ELSEVIER

Journal of Neuroscience Methods 83 (1998) 185–194

JOURNAL OF NEUROSCIENCE METHODS

Estimating stimulus response latency

Howard S. Friedman [a], Carey E. Priebe [b,*]

[a] *Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*
[b] *Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218, USA*

Received 31 December 1997; received in revised form 3 April 1998; accepted 5 April 1998

Abstract

Stimulus response latency is the delay in the onset of stimulus-evoked neuronal activity. We develop maximum likelihood and least squares estimators of stimulus response latency and present a comparison of the performance of these methods with estimators commonly used in the neuroscience literature. The formal statistical change-point estimation problem is nontrivial due to the inclusion of a 'nuisance parameter', the end of stationarity in the stimulus-evoked activity. Our results suggest that the automation of the estimation of stimulus response latency will benefit from the use of the maximum likelihood estimator. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Latency; Peri-stimulus histogram; Change-point estimation; Maximum likelihood estimation; Least squares estimation; Nuisance parameter

1. Introduction

Information is transmitted in the nervous system through the firing of action potentials, known as spikes, by neurons (see Kandel and Schwartz, 1985 for an introduction). Factors affecting the temporal firing patterns allow neurophysiologists to assess the functional role of the neuron. Temporal aspects of the firing patterns which are typically examined include the mean firing rate, the autocorrelation, and the stimulus response latency.

Since neurons have a finite transmission velocity and a synaptic delay, a lag exists between the stimulus onset and the evoked modulation in neural activity. This delay, known as the stimulus response latency, provides information concerning hierarchal processing and functionality (Bullier and Nowak, 1995; Gawne et al., 1996). For example, in the primary visual cortex neurons are often categorized as belonging to a magnocellular or parvocellular system (Livingston and Hubel, 1984). These systems are believed to perform different functional roles (Livingston and Hubel, 1988; Fellman and Van Essen, 1991) and a significant difference in response latency has been reported for these two populations (Nowak et al., 1995). Neurons often have a non-stimulus evoked firing rate, known as the spontaneous firing rate. Consequently, response latency detection reduces to the determination of a change-point in the neural firing rate from the spontaneous activity rate to the stimulus-evoked response rate.

In a typical neural recording session, a stimulus is presented a number of times and the spike arrival times from stimulus onset are binned to form a peri-stimulus histogram (see Fig. 1). A number of techniques which have been utilized to measure stimulus response latency involve computations based on this histogram.

These include assuming the spontaneous activity is Poisson distributed and searching for a combination of consecutive bins which exceed a fixed threshold determined by this distribution (Maunsell and Gibson, 1992), smoothing the peri-stimulus histogram and de-

* Corresponding author. Present address: 104 Whitehead Hall, 3400 N. Charles St., John Hopkins University, Baltimore, MD 21218-2682, USA. Tel.: +1 410 5167200; fax: +1 410 5167459; e-mail: cep@jhu.edu

PII S0165-0270(98)00075-2

termining the time corresponding to half-height of the peak (Gawne et al., 1996), and determination by visual inspection of the spike arrival trains (for example, Celebrini et al., 1993; Mazzoni et al., 1996).

While these techniques have been effective in determining a value for the latency, they typically provide only a local examination of the neural firing pattern and do not invoke the mathematical formulations of change-point estimation (Carlstein, 1988; Larson, 1992; Muller, 1992). Since maximum likelihood and least squares estimation methodologies have advantageous theoretical properties (see Rice, 1995 for an introduction) one may anticipate success when using related approaches to detect response latencies. Previous research has demonstrated that maximum likelihood estimation of change-points applied to the neural spike train can be an effective latency estimator (Seal et al., 1983; Commenges and Seal, 1985). This technique is distinguished from the techniques we present in that it operates on the neural spike train rather than the peri-stimulus histogram.

The goal of our research is to derive latency estimation techniques which use developments from mathematical statistics and to compare the efficacy of these latency detection techniques to some of the techniques used in the literature. The results of this comparison indicate that maximum likelihood is the preferred technique for estimating stimulus response latency under a criterion of minimum mean squared error.

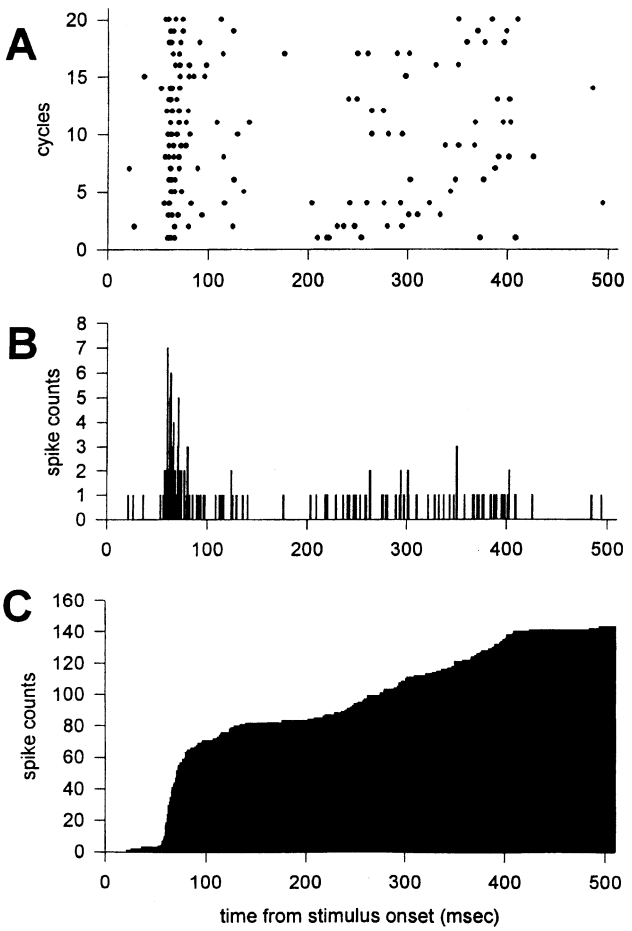

Fig. 1. Spike arrival times of a neuron from area V1 of an awake, fixating monkey for a flashing stimulus (1 Hz) are represented as tick marks in (A). The stimulus was presented for 500 ms, starting at time $t = 0$, for a total of 20 presentations. Using a bin width of 1 ms the histogram of spike arrival times (B) clearly demonstrates that there is a burst of activity approximately 60 ms from the time of stimulus onset. Summing the peri-stimulus histogram (B) produces the cumulative peri-stimulus histogram (C). The neural response as a function of time may be categorized into three response periods: (1) nonstimulus evoked response rate, (2) initial stimulus evoked response rate, and (3) terminal stimulus evoked response. Transitions between response periods are changepoints in the neural firing rate. The time from stimulus onset ($t = 0$) to the change-point between the non-stimulus evoked rate and the initial stimulus evoked response rate is defined as the neural response latency θ.

