Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 49 (2005) 243–263
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Statistical methods and software for risk assessment: applications to a neurophysiological data set☆

G.J. Karavasilis, V.K. Kotti, D.S. Tsitsis, V.G. Vassiliadis, A.G. Rigas*

Department of Electrical and Computer Engineering, Democritus University of Thrace, V. Sofias 12, GR-67100 Xanthi, Greece

Received 5 December 2003; accepted 13 May 2004
Available online 15 June 2004

Abstract

Different statistical methods for the assessment of potential risk factors are discussed, in the case of a complex neurophysiological system, involving binary observations. The first approach describes the use of non-parametric methods, which are based on the cross-product ratio (CPR). The estimates of the CPR are given both in time and frequency domains and their asymptotic distributions are discussed. The second approach is a parametric one and is based on the formulation of a logistic regression model. The estimated parameters and the corresponding odds ratio (OR) can be used to evaluate the behaviour of the system. These methods can be implemented using different statistical software, e.g. S-PLUS, GENSTAT, SPSS, LOGXACT, GLIM. Some of the computational procedures are provided, and the results obtained are clearly displayed. A discussion about the statistical packages is presented.

Keywords: Cross-product ratio; Risk assessment; Logistic regression; Muscle spindle

☆ The source code is available in the Appendix at http://utopia.duth.gr/∼rigas/ and in the electronic version of the paper at http://www.sciencedirect.com.

* Corresponding author. Tel.: +30-25410 79589; fax: +30-25410 79590
E-mail address: rigas@ee.duth.gr (A.G. Rigas).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.05.010

1. Introduction

The aim of this work is to present a discussion on statistical methods and software packages (S-PLUS, GENSTAT, SPSS, LOGXACT, GLIM) used in the study of risk assessment of a complex neurophysiological system. Such packages are very useful in handling large data sets, implement techniques in time and frequency domain of analyzing binary time series and apply the popular method of logistic regression. The field of neurophysiology is a rich source of binary type data, where both parametric and non-parametric methods can be used in order to assess the interconnections between neural networks and to identify the behaviour of a complex neurophysiological system of particular interest (see Brillinger, 1976, 1988; Kotti and Rigas, 2003a; Marmarelis and Marmarelis, 1978; Rigas, 1996; Rigas and Liatsis, 2000).

In the first part of this work, the cross-product ratio (CPR) is defined as the odds ratio (OR) in a 2×2 contingency table between two binary time series. Estimates of the CPR are presented by using methods of analyzing binary time series both in time and frequency domain. An example from the field of neurophysiology is described and computation of the estimates for the CPR is carried out in S-PLUS. In the second part, the logistic regression is used for assessing the risk factors in the function of the neurophysiological system. The estimates of the parameters of the logistic model are obtained by using the maximum likelihood function. This can be done in each of the three packages GENSTAT, S-PLUS or SPSS. The advantages and disadvantages on the use of these packages are discussed. Finally, a comparison between the two methods shows that the logistic regression provides extra information and reveals more characteristics about the function of the neurophysiological system.

2. Measuring association in binary time series

2.1. The bivariate point process

Let $N(t) = \{N_1(t), N_2(t)\}$, $-\infty < t < \infty$, be a bivariate point process where $N_1(t)$ denotes the number of events of type 1 that occurred in the time interval $(0, t]$ and $N_2(t)$ the number of events of type 2 occurred in the same time interval. By $\{\mathrm{d}N_a(t) = N_a(t, t + \mathrm{d}t]\}$, we denote the increments of $N_a(t)$, $a = 1, 2$.

Suppose that the process satisfies the following conditions:

(a) It is stationary. This means that the distribution of the variate $\{N_1(I_1), N_2(I_2)\}$ is the same with the distribution of the variate $\{N_1(I_1 + \tau), N_2(I_2 + \tau)\}$, where $I_i = (a_i, b_i)$ and $I_i + \tau = (a_i + \tau, b_i + \tau)$, $i = 1, 2$.
(b) It is orderly. This means that the points of the process are isolated with probability one, that is $\Pr(N_a(t, t+h] > 1) = o(h)$, $a = 1, 2$. The condition of orderliness prohibits events of the process to occur simultaneously and permits the point process to be considered as binary time series.
(c) It is strong mixing. This means that increments of the point process well-separated in time are independent.

More details about stationary point processes which are orderly and satisfy the condition of strong mixing can be found in Brillinger (1975), Cox and Isham (1980) and Daley and Vere-Jones (1988). Examples of point processes include the times of firing of nerve cells, the times of earthquakes in different places and the times of beats of human hearts.

An example from the field of neurophysiology is presented in this work. Our interest is to examine the effect of a gamma motoneurone on the behaviour of the neurophysiological system called muscle spindle. Some basic characteristics of the muscle spindle are discussed in Kotti and Rigas (2003a). The bivariate point process $N(t) = \{N_1(t), N_2(t)\}$ is observed for a time interval $T = 15866$ ms and the events of type 1, $N_1(T) = 1010$, represent the pulses of the gamma motoneurone imposed on the muscle spindle while the events of the type 2, $N_2(T) = 538$, represent the response of the system to the effect of the gamma motoneurone. The response of the muscle spindle is recorded from the Ia sensory axon and the information of the muscle spindle to the spinal cord is transferred through this axon.

2.2. Parameters of the bivariate point process

Certain parameters of a bivariate point process can be defined both in the time and in the frequency domain. The mean intensity of type a events is defined by

$$p_a = \lim_{h\to 0} \Pr\{\text{type } \alpha \text{ event occurs in } (t, t+h]\}/h \tag{1}$$

for $a = 1, 2$. The mean intensity does not depend on t because of stationarity. It follows from the condition of orderliness that

$$E[\mathrm{d}N_a(t)] = p_a\,\mathrm{d}t. \tag{2}$$

In the survival literature the quantity $p_a\,\mathrm{d}t$ corresponds to the unconditional probability of hazard rate (see Johnson and Johnson, 1980, p. 51). The second-order product density of type a events with type b events is defined by

$$p_{ab}(u) = \lim_{h_1\to 0, h_2\to 0} \Pr\{\text{type } \alpha \text{ event in } (t+u, t+u+h_1] \text{ and type } b \text{ event in } (t, t+h_2]\}/h_1h_2 \tag{3}$$

for $a, b = 1, 2$ and u is the lag time between the events a and b ($u \neq 0$). It also holds, because of orderliness, that

$$E[\mathrm{d}N_a(t+u)\,\mathrm{d}N_b(t)] = p_{ab}(u)\,\mathrm{d}u\,\mathrm{d}t. \tag{4}$$

The function that relates the events of type a with events of type b is called cumulant density $q_{ab}(u)$ and is defined by

$$q_{ab}(u) = p_{ab}(u) - p_a p_b \quad (u \neq 0 \text{ and } a, b = 1, 2). \tag{5}$$

The condition of strong mixing implies that $p_{ab}(u) \to p_a p_b$, or equivalently $q_{ab}(u) \to 0$ as $u \to \infty$.

