sum_{j_1}^{p} \cdot\cdot \sum_{j_q}^{p} \exp\left\{ -\frac{\left\| \mathbf{y}_i - f\left\{\mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \sigma\mathbf{D}^{\frac{t}{2}}\mathbf{x}_{j_1,\dots,j_q}\right)\right\}\right\|^2}{2\sigma^2} \right\} \prod_{k=1}^{q} w_{j_k} \right] \end{aligned}$$

In addition to the standard Gaussian quadrature approximation described above, *Pinheiro* and *Bates (1995)* (see also *Liu* and *Pierce, 1994*), suggested an adaptive[5] Gauss-Hermite quadrature procedure to calculate integral $\int p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right) p\left(\mathbf{u}_i; \mathbf{D}\right) d\mathbf{u}_i$. According to this modified Gauss-Hermite procedure, the integral is centered on the empirical Bayes estimate of $\mathbf{u}_i$ [recall from section 6.5 that estimator $\hat{\mathbf{u}}_i$ is defined as the value that maximizes function (6.35)]. Further, matrix $\mathbf{G}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}$ is used in place of $\mathbf{D}$ for the scaling of the quadrature abscissas, where $\mathbf{G}$ is given by:

$$\mathbf{G} = \left[\frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \Big|_{\mathbf{u}_i = \hat{\mathbf{u}}_i} \cdot \left[\frac{\partial}{\partial \mathbf{u}_i} f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] \Big|_{\mathbf{u}_i = \hat{\mathbf{u}}_i} + \mathbf{D}^{-1},$$

or, alternatively, due to that $\mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \equiv \partial / \partial \mathbf{u}_i \left[f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]$, by:

$$\mathbf{G} = \mathbf{Z}_i^t\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \Big|_{\mathbf{u}_i = \hat{\mathbf{u}}_i} \cdot \mathbf{Z}_i\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\} \Big|_{\mathbf{u}_i = \hat{\mathbf{u}}_i} + \mathbf{D}^{-1}. \tag{6.58}$$

(In fact, $\mathbf{G}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}$ corresponds to the approximate second-order partial derivative of $\left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^t \left[\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right] + \mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i$, with respect to $\mathbf{u}_i$ evaluated at the estimator $\hat{\mathbf{u}}_i$ of $\mathbf{u}_i$). Thus, by the above it is suggested to use the transformation $\mathbf{u}_i = \hat{\mathbf{u}}_i + \sigma\left[\mathbf{G}\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right]^{-\frac{1}{2}} \mathbf{x}$. With this modification under consideration, the adaptive Gauss-Hermite quadrature approximation to the integral over $\mathbf{u}_i$ is then derived as follows:

$$\begin{aligned}
p\left(\mathbf{y}_i; \mathbf{b}, \mathbf{D}, \mathbf{R}_i\right) &= \int p\left(\mathbf{y}_i \mid \mathbf{u}_i; \mathbf{b}, \mathbf{R}_i\right) p\left(\mathbf{u}_i; \mathbf{D}\right) d\mathbf{u}_i \\
&= \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\sigma^2\right)^{-\frac{q}{2}} \mid \mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} d\mathbf{u}_i \\
&= \left(2\pi\sigma^2\right)^{-\frac{n_i}{2}} \int \left(2\pi\right)^{-\frac{q}{2}} \mid \mathbf{G}\mathbf{D} \mid^{-\frac{1}{2}} \exp\left\{-\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)\right\}\right\|^2}{2\sigma^2} - \frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2} + \frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} \\
&\quad \times \exp\left\{-\frac{\mathbf{u}_i^t \mathbf{D}^{-1} \mathbf{u}_i}{2\sigma^2}\right\} d\mathbf{u}_i
\end{aligned}$$

[5] Standard quadrature rules are all based on n subintervals of *equal* size. Quadrature rules which adapt the length of the subintervals to the local behavior are called *adaptive* (*Conte* and *de Boor, 1981*).

$$= (2\pi\sigma^2)^{-\frac{n_i}{2}} \int (2\pi)^{-\frac{q}{2}} \mid \mathbf{GD} \mid^{-\frac{1}{2}} \exp\left\{ -\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}\right)\right\}\right\|^2}{2\sigma^2} - \frac{\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}\right)^t \mathbf{D}^{-1}\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}\right)}{2\sigma^2} + \frac{\|\mathbf{x}\|^2}{2}\right\} \times \exp\left\{-\frac{\|\mathbf{x}\|^2}{2}\right\} dx$$

$$\simeq (2\pi\sigma^2)^{-\frac{n_i}{2}} \sum_{j_1}^{p} \cdots \sum_{j_q}^{p} \exp\left\{ -\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)\right\}\right\|^2}{2\sigma^2} - \frac{\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)^t \mathbf{D}^{-1}\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)}{2\sigma^2} + \frac{\left\|\mathbf{x}_{j_1,\ldots,j_q}\right\|^2}{2}\right\} \prod_{k=1}^{q} w_{j_k}$$

and, via similar to the standard Gauss-Hermite approximation manipulations, the approximation to log-likelihood function $\lambda = \ln L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y})$ is readily calculated as (see *Pinheiro & Bates, 1995,* p. 9):

$$\lambda(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y}) = \ln L(\mathbf{b}, \mathbf{D}, \mathbf{R}_i; \mathbf{y})$$

$$= -\frac{1}{2}\left[N \ln(2\pi\sigma^2) + m \ln|\mathbf{D}| + \sum_{i=1}^{m} \ln|\mathbf{G}|\right] +$$

$$+ \sum_{i=1}^{m} \ln\left[\sum_{j_1}^{p} \cdot\cdot \sum_{j_q}^{p} \exp\left\{ -\frac{\left\|\mathbf{y}_i - f\left\{\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)\right\}\right\|^2}{2\sigma^2} - \frac{\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)^t \mathbf{D}^{-1}\left(\hat{\mathbf{u}}_i + \sigma\mathbf{G}^{-\frac{1}{2}}\mathbf{x}_{j_1,\ldots,j_q}\right)}{2\sigma^2} + \frac{\left\|\mathbf{x}_{j_1,\ldots,j_q}\right\|^2}{2}\right\} \prod_{k=1}^{q} w_{j_k}\right].$$

Clearly, as is the case with the preceeding approximation methods, the usual approach once the (standard/adaptive) Gauss-Hermite quadrature for the evaluation of the log-likelihood function of the (complete) data is obtained, is to invoke the maximum likelihood (ML) or the restricted maximum likelihood (REML) method for the estimation of the parameters of interest in the nonlinear mixed-effects model.

6.7 Nonparametric/Semiparametric Methods

All approximate methods for parameter estimation, considered up to this point, are appropriate under fully parametric model specifications. In particular, the fully parametric model that imposes Gaussian distributions for the random parameters $\mathbf{u}_i$ and ε_i defined in equations (6.3) and (6.4), is essential for estimation and inference via the preceding methodology.

In general, with parametric models in statistics, each parameter of the model is assumed to arise from a specific parametric family of distributions. A commonly used distribution, as already seen, is the Normal distribution. In many settings, however, it is unrealistic to assume that the distribution of the parameters belongs to a specific parametric distributional family. For this reason, alternatively, one may wish to avoid any distributional assumptions.

In the nonparametric approach, no parametric assumptions about the assumed shape of a parameter distribution are made. In this sense, nonparametric (as well semiparametric) model specifications provide a more flexible framework for estimation and inference. On the other hand, the main disadvantages are that such approaches are usually computationally intensive, and result in a discrete estimate of a possibly continuous distribution.

In an effort to reach a compromise between the very restrictive parametric models and the rather loose and very general nonparametric models (as concerns their distributional assumptions), semiparametric models that essentially borrow features from both parametric and nonparametric specifications have been proposed. In many settings, however, it is unrealistic to assume that the distribution of the parameters belongs to a specific parametric distributional family. For this reason, alternatively, one may wish to avoid any distributional assumptions. In the sequel, we briefly review the existing literature on these frameworks, appearing in the longitudinal data context, in turn.