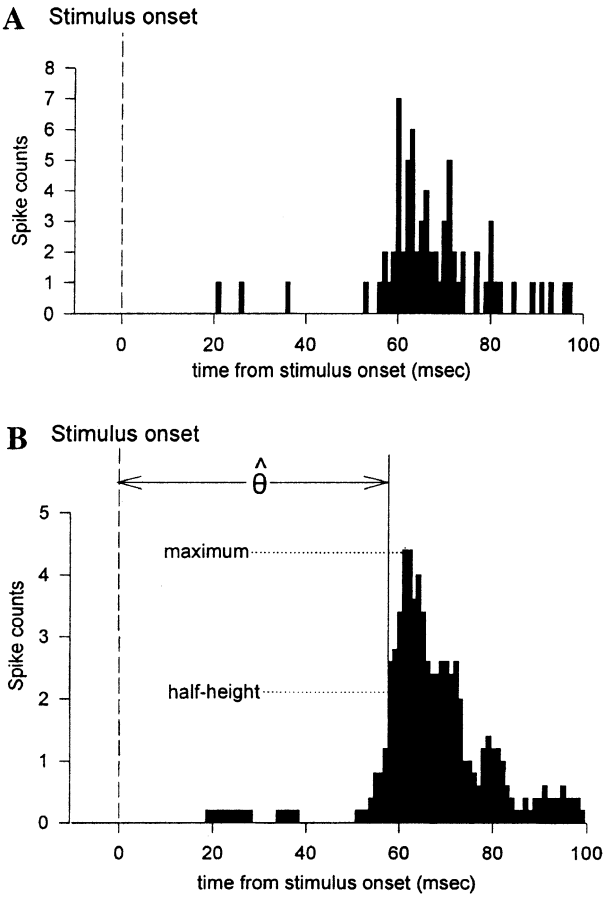

Fig. 2. A common nonparametric technique for estimating the latency θ involves smoothing the peri-stimulus histogram (same as Fig. 1B) to obtain a smoothed version (box smooth, 5 ms bandwidth) of the stimulus evoked neural response (2B). The maximum and minimum values of the smoothed histogram are then determined. The first time from stimulus onset in which the histogram exceeds the average of the minimum and maximum of the smoothed histogram is the estimated latency $\hat{\theta}_{HH}$.

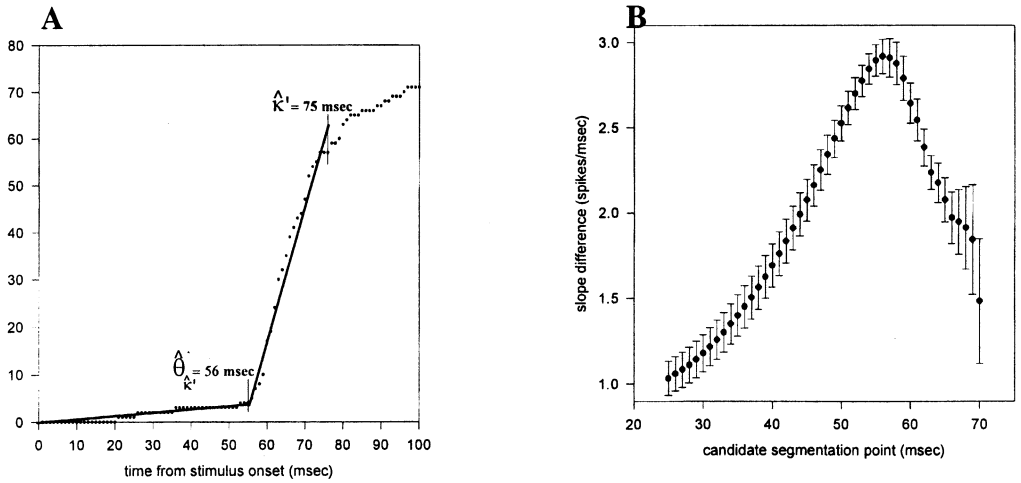

Fig. 3. The maximum likelihood and least squares techniques for estimating the latency θ involve estimating the nuisance parameter κ. As depicted in A, this is accomplished by truncating the cumulative peri-stimulus histogram at candidate cutoff value $\hat{\kappa}'$ so that the curve can be considered a single-knot linear spline on $[0, \hat{\kappa}']$. The histogram may then be segmented into two parts, the first section extending from the time of stimulus onset to the candidate segmentation point $\hat{\theta}'_{\hat{\kappa}'}$, the second segment from $\hat{\theta}'_{\hat{\kappa}'}$ to the candidate cutoff $\hat{\kappa}'$. Least squares fits for each segment result in an estimated slope and intercept for each line segment. The candidate segmentation point $\hat{\theta}'_{\hat{\kappa}'}$ which maximizes the difference in the slopes of the two segments (B) corresponds to the candidate estimated latency $\hat{\theta}'_{\hat{\kappa}'}$ for the candidate cutoff $\hat{\kappa}'$. For the cell depicted in Fig. 1, Fig. 3 indicates that the candidate estimated latency is $\hat{\theta}'_{\hat{\kappa}'} = 56$ ms for $\hat{\kappa}' = 75$ ms.

2. Methods

2.1. Estimation techniques

A common nonparametric technique which has been used to estimate the neural response latency θ involves smoothing the peri-stimulus histogram. The first time from stimulus onset in which the histogram exceeds the half-height of the peak of this smoothed function is used as the estimated latency $\hat{\theta}_{\mathrm{HH}}$ (e.g. Gawne et al., 1996). In our simulations, the peri-stimulus histogram is smoothed with a simple box smoother using the optimal bandwidth. Fig. 2 depicts this estimation, in which the delay between stimulus onset and the time corresponding to the midpoint between the minimum and maximum of the smoothed peri-stimulus histogram is the estimate $\hat{\theta}_{\mathrm{HH}}$.

A common parametric technique for latency detection assumes that the spontaneous activity has a Poisson distribution. The spontaneous activity rate λ_1 is estimated by fitting a Poisson distribution to the 250 bins before stimulus presentation in the peri-stimulus histogram. For this technique the estimated latency $\hat{\theta}_{\mathrm{MG}}$ is defined as the time from stimulus onset to the first bin that exceeds a level corresponding to a significance of $p = 0.01$ for the background distribution Poisson $(\hat{\lambda}_1)$ and is immediately followed, in sequence, by a bin that exceeds the 0.01 level and a bin that exceeds the 0.05 level (Maunsell and Gibson, 1992; Nowak et al., 1995).

We write the peri-stimulus histogram as $\hat{f}(t)$ and the cumulative peri-stimulus histogram as $\hat{F}(t)$ for discrete values of $t = 0, 1, \ldots, \tau$ representing bin numbers starting from stimulus onset. We derive maximum likelihood and least squares estimators for the latency by assuming that the spontaneous activity from stimulus onset $(t = 0)$ to latency $(t = \theta)$ has a Poisson (λ_1) distribution and the response activity from latency $(t = \theta)$ to cutoff $(t = \kappa)$ has a Poisson (λ_2) distribution. Thus the stochastic process ξ representing neural activity may be written as

$$\xi(t; \lambda_1, \lambda_2, \theta, \kappa) = \begin{cases} \text{Poisson}\,(\lambda_1) & \text{on} \quad [0, \theta) \\ \text{Poisson}\,(\lambda_2) & \text{on} \quad [\theta, \kappa] \\ \text{unknown} & \text{on} \quad (\kappa, \tau] \end{cases} \tag{1}$$

where Poisson rates are in units of spikes/bin. While the Poisson assumption for spikes/bin is by no means universally accepted, it is nonetheless a common approach for modeling cortical neural spike trains (see, for instance, Gerstein and Mandelbrot, 1964; Shadlen and Newsome, 1994). Additionally, we have assumed a step change in the neural firing rate (Seal et al., 1983).