Table 1
2×2 contingency table involving two point processes

		$dN_a(t+u)$		
		0	**1**	Total
$dN_b(t)$	**0**	1	$p_a du$	1
	1	$p_b dt$	$p_{ab}(u)dtdu$	$p_b dt$
	Total	1	$p_a du$	

We assume further that the point process has the first two moments finite. Then the cross-spectral density is the Fourier transform of $q_{ab}(u)$ defined by

$$f_{ab}(\lambda) = (2\pi)^{-1} \int_{-\infty}^{\infty} q_{ab}(u) e^{-i\lambda u}\, du, \quad -\infty < \lambda < \infty \ (a \neq b). \tag{6}$$

It is known from the Fourier analysis (see Brillinger, 1981) that by taking the inverse transform of (6) we have

$$q_{ab}(u) = \int_{-\infty}^{\infty} f_{ab}(\lambda) e^{i\lambda u}\, d\lambda \quad (a \neq b) \tag{7}$$

and therefore

$$p_{ab}(u) = p_a p_b + \int_{-\infty}^{\infty} f_{ab}(\lambda) e^{i\lambda u}\, d\lambda \quad (a \neq b). \tag{8}$$

Finally, we define the function $\hat{d}_a^{(T)}(\lambda)$ as the modified finite Fourier–Stieltjes transform of the increments $[dN_a(t) - p_a\, dt]$ on $(0, T]$, given by

$$\hat{d}_a^{(T)}(\lambda) = \int_0^T e^{-i\lambda t}[dN_a(t) - p_a\, dt]. \tag{9}$$

The definitions of the above functions both in the time and in the frequency domain are discussed in Brillinger (1975) and Rigas (1996).

2.3. The cross-product ratio

Table 1 presents a 2×2 contingency table which is proposed for measuring the association between the increments $dN_a(t+u)$ and $dN_b(t)$ of the bivariate point process. This table suggests the following quantity:

$$\theta_{ab}(u) = \frac{p_{ab}(u)\, du\, dt}{p_a\, du p_b\, dt} = \frac{p_{ab}(u)}{p_a p_b} \tag{10}$$

($a, b = 1, 2$ and $a \neq b$) as a measure of the degree of association at lag u between the events of type a with the events of type b. This measure is called CPR and corresponds to the OR for an ordinary 2×2 contingency table. In the case of independence it is $\theta_{ab}(u) = 1$. In the case of dependence CPR tends to 1 due to the strong mixing condition. Such an approach is described in Brillinger (1976).

3. Two methods of estimating the CPR

In this section, two methods of estimating the CPR are discussed: the first in the time domain and the second in the frequency domain.

3.1. Histogram-based estimate

Suppose that the process $\{N_1(t), N_2(t)\}$ is observed for $0 < t \leqslant T$ and the times of events of types 1 and 2 are recorded. Let the times of events of type 1 be $s_1 < s_2 < \cdots < s_{N_1(T)}$ and the times of events of type 2 be $t_1 < t_2 < \cdots < t_{N_2(T)}$. Then the estimate of second-order product density (see Cox and Lewis, 1972) is based on the statistic

$$J_{ab}^{T}(u) = \#\left\{(s_j, t_k) \text{ such that } u - \frac{h}{2} < s_j - t_k < u - \frac{h}{2} \text{ and } s_j \neq t_k\right\}, \tag{11}$$

where "#" counts the number of events of type a which fall in a cell of bin width h and midpoint u time units along from a type b event ($a, b = 1, 2$). The $J_{ab}^{T}(u)$ is a histogram-type statistic. The estimate of the second-order product density is now given by

$$\hat{p}_{ab}(u) = \frac{\sum_{i=-m}^{m} w_i J_{ab}^{T}(u - ih)}{hT}, \tag{12}$$

where w_i are weights such that $\sum_{i=-m}^{m} w_i = 1$. The use of weights in (12) improves the properties of the estimate. It can be proved that $\sqrt{\hat{p}_{ab}(u)}$ tends to Normal distribution with mean $\sqrt{p_{ab}(u)}$ and variance $(1/4hT)\sum_{i=-m}^{m} w_i^2$ (see Brillinger, 1976). An estimate of CPR can now be obtained by substituting the estimates of the first- and second-order product densities in (10) as follows:

$$\hat{\theta}_{ab}(u) = \frac{\hat{p}_{ab}(u)}{\hat{p}_a \hat{p}_b}, \tag{13}$$

where $\hat{p}_a = N_a(T)/T$ and $\hat{p}_b = N_b(T)/T$ $(a, b = 1, 2)$.

From the above results it follows that $\sqrt{\hat{\theta}_{ab}(u)}$ tends asymptotically to a Normal distribution with mean $\sqrt{\theta_{ab}(u)}$ and variance $\sum_{i=-m}^{m} w_i^2/(4hTp_a p_b)$. Thus a 95% point-wise approximate confidence interval for $\sqrt{\theta_{ab}(u)}$ is given by

$$\sqrt{\hat{\theta}_{ab}(u)} \pm 1.96 \sqrt{\sum w_i^2} \Big/ \left(2\sqrt{hT\hat{p}_a \hat{p}_b}\right). \tag{14}$$

Fig. 1 shows the square root of the estimated CPR. The confidence intervals are obtained from (14) using the weights $(0.25, 0.5, 0.25)$ and $h = 1$ ms. It is clear that the system is more likely to fire in the interval between 11 and 25 ms. This is an assessment of the muscle spindle's normal function. Deviations from this behaviour may indicate the existence of possible risk factors for abnormal firing of the system.

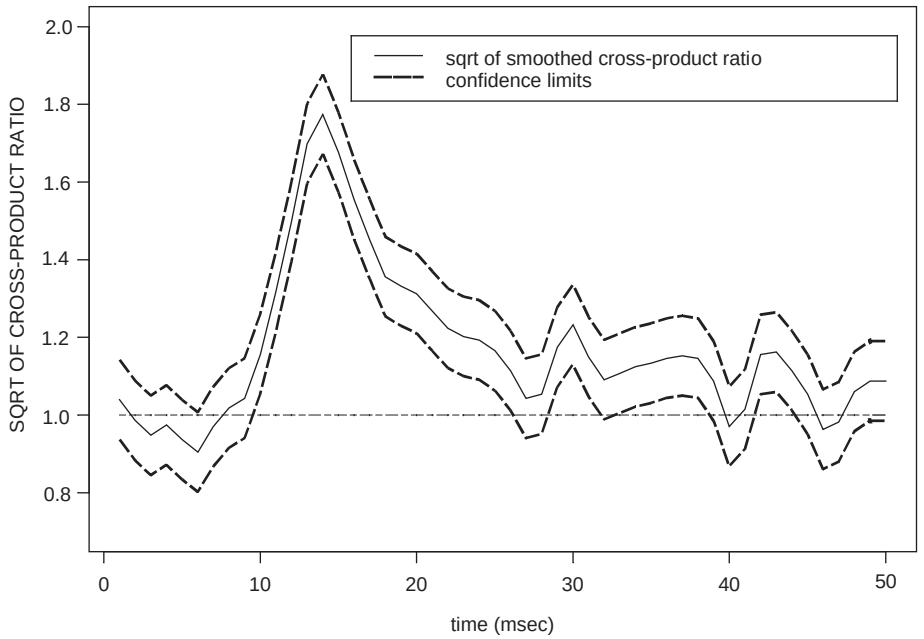

Fig. 1. Square root of the histogram-based estimate of the CPR for the neurophysiological data set. The solid line in the middle corresponds to the smoothed estimate and the dotted lines above and below it correspond to the 95% point-wise approximate confidence limits.