6.7.1 Nonparametric Model Specification

As previously mentioned, with methods based on nonparametric specifications, no parametric assumptions (e.g. normality) about the form of the parameter distributions are made. In particular, the only assumption made is that parameters are random variables with common distribution function $\mathcal{H}$, where $\mathcal{H}$ is a completely unspecified distribution. $\mathcal{H}$ has come to be known as the mixing distribution. Inference for fixed/random parameters is based on the (marginal) likelihood of the data, which in turn is associated with the unknown distribution $\mathcal{H}$. Thus, the obvious consequence is that $\mathcal{H}$ need to be somehow estimated. This is accomplished by the distribution that yields the highest likelihood of all distributions. From this estimated distribution, the means, standard deviations and covariances can be derived along with any other statistics of the distribution. The specific method is referred to as the nonparametric maximum likelihood (NPML in abbreviation) method of the mixing distribution [we refer to *Laird (1978), Kiefer* and *Wolfowitz (1956), Lindsay (1983; 1995), McLachlan* and *Basford (1988), McLachlan* and *Peel (2000)* and *Böhning (2000)* for a discussion of mixture distributions and their estimation through NPML method in a general context].

The nonparametric ML approach in the context of longitudinal studies, and especially in the population pharmacokinetic modeling, was initially introduced by *Mallet (1986)*. Mallet proposes a model formulation that is completely nonparametric, in the sense that no parametric form is assumed for the random parameters vectors $\boldsymbol{\beta}_i$ in terms of fixed effects $\mathbf{b}$, nor is any assumption made about their distribution. In other words, the distribution of each $\boldsymbol{\beta}_i$ $(i = 1, 2, ..., m)$ is assumed to lie in a wide class of probability functions which must be determined. Specifically, $\boldsymbol{\beta}_i$ is assumed to follow an $\mathcal{H}$ distribution, where $\mathcal{H}$ remains completely unrestricted. Estimation of $\mathcal{H}$ is achieved by (nonparametric) maximum likelihood.

As concerns now the other remaining random part of the NLME model, under the nonparametric specification of $\boldsymbol{\beta}_i$, within-subject error term $\boldsymbol{\varepsilon}_i$ is specified by a parametric form [e.g. take $\boldsymbol{\varepsilon}_i$ to be normally distributed with $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$]. Thus, in this respect, the (two-stage) NLME model under the nonparametric specification is as follows:

$$\underline{stage\ 1}: \quad (describes\ the\ within - subject\ variation)$$
$$\mathbf{y}_i = f\left(\boldsymbol{\beta}_i\right) + \boldsymbol{\varepsilon}_i \quad (i = 1, 2, ..., m)\,,$$
$$\boldsymbol{\varepsilon}_i \sim N_{n_i}\left(\mathbf{0}, \mathbf{R}_i\right),$$

and

$$\underline{stage\ 2}: \quad (describes\ the\ between - subject\ variation)$$
$$\boldsymbol{\beta}_i \sim \mathcal{H},$$

where, as usual, $\mathbf{y}_i = \left(y_{i1}, y_{i2}, ..., y_{in_i}\right)^t$, $\boldsymbol{\beta}_i$ is an unobservable $(t \times 1)$ vector of random parameters specific to the ith subject and associated to the covariates, $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i}\right)^t$ is the $(n_i \times 1)$ error vector, and $\mathbf{R}_i$ denotes the variance-covariance matrix of $\boldsymbol{\varepsilon}_i$ under the Gaussian assumption. Also, as already stated, $\mathcal{H}$ is an entirely unrestricted distribution function. Note that the between-subject variation (stage 2) is uniquely accommodated through the distribution $\mathcal{H}$.

Only a few basic nonparametric maximum likelihood methods for the estimation of distribution $\mathcal{H}$ (and consequently the estimation of the complete data log-likelihood), have appeared in the literature; one is the sequential algorithm of *Fedorov (1972),* also known as the Basic Algorithm. Another one is the NPML method of *Mallet (1986),* which basically constitutes a refinement of the Basic Algorithm of Fedorov. An alternative approach, namely the nonparametric expectation-maximization (NPEM) method is described by *Schumitzky (1991, 1993).* As its name suggests, the NPEM method is a maximum likelihood approach based on the nonparametric EM algorithm. The specific method has now been implemented as a computer program [cf. *Schumitzky et al., (in preparation)*]. A good review of the preceding approaches is given in *Davidian* and *Giltinan (1995).*

Modifications to the above methods suggest smoothing the obtained discrete estimate of distribution $\mathcal{H}$, due to that in many occasions in pharmacokinetics it is fairly possible for distribution $\mathcal{H}$ to be smooth. Thus, a common approach is to smooth the discrete estimate, for example, using a normal kernel function (see, e.g. *Mallet, 1986; Schumitzky, 1993*). Furthermore, *Mallet (1988)* suggested a modification of the NPML approach of *Mallet (1986)*. The consequence is that a separate maximization is required to estimate unknown fixed parameters, resulting though in an increase in computational effort and consuming time.

6.7.2 Semiparametric Model Specification

In the sense that the nonparametric approach allows the random parameters to arise from virtually any distribution and the fully parametric approach determines specific distributional behavior for the random parameters, semiparametric specification may be seen as a compromise between these two distinct methods. According to the semiparametric specification for the NLME model, a parametric model is assumed for the $\boldsymbol{\beta}_i$'s $(i = 1, 2, ..., m)$, while for the random effects $\mathbf{u}_i$ $(i = 1, 2, ..., m)$ a more flexible distributional form is typically chosen. Specifically, it is assumed that the random effects $\mathbf{u}_i$ arise from a class of probability densities that includes the normal density, densities with multiple modes, skewed densities, but excludes densities that are unlikely to be suitable for real-world population pharmacokinetic experiments. In addition, as in the nonparametric approach, the normal distribution is mostly used to parameterize the random errors $\boldsymbol{\varepsilon}_i$, but it is likely to use other parametric forms as well.

Davidian and *Gallant (1992; 1993)* following ideas of *Gallant* and *Nychka (1987)*, originated this approach proposing a particular inferential method, which is referred to as the smooth nonparametric maximum likelihood (SNP) method in the pharmacokinetics literature. In this setting, the semiparametric specification for the (two-stage) NLME model is given as follows:

$$\underline{stage\ 1}: \ (describes\ the\ within-subject\ variation)$$

$$\mathbf{y}_i = f(\boldsymbol{\beta}_i) + \boldsymbol{\varepsilon}_i \quad (i = 1, 2, ..., m),$$

$$\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i),$$

and

$$\underline{stage\ 2}: \quad (describes\ the\ between - subject\ variation)$$

$$\boldsymbol{\beta}_i = \mathrm{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right),$$

$$\mathbf{u}_i \sim h, \; h \in \mathcal{H}$$

with $\mathbf{d}\left(\mathbf{T}_i, \mathbf{b}, \mathbf{u}_i\right)$ denoting a parametric model for the $\boldsymbol{\beta}_i$'s, and with h being a density belonging to a class $\mathcal{H}$ of 'smooth' densities. In defining the class of densities $\mathcal{H}$, a number of various alternatives exist. See *Davidian* and *Gallant (1992; 1993)* for more details on the topic. In the following years, further papers concerning nonparametric and semiparametric models were published. For instance, other non- and semiparametric approaches appeared in the recent literature include *Zeger* and *Diggle (1994)* and *Mandema et al. (1992)*.

6.8 Bayesian Approaches to the NLME Model

By far, the common approaches to fit nonlinear longitudinal data are, primarily, classical parametric and secondarily nonparametric/semiparametric approaches, previously illustrated. However, to circumvent the integration problem occurring in NLME models, the Bayesian approximate methodology may be alternatively used. In fact, due to the recent advances in Bayesian computational techniques, NLME models are nowadays very naturally implemented within the Bayesian framework.