Our goal is to estimate θ in the presence of the nuisance parameter κ and under the assumptions $\lambda_2 > \lambda_1$ and $0 < \theta < \kappa < \tau$. The cutoff κ is a nuisance parameter; its value is incidental but its estimation accuracy nonetheless impacts the accuracy of the subsequent estimate of the parameter of interest θ.

Given an estimate $\hat{\kappa}$ for the nuisance parameter κ, the maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}(\hat{\kappa})}$ for the latency is the value of $\theta \in \{1, \ldots, \hat{\kappa} - 1\}$ which maximizes the discrete log-likelihood function

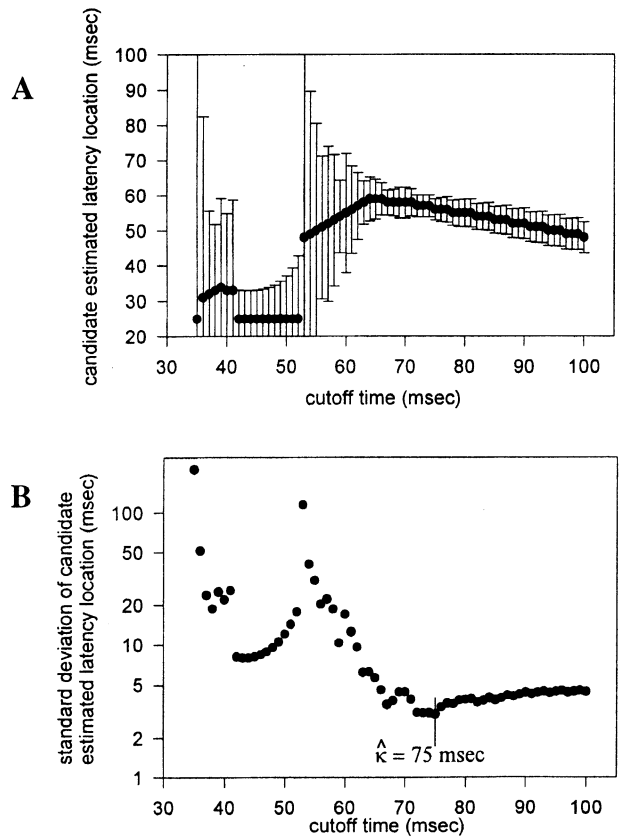

Fig. 4. The candidate estimated latency $\hat{\theta}'_{\hat{\kappa}'}$ reported by the technique of Fig. 3 is dependent on the choice of candidate cutoff value $\hat{\kappa}'$. A depicts the latency estimate and the uncertainty in the location of this estimate for each choice of cutoff $\hat{\kappa}'$. B depicts the uncertainty directly. The estimated cutoff $\hat{\kappa}$ is the point at which the single-knot linear spline is most accurate, in terms of uncertainty in candidate estimated latency $\hat{\theta}'_{\hat{\kappa}'}$. As shown in B, this is determined by finding the uncertainty in the intersection of the two segments at the candidate estimated latency $\hat{\theta}'_{\hat{\kappa}'}$ for each candidate cutoff value $\hat{\kappa}'$. This results in an estimate of $\hat{\kappa} = 75$ ms for the cell depicted in Fig. 1.

$$l(\theta|\hat{\kappa}) = -\hat{\lambda}_1(\theta+1) + (\log\hat{\lambda}_1)\sum_{t=0}^{\theta}\hat{f}(t) - \sum_{t=0}^{\theta}\log(\hat{f}(t)!) - \hat{\lambda}_2(\hat{\kappa}-\theta) + (\log\hat{\lambda}_2)\sum_{t=\theta+1}^{\hat{\kappa}}\hat{f}(t) - \sum_{t=\theta+1}^{\hat{\kappa}}\log(\hat{f}(t)!) \quad (2)$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given by

$$\hat{\lambda}_1 = \left(\sum_{t=0}^{\theta}\hat{f}(t)\right)/(\theta+1)$$

and

$$\hat{\lambda}_2 = \left(\sum_{t=\theta+1}^{\hat{\kappa}}\hat{f}(t)\right)/(\hat{\kappa}-\theta) \quad (3)$$

That is, $\hat{\theta}_{\mathrm{ML}(\hat{\kappa})} = \arg\max_{\theta\in\{1,\ldots,\hat{\kappa}-1\}} l(\theta|\hat{\kappa})$.

To obtain a least squares estimate $\hat{\theta}_{\mathrm{LS}}$ for the latency given an estimate $\hat{\kappa}$ for the nuisance parameter κ note that the model described in Eq. (1) implies that, if the nuisance parameter is known, then the cumulative peri-stimulus histogram is a single-knot linear spline on $[0, \kappa]$. Under the assumption that the estimate $\hat{\kappa}$ is close to the true value κ, we proceed by assuming that the cumulative peri-stimulus histogram is a single-knot linear spline on $[0, \hat{\kappa}]$. The least squares estimate $\hat{\theta}_{\mathrm{LS}(\hat{\kappa})}$ for the latency is the value of $\theta\in\{1,\ldots,\hat{\kappa}-1\}$ which minimizes the discrete residual sum of squares function

$$\mathrm{RSS}(\theta|\hat{\kappa}) = \sum_{t=0}^{\theta}(\hat{F}(t) - \hat{\lambda}_1 t)^2 + \sum_{t=\theta+1}^{\hat{\kappa}}(\hat{F}(t) - \hat{\lambda}_1\theta + \hat{\lambda}_2(\theta - t))^2 \quad (4)$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are given by

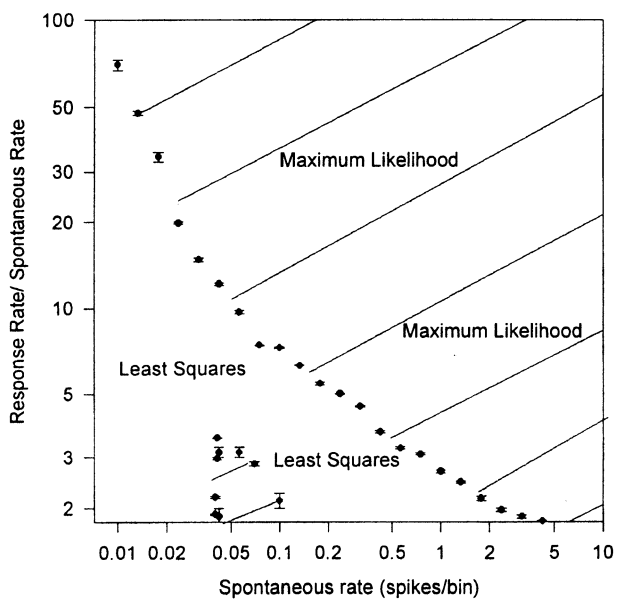

Fig. 5. Monte Carlo Simulation #1 (κ known). The estimator of stimulus response latency with the minimum MSE over the class $\{HH, MG, ML, LS\}$ depends on the spontaneous rate λ_1 and the response rate λ_2. For simulations in which these rates are stationary the best estimator is either the maximum likelihood estimator or the least squares estimator. No point in the parameter space is found where either the Maunsell-Gibson or half-height technique have a smaller MSE than both the least squares and maximum likelihood techniques. From this experiment we conclude that when κ is known $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$ are admissible estimators of stimulus response latency θ relative to the class $\{HH, MG, ML, LS\}$ while $\hat{\theta}_{\mathrm{MG}}$ and $\hat{\theta}_{\mathrm{HH}}$ are inadmissible.