3.2. Periodogram-based estimate

An estimate of the CPR in the frequency domain is based on the modified periodogram statistic, which is defined by

$$\hat{I}_{ab}^{(T)}(\lambda) = \frac{1}{2\pi T}\, \hat{d}_a^{(T)}(\lambda)\overline{\hat{d}_b^{(T)}(\lambda)} \quad (a \neq b). \tag{15}$$

By $\overline{\hat{d}_b^{(T)}(\lambda)}$ we denote the conjugate function of $\hat{d}_b^{(T)}(\lambda)$. More details about the modified periodogram statistic are given in Rigas (1996). An estimate for the second-order product density can be obtained from (9) as follows:

$$\hat{p}_{ab}(u) = \hat{p}_a \hat{p}_b + \frac{2\pi}{T} \sum_j W^{(T)}(\lambda_j)\, \hat{I}_{ab}^{(T)}(\lambda_j)\, \mathrm{e}^{\mathrm{i}\lambda_j u}, \tag{16}$$

where $\lambda_j = 2\pi j/T$, $j = \pm 1, \ldots, \pm(T-1)/2$. By $W^{(T)}(\lambda) = W(b_T\lambda)$ we denote a convergence factor or taper (see Brillinger, 1981) and b_T is the bandwidth of the taper. We choose b_T such that, $b_T T \to \infty$ as $T \to \infty$ and $b_T \to 0$. The convergence factor in (16) reduces the variance of the estimate and improves its characteristics. The estimate

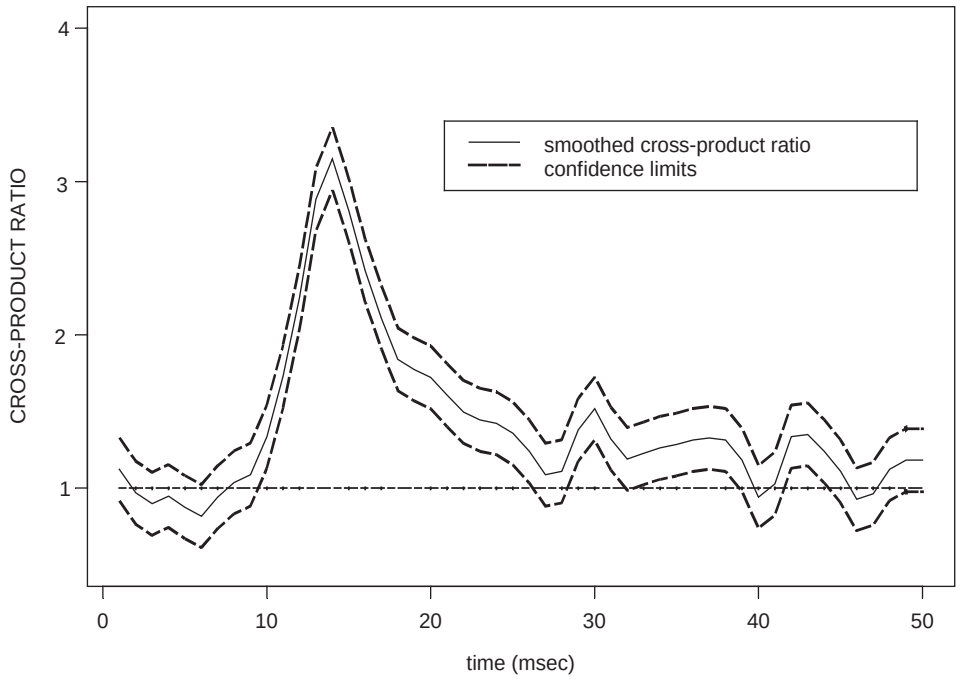

Fig. 2. Periodogram-based estimate of the CPR for the neurophysiological data set. The solid line in the middle corresponds to the smoothed estimate and the dotted lines above and below it correspond to the 95% point-wise approximate confidence limits.

$\hat{p}_{ab}(u)$ is asymptotically Normal with mean $p_{ab}(u)$ and variance

$$Var[\hat{p}_{ab}(u)] \cong \frac{(2\pi)^{-1} p_a p_b}{b_T T} \int W^2(\lambda)\, \mathrm{d}\lambda. \tag{17}$$

The asymptotic properties of $\hat{p}_{ab}(u)$ derive from the work of Rigas (1996), while a more detailed proof is given in Rigas and Tsitsis (1996). This suggests that the estimate $\hat{\theta}_{ab}(u)$, of the CPR is asymptotically Normal with mean $\theta_{ab}(u)$ and variance

$$Var[\hat{\theta}_{ab}(u)] \cong \frac{(2\pi)^{-1}}{b_T T p_a p_b} \int W^2(\lambda)\, \mathrm{d}\lambda. \tag{18}$$

Thus, a 95% point-wise approximate confidence interval for $\theta_{ab}(u)$ is given by

$$\hat{\theta}_{ab}(u) \pm 1.96 \sqrt{\frac{(2\pi)^{-1}}{b_T T \hat{p}_a \hat{p}_b} \int W^2(\lambda)\, \mathrm{d}\lambda}. \tag{19}$$

Fig. 2 shows the estimates of the CPR for the neurophysiological data set. The estimate is obtained by using (16). A Tukey convergence factor was used with $b_T = 0.31$. In this case

$\int W^2(\lambda)\,\mathrm{d}\lambda = 0.75$. The characteristics of the estimated CPR are almost identical with the previous estimate involving the square root of CPR. Therefore, both methods show that the system is likely to fire in the interval between 11 and 25 ms.

The computation of the estimates of the CPR both in the time and in the frequency domain was carried out on the statistical package S-PLUS. More details about the computations involved are given in Section 5 and in the appendix.

4. Formulation of a logistic regression model

In this section, a logistic regression model is formulated, in order to describe the behaviour of the muscle spindle, when it is affected by the presence of a gamma motoneurone. The effects of the imposed stimulus are transmitted to the spinal cord by the axon of a sensory nerve closely associated with the muscle spindle. The axon of the sensory nerve fires when the membrane's potential exceeds a critical level called threshold. The membrane's potential is influenced both by external and internal processes. The model which is used for the description of this system is discussed in Brillinger (1988) and involves three parameters: the threshold, the recovery and the summation function.

Let Y_t denote the firing process of the sensory nerve which is associated with the muscle spindle. By choosing the time sampling interval h, the observations of the output can be written as follows:

$$y_t = \begin{cases} 1 & \text{when a spike occurs in } (t, t+h], \\ 0 & \text{otherwise,} \end{cases}$$

where $t = h, \ldots, Nh$ and $T = Nh$ is the time interval in which the time series is observed. In our case we choose the sampling interval $h = 1$ ms. The input X_t imposed by the gamma motoneurone on the system consists of the observations x_t defined similarly.

Let θ_t denote the threshold level at the trigger zone at time t, given by: $\theta_t = \theta_t^* + \varepsilon_t$, where ε_t is the noise process, which includes contributions of unmeasured terms that influence the firing of the system. There is experimental evidence and theoretical verification that ε_t follows a normal distribution (see Holden, 1976). θ_t^* is a function of t, which represents the form of threshold at time t. We assume in this case that $\theta_t^* = \theta_0$, where θ_0 is an unknown constant.

The function representing the external processes that influence the membrane's potential at the trigger zone is the summation function. This function represents the effect of the input on the output of the system at any given time t and is described by a set of coefficients $\{a_u\}$. Thus, the membrane's potential at any given time t, due to external stimuli is defined by

$$SF_t = \sum_{u \leqslant t} a_u x_{t-u},$$

where x_{t-u} is the observation of the input at time $t - u$.

The internal processes are responsible for the spontaneous firing of the system. This is an ability of the system to produce a series of nerve pulses on its own by increasing

the membrane's resting potential to the level of threshold. The recovery function can be described by a polynomial function of order k which is given by

$$V_t = \sum_{i=1}^{k} \theta_i \gamma_t^i,$$

where $\{\theta_i\}$ are the coefficients of the recovery function and γ_t is the time elapsed since the system last fired.