In the current section, we briefly overview Bayesian methodology that finds application to longitudinal data (and more specifically pharmacokinetic data), appeared in the literature. Before proceeding, however, it would be instructive to express the nonlinear mixed-effects model from a Bayesian framework. Specifically, as already has been seen, parametric as well as nonparametric/semiparametric specifications of NLME model include the formulation of two-stage models (where stage 1 is used to describe the within-subject variability, and stage 2 addresses between-subject variation). To follow a Bayesian framework, a third stage, at which prior distributions are specified for the (population) parameters from the first and second stages, is incorporated. In this setting, a parametric general (three stage) Bayesian NLME model is expressed as (see, e.g., *Davidian* and *Giltinan, 1995*):

$$\underline{stage\ 1}: \ \ (describes\ the\ within - subject\ variation)$$

$$\mathbf{y}_i = f\left(\boldsymbol{\beta}_i\right) + \varepsilon_i \quad (i = 1, 2, ..., m)\,,$$

$$\boldsymbol{\varepsilon}_i \sim \left(\mathbf{0}, \mathbf{R}_i\right),$$

$$\underline{stage\ 2}: \ \ (describes\ the\ between - subject\ variation)$$

$$\boldsymbol{\beta}_i = \mathrm{d}\left(\mathrm{T}_i, \mathbf{b}, \mathbf{u}_i\right),$$

$$\mathbf{u}_i \sim \left(0, \mathrm{D}\right),$$

and

$$\underline{stage\ 3}: \ \ (hyperprior\ distribution)$$

$$\left(\mathbf{b}, \mathbf{R}_i, \mathbf{D}\right) \sim p\left(\mathbf{b}, \mathbf{R}_i, \mathbf{D}\right).$$

whereby, stage 3 essentially specifies a prior distribution p for all parameters (i.e. b, R_i and $\mathbf{D}$) in stages 1 and 2. The most common distributional choice for both stages 1 and 2 is that of the multivariate Normal distribution, while the prior distribution is chosen to be noninformative. Generally, the prior distribution is assumed to be the product of independent conjugate priors for each of these parameters.

In recent years several authors have been employed Bayesian analysis for the estimation of parameters in pharmacokinetic models involving longitudinal data. A fully parametric Bayesian approach, for instance, was used by *Berkey (1982); Racine-Poon (1985); Wakefield* and *Racine-Poon (1985); Wakefield (1996); Wakefield* and *Bennett (1996); Tierney* and *Kadane (1986); Geweke (1989)* and *Wakefield et al. (1994)* among others. Most of these population pharmacokinetic analyses concern nonlinear hierarchical population models (as described, for example, in *Longford; 1993* or *Goldstein; 1995*), which naturally extend the NLME models in the sense that mixed models may be viewed as hierarchical models with a single level of grouping.

As already emphasized in the previous sections, estimation of the models' parameters is not straightforward due to the integration problem caused by the nonlinearity of response function $f(\cdot)$. In this context, Markov chain Monte Carlo (MCMC) methods have been shown to be useful. In particular, a special Markov chain technique, namely the Gibbs sampling algorithm introduced in *Geman* and *Geman (1984)* and brought to the attention of statistical community by *Gelfand* and *Smith (1990)*, has proven to be extremely useful with Bayesian nonlinear hierarchical models. As an example, consider *Wakefield et al. (1994)* and *Wakefield (1996)* who provide descriptions of Markov chain approaches to Bayesian calculations for hierarchical models. *Wakefield et al. (1994)*, illustrate an application of a Gibbs sampler variant to analyze the pharmacokinetic data on the plasma concentration of the drug Cadralazine measured in 10 cardiac failure patients at various times, using a nonlinear population model (they also analyze the famous *Potthoff* and *Roy, 1964* data, considering a Normal-linear population model). In a similar fashion, *Wakefield (1996)* has used the general Hastings-Metropolis algorithm (*Hastings, 1970*) to implement Bayesian inference for the drug quinidine data.

Another interesting work on the specific area was done by *Geweke (1989)* who proposed the use of importance sampling (IS) for Bayes models. IS provides a simple and efficient way of performing Monte Carlo integration. It relies on much the same calculations as the Gibbs sampler, however, it does not rely on an underlying Markov chain as Gibbs sampling algorithm. Instead, many independent and identically distributed replicates are run in order to create an importance sample. See also *Pinheiro* and *Bates (1995)* for a detailed description of the IS approximate method.

Two alternative approaches for the analysis of nonlinear mixed(-effects) models within the general context of Bayesian inference, are the EM-type approximations [see, e.g., *Racine-Poon (1985)*; *Racine-Poon* and *Smith (1990)*], and the Laplace approximation method [cf. *Tierney* and *Kadane (1986)*; *Kass* and *Steffey (1989)*]. More recently, *Achcar* and *Smith (1990)* attempted on improving the Laplace method of Tierney and Kadane by suggesting suitable parameter transformations. Finally, in another context, *Berkey (1982)* fits a well-known nonlinear growth model, namely the Jenss model (*Jenss* and *Bayley; 1937*), to child repeated measurement data and estimate parameters, by means of an empirical Bayes approach (following methodology of *Lindlay* and *Smith (1972)* for linear mixed models).

In summary, as a final remark, it is important to note that Markov chain and other related Bayesian sampling techniques have generally shown to be very useful and appealing for the analysis of complicated nonlinear models, caution however is required when applying the latter methods due to their complexity and computational extensiveness.

6.9 Software for Nonlinear Mixed-Model Analysis

The development of statistical procedures for implementing nonlinear mixed effects models has been an active area of research in the past two decades, mainly due to the significant advances in computing hardware and software. As a result, a variety of software is currently available that enables researchers to analyze longitudinal/repeated measures data, (especially population pharmacokinetic/pharmacodynamic data), using nonlinear mixed model methodology.

For instance, SAS (*Littell et al., 1996*) procedure `NLMIXED`, fits nonlinear mixed models using likelihood based methods. In particular, `PROC NLMIXED` maximizes an approximation to the, difficult to calculate, likelihood of the data. Different methods for approximating the likelihood are available, the principal ones being adaptive Gaussian quadrature (see section 6.6), and the first-order linearization method (see section 6.3) of *Beal* and *Sheiner (1982; 1988; 1992)* and *Sheiner* and *Beal (1980; 1985)*. The default method in `PROC NLMIXED` is adaptive Gauss-Hermite quadrature. However, the procedure enables the user to implement the ordinary Gaussian quadrature in request. Also, as already mentioned, the well-known first-order method of Beal and Sheiner is optionally available in `PROC NLMIXED`. The estimation method of *Lindstrom* and *Bates (1990)* (Section 6.4), is not available. However, the closely related Laplacian approximation is an option.

As is known, the need for maximization of the approximated likelihood requires implementation of numerical optimization techniques. Several iterative optimization algorithms are currently available in `PROC NLMIXED` for this purpose. The default is a dual quasi-Newton algorithm. Successful convergence of the optimization algorithm results in parameter estimates along with their approximate standard errors computed from the second derivative matrix of the likelihood function. Thereby, `PROC NLMIXED` readily computes efficient estimates of the model's parameters and valid standard errors of the latter estimates. Finally, we note that due to the general nonlinear formulation, no direct analogue to the REML method is available in the `NLMIXED` procedure. Only standard maximum likelihood methods are used.

An alternative for the analysis of NLME models as concerns commercial packages, is provided by the S-PLUS (*Mathsoft Inc., 1997*) function `nlme`, written by *Pinheiro et al. (1993)*. The `nlme` function fits nonlinear mixed-effects models using the two-stage algorithm, as is defined in *Lindstrom* and *Bates (1990)* in the special case where $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$ [with the modification though, that for the PNLS step of the two-stage algorithm the loosely coupled structure of the nonlinear least squares minimization problem as described in *Soo* and *Bates (1992)*, is used]. Either maximum likelihood (ML) or restricted maximum likelihood (REML) may be used as the plausible estimation method. More precisely, a list of starting values for the fixed-effects parameters are required for the two-stage iterative algorithm. The default starting values for the random effects are zero. Also, starting values for the variance parameters are automatically generated using a formula from *Laird et al. (1987)*, in the case they are not supplied. As is the case with the linear relative of `nlme` function, namely the `lme` set of functions, various structures may be used for the parameterization of the between-subject variance (i.e. parameterization of matrix $\mathbf{D}$). A useful reference describing the different variance-covariance parameterizations is *Pinheiro* and *Bates (1996)*.