$$\hat{\lambda}_1 = \frac{-\left(\sum_{t=0}^{\theta} t\hat{F}(t)\right) - \theta\left(\sum_{t=\theta+1}^{\hat{\kappa}} \hat{F}(t)\right) + \theta\left(\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)\right)\left[\frac{\sum_{t=\theta+1}^{\hat{\kappa}} \hat{F}(t)(\theta - t)}{\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)^2}\right]}{\theta^2\left(\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)\right)\left[\frac{\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)}{\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)^2}\right] - \left(\sum_{t=0}^{\hat{k}} t^2\right) - \theta^2(\hat{\kappa} - \theta)}$$

and

$$\hat{\lambda}_2 = \frac{\hat{\lambda}_1\theta \sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t) - \sum_{t=\theta+1}^{\hat{\kappa}} \hat{F}(t)(\theta - t)}{\sum_{t=\theta+1}^{\hat{\kappa}} (\theta - t)^2} \quad (5)$$

That is, $\hat{\theta}_{\mathrm{LS}(\hat{\kappa})} = \arg\min_{\theta \in \{1, \ldots, \hat{\kappa}-1\}} \mathrm{RSS}(\theta|\hat{\kappa})$.

In order to obtain an estimate $\hat{\kappa}$ for the cutoff as required for $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$, it is necessary to select the point at which the assumption of stationarity in the response rate is best supported. For a fixed candidate cutoff value $\hat{\kappa}'$, we segment the data sequence into two parts and perform a least squares fit to both segments. All possible candidate segmentation points $\hat{\theta}_{\hat{\kappa}'}$ are examined in this manner and the candidate which maximizes the difference in the slopes of the two lines is the candidate estimated latency $\hat{\theta}_{\hat{\kappa}'}$ for the candidate cutoff $\hat{\kappa}'$; $\hat{\theta}_{\hat{\kappa}'} = \arg\max_{\hat{\theta}_{\hat{\kappa}'}}$ (slope difference). As seen in Fig. 3, the difference in the slopes of the lines has a clear peak around $\hat{\theta}_{\hat{\kappa}'} = 56$ ms, when the candidate cutoff is $\hat{\kappa}' = 75$ ms for the example data set depicted in Fig. 1. The error in the location of the intersection of the two lines is considered to be an estimate of the uncertainty in the candidate estimated latency $\hat{\theta}_{\hat{\kappa}'}$. To determine the final estimate $\hat{\kappa}$ for the cutoff, we consider all values for the candidate cutoff $\hat{\kappa}'$ between 35 ms after stimulus onset and the final bin. We select as our estimate $\hat{\kappa}$ of κ the candidate cutoff which results in the smallest uncertainty in the candidate estimated latency; $\hat{\kappa} = \arg\min_{\hat{\kappa}'}$ (uncertainty in $\hat{\theta}_{\hat{\kappa}'}$) (see Fig. 4).

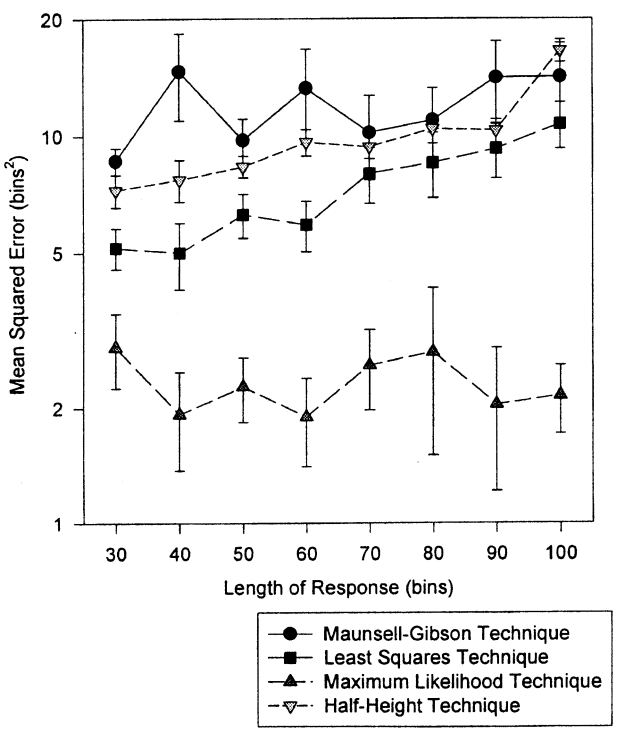

Fig. 6. Monte Carlo Simulation #2 (κ known, varying response length). For $\lambda_1 = 1$ spikes/bin and $\lambda_2 = 4$ spikes/bin, the MSE performance of the latency estimators is independent of response length $\kappa - \theta$.

2.2. Monte Carlo comparison methodology

The Poisson assumption technique of Maunsell and Gibson ($\hat{\theta}_{\mathrm{MG}}$) differs from the other estimators in that it may report that a latency does not exist, whereas the other three estimators always provide a latency location. In order to fairly compare all four estimators, we select from the estimated latencies only values which fell into a pre-defined 'accepting region' R_a. The probability that an estimated latency falls into the accepting region is defined to be the efficiency. For each estimator $Z \in \{HH, MG, ML, LS\}$ we obtain an estimate of the efficiency $\hat{e}_Z$ as the ratio of the number of times $\hat{\theta}_Z \in R_a$ to the number of Monte Carlo replicates. The estimated mean squared error $\hat{MSE}_Z$ for each estimator is calculated from the subset of the 500 Monte Carlo replicates for which $\hat{\theta}_Z \in R_a$, and the standard error of this mean squared error is estimated using a bootstrap technique.

To directly compare the performance of the four estimators described above, we first consider the case in which the nuisance parameter κ is known. Thus for Monte Carlo Simulation #1 we simulate vectors containing 100 random variables in which the first 50 component samples of the simulated vector are independent and identically distributed (i.i.d.) from a Poisson (λ_1) distribution, representing the spontaneous activity, and the last 50 components are i.i.d. Poisson (λ_2), representing the initial response activity. We consider the subset of parameter space defined by $0.01 \leq \lambda_1 \leq 10$ and $2 \leq \lambda_2 \leq 10$. In this case the assumptions of both the least squares and maximum likelihood models are correct, with a latency of $\theta = 50$ and a (known) cutoff of $\kappa = 100$. All four estimators search for latencies from the 10th bin ($t = 10$) to the 90th bin ($t = 90$).