The logistic regression model that describes the behaviour of the muscle spindle under the influence of a gamma motoneurone at any given time t, incorporates the threshold, the summation, the recovery function and can be expressed in the form

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \sum_{u \leqslant t} a_u x_{t-u} + \sum_{i=1}^{k} \theta_i \gamma_t^i - \theta_0, \tag{20}$$

where π_t is the probability of an output spike to occur, i.e. $\pi_t = \Pr(Y_t = 1)$. The unknown parameters are the summation function coefficients $\{a_u\}$, the recovery function coefficients $\{\theta_i\}$ and the constant threshold θ_0. The minus before the constant level of the threshold indicates that the strength of the external and the internal processes must exceed the level of threshold in order to get an output spike. Such a model is discussed in Brillinger (1988), but instead of the logistic link function he uses the probit link function. It is known by Cox and Snell (1989) that these two link functions produce similar results. The only difference we notice is the scaling in the graphs of the estimated functions.

The maximum-likelihood estimates for the unknown parameters included in the model can be obtained by one of the most commonly used statistical packages, like GENSTAT, S-PLUS or SPSS. Table 2 gives the output of GENSTAT, Fig. 3 depicts the estimated functions of the fitted model and Fig. 4 presents the OR of the estimated coefficients $\{a_u\}$, together with the 95% pointwise confidence intervals. The estimated coefficients of the summation function are positive and far from zero in the interval 11–30 ms, which implies that the covariates $x_{t-11}, \ldots, x_{t-30}$ can be considered as risk factors because they significantly increase the OR of an output spike. The overall behaviour of the summation function is excitatory, which means that the input spikes from a gamma motoneurone accelerate the firing of the system. The recovery function is described by a second-order polynomial since $\theta_i; i \geqslant 3$ are not statistically significant. It is clear that the recovery function does not cross the estimated threshold level, which indicates that the presence of the gamma motoneurone prohibits the spontaneous firing of the system. However, the contribution of the recovery function to the firing of the system is not negligible, as it is indicated in Table 3 .

Since the logistic regression model is based on a binomial distribution and involves binary data, we can rely on the difference of deviance between successive models, in order to decide which of them gives the best fit (see McCullagh and Nelder, 1989). The difference of deviance is asymptotically distributed as a chi-square (χ_p^2) with p degrees of freedom, where p is the difference in the number of parameters between two successive models. The possible models, the values of their deviance and the degrees of freedom are shown in Table 3. The change in deviance on adding a second-order recovery function to the model

Table 2
Output of GENSTAT for the logistic regression model

	Estimate	s.e.	t (*)	t pr.	Antilog of estimate
Constant	−7.444	0.229	−32.53	< 0.001	0.000585
th[1]	0.225	0.0161	13.98	< 0.001	1.252
th[2]	−0.004	0.000353	−11.32	< 0.001	0.996
A[1]	0.366	0.173	2.11	0.035	1.442
A[2]	−0.11	0.203	−0.54	0.587	0.8956
A[3]	−0.187	0.209	−0.9	0.369	0.8292
A[4]	−0.003	0.194	−0.02	0.986	0.9967
A[5]	−0.104	0.201	−0.52	0.606	0.9013
A[6]	−0.477	0.228	−2.09	0.036	0.6208
A[7]	0.048	0.19	0.25	0.799	1.05
A[8]	−0.008	0.192	−0.04	0.966	0.9918
A[9]	−0.025	0.194	−0.13	0.898	0.9754
A[10]	0.312	0.175	1.78	0.075	1.366
A[11]	0.83	0.15	5.53	< 0.001	2.293
A[12]	1.011	0.144	7.03	< 0.001	2.75
A[13]	1.673	0.127	13.2	< 0.001	5.328
A[14]	1.938	0.126	15.44	< 0.001	6.946
A[15]	1.826	0.135	13.55	< 0.001	6.21
A[16]	1.595	0.144	11.07	< 0.001	4.927
A[17]	1.732	0.145	11.94	< 0.001	5.652
A[18]	1.115	0.169	6.61	< 0.001	3.05
A[19]	1.404	0.155	9.04	< 0.001	4.072
A[20]	1.162	0.164	7.08	< 0.001	3.198
A[21]	1.173	0.162	7.23	< 0.001	3.232
A[22]	0.978	0.17	5.76	< 0.001	2.659
A[23]	0.818	0.173	4.74	< 0.001	2.266
A[24]	0.89	0.168	5.3	< 0.001	2.435
A[25]	0.532	0.18	2.96	0.003	1.702
A[26]	0.681	0.171	3.99	< 0.001	1.975
A[27]	0.146	0.2	0.73	0.465	1.158
A[28]	0.344	0.182	1.89	0.059	1.411
A[29]	0.383	0.177	2.16	0.031	1.467
A[30]	0.917	0.152	6.02	< 0.001	2.502
A[31]	0.144	0.188	0.77	0.443	1.155
A[32]	0.316	0.177	1.78	0.074	1.372
A[33]	0.193	0.182	1.06	0.289	1.213
A[34]	0.315	0.173	1.82	0.068	1.371
A[35]	0.232	0.177	1.31	0.19	1.262
A[36]	0.306	0.173	1.77	0.077	1.358
A[37]	0.36	0.169	2.13	0.033	1.433
A[38]	0.203	0.176	1.15	0.249	1.225
A[39]	0.4	0.168	2.38	0.017	1.492
A[40]	−0.388	0.217	−1.78	0.074	0.6783
A[41]	−0.11	0.194	−0.57	0.571	0.8958
A[42]	0.472	0.161	2.93	0.003	1.603
A[43]	0.323	0.167	1.93	0.054	1.381
A[44]	0.061	0.182	0.34	0.737	1.063
A[45]	0.316	0.172	1.84	0.066	1.372
A[46]	−0.434	0.226	−1.92	0.055	0.6482
A[47]	0.044	0.186	0.24	0.814	1.045
A[48]	0.009	0.185	0.05	0.963	1.009
A[49]	0.252	0.172	1.47	0.142	1.287
A[50]	0.042	0.185	0.23	0.822	1.043