A program, entirely devoted to nonlinear mixed model analysis, is the software package NONMEM (*Beal* and *Sheiner, 1992*). The specific program has been widely used by practitioners in the area of pharmacokinetic and pharmacodynamic analysis. NONMEM performs maximum likelihood estimation based either on the first-order linearization method of Beal and Sheiner, or on the conditional first-order linearization method of Lindstrom and Bates. It has the characteristic of supporting a quite general modeling of the within-subject covariance structures. The core of the NONMEM program is a set of subroutines written in the FORTRAN programming language, thus NONMEM can run on any platform supporting a FORTRAN compiler.

As concerns the fit of nonparametric NLME models (see section 6.7.1), nonparametric maximum likelihood estimation using a continuous version of the EM algorithm of *Schumitzky (1991; 1993)* is implemented in the NPEM program, which is available as part of the USC*PACK suite of PC programs (for more details on the USC*PACK collection of PC programs the interested reader is referred to *Jelliffe et al., 1994*).

Similarly, the NPML program, using the Basic Algorithm of Section 6.7.1 as described by *Mallet (1986)*, computes the nonparametric ML estimate of the unknown distribution density $\mathcal{H}$. NPML was the first program to compute the entire discrete distribution function $\mathcal{H}$, without making any parametric assumptions.

A program for analyzing longitudinal data using the ideas of semiparametric modeling (see section 6.7.2) is NLMIX. NLMIX is a FORTRAN program that implements the smooth nonparametric maximum likelihood method (SNP) of *Davidian* and *Gallant (1992; 1993)*. The interested reader may be referred to *Davidian* and *Gallant (1994)* for a detailed description of NLMIX program.

Finally, as far as Bayesian analysis of NLME models is concerned, many standard nonlinear Bayesian models can be implemented in the software package BUGS (*Spiegelhalter et al., 1995*).

APPENDIX

Table A1

The Potthoff and Roy (1964) data

	age			
	8	10	12	14
female	21.0	20.0	21.5	23.0
female	21.0	21.5	24.0	25.5
female	20.5	24.0	24.5	26.0
female	23.5	24.5	25.0	26.5
female	21.5	23.0	22.5	23.5
female	20.0	21.0	21.0	22.5
female	21.5	22.5	23.0	25.0
female	23.0	23.0	23.5	24.0
female	20.0	21.0	22.0	21.5
female	16.5	19.0	19.0	19.5
female	24.5	25.0	28.0	28.0
male	26.0	25.0	29.0	31.0
male	21.5	22.5	23.0	26.5
male	23.0	22.5	24.0	27.5
male	25.5	27.5	26.5	27.0
male	20.0	23.5	22.5	26.0
male	24.5	25.5	27.0	28.5
male	22.0	22.0	24.5	26.5
male	24.0	21.5	24.5	25.5
male	23.0	20.5	31.0	26.0
male	27.5	28.0	31.0	31.5
male	23.0	23.0	23.5	25.0
male	21.5	23.5	24.0	28.0

(continued)

Table A1 (continued)

	age			
	8	10	12	14
male	17.0	24.5	26.0	29.5
male	22.5	25.5	25.5	26.0
male	23.0	24.5	26.0	30.0
male	22.0	21.5	23.5	25.0

Table A2

The rat body weight data of Box (1950)

		week		
1	2	3	4	5
57	86	114	139	172
60	93	123	146	177
52	77	111	144	185
49	67	100	129	164
56	81	104	121	151
46	70	102	131	153
51	71	94	110	141
63	91	112	130	154
49	67	90	112	140
57	82	110	139	169
59	85	121	146	181
54	71	90	110	138
56	75	108	151	189
59	85	116	148	177
57	72	97	120	144
52	73	97	116	140
52	70	105	138	171
61	86	109	120	129
59	80	101	111	122
53	79	100	106	133
59	88	100	111	122
51	75	101	123	140
51	75	92	100	119

(continued)

Table A2 (continued)

week				
1	2	3	4	5
56	78	95	103	108
58	69	93	116	140
46	61	78	90	107
53	72	89	104	122

Table A3

The theophylline data of Pinheiro and Bates (1995)

patient 1	0.74	2.84	6.57	10.50	9.66	8.58	8.36	7.47	6.89	5.94	3.28
patient 2	0.00	1.72	7.91	8.31	8.33	6.85	6.08	5.40	4.55	3.01	0.90
patient 3	0.00	4.40	6.90	8.20	7.80	7.50	6.20	5.30	4.90	3.70	1.05
patient 4	0.00	1.89	4.60	8.60	8.38	7.54	6.88	5.78	5.33	4.19	1.15
patient 5	0.00	2.02	5.63	11.4	9.33	8.74	7.56	7.09	5.90	4.37	1.57
patient 6	0.00	1.29	3.08	6.44	6.32	5.53	4.94	4.02	3.46	2.78	0.92
patient 7	0.15	0.85	2.35	5.02	6.58	7.09	6.66	5.25	4.39	3.53	1.15
patient 8	0.00	3.05	3.05	7.31	7.56	6.59	5.88	4.73	4.57	3.00	1.25
patient 9	0.00	7.37	9.03	7.14	6.33	5.66	5.67	4.24	4.11	3.16	1.12
patient 10	0.24	2.89	5.22	6.41	7.83	10.2	9.18	8.02	7.14	5.68	2.42
patient 11	0.00	4.86	7.24	8.00	6.81	5.87	5.22	4.45	3.62	2.69	0.86
patient 12	0.00	1.25	3.96	7.82	9.72	9.75	8.57	6.59	6.11	4.57	1.17

Bibliography

[1] **Abramowitz, M. and Stegun, I. (1964).** *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards Applied Mathematics Series No. 55. Washington, D.C.: U.S. Government Printing Office

[2] **Achcar, J.A. and Smith, A.F.M. (1990).** Aspects of Reparametrization in Approximate Bayesian Inference. *In Essays in Honor of George A. Barnard, ed. J. Hodges, Amsterdam: North-Holland*

[3] **Agresti, A. (1996).** *An Introduction to Categorical Data Analysis.* Wiley, New York

[4] **Akaike, H. (1974).** A New Look at the Statistical Model Identification. *IEEE Transaction on Automatic Control,* AC-19, 716-723

[5] **Anderson, S.J. and Jones, R.H. (1995).** Smoothing Splines for Longitudinal Data. *Statistics in Medicine,* 14, 1235-1248

[6] **Beal, S.L. (1984).** Population Pharmacokinetic Data and Parameter Estimation Based on their First Two Statistical Moments. *Drug and Metabolism Reviews,* 15, 173-193

[7] **Beal, S.L. and Sheiner, L.B. (1982).** Estimating Population Kinetics. *CRC Crit. Rev. Biomed. Eng.,* 8, 195-222

[8] **Beal, S.L. and Sheiner, L.B. (1988).** Heteroskedastic Nonlinear Regression. *Technometrics,* 30, 327-338

[9] **Beal, S.L. and Sheiner, L.B. (1992).** NONMEM User's Guide. *University of California, San Francisco, NONMEM Project Group*

[10] **Berkey, C.S. (1982).** Bayesian Approach for a Nonlinear Growth Model. *Biometrics,* 38, 953-961

[11] **Berkey, C.S. and Laird, N.M. (1986).** Nonlinear Growth Curve Analyses: Estimating Population Parameters. *Annals of Human Biology,* 13, 111-128

[12] **BMDP Statistical Software (1990).** *BMDP Statistical Software Manual, Vol. 2.* Los Angeles: BMDP Statistical Software

[13] **Boeckmann, A.J., Sheiner, L.B. and Beal, S.L. (1992).** NONMEM User's Guide, Part V, Introductory Guide. *University of California, San Francisco*

[14] **Böhning, D. (2000).** *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others.* Chapman & Hall/CRC, Boca Raton

[15] **Böhning, D. and Seidel, W. (2003).** Recent Developments in Mixture Models. *Computational Statistics and Data Analysis*, 41, 349-357

[16] **Box, G.E.P. (1950).** Problems in the Analysis of Growth and Wear Curves. *Biometrics,* 6, 362-387

[17] **Box, G.E.P. and Jenkins, G.M. (1976).** *Time Series Analysis: Forecasting and Control, Revised Edition.* Holden-Day, San Francisco

[18] **Bozdogan, H. (1987).** Model Selection and Akaike's Information Criterion (AIC): the General Theory and its Analytical Extensions. *Psychometrika,* 52, 345-370

[19] **Breslow, N.E. and Clayton, D.G.** (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 88, 9-25