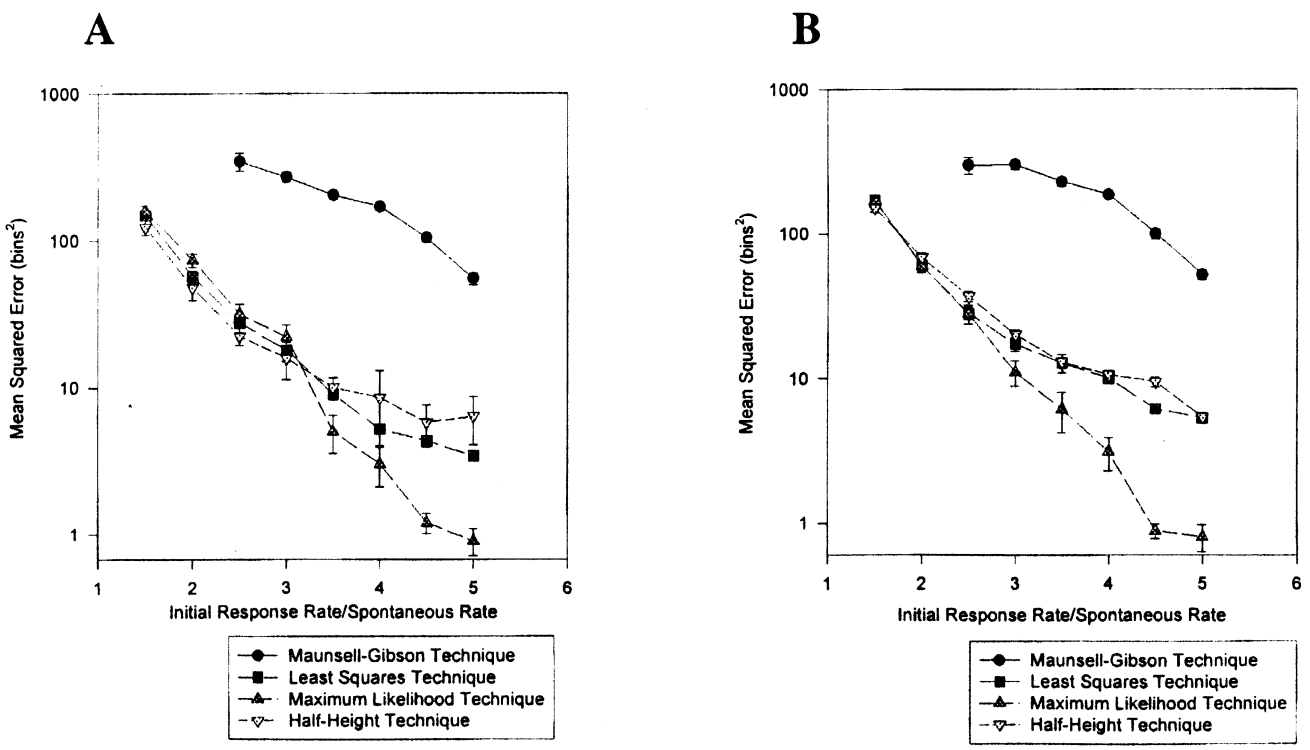

Fig. 7. Monte Carlo Simulation # 3. Plots are presented for a spontaneous rate $\lambda_1 = 1$ spike/bin and for (A) $\lambda_3 = 1$ spike/bin and (B) $\lambda_3 = 0.9 \cdot \lambda_2$ spikes/bin. The maximum likelihood estimator $\hat{\theta}_{ML}$ has the smallest MSE when $\lambda_2 > 3$ (this agrees with Fig. 5) and none of the estimators exhibits a statistically significant superiority for $\lambda_2 \leq 3$. For large values of λ_1 (e.g. $\lambda_1 > 2$) $\hat{\theta}_{ML}$ is always the superior estimator. For fixed λ_1 the MSEs decrease as λ_2 increases, as intuition suggests. This simulation requires the estimation of the nuisance parameter κ for $\hat{\theta}_{ML}$ and $\hat{\theta}_{LS}$.

This is repeated 100 times for each estimator. From the resulting vectors of estimated latencies, only those in the acceptance region $R_a = [20, 80]$ are used to estimate the MSE.

In order to investigate the dependence of estimation performance on the length of the response, we again fix the latency at $\theta = 50$ and assume κ is known. For Monte Carlo Simulation # 2 we vary the length of the simulated vector representing the initial response activity. For this simulation the parameters are the spontaneous rate λ_1, the response rate λ_2, and the response length $\kappa - \theta$. Each technique scanned the region between $t = 10$ and $t = \kappa - 10$ and reported an estimated latency. This was repeated 100 times for each estimator and each value of response length. In this case $R_a = [35, 65]$.

Monte Carlo Simulation # 3 is designed to compare the full latency estimation techniques, including the estimation of the nuisance parameter. We simulate vectors containing 150 random variables. The first 50 component samples are i.i.d. from a Poisson (λ_1) distribution, representing the spontaneous activity. Observations 51 through 100 are i.i.d. Poisson (λ_2), representing the initial response activity. The last 50 samples are i.i.d. Poisson (λ_3), representing nonstationarity in the response activity. Hence, the latency is fixed at $\theta = 50$ throughout the simulation and the cutoff is fixed at $\kappa = 100$. We vary λ_1 between 0.1 spikes/bin and 10 spikes/bin. Various ratios of λ_2/λ_1 are considered in a range from 2 to 10. For fixed values of λ_1 and λ_2, λ_3 is varied over a range from λ_1 to $0.9 \cdot \lambda_2$. Since the half-height technique and the Maunsell and Gibson technique do not require stationarity in the response rate, these techniques scan the range between $t = 10$ and $t = 140$ and report an estimated latency for the observed sequence. For least squares and maximum likelihood we first estimate the cutoff κ. These two techniques then scan the range from $t = 10$ to $t = \hat{\kappa} - 5$, five bins less than the estimated cutoff $\hat{\kappa}$. We consider $R_a = [20, 80]$ and report Monte Carlo results which are obtained by repeating this process 500. For each of the four techniques and for each choice of λ_1, λ_2, λ_3 there are 500 estimated latencies upon which to perform statistical inference in order to distinguish between the performance of the four estimators $\hat{\theta}_{HH}$, $\hat{\theta}_{MG}$, $\hat{\theta}_{ML}$, $\hat{\theta}_{LS}$.

2.3. Application to neural data

To test our techniques on neural data, we use two data sets of spike arrival times recorded from neurons in the primary visual cortex of an awake, fixating monkey. The animal performed a fixation task during which time a figure was continuously flashed on for 500 ms and off for 500 ms. The trials ranged in length from

800 to 4700 ms. Trials were segmented into one second cycles, where the spike arrival time was with respect to the most recent figure onset. Spike arrival times were placed in 1 ms bins to form a peri-stimulus histogram. Response latencies are estimated for each stimulus presentation as well as for the entire data set. The Maunsell-Gibson and half-height techniques scan the range between 10 and 100 ms. For maximum likelihood and least squares, the cutoff is estimated between 10 and 100 ms; candidate latencies between 10 and $\hat{\kappa}-3$ are then considered. A mean squared error is obtained from the distribution of latencies from individual presentations where the bias term is estimated using the latency obtained for the entire data set.