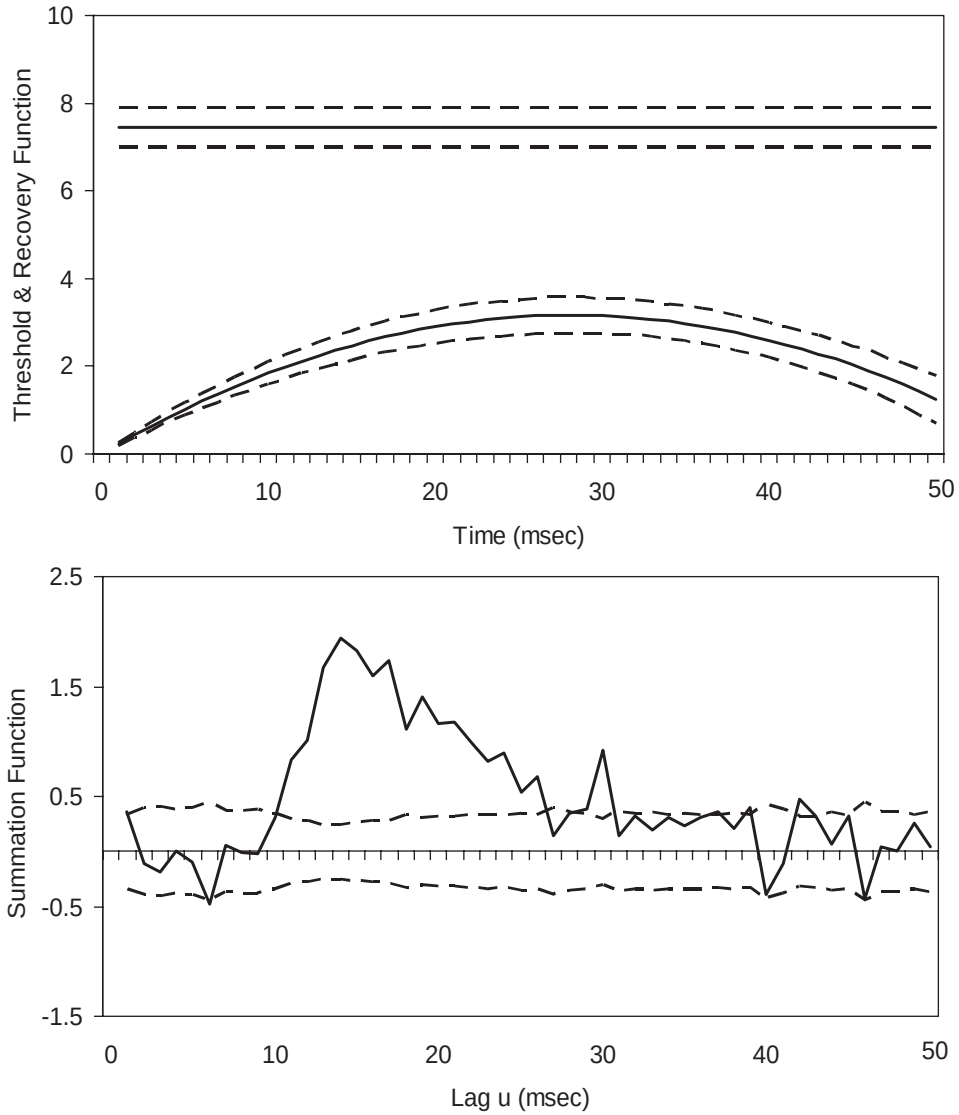

Fig. 3. Estimates of the threshold, the recovery and the summation functions, obtained by the logistic regression model given by (20). The dotted lines correspond to the 95% confidence limits. The recovery function does not cross the level of threshold. Therefore, the system does not fire spontaneously. The summation function is outside the confidence interval between 11–25 ms.

that includes only a constant term alone is $4699 - 4634 = 65$ on 2 d.f. Since the upper 5% point of the χ^2 -distribution on 2 d.f. is 5.99, this is very significant at the 5% level. If the summation function is added to a model that includes only a constant term, the deviance is reduced by 687 on 50 d.f. which is highly significant. However, when both the recovery

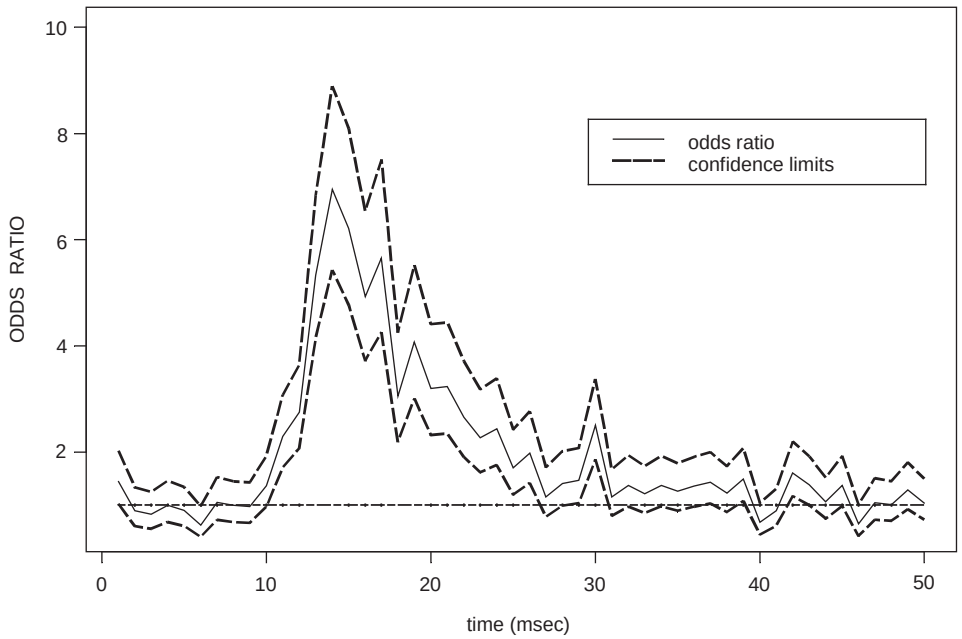

Fig. 4. The estimate of the OR by the logistic regression model. The solid line in the middle corresponds to the OR of the estimated coefficients $\{\hat{a}_u\}$. The dotted lines above and below the OR correspond to the 95% point-wise confidence limits.

Table 3
Comparison of the deviance between successive logistic regression models, for the assessment of the goodness of fit

Model	Deviance	d.f.
Null	21995	15866
$\eta_t = -\theta_0$	4699	15865
$\eta_t = \theta_1\gamma_t + \theta_2\gamma_t^2 - \theta_0$	4634	15863
$\eta_t = \sum_{u\leqslant t} a_u x_{t-u} - \theta_0$	4012	15815
$\eta_t = \theta_1\gamma_t + \theta_2\gamma_t^2 + \sum_{u\leqslant t} a_u x_{t-u} - \theta_0$	3708	15813

and the summation function are included in the model the change in deviance becomes $4699 - 3708 = 991$ on 52 d.f. which proves that the best-fitted model should include both the recovery and the summation function.

The number of the explanatory variables x_{t-u} can be reduced by choosing a smaller interval u for studying the behaviour of the summation function. In fact, fitting the model with 35, 40, 45 and 50 explanatory variables x_{t-u}, causes successive changes in deviance by 16, 15, and 7 units, respectively. These changes imply that some of the explanatory variables between x_{t-35} and x_{t-45} are significant risk factors and cannot be eliminated from the

model. The explanatory variables $x_{t-46}, \ldots, x_{t-50}$ can be excluded from the model since the upper 5% point of the χ^2-distribution on 5 d.f. is 11.07.

The validity of the specific model can be checked by carrying out a graphical comparison between the theoretical and the empirical probability (see Brillinger, 1988), as follows: selected values of the linear predictor η are obtained, by dividing the range of the estimated linear predictor of the model $\hat{\eta}_t$, into a number of small intervals, usually of equal width. The center value of each interval is considered as a selected value of the linear predictor. The theoretical probability is given by

$$\pi(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

The estimated probability of firing $\hat{\pi}(\eta)$ for a given linear predictor η, can be defined as the ratio of the number of firings to the total number of possible firings into the small interval $(\eta - h, \eta + h)$

$$\hat{\pi}(\eta) = \frac{\sum_t \{Y_t = 1, \eta - h < \hat{\eta}_t < \eta + h\}}{\sum_t \{t, \eta - h < \hat{\eta}_t < \eta + h\}}.$$

Fig. 5depicts the goodness-of-fit plot for the logistic regression model given by (20). The validity of this model depends on the closeness of the estimated probability to the theoretical probability. There seems to be a reasonable fit for this model.