[20] **Brockwell, P.J. and Davis, R.A. (1996).** *Introduction to Time Series and Forecasting.* Springer-Verlag, New York

[21] **Brown, H. and Prescott**, R. (1999). *Applied Mixed Models in Medicine.* Wiley, New York

[22] Bryk, A.S., **Raudenbush, S.W. and Congdon, R.T. (1996).** *Hierarchical Linear and Nonlinear Modeling with HLM/2L and HLM/3L Programs.* Scientific Software International, Inc., Chicago

[23] **Butler, S.M. and Louis, T.A. (1992).** Random Effects Models with Nonparametric Priors. *Statistics in Medicine,* 11, 1981-2000

[24] **Carlin, B.P. (1986).** Hierarchical Longitudinal Modeling. *In Markov Chain Monte Carlo in Practice, eds. W.R. Wilks, S. Richardson, and D.J. Spiegelhalter.* London: Chapman & Hall

[25] **Carlin, B.P. and Louis, T.A. (1996).** *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman & Hall

[26] **Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983).** *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth

[27] **Chib, S. and Carlin, B.P. (1999).** On MCMC Sampling in Hierarchical Longitudinal Models. *Statistics and Computing,* 9, 17-26

[28] **Cnaan, A., Laird, N.M. and Slasor, P. (1997).** Using the General Linear Mixed Model to Analyze Unbalanced Repeated Measures and Longitudinal Data. *Statistics in Medicine,* 16, 2349-2380

[29] **Conte, S.D. and de Boor, C. (1981).** *Elementary Numerical Analysis, an Algorithmic Approach.* Mc Graw-Hill International Editions

[30] **Couvreur, C. (1996).** The EM Algorithm: A Guided Tour. *To appear in the Proceedings of the 2nd IEEE European Workshop on Computationaly Intensive Methods in Control and Signal Processing, Pragues, Czech Rep., August 28-30, 1996*

[31] **Crump, S.L. (1951).** The Present Status of Variance Component Analysis. *Biometrics,* 7, 1-16

[32] **Davidian, M. (2000).** Applied Longitudinal Data Analysis, *Lecture Notes,* Department of Statistics. North Carolina State University

[33] **Davidian, M. and Gallant, R.A. (1992).** Smooth Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, with Application to Quinide. *Journal of Pharmacokinetics and Biopharmaceutics,* 20, 529-556

[34] **Davidian, M. and Gallant, R.A. (1993).** The Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Biometrika,* 80, 475-488

[35] **Davidian, M. and Gallant, R.A. (1994).** Nlmix: a Program for Maximum Likelihood Estimation of the Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Unpublished Technical Report*

[36] **Davidian, M. and Giltinan, D.M. (1995).** *Nonlinear Models for Repeated Measurement Data.* New York: Chapman & Hall

[37] **Davis, P.J. and Rabinowitz, P. (1984).** *Methods of Numerical Integration.* Academic Press, California

[38] **Dawson, K.S., Gennings, C. and Carter, W.H. (1997).** Two Graphical Techniques Useful in Detecting Correlation Structure in Repeated Measures Data. *The American Statistician,* 51, 275-283

[39] **De Bruijn, N.G. (1961).** *Asymptotic Methods in Analysis.* Amsterdam: North-Holland

[40] **Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977).** Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1-38

[41] **Dennis, J.E. and More, J.J. (1977).** Quasi-Newton Methods, Motivation and Theory. *SIAM rev,* 19, 46-89

[42] **Diggle, P.J. (1988).** An Approach to the Analysis of Repeated Measurements. *Biometrics*, 44, 959-971

[43] **Diggle, P.J. (1990).** *Time Series: A Biostatistical Introduction.* Oxford University Press, Oxford

[44] **Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994).** *Analysis of Longitudinal Data.* Clarendon Press, Oxford

[45] **Diggle, P.J. and Verbyla, A.P. (1998).** Nonparametric Estimation of Covariance Structure in Longitudinal Data. *Biometrics,* 54, 401-415

[46] **Eisenhart, C. (1947).** The Assumptions Underlying the Analysis of Variance. *Biometrics,* 3, 1-21

[47] **Everitt, B.S. and Hand, D.J.** (1981). *Finite Mixture Distributions.* New York: Chapman & Hall

[48] **Fai, A.H. and Cornelius, P.L. (1996).** Approximate F-tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-plot Experiments. *Journal of Statistical Computing and Simulation,* 54, 363-378

[49] **Fearn, T. (1975).** A Bayesian Approach to Growth Curves. *Biometrika,* 62, 89-100

[50] **Federer, W.T. (1968).** Non-negative Estimators for Components of Variance. *Applied Statistics,* 17, 171-174

[51] **Fedorov, V.V. (1972).** *Theory of Optimal Experiments.* Academic Press, New York

[52] **Gallant, A.R. and Nychka, D.W. (1987).** Seminonparametric Maximum Likelihood Estimation. *Econometrica,* 55, 363-390

[53] **Gauss, C.F. (1816).** Methodus Nova Integralium Valores per Approximationem Inveniendi. *Comment. Soc. Regiae Sci. Gottingensis Rec. Vol. III, Göttingen*

[54] **Geisser, S.** (1970). Bayesian Analysis of Growth Curves. *Sankhya, Series A,* 32, 53-64

[55] Gelfand, **A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M.** (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association*, 85, 972-985

[56] **Gelfand A.E., Sahu, S.K. and Carlin, B.P. (1995).** Efficient Parametrizations for Normal Linear Mixed Models. *Biometrika,* 82, 479-488

[57] **Gelfand A.E., Sahu, S.K. and Carlin, B.**P. (1996). Efficient Parametrizations for Generalized Linear Mixed Models (with discussion). *In Bayesian Statistics 5, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith.* Oxford: Oxford University Press

[58] **Gelfand, A.E. and Smith, A.F.M. (1990).** Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398-409

[59] **Geman, S. and Geman, D. (1984).** Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741

[60] **Gentle, J.E. (1988).** Cholesky Factorization. *In Numerical Linear Algebra for Applications in Statistics.* Berlin: Springer-Verlag

[61] **Geweke, J. (1989).** Bayesian Inference in Econometrics Models Using Monte Carlo Integration. *Econometrica,* 57, 1317-1339

[62] **Goldberger, A.S. (1962).** Best Linear Unbiased Prediction in the Generalized Linear Regression Model. *Journal of the American Statistical Association,* 57, 369-375

[63] **Goldstein,** H. (1995). *Multilevel Statistical Models.* Halstead Press, New York

[64] **Golub, G.H. (1973).** Some Modified Matrix Eigenvalue Problems. *SIAM Review,* 15, 318-334

[65] **Golub, G.H. and Welsch, J.H. (1969).** Calculation of Gaussian Quadrature Rules. *Math. Comp.,* 23, 221-230

[66] **Grady, J. and Helms, R.W. (1995).** Model Selection Techniques for the Covariance Matrix for Incomplete Longitudinal Data. *Statistics in Medicine,* 1397-1416

[67] **Graybill, F.A. and Hultquist, R.A. (1961).** Theorems Concerning Eisenhart's Model II. *Annals of Mathematical Statistics,* 32, 261-269

[68] **Grenander, U. and Szego, G. (1958).** *Toeplitz Forms and their Applications.* Berkeley, University of California Press

[69] **Gumpertz, M.L. and Pantula, S.G. (1992).** Nonlinear Regression with Variance Components. *Journal of the American Statistical Association*, 87, 201-209

[70] **Hand, D. and Crowder, M. (1990).** *Analysis of Repeated Measures.* Chapman and Hall

[71] **Hand, D. and Crowder, M. (1996).** *Practical Longitudinal Data Analysis.* Chapman and Hall

[72] **Hand, D. and Taylor, C.C. (1986).** *Multivariate Analysis of Variance and Repeated Measures, a Practical Approach for Behavioural Scientists.* Chapman and Hall

[73] **Hartley, H.O. and Rao, J.N.K. (1967).** Maximum Likelihood Estimation for the Mixed Analyses of Variance Model. *Biometrika,* 54, 93-108

[74] **Harville, D.A. (1976).** Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics,* 4, 384-395

[75] **Harville, D.A. (1977).** Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association,* 72, 320-338