3. Results

3.1. Monte Carlo comparison results

Fig. 5 presents the results of Monte Carlo Simulation #1, in which the nuisance parameter κ is known. As indicated in the figure, the parameter space divides into regions for which either the least squares estimator or the maximum likelihood estimator has the minimum estimated MSE of the four estimators considered. That is, $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$ are admissible estimators of stimulus response latency θ relative to the class $\{HH, MG, ML, LS\}$ for this experiment. We do not observe any area of the parameter space in which either the Maunsell-Gibson technique or the half-height technique is admissible; $\hat{\theta}_{\mathrm{MG}}$ and $\hat{\theta}_{\mathrm{HH}}$ are inadmissible estimators. While the minimum MSE estimator is a function of both the spontaneous rate λ_1 and the ratio λ_2/λ_1, the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$ is generally superior for larger values of λ_1.

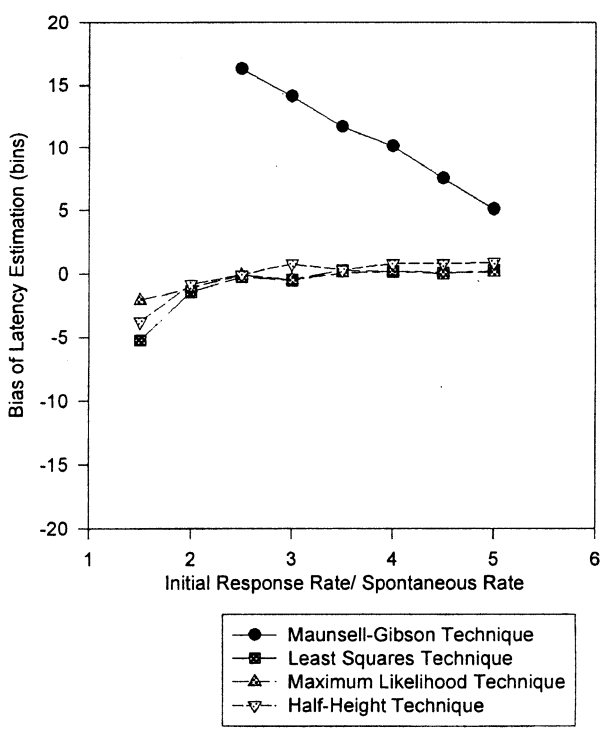

Fig. 8. Monte Carlo Simulation #3. The Maunsell-Gibson estimator $\hat{\theta}_{\mathrm{MG}}$ is biased, reporting latencies which tend to be larger than the true latency. The other estimators demonstrate little or no bias for larger values of λ_2/λ_1.

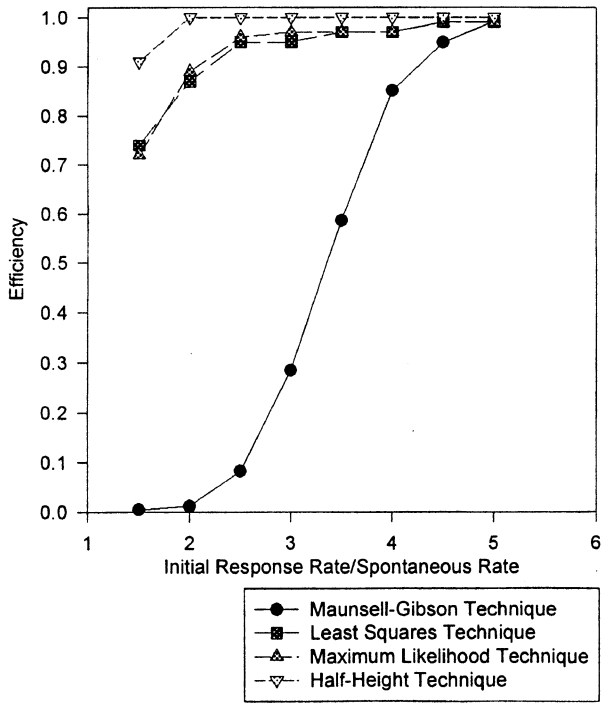

Fig. 9. Monte Carlo Simulation #3. The estimated efficiency $\hat{e}_Z$ is similar for all values of initial response rate λ_2 for $\hat{\theta}_{\mathrm{HH}}$, $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$, with the half-height technique $\hat{\theta}_{\mathrm{HH}}$ exhibiting slight superiority. That is, these techniques detected the latency within the acceptance region R_a most of the time. For small values of λ_2/λ_1, the Maunsell-Gibson technique $\hat{\theta}_{\mathrm{MG}}$ failed to detect the latency within R_a a significant proportion of the time. For the least squares and maximum likelihood estimation more than 50% of the inefficiency may be attributed to the cutoff detection.

As seen in Fig. 6, Monte Carlo Simulation #2 suggests that for $\lambda_1 = 1$ spikes/bin and $\lambda_2 = 4$ spikes/bin the maximum likelihood estimate $\hat{\theta}_{\mathrm{ML}}$ has the smallest MSE of all the techniques. Furthermore, the MSEs for all four techniques appear to be independent of the response length $\kappa - \theta$.

Fig. 7 investigates the performance of the four estimators when the nuisance parameter κ must be estimated. Fig. 7 indicates that when the spontaneous rate $\lambda_1 = 1$ spike/bin and (A) $\lambda_3 = 1$ spike/bin and (B) $\lambda_3 = 0.9 \cdot \lambda_2$ spikes/bin the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$ has the smallest MSE when $\lambda_2 > 3$ (this agrees with Fig. 5) and none of the estimators exhibits a statistically significant superiority for $\lambda_2 < 3$. For large values of λ_1 maximum likelihood is always the superior estimator. This result is supported by the results presented in Fig.

5 for $\lambda_1 > 2$. As intuition suggests, for fixed λ_1 the MSEs decrease as λ_2 increases.

Fig. 8 indicates that the Maunsell-Gibson estimator $\hat{\theta}_{MG}$ is biased, reporting latencies which tend to be larger than the true latency. The other estimators demonstrate little or no bias for larger values of λ_2/λ_1. As seen in Fig. 9, the estimated efficiency $\hat{e}_Z$ is similar for all values of initial response rate λ_2 for $\hat{\theta}_{HH}$, $\hat{\theta}_{ML}$ and $\hat{\theta}_{LS}$, with the half-height technique $\hat{\theta}_{HH}$ exhibiting slight superiority. That is, these techniques detected the latency within the acceptance region R_a most of the time. For small values of λ_2/λ_1, the Maunsell-Gibson technique $\hat{\theta}_{MG}$ failed to detect the latency within R_a a significant proportion of the time. For $\hat{\theta}_{ML}$ and $\hat{\theta}_{LS}$ more than 50% of the inefficiency can be attributed to the estimation of the nuisance parameter κ.

3.2. Application to neural data

Of course, assumptions made in Monte Carlo analysis may or may not be appropriate for real neural responses. Fig. 10 depicts the results of automated stimulus response latency estimation applied to data sets obtained from two neurons in the primary visual cortex. Fig. 10 suggests that in both cases the estimate $\hat{\kappa}$ for the nuisance parameter κ is reasonable. Using the estimated cutoff $\hat{\kappa}$, response latency estimates are successfully obtained using least squares and maximum likelihood for the data sets. These values agree closely with those obtained via the Maunsell-Gibson and half-height techniques.