5. Discussion on the use of different statistical packages

S-PLUS see (http://www.insightful.com) is a very powerful statistical package, which has been used for the computation of the CPR. S-PLUS is programmable, because it is based on S-language. This feature offers great flexibility, since it allows the development of new functions which can then be incorporated into any application. R-language is an open source project see (http://www.r-project.org) which is also based on S-language. An additional characteristic is that S is an object-oriented language. Every data set or graph like vectors, matrices, lists, data frames, etc. can be considered as an S-object. In S-language it is preferable to work with S-objects rather than with individual elements of a vector or a matrix (see Venables and Ripley , 2002). Specifically a loop with upwards of 50000 iterations is not very workable (see Insightful Corporation, 2001). An attempt to compute $Jab\ [u]$ in Program 1 (see Appendix) using for commands and a counter, such as

```
for (i in 1 : NaT) {
    for (j in 1 : NbT) {
        for (u in 1 : n) { ...
            Jab [u] < − Jab [u] + 1
            ...} } }
```

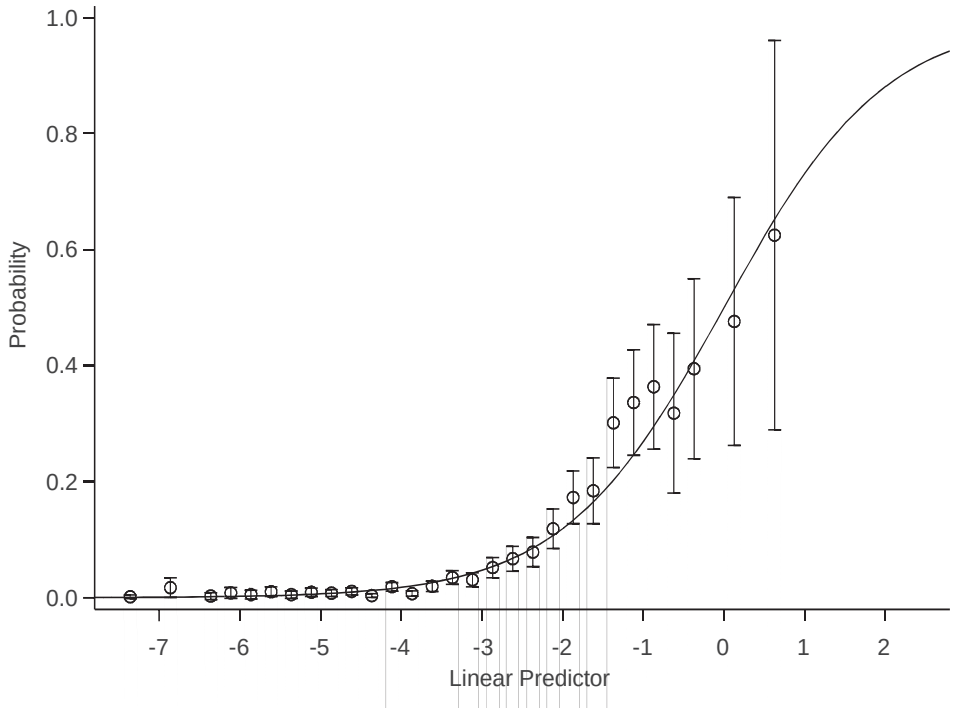

Fig. 5. Goodness of fit plot for the validation of the logistic regression model given by (20). The empirical probability is depicted as a small circle, while the vertical bars provide the estimated standard error limits around its value. The theoretical probability is the smooth curve that corresponds to the logistic link function.

results in a very slow program. This would be the usual approach with a traditional programming language like FORTRAN. This set of commands can be replaced by the construction of the matrix dif and the command:

$$Jab\ [u] <- sum\ (abs\ (dif - u) < (h/2))$$

The expression $abs\ (dif - u) < (h/2)$ gives 0 when it is false or 1 if it is true. Therefore $sum\ (abs\ (dif - u) < (h/2))$ counts the number of elements in the matrix dif which satisfy the property $abs\ (dif - u) < (h/2))$. This is an operation on the object-matrix $Jab\ [u]$. In practice this operation gives very fast results using S-PLUS. However, S-PLUS is not as easy to learn as other statistical software, but the combination of high-quality graphics and modern statistical procedures makes it very attractive for the research worker. The computations involved in the estimation of the CPR could be performed equally well on GENSTAT (see Baird et al., 2002), because it provides all the directives required for the computations, both in the time and in the frequency domain. Furthermore, the programs developed by the users for complicated tasks can be formed as procedures and be included in a GENSTAT library for later use. In fact GENSTAT is a powerful, extensible and very

well-designed language for computing (see Payne, 2002). It provides good facilities for generalized linear models and the major user group is statisticians working in biological research. The graphics produced are not of the highest quality but they are acceptable. The popular statistical package SPSS is menu-driven and cannot be used for the estimation of the CPR, because the main menus do not include built-in tools for Fourier transforms.

The estimates of the unknown parameters in the logistic regression model were obtained by GENSTAT. These results could also be obtained by using either S-PLUS or SPSS. A friendly menu-driven user interface is available in all packages and allows the use of logistic regression without having to learn the command language. The commands used by the menus are retained in the session log of GENSTAT or in the syntax editor of SPSS so that they are available for later use. A guide to the various statistical techniques available with SPSS for regression models, including binary logistic regression (see SPSS 12.0, 2004) explains how to obtain the appropriate statistical analyses with the dialog box interface.

GENSTAT, S-PLUS or SPSS are general purpose statistical software that can perform a large variety of tasks. A special software that can be used mainly for logistic regression is LOGXACT, which has the advantage of using exact likelihood analysis and applying Monte Carlo techniques. Both methods can be employed in the case that the asymptotic maximum-likelihood method fails to converge (see LogXact-5 for Windows, 2002). An example is given by Kotti and Rigas (2005) where all the other packages fail to calculate the estimates of the logistic regression except the LOGXACT package.

Finally, the statistical package GLIM can be used for the analysis of the logistic regression models, because it has been designed for analysing generalised linear models. However GLIM has two disadvantages: (a) It uses a non-user-friendly interface, which requires written directives given from the user and (b) It can handle up to 30 covariates. Details about the GLIM package are given in Christensen (1997).

An attempt has been made in connection with MATLAB to improve the user-friendly ability of GLIM, through GLMLAB. This is a set of m-files for using MATLAB for analysing generalised linear models. It is developed by Peter Dunn originally to replace GLIM in a small way at University of Southern Queensland, Australia. Nowadays, it has grown and contains quite a lot of the features found in GLIM. It enables models such as multiple regression, probit models, logistic regression and log-linear models (among others) to be fitted. GLMLAB is very useful especially for people who work with similar problems in the field of engineering and can be obtained from the MathWorks user-contributed files see (http://www.mathworks.com/support/ftp/statv4.shtml), (http://www.sci.usq.edu.au/staff/dunn/glmlab/glmlab.html).

6. Summary and discussion of the results

In this work, two alternative methods were presented for the identification of a complex neurophysiological system. A data set which describes the response of the muscle spindle to the effect of a gamma motoneurone is used for the estimation of the parameters involved in both cases. The first method is a non-parametric approach and the estimate of the CPR, which corresponds to the OR between two binary time series, was obtained either as a histogram-based estimate in the time domain or as a periodogram-based estimate in the

frequency domain. The second method is a parametric approach and the estimates for the parameters of the logistic regression model were obtained by using the maximum-likelihood function.

The results obtained by using the non-parametric methods are restricted because they provide information only about the relation between the input and the output of the system. The advantage of using the logistic regression approach is the flexibility of the proposed model, since it allows the incorporation of physiologically meaningful parameters which are responsible for the firing of the system: the threshold, the recovery and the summation function. The parametric approach therefore provides extra information and reveals more characteristics about the system we examine. The threshold and the recovery function taken together seem to describe the time course of the intrinsic membrane properties, whereas the summation function reveals a direct relationship between the input and the output of the system. Furthermore, alternative forms for the threshold and the summation function can be used. An exponentially threshold model can be used instead of the constant threshold, which seems to be more realistic from a physiological point of view (see Kotti and Rigas, 2003b). In the present model, the summation function takes into consideration all the previous input spikes. A summation function that takes into account only the input spikes that have occurred after the time of the previous output spike can also be considered. In this case, a carry over effect function can be incorporated into the model that describes the effect on the membrane's potential of the input spikes that have occurred before the time of the previous output spike (see Kotti and Rigas, 2003a).