[76] **Harville, D.A. (1979).** Some Useful Representations for Constrained Mixed-Model Estimation. *Journal of the American Statistical Association,* 74, 200-206

[77] **Harville, D.A. (1997).** *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York

[78] **Hastie, T.J. and Tibshirani, R.J.** (1990). *Generalized Additive Models.* Chapman and Hall

[79] **Hastings, W.K.** (1970). Monte Carlo Sampling-Based Methods Using Markov Chains and their Applications. *Biometrika,* 57, 97-109

[80] **Henderson, C.R. (1949).** Estimates of Changes in Heard Environment. *Journal Dairy Sci.,* 32

[81] **Henderson, C.R. (1950).** Estimation of Genetic Parameters. *The Annals of Mathematical Statistics,* 21, 309-310

[82] **Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.M. (1959).** The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics,* 15, 192-218

[83] **Henderson, C.R. (1975).** Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics,* 31, 423-447

[84] **Hertzog, C. and Rovine, M. (1985).** Repeated-measures Analysis of Variance in Developmental Research: Selected Issues. *Child Development,* 56, 787-809

[85] **Hill, B.M. (1965).** Inference About Variance Components in the One-way Model. *Journal of the American Statistical Association*, 60, 806-825

[86] **Hill, B.M. (1967).** Correlated Errors in the Random Model. *Journal of the American Statistical Association*, 62, 1387-1400

[87] **Hirst, K., Boyle, D.W., Zerbe, G.O. and Wilkening, R.B.** (1991). On Nonlinear Random Effects Models for Repeated Instruments. *Communications in Statistics-Simulation,* 20, 463-478

[88] **Huynh, H. and Feldt, L.S. (1970).** Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions. *Journal of the American Statistical Association*, 65, 1582-1589

[89] **Inselberg, A. (1985).** The Plane with Parallel Coordinates. *The Visual Computer,* 1, 69-91

[90] **Jelliffe, R.W., Shumitzky, A. and Van Guilder, M. (1994).** User Manual for Version 10.0 of the USC*PACK Collection of PC Programs. *Laboratory of Applied Pharmacokinetics, University of Southern California*

[91] **Jennrich, R.I. and Schluchter, M.D. (1986).** Unbalanced Repeated-measures Models with Structured Covariance Matrices. *Biometrics,* 42, 805-820

[92] **Jenss, R.M. and Bayley, N. (1937).** A Mathematical Method for Studying the Growth of a Child. *Human Biology,* 9, 556-563

[93] **Johnson, R.A. and Wichern, D.W. (1992).** *Applied Multivariate Statistical Analysis, Third Edition.* Englewood Cliffs, New Jersey: Prentice Hall

[94] **Jones, R.H. (1985).** Repeated Measures, Interventions and Time Series Analysis. *Psychoneuroendocrinology,* 10, 5-14

[95] **Jones, R.H. (1993).** *Longitudinal Data with Serial Correlation: A State-space Approach.* Chapman and Hall

[96] **Jones, R.H. and Boati-Boateng, F. (1991).** Unequally Spaced Longitudinal Data with AR(1) Serial Correlation. *Biometrics,* 47, 161-175

[97] **Jowett, G.H. (1952).** The Accuracy of Systematic Sampling from Conveyor Belts. *Applied Statistics,* 1, 50-59

[98] **Kalton, G. (1987).** *Introduction to Survey Sampling.* Quantitative Applications in the Social Sciences

[99] **Kass, R. and Raftery, A. (1995).** Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795

[100] **Kass, R. and Steffey, D. (1989).** Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *Journal of the American Statistical Association*, 84, 717-726

[101] **Keselman, H.J., Algina, J., Kowelchuk, R.K. and Wolfinger, R.D. (1999).** A Comparison of Recent Approaches to the Analysis of Repeated Measurements. *British Journal of Mathematical and Statistical Psychology,* 52, 63-78

[102] **Keselman, H.J., Algina, J., Kowelchuk, R.K. and Wolfinger, R.D. (1998).** A Comparison of two Approaches for Selecting Covariance Structures in the Analysis of Repeated Measurements. *Commun. Statistics-Simulation,* 27, 591-604

[103] **Khatri, C.G. and Srivastava, M.S. (1971).** On Exact Non-null Distributions of Likelihood Ratio Criteria for Sphericity Test and Equality of Two Covariance Matrices. *Sankhya A,* 33, 201-206

[104] **Kiefer, J. and Wolfowitz, J. (1956).** Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics,* 27, 887-906

[105] **Kloesgen, W. (1999).** Types and Forms of Data. *German National Research Center for Information Technology (unpublished manuscript)*

[106] **Kolmogorov, A.N. (1941).** The Local Structure of Turbulence in an Incompressible Fluid at Very Large Reynolds Numbers. *Doklady Akademii Nauk SSSR,* 30, 229-303

[107] **Laird, N.M. (1978).** Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association*, 73, 805-811

[108] **Laird, N.M., Lange, N. and Stram, D. (1987).** Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association*, 82, 97-105

[109] **Laird, N.M. and Louis, T.A. (1982).** Approximate Posterior Distributions for Incomplete Data Problems. *Journal of the Royal Statistical Society, Series B,* 44, 190-200

[110] **Laird, N.M. and Ware, J.H. (1982).** Random-Effects Models for Longitudinal Data. *Biometrics,* 38, 963-974

[111] **Lange, N., Carlin, B.P. and Gelfand, A.E. (1992).** Hierarchical Bayes Models for the Progression of HIV Infection Using Longitudinal CD4 T-cell Numbers (with discussion). *Journal of the American Statistical Association*, 87, 615-632

[112] **Lange, N. and Laird, N.M. (1989).** The Effect of Covariance Structure on Variance Estimation in Balanced Growth-curve Models with Random Parameters. *Biometrics,* 84, 241-247

[113] **Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989).** Robust Statistical Modeling Using the t-distribution. *Journal of the American Statistical Association*, 84, 881-896

[114] **Leonard, T., Hsu, J.S.J. and Tsui, K.W. (1989).** Bayesian Marginal Inference. *Journal of the American Statistical Association*, 84, 1051-1058

[115] **Lesaffre, E., Todem, D., Verbeke, G. and Kenward, M. (2000).** Flexible Modelling of the Covariance Matrix in a Linear Random Effects Model. *Biometrical Journal,* 42, 807-822

[116] **Liang, K.Y. and Zeger, S.L. (1986).** Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika,* 73, 13-22

[117] **Lindley, D.V. and Smith, A.F.M. (1972).** Bayes Estimates for the Linear Model (with discussion). *Journal of the Royal Statistical Society, Series B,* 34, 1-42

[118] **Lindsay, B.G. (1983).** The Geometry of Mixture Likelihoods: a General Theory. *The Annals of Statistics,* 11, 86-94

[119] **Lindsay, B.G. (1995).** Mixture Models: Theory, Geometry and Applications. *NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5 Institute of Mathematical Statistics, Hayward*

[120] **Lindstrom, M.J. and Bates, D.M. (1988).** Newton-Raphson and EM algorithms for Linear Mixed-Effects Models for Repeated Measures Data. *Journal of the American Statistical Association,* 83, 1014-1022

[121] **Lindstrom, M.J. and Bates, D.M. (1990).** Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics,* 46, 673-687

[122] **Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996).** *SAS System for Mixed Models.* SAS Institute, Inc., Cary, NC

[123] **Littell, R.C., Pendergast, J. and Natarajan, R. (2000).** Modelling Covariance Structure in the Analysis of Repeated Measures Data. *Statistics in Medicine,* 19, 1793-1819

[124] **Liu, Q. and Pierce, D.A. (1994).** A Note on Gauss-Hermite Quadrature. *Biometrika,* 81, 624-629

[125] **Liu, C. and Rubin, D.B. (1994).** The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika,* 81, 633-648

[126] **Liu, C. and Rubin, D.B. (1995).** Application of the ECME Algorithm and the Gibbs Sampler to General Linear Mixed Models. *Proceedings of the 17th International Biometric Conference,* 1, 97-107

[127] **Longford, N. (1993).** *Random Coefficient Models.* Clarendon Press, Oxford