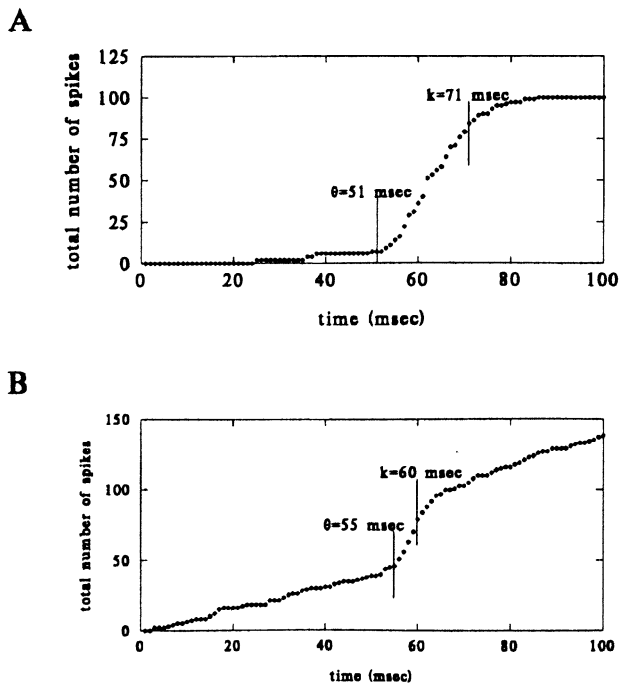

Fig. 10. Application to neural data. Cumulative peri-stimulus histograms of two neurons recorded from area V1 of an awake, behaving monkey. The estimates of the cutoff κ (the point at which response rate is no longer stationary) are $\hat{\kappa} = 71$ ms in A and $\hat{\kappa} = 60$ ms in B. For cell A, the estimates for the latency are: $\hat{\theta}_{ML} = 51$, $\hat{\theta}_{LS} = 51$, $\hat{\theta}_{HH} = 54$, $\hat{\theta}_{MG} = 52$. For cell B, the estimates for the latency are: $\hat{\theta}_{ML} = 55$, $\hat{\theta}_{LS} = 54$, $\hat{\theta}_{HH} = 56$, $\hat{\theta}_{MG} = 56$.

Fig. 11 indicates that no significant differences are observed in the distribution of latencies obtained for cell A. Furthermore, the correlation in reported latencies is greater than 0.99 for the three pairwise comparisons. For cell B, the mean squared error of the maximum likelihood technique is significantly smaller than that of the other techniques due to the difference in the bias term. The mean value of maximum likelihood estimation for this cell is 47.3 ms, whereas the mean value for the half-height technique (using a normal smoother with a 5-ms band width; see Gawne et al., 1996) is 42 ms. For least squares the mean value is 33.2 ms. The Maunsell-Gibson technique fails to obtain an estimated latency for these data because the response rate is too low; it never occurs that three consecutive bins contain a spike.

To further investigate the performance of the techniques on the cell B data, simulations designed specifically to relate to this data are performed. The response rates are fixed at the estimated rates for a single presentation ($\lambda_1 = 0.018$ spikes/ms, $\lambda_2 = 0.137$ spikes/ms, $\lambda_3 = 0.022$ spikes/ms) and the relevant durations are obtained from Fig. 10b ($n_1 = 55$ ms, $n_2 = 6$ ms, $n_3 = 39$ ms). The results indicate that the maximum likelihood technique has the smallest mean squared error due to a smaller bias term. The mean value for the maximum likelihood estimator is 46.5 ms (standard error = 0.7 ms), compared to 42.1 ms (standard error = 0.7 ms) for half-height and 28.5 ms (standard error = 0.9 ms) for least squares.

4. Discussion

The major conclusions to be drawn from the investigations presented herein are two-fold. First, when κ is known, for the portion of the λ_1, λ_2 parameter space studied, $\hat{\theta}_{ML}$ and $\hat{\theta}_{LS}$ are admissible estimators of stimulus response latency θ relative to the class $\{HH, MG, ML, LS\}$ while $\hat{\theta}_{MG}$ and $\hat{\theta}_{HH}$ are inadmissible. That is, with respect to the mean squared error criterion, there is no situation in which statistical analysis suggests the use of $\hat{\theta}_{MG}$ and $\hat{\theta}_{HH}$. Second, in the more realistic scenario in which the estimation of θ must be done in the presence of the nuisance parameter κ, the maximum likelihood estimator $\hat{\theta}_{ML}$ is the superior estimator for large values of the ratio λ_2/λ_1 or large values of λ_1. The statistical problem of estimating a change-point in the presence of a nuisance parameter is a difficult one, and our method of obtaining the estimate $\hat{\kappa}$ for this particular application is but one option.