Furthermore, the estimated coefficient $\hat{a}_u$ of the logistic regression model can be used for the direct estimation of its OR, $\psi = \exp(\hat{a}_u)$, while the $100(1-\gamma)\%$ confidence interval for ψ is given by the expression

$$\exp[\hat{a}_u \pm z_{\gamma/2} \times SE(\hat{a}_u)],$$

where $z_{\gamma/2}$ denotes the standard normal deviate with a tail area of $\gamma/2$. Thus, the estimated coefficients allow the identification of the potential risk factors, since values of $\{\hat{a}_u\}$ greater than zero increase the odds of success, while values of $\{\hat{a}_u\}$ less than zero decrease the odds of success. This corresponds to estimated OR $\psi < 1$ and $\psi > 1$, respectively. Fig. 6 depicts the OR for the estimates of $\{\hat{a}_u\}$ obtained by the logistic regression method, plotted together with the estimate of the unsmoothed CPR obtained by the periodogram-based method. Although the true OR for this data set is not known, the comparison between the two estimates shows that the OR gives a stronger indication of the system's excitatory behaviour than the CPR. The higher values of the OR compared with the values of the CPR emphasize on the importance of the internal processes of the system, which are described by the extra parameters included in the logistic regression model.

However, there are cases where the method of logistic regression fails completely or produces very poor results. This problem is caused by certain structures in the data, which occur when we deal with data sets that are small or data sets that are large but sparse. In this case the maximum-likelihood estimation fails because some of the covariates are considered as perfect predictors and their estimates may then need to be infinite to maximize the likelihood function. This problem can be identified by extreme estimates and standard errors which are noticeable and indicate a problem in the maximum-likelihood estimation. The problem in this case can be overcome by applying the alternative non-parametric approach,

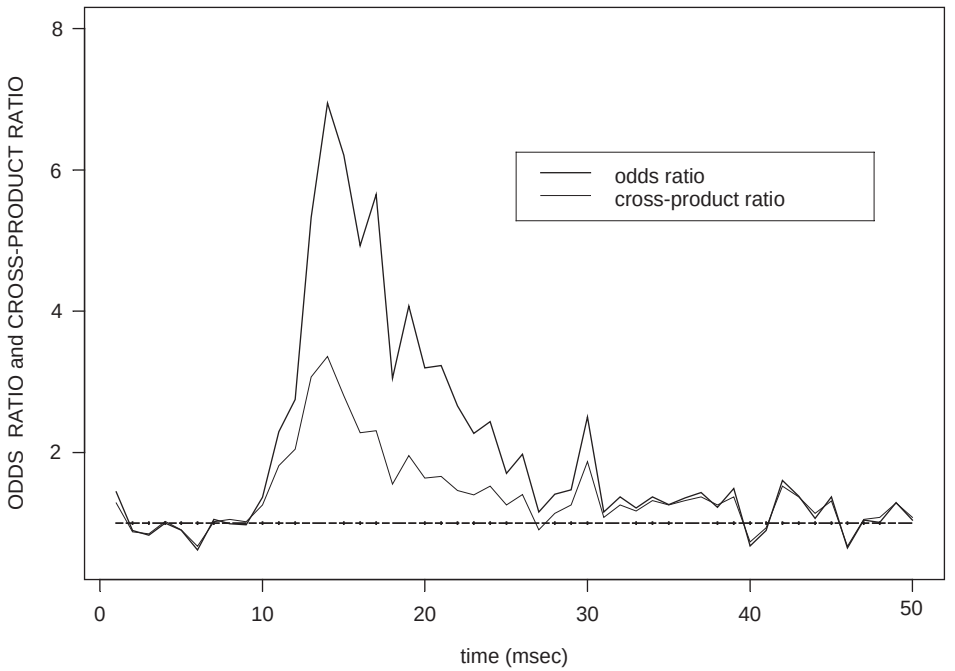

Fig. 6. Comparison between the OR of the estimates $\{\hat{a}_u\}$ derived from the logistic regression model and the estimate of CPR. The unsmoothed periodogram-based estimate for CPR is used.

by performing exact logistic regression (see Kotti and Rigas, 2005), or by an alternative method that is based on penalized maximum likelihood (see Heinze and Schemper, 2002).

The logistic regression approach is a highly flexible method that allows the incorporation of many biologically meaningful parameters, which comply with the physiological phenomena and explain most of the underlying variability. The results obtained are then supposed to be less biased and more precise. A goodness of fit test based on the successive changes in deviance can be used for the selection of the best fitted model. Once the unknown parameters have been estimated, derived quantities such as the OR can be estimated as well. The implementation of the logistic regression is very simple and it is available on most of the statistical packages. When the maximum likelihood estimation, which is employed in the logistic regression, fails to converge it is worth considering the alternative non-parametric approach.

Acknowledgements

The authors are very grateful to an Associate Editor and three anonymous referees for their helpful and constructive comments which led to a substantial improvement of the paper. Thanks are also expressed to Professors G.P. Moore and J.R. Rosenberg for their

kindness to provide the neurophysiological data set. The source code of the programs listed in the Appendix, is available at http://utopia.duth.gr/∼rigas/ and the electronic version of the paper at http://www.sciencedirect.com.

Appendix A. List of programs used in the analysis of the neurophysiological example

Program 1 in S-PLUS—computation of the square root of cross-product ratio in the time domain

```
cpr.td <- function(a, b, h=1, Time, n=50)
{
    NaT <- length(a[a != 0])
    NbT <- length(b[b != 0])

    Pa <- NaT/Time ## estimation of Pa
    Pb <- NbT/Time ## estimation of Pb

    dif <- matrix(0, nrow=NbT, ncol=NaT)
    for (i in 1:NaT)   {
        for (j in 1:NbT)   {
            if (a[i] != b[j])   dif[j,i] <- a[i]-b[j]    }}

    Jab <- vector("numeric", length=n)
    for (u in 1:n) {    Jab[u] <- sum(abs(dif-u)<(h/2))  }

    Pab <- Jab/(h*Time)
    cprab <- Pab/(Pa*Pb)

    cprab.smo <- vector(mode="numeric", length=n)
    for (i in 2:(n-1))
    {   cprab.smo[i] <- sum( 0.25*cprab[i-1]+0.5*cprab[i]+0.25*cprab[i+1])  }

    cprab.smo[1] <-0.5*cprab[1]+0.5*cprab[2]
    cprab.smo[n] <-0.5*cprab[n]+0.5*cprab[n-1]

    one<- vector("numeric", length=n)
    one[1:n] <- 1
    sum.wisq <- sum(c(0.25, 0.5, 0.25)^2)

    plot(1:n, sqrt(cprab.smo[1:n]), type="l", lwd=3, xlab="time (msec)", ylab="SQRT OF
    CROSS-PRODUCT RATIO", ylim=c(0.7, 2))
    lines(1:n, one, type="l", lwd=1, lty=6)

    climup <- sqrt(cprab.smo)+(1.96*sqrt(sum.wisq))/(2*sqrt(h*Time*Pa*Pb))
    par(new=T, xaxs="d", yaxs="d")
    plot(1:n, climup, type="l",axes=F,  lwd=3, lty=4, xlab="", ylab="")

    climlo <- sqrt(cprab.smo)-(1.96*sqrt(sum.wisq))/(2*sqrt(h*Time*Pa*Pb))
    par(new=T, xaxs="d", yaxs="d")
    plot(1:n, climlo, type="l", axes=F, lwd=3, lty=4, xlab="", ylab="")

    legend(locator(1), c("sqrt of smoothed cross-product ratio", "confidence limits"),
    lwd=c(3, 3), col=c(1, 1), lty=c(1, 4))

}
```