[128] **Louis, T.A. (1991).** Using Empirical Bayes Methods in Biopharmaceutical Research. *Statistics in Medicine,* 10, 811-829

[129] **Mackay, D.J.C. (1996).** Choice of Basis for Laplace Approximation. (*Unpublished manuscript)*

[130] **Magder, L.S. and Zeger, S.L. (1996).** A Smooth Nonparametric Estimate of a Mixing Distribution Using Mixtures of Gaussians. *Journal of the American Statistical Association*, 91, 1141-1152

[131] **Mallet, A. (1986).** A Maximum Likelihood Estimation Method for Random Coefficient Regression Models. *Biometrika,* 73, 645-656

[132] **Mallet, A., Mentre, F., Steimer, J.L. and Lokiek, F. (1988).** Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, with Applications to Cyclosporine. *Journal of Pharmacokinetics and Biopharmaceutics,* 16, 311-327

[133] **Mandema, J., Verotta, D. and Sheiner, L.B. (1992).** Building Population Pharmacokinetic-Pharmacodynamic Models. I. Models for Covariate Effects. *Journal of Pharmacokinetics and Biopharmaceutics,* 20, 511-528

[134] **Mansour, H., Nordheim, E.V. and Rutledge, J.J. (1985).** Maximum Likelihood Estimation of Variance Components in Repeated Measures Designs Assuming Autoregressive Errors. *Biometrics,* 41, 287-294

[135] **Mathai, A.M. and Rathie, P.N. (1970).** The Exact Distribution for the Sphericity Test. *ournal of Statistical Research,* 4, 10-159

[136] **Matheron, G. (1963).** Principles of Geostatistics. *Economic Geology,* 58, 1246-1266

[137] **MathSoft, Inc. (1997).** *S-PLUS User's Guide.* Data Analysis Product Division, MathSoft, Seattle, WA

[138] **Mauchly, J.W. (1940).** Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics,* 11, 204-209

[139] **McLachlan, G.J. and Basford, K.E. (1988).** *Mixture Models. Inference and Applications to Clustering.* Marcel Dekker, New York

[140] **McLachlan, G.J. and Krishnan, T. (1997).** *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics

[141] **McLachlan, G.J. and Peel, D. (2000).** *Finite Mixture Models.* Wiley Series in Probability and Statistics

[142] **McLachlan, G.J., Peel, D. and Bean, R.W. (2003).** Modelling High-dimensional Data by Mixtures of Factor Analyzers. *Computational Statistics and Data Analysis*, 41, 379-388

[143] **McLean, R.A., Sanders, W. and Stroup, W. (1991).** A unified Approach to Mixed Linear Models. *Journal of the American Statistical Association*, 79, 853-862

[144] **Meng, X.L. and Rubin, D.B. (1991).** Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm. *Journal of the American Statistical Association*, 86, 899-909

[145] **Meng, X.L. and Van Dyk, D. (1998).** Fast EM-type Implementations for Mixed Effects Models. *Journal of the Royal Statistical Society, Series B,* 59, 559-578

[146] **Montgomery, D.C. (1997).** *Design and Analysis of Experiments.* Wiley, New York

[147] **Nagarsenker, B.N. and Pillai, K.C.S. (1972).** The Distribution of the Sphericity Test Criterion. *NTIS Report No AD754-232, Washington, DC*

[148] **Nagarsenker, B.N. and Pillai, K.C.S. (1973).** The Distribution of the Sphericity Test Criterion. *Journal of Multivariate Analysis,* 3, 226-235

[149] **Nelder, J.A. and Wedderburn, R.W.M. (1972).** Generalized Linear Models. *Journal of the Royal Statistical Society, Series A,* 135, 370-384

[150] **O'Brien, R.G. and Kaiser, M.K. (1985).** MANOVA Method for Analyzing Repeated Measures Designs: an Extensive Primer. *Psychological Bulletin,* 97, 316-333

[151] **Olson, C.L. (1976).** On Choosing a Test Statistic in Multivariate Analysis of Variance. *Psychological Bulletin,* 83, 574-586

[152] **Pantula, S.G. and Pollock, K.H. (1985).** Nested Analysis of Variance with Autocorrelated Errors. *Biometrics,* 41, 909-920

[153] **Parker, J.C. (ed) (1988).** SAS Users Guide: Statistics, Cary, NC: SAS Institute Inc

[154] **Patterson, H.D. (1964).** Theory of Cyclic Rotation Experiments. *Journal of the Royal Statistical Society, Series B,* 26, 1-45

[155] **Patterson, H.D. and Thompson, R. (1971).** Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrika,* 58, 545-554

[156] **Pendergast, J.F. and Broffitt, J.D. (1986).** Robust Estimation in Growth Curve Models. *Communications in Statistics: Theory and Methods,* 14, 1919-1939

[157] **Pinheiro, J.C., Bates, D.M. and Lindstrom, M. (1993).** Nonlinear Mixed Effects Classes and Methods for S. *Technical Report No. 906, Department of Statistics, University of Wisconsin*

[158] **Pinheiro, J.C. and Bates,** D.M. (1995). Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *Journal of Computational and Graphical Statistics,* 4, 12-35

[159] **Pinheiro, J.C. and Bates, D.M. (1996).** Unconstrained Parametrizations for Variance-Covariance Matrices. *Statistics and Computing,* 6, 289-296

[160] **Pinheiro, J.C., Liu, C. and Wu, Y.N. (2001).** Efficient Algorithms for Robust Estimation in Linear Mixed-Effects Models Using the Multivariate t-Distribution. *Journal of Computational and Graphical Statistics.* To appear

[161] **Pocock, S.J., Cook, D.G. and Beresford, S.A.A. (1981).** Regression of Area Mortality Rates on Explanatory Variables: What Weighting is Appropriate? *Applied Statistics,* 30, 286-295

[162] **Potthoff, R.F.** and Roy, **S.N. (1964).** A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems. *Biometrika,* 51, 313-326

[163] **Prosser, R., Rasbash, J. and Goldstein, H. (1991).** *ML3 Software for Three-level Analysis.* User's Guide for V.2, Institute of Education, University of London

[164] **Racine-Poon, A. (1985).** A Bayesian Approach to Nonlinear Random Effects Models. *Biometrics,* 41, 1015-1023

[165] **Racine-Poon, A. (1992).** Saga: Samples Assisted Graphical Analysis (disc: P401-404). *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting,* 389-401

[166] **Racine-Poon, A. and** Smith, **A.F.M. (1990).** Population Models. *In Statistical Methodology in the Pharmaceutical Sciences (ed. D. Berry).* New York

[167] **Rao, C.R. (1971a).** Estimation of Variance and Covariance Components-MINQUE Theory. *Journal of Multivariate Analysis,* 1, 257-275

[168] **Robinson, G.K. (1991).** That BLUP is a good thing: the Estimation of Random Effects. *Statistical Science*, 6, 15-51

[169] **Rouanet, H. and Lepine, D. (1970).** Comparison Between Treatments in a Repeated-measures Design: ANOVA and Multivariate Methods. *British Journal of Mathematical and Statistical Psychology,* 23, 147-163

[170] **Rutter, C.M. and Elashoff, R.M. (1994).** Analysis of Longitudinal Data: Random Coefficient Regression modelling. *Statistics in Medicine,* 13, 1211-1231

[171] **Schumitzky, A. (1991).** Nonparametric EM Algorithms for Estimating Prior Distributions. *Applied Math. Comput.,* 45, 143-157

[172] **Schumitzky, A. (1993).** The Nonparametric Maximum Likelihood Approach to Pharmacokinetic Population Analysis. *Proceedings of the 1993 Western Simulation Multiconference: Simulation for Health Care. San Diego Society for Computer Simulation,* pp. 95-100

[173] **Schumitzky, A., Van Guilder, M. and Jelliffe, R.** A PC Computer Program for Estimating Population Pharmacokinetics. In preparation

[174] **Schwarz, G. (1978).** Estimating the Dimension of a Model. *The Annals of Statistics,* 6, 461-464

[175] **Searl**e, S.R. (1971). *Linear Models.* Wiley, New York

[176] **Searle, S.R. (1971).** Topics in Variance Component Estimation. *Biometrics,* 27, 1-76