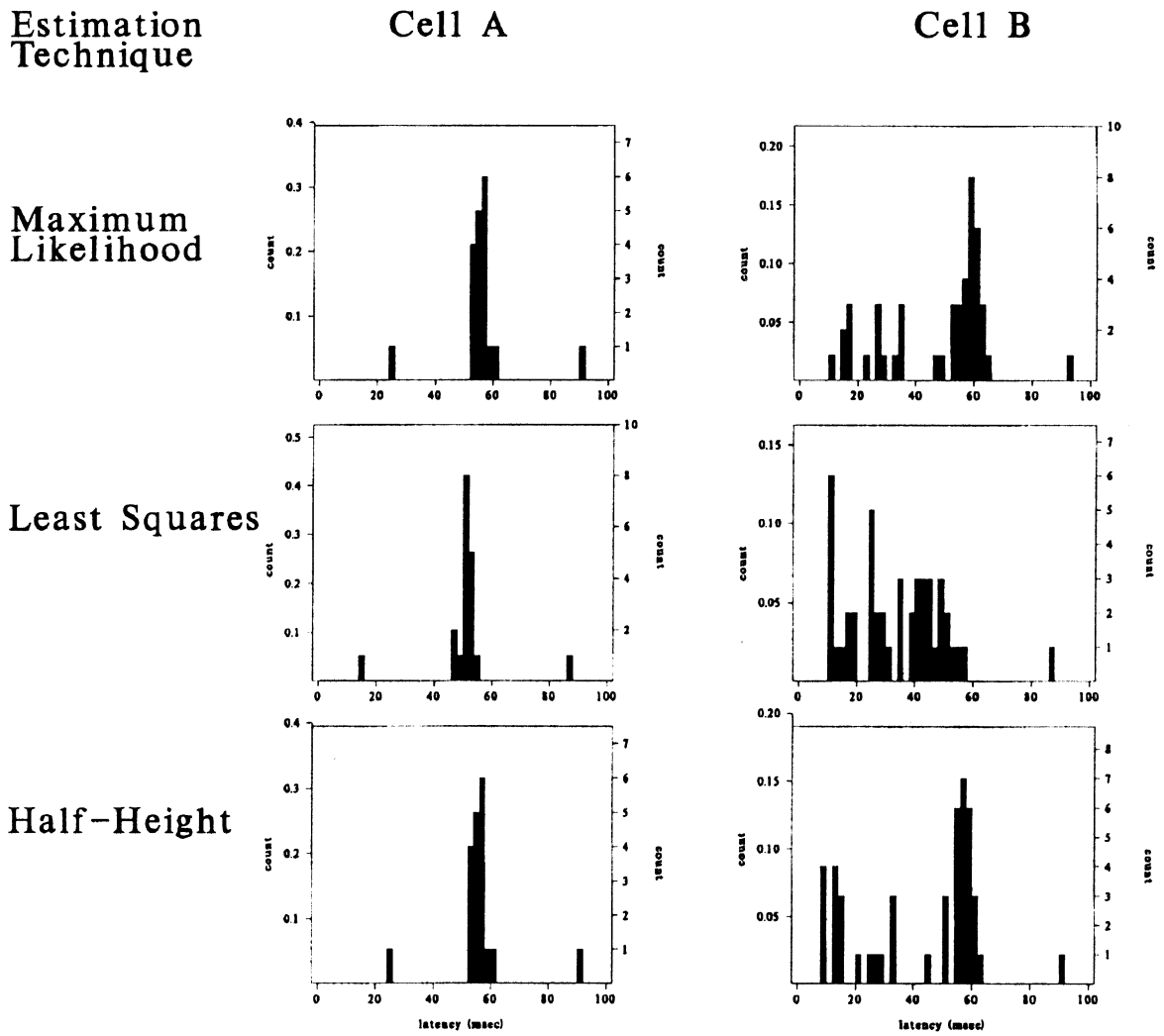

Fig. 11. Application to neural data. Depicted are the distributions of estimated latencies via maximum likelihood, least squares, and half-height, for the individual stimulus presentations which constitute the data for the two cells considered in Fig. 10. (Maunsell-Gibson never detects a latency for these individual stimulus presentations). The cell A data consists of 19 presentations, and the cell B data consists of 46 presentations. For cell A, the distributions are nearly identical. For cell B maximum likelihood is superior.

We emphasize the importance of choosing the optimal smoothing bandwidth for the popular half-height estimate $\hat{\theta}_{\mathrm{HH}}$. This choice is a function of λ_1, λ_2, the length of initial response rate, and the latency itself. The performance of the estimator is quite sensitive to suboptimal smoothing. Indeed, a choice of smoothing bandwidth off by as few as five bins from the optimal can result in an estimator that is a full order of magnitude less accurate. We expended significant computational effort to present best-case results for $\hat{\theta}_{\mathrm{HH}}$, but warn the practitioner of the potential for significantly poorer performance. The superior performance in terms of efficiency for $\hat{\theta}_{\mathrm{HH}}$ compared to $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$ should not be disregarded. The majority of this difference can be attributed to error in estimating the nuisance parameter κ for $\hat{\theta}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{LS}}$.

The Maunsell-Gibson estimator $\hat{\theta}_{\mathrm{MG}}$ has the advantage of being robust to non-stationarities in the response rate, but has some notable limitations as well. In particular, Figs. 7 and 8 indicate that the bias-dominated mean squared error of $\hat{\theta}_{\mathrm{MG}}$ makes it inadmissible. Nonetheless, the Maunsell-Gibson technique has the advantageous property of determining whether there is a response at all. For the maximum likelihood and least squares techniques, this information may be obtained following detection of the cutoff. At this point, the experimenter has an estimate of the spontaneous activity rate, the initial response rate, and the standard errors of those rates. Using a t-test the experimenter may decide whether the initial response rate is significantly larger than the spontaneous activity rate, and, consequently whether there is any modulation in neural activity.

A comment is in order regarding the parameters λ_1 and λ_2 and our decision to report results in terms of λ_2/λ_1. For a fixed cell and stimulus paradigm (and stationarity in response over the time of recording) increasing the

number of repetitions of a stimulus increases λ_1 while keeping λ_2/λ_1 fixed. Thus, for instance, Fig. 5 suggests that we can always move into the region of parameter space for which $\hat{\theta}_{ML}$ is the superior estimator, while Fig. 7 indicates that if λ_2/λ_1 is large then the experimenter should use $\hat{\theta}_{ML}$.

It is noteworthy that the maximum likelihood and least squares approaches presented herein are general techniques, suitable for estimating the latency of either excitation or inhibition. In order to search for the moment of inhibition, the cutoff detection must now minimize the difference between the slopes of the two lines, rather than maximize. Once the cutoff has been obtained, the maximum likelihood and least squares approaches detect the onset of inhibition with similar efficiency as for the onset of excitation.

The results obtained for cell A indicate that the choice of latency estimation technique is not critical for neurons with very low spontaneous activity (in 18 of 19 trials, no spikes occurred in the first 50 ms of stimulus presentation). For cell B, where the spontaneous activity is estimated to be 18 spikes/s, the maximum likelihood estimator is superior. For typical ranges of neural response, the half-height technique selects the first spike following a stimulus presentation. Consequently, the half-height technique often selects latencies that are smaller than the true latency due to the high spontaneous activity.

The remarkable agreement between the simulation results and the distribution obtained from the neural data set, in spite of the simplifying assumptions of Poisson activity and a step change in firing rate, support the notion that our model characterizes the neural data sufficiently well to have strong predictive power.

Acknowledgements

The authors thank two anonymous referees for comments which led to an improved manuscript. We thank Rudiger von der Heydt for providing the experimental data. This work was partially supported by the Whitaker Foundation and by NIH EY02966.

References

Bullier J, Nowak LG. Parallel versus serial processing: new vistas on the distributed organization of the visual system. Curr Opin Neurobiol 1995;5:497–503.

Carlstein E. Nonparametric change-point estimation. Ann Stat 1988;16:188–97.

Celebrini S, Thorpe S, Trotter Y, Imbert M. Dynamics of orientation coding in area V1 of the awake primate. Vis Neurosci 1993;10(81):18–25.

Commenges D, Seal J. The analysis of neuronal discharge sequences. Stat Med 1985;4(1):91–104.

Fellman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1991;1:1–47.

Gawne TJ, Kjaer TW, Richmond BJ. Latency: another potential code for feature binding in striate cortex. J Neurophysiol 1996;76:1356–60.

Gerstein G, Mandelbrot B. Random walk models for the spike activity of a single neuron. Biophys J 1964;4:41–68.

Kandel ER, Schwartz JH. Principles of Neural Science. Elsevier Science: New York, 1985.

Larson HJ. Least squares estimation of linear splines with unknown knot locations. Comput Stat Data Anal 1992;13:1–8.

Livingston MS, Hubel DH. Anatomy and physiology of a color system in the primate visual cortex. J Neurosci 1984;4:309–56.

Livingston MS, Hubel DH. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science 1988;240:740–9.

Maunsell JHR, Gibson JR. Visual response latencies in striate cortex of the macaque monkey. J Neurophysiol 1992;68:1332–44.

Mazzoni P, Bracewell MR, Barash S, Andersen RA. Spatially tuned auditory responses in area LIP of macaques performing delayed memory saccades to acoustic targets. J Neurophysiol 1996;75:1233–41.

Muller HG. Change-points in nonparametric regression analysis. Ann Stat 1992;20:737–61.

Nowak LG, Munk MHJ, Girard P, Bullier J. Visual latencies in areas V1 and V2 of the macaque monkey. Vis Neurosci 1995;12:371–84.

Rice JA. Mathematical Statistics and Data Analysis. Duxbury Press: Belmont, CA, 1995.

Seal J, Commenges D, Salamon R, Bioulac B. A statistical method for the estimation of neural response latency and its functional interpretation. Brain Res 1983;278:382–6.

Shadlen MN, Newsome WT. Noise, neural codes and cortical organization. Curr Opin Neurobiol 1994;4:569–79.