Program 2 in S-PLUS—computation of cross-product ratio in the frequency domain

```
cpr.fd <- function(dNa, dNb, Time, n=50)
{
    NaT <- length(dNa[dNa != 0])
    NbT <- length(dNb[dNb != 0])

    pa <- NaT/Time     ## estimation of pa
    pb <- NbT/Time     ## estimation of pb

    daT <- fft(dNa-pa)     ##  pa=mean(dNa)
    dbT <- fft(dNb-pb)     ##  pb=mean(dNb)

    if (Time%%2==0) QT <- Time/2  else  QT <- (Time+1)/2

    Iab <- daT*Conj(dbT)/(2*pi*Time)

    if (Time%%2==0)    Iab <- c( Iab[ (QT+2):Time ], Iab[2:QT] )
    else  Iab <- c( Iab[ (QT+1):Time ] , Iab[2:QT] )

    w <- 0.5*(1+cos(  ( ( -(QT-1):(QT-1) )*pi ) / (QT-1) ))
    ## Tukey-Hanning convergence factor
    w <- w[-QT]

    bT <- Time/(2*pi*QT)              ## the bandwidth of the convergence factor

    freq <- c( 2*pi*((-(QT-1)):(-1))/Time , 2*pi*(1:(QT-1))/Time  )
    ## frequencies without 0 and pi

    qab <- vector(mode="complex", length=n)
    for (u in 1:n) {   qab[u] <- sum(w*Iab*(exp((0+1i)*freq*u)))    }

    pab <- Re(((2*pi)/Time)*qab)+pa*pb
    cprab <- pab/(pa*pb)

    varcpr <- (3/4)/(2*pi*bT*Time*pa*pb)

    x <- vector(mode="numeric", length=n)
    x[1:n] <- 1

    y<- vector(mode="numeric", length=n)
    y[1:n]<- 1.96*sqrt(varcpr)

    plot(1:n, cprab[1:n], type="l",ylim=c(0.5,3.5), xlab="time (msec)", ylab="CROSS-
    PRODUCT RATIO")
    lines(1:n, x[1:n], type="l", lwd=2, lty=6)

    climup <- cprab+y
    par(new=T, xaxs="d", yaxs="d")
    plot(1:n, climup, type="l",axes=F,  lwd=3, lty=4, xlab="", ylab="")

    climlo <- cprab-y
    par(new=T, xaxs="d", yaxs="d")
    plot(1:n, climlo, type="l", axes=F, lwd=3, lty=4, xlab="", ylab="")

    legend(locator(1), c("smoothed cross-product ratio", "confidence limits"),
    lwd=c(1, 3), lty=c(1, 4))

}
```

Program 3 in GENSTAT—computation of the parameters of the logistic model

```
VARIATE [NVAL=538] y
OPEN NAME='y.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] y
CLOSE CHANNEL=2

CALCULATE Y=EXPAND(y;15866)

VARIATE [NVAL=15866] A[1...50]

OPEN NAME='a1-10.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] A[1...10]
CLOSE CHANNEL=2

OPEN NAME='a11-20.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] A[11...20]
CLOSE CHANNEL=2

OPEN NAME='a21-30.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] A[21...30]
CLOSE CHANNEL=2

OPEN NAME='a31-40.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] A[31...40]
CLOSE CHANNEL=2

OPEN NAME='a41-50.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] A[41...50]
CLOSE CHANNEL=2

VARIATE [NVAL=15866] g[1]
OPEN NAME='c.dat'; CHANNEL=2; FILETYPE=input
READ [CHANNEL=2; END=*] g[1]
CLOSE CHANNEL=2

CALCULATE th[1]=(g[1].le.50)*g[1]
CALCULATE th[2]=th[1]**2

MODEL [DIST=BINOMIAL; LINK=logit] Y; NBINOMIAL=1
FIT  [PRINT=model,deviance,estimates;CONSTANT=esti; TPROB=yes]  th[1,2], A[1...50]
```

References

Baird, D.B., Harding, S.A., Lane, P.W., Murray, D.A., Payne, R.W., Soutar, D.M., 2002. Introduction Genstat for Windows. 6th Edition. VSN International, Oxford.

Brillinger, D.R., 1975. The identification of point process systems. Ann. Probab. 3, 909–929.

Brillinger, D.R., 1976. Measuring the association of point processes: a case history. Ann. Math. Monthly 86, 16–22.

Brillinger, D.R., 1981. Time Series: Data Analysis and Theory,. Expanded Edition. Holden-Day, San Francisco.

Brillinger, D.R., 1988. Maximum likelihood analysis of spike trains of interacting nerve cells. Biol. Cybern. 59, 189–200.

Christensen, R., 1997. Log-Linear Models and Logistic Regression. Springer, New York.

Cox, D.R., Isham, V., 1980. Point Processes. Chapman & Hall, London.

Cox, D.R., Lewis, P.A.W., 1972. Multivariate point processes. Proceedings of the Sixth Berkeley Symposium, Math. Statist. Prob. 2, 401–448.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data. Chapman & Hall/CRC, New York.

Daley, D.J., Vere-Jones, D., 1988. An Introduction to the Theory of Point Processes. Springer, Berlin.

Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Statist. Med. 21, 2409–2419.

Holden, A.V., 1976. Models for the Stochastic Activity of Neurons. Springer, Berlin.
http://www.insightful.com
http://www.mathworks.com/support/ftp/statv4.shtml
http://www.r-project.org
http://www.sci.usq.edu.au/staff/dunn/glmlab/glmlab.html
Insightful Corporation, 2001. S-PLUS 6: Programmer's Guide for Windows, Washington.
Johnson, R.E., Johnson, N., 1980. Survival Models and Data Analysis. Wiley, New York.
Kotti, V.K., Rigas, A.G., 2003a. Identification of a complex neurophysiological system using the maximum likelihood approach. J. Biol. Systems 11 (2), 189–204.
Kotti, V.K., Rigas, A.G., 2003b. A nonlinear stochastic model used for the identification of a biological system. In: Capasso,V. (Ed.), Mathematical Modeling & Computing in Biology and Medicine. The Miriam Project Series Escapulario Co., Italy, pp. 587–592.
Kotti, V.K., Rigas, A.G. Logistic regression methods and their implementation. In: Edler, L., Kitsos, C.P. (Eds.), Quantitative Methods for Cancer and Human Risk Assessment. Wiley, New York, 2005.
LogXact-5 for Windows, 2002. User Manual. Cytel Software Corporation, Cambridge, MA.
Marmarelis, P.Z., Marmarelis, V.Z., 1978. Analysis of Physiological Systems. Plenum Press, New York.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models,. Chapman & Hall, London.
Payne, R.W. (Ed.), 2002. The Guide to GenStat Release 6.1, Part 1: Syntax and Data Management. Part 2: Statistics VSN International, Oxford.
Rigas, A.G., 1996. Spectral analysis of a stationary bivariate point process with applications to neurophysiological problems. J. Time Ser. Anal. 17 (2), 171–186.
Rigas, A.G., Liatsis, P., 2000. Identification of a neuroelectric system involving a single input and a single output. Signal Process. 80, 1883–1894.
Rigas, A.G., Tsitsis, D.S., 1996. Spectral analysis techniques of stationary point processes: extensions and applications to neurophysiological problems. Comput. Math. Appl. 32 (11), 93–99.
SPSS 12.0, 2004. Regression Models. Prentice-Hall, Englewood Cliffs, NJ.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. 4th Edition. Springer, New York.