[177] **Searle, S.R. (1979).** Notes on Variance Component Estimation: A Detailed Account of Maximum Likelihood and Kindred Methodology. *Biometrics Unit, Warren Hall, Cornell University, Ithaca, NY*

[178] **Searle, S.R. (1995).** The Matrix Handling of BLUE and BLUP in the Mixed Linear Model. *Invited Paper for the Fourth International Workshop on Matrix Methods for Statistics, Montreal, Quebec*

[179] **Searle, S.R., Casella, G. and McCulloch,** C.E. (1992). *Variance Components.* Wiley, New York

[180] **Sheiner, L.B. and Beal, S.L. (1980).** Evaluation of Methods for Estimating Population Pharmacokinetic Parameters. I. Michaelis-Menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics,* 8, 553-571

[181] **Sheiner, L.B. and Beal, S.L. (1985).** Pharmacokinetic Parameter Estimates from Several Least Squares Procedures: Superiority of Extended Least Squares. *Journal of Pharmacokinetics and Biopharmaceutics,* 13, 185-201

[182] **Shi, M., Weiss, R.E. and Taylor, J.M.G. (1996).** An Analysis of Paediatrics CD4 Counts for Acquired Immune Deficiency Syndrome Using Flexible Random Curves. *Applied Statistics,* 45, 151-163

[183] **Smith, A.F.M. (1973).** A General Bayesian Linear Model. *Journal of the Royal Statistical Society, Series B,* 35, 67-75

[184] **Smith, D.M. and Diggle, P.J. (1994).** Oswald: Object-oriented Software for the Analysis of Longitudinal Data in S. *Technical Report MA94/95, Lancaster University Department of Mathematics and Statistics*

[185] **Smith, D.M., Robertson, B. and Diggle, P.J. (1997).** *Object-oriented Software for the Analysis of Longitudinal Data in S.* Technical Report MA 96/192. Department of Mathematics and Statistics, University of Lancaster

[186] **Solomon, P.J. and Cox, D.R. (1992).** Nonlinear Component of Variance Models. *Biometrika,* 79, 1-11

[187] **Soo, Y.W. and Bates, D.M. (1992).** Loosely Coupled Nonlinear Least Squares. *Computational Statistics and Data Analysis,* 14, 249-259

[188] **Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W.** (1995). BUGS 0.5 Bayesian Inference Using Gibbs Sampling Manual. *MRC Biostatistics Unit, Institution of Public Health, Robinson Way, Cambridge*

[189] **Stata Corporation** (1997). *Stata Reference Manual.* College Station, TX: Stata Press

[190] **Stiratelli, R., Laird, N.M. and Ware, J.H. (1984).** Random Effects Models for Serial Observations with Binary Response. *Biometrics,* 40, 961-971

[191] **Strenio, J.F., Weisberg, H.J. and Bryk, A.S. (1983).** Empirical Bayes Estimation of Individual Growth-curve Parameters and their Relationship to Covariates. *Biometrics,* 39, 71-86

[192] **Thompson, W.A. (1962).** The Problem of Negative Estimates of Variance Components. *The Annals of Mathematical Statistics*, 33, 273-289

[193] **Tiao, G.C. and Box, G.E.P. (1967).** Bayesian Analysis of a Three-component Hierarchical Design Model. *Biometrika,* 54, 109-125

[194] **Tierney, L. and Kadane, J.B. (1986).** Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81, 82-86

[195] **Tierney, L., Kass, R.E. and Kadane, J.B. (1989).** Approximate Marginal Densities of Non-linear Functions. *Biometrika,* 76, 425-433

[196] **Verbeke, G. and Lesaffre, E. (1996).** A Linear Mixed-effects Model with Heterogeneity in the Random-effects Population. *Journal of the American Statistical Association*, 91, 217-221

[197] **Verbeke, G., Lesaffre, E. and Brant, L.J. (1998).** The Detection of Residual Serial Correlation in Linear Mixed Models. *Statistics in Medicine,* 17, 1391-1402

[198] **Verbeke, G. and Molenbergs, G.** (1997). *Linear Mixed Models in Practice, a SAS-oriented Approach.* Springer-Verlag, New York

[199] **Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999).** The Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines. *Applied Statistics,* 48, 269-312

[200] **Vines, S.K., Gilks, W.R. and Wild, P. (1996).** Fitting Bayesian Multiple Random Effects Models. *Statistics and Computing,* 6, 337-346

[201] **Vonesh, E.F. (1992).** Nonlinear Models for the Analysis of Longitudinal Data. *Statistics in Medicine,* 11, 1929-1954

[202] **Vonesh, E.F. (1996).** A Note on Laplace's Approximation in Nonlinear Mixed Effects Models. *Biometrika,* 83, 447-452

[203] **Vonesh, E.F. and Carter, R.L. (1987).** Efficient Inference for a Random Coefficient Growth Curve Model with Unbalanced Data. *Biometrics,* 43, 617-628

[204] **Vonesh, E.F. and Carter, R.L. (1992).** Mixed-effects Nonlinear Regression for Unbalanced Repeated Measures. *Biometrics,* 48, 1-17

[205] **Vonesh, E.F. and Chinchilli, V.M.** (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* New York: Marcel Dekker

[206] **Wakefield, J.C. and Bennett, J. (1996).** The Bayesian Modeling of Covariates for Population Pharmacokinetic. *Journal of the American Statistical Association,* 91, 917-927

[207] **Wakefield, J.C. and Racine-Poon, A. (1985).** An Application of Bayesian Population Pharmacokinetics/Pharmacodynamic Models to Dose Recommendation. *Statistics in Medicine,* 14, 971-986

[208] **Wakefield, J.C., Smith, A.F.M., Racine-Poon, A. and Gelfand, A.E. (1994).** Bayesian Analysis of Linear and Nonlinear Population Models Using the Gibbs Sampler. *Applied Statistics,* 43, 201-221

[209] **Wakefield, J.C. (1996).** The Bayesian Analysis of Population Pharmacokinetic Models. *Journal of the American Statistical Association,* 91, 62-75

[210] **Wahba, G.** (1990). *Spline Models for Observational Data.* Philadelphia, SIAM

[211] **Wald, A. (1943).** Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large. *Transactions of the American Mathematical Society,* 54, 426-482

[212] **Wang, Y. (1998).** Mixed-Effects Smoothing Spline ANOVA. *Journal of the Royal Statistical Society, Series B,* 60, 159-174

[213] **Wang, Y. and Taylor, J.M.G. (1995).** Inference for Smooth Curves in Longitudinal Data with Application to an AIDS Clinical Trial. *Statistics in Medicine,* 14, 1205-1218

[214] **Ware, J.H. (1985).** Linear Models for the Analysis of Longitudinal Studies. *The American Statistician,* 39, 95-101

[215] **Wegman, E.J. (1990).** Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association,* 85, 664-675

[216] **Wegman, E.J. and Qiang, L.** High Dimensional Clustering Using Parallel Coordinates and the Grand Tour. (*Unpublished manuscript)*

[217] **Weiss, R.E. and Lazaro, C.G. (1992).** Residual Plots for Repeated Measures. *Statistics in Medicine,* 11, 115-124

[218] **Weiss, R.E. (1997).** Exploratory Data Graphics for Repeated Measures Data. *(Unpublished manuscript)*

[219] **Wolfinger**, R.D. **(1993).** Laplace's Approximation for Nonlinear Mixed Models. *Biometrika,* 80, 791-795

[220] **Wolfinger, R.D. and Lin, X. (1997).** Two Taylor-series Approximation Methods for Nonlinear Mixed Models. *Computational Statistics and Data Analysis,* 25, 465-490

[221] **Wu, C.F. (1983).** On the Convergence Properties of the EM Algorithm. *The Annals of Statistics,* 11, 95-103

[222] **Wypij, D., Pugh, M. and Ware,** J.H. (1993). Modeling Pulmonary Function Growth with Regression Splines. *Statistica Sinica,* 3, 329-350

[223] **Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K., Vonesh, E. and Wolfinger, R. (1994).** Population Pharmacokinetic/Pharmacodynamic Methodology and Applications: a Bibliography. *Biometrics,* 50, 566-575

[224] **Zeger, S.L. and Diggle, P.J. (1994).** Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters. *Biometrics,* 50, 689-699