J Neurophysiol 94: 8–25, 2005;
doi:10.1152/jn.00648.2004.

Statistical Issues in the Analysis of Neuronal Data

Robert E. Kass,[1] Valérie Ventura,[1] and Emery N. Brown[2]

[1]*Department of Statistics and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania; and*
[2]*Neuroscience Research Laboratory, Department of Anesthesia and Critical Care, Massachusetts General Hospital, and the Division of Health Sciences and Technology, Harvard Medical School, Massachusetts Institute of Technology, Boston, Massachusetts*

Submitted 25 June 2004; accepted in final form 19 February 2005

Kass, Robert E., Valérie Ventura, and Emery N. Brown. Statistical issues in the analysis of neuronal data. *J Neurophysiol* 94: 8–25, 2005; doi:10.1152/jn.00648.2004. Analysis of data from neurophysiological investigations can be challenging. Particularly when experiments involve dynamics of neuronal response, scientific inference can become subtle and some statistical methods may make much more efficient use of the data than others. This article reviews well-established statistical principles, which provide useful guidance, and argues that good statistical practice can substantially enhance results. Recent work on estimation of firing rate, population coding, and time-varying correlation provides improvements in experimental sensitivity equivalent to large increases in the number of neurons examined. Modern nonparametric methods are applicable to data from repeated trials. Many within-trial analyses based on a Poisson assumption can be extended to non-Poisson data. New methods have made it possible to track changes in receptive fields, and to study trial-to-trial variation, with modest amounts of data.

INTRODUCTION

Technical advances have made available new methods for collecting, storing, and manipulating electrophysiological data. Investigations may now not only characterize neuronal activity in anatomically well defined regions, but they can also examine dynamics of neuronal response and their relationship to behavior. Although elementary methods of data analysis [such as *t*-tests or visual examination of the peristimulus time histogram (PSTH)] remain useful for many purposes, the growing complexity of neuroscientific experiments, often examining subtle changes on a comparatively fine timescale, requires careful attention to statistical methods for data analysis. In this overview we discuss some of the fundamental data analytical issues that face researchers in neurophysiology, illustrating the general points with the problems of describing the evolution of a neuron's firing rate across time, finding accurate population codes, and assessing time-varying correlation between 2 neurons. In each case recent work has provided a statistical technique that outperforms previous methodology, boosting the scientific information as effectively as if the number of experimental trials, or the number of neurons, had been increased by a substantial factor. We also indicate some of the ways modern statistical procedures can accommodate important complexities, such as dynamic changes in temporal and spatial aspects of hippocampal place cell firing and trial-to-trial variability in cortical neurons recorded from behaving animals. Our review supplements the brief and general guidance offered by Curran-Everett and Benos (2004), and may be regarded as an update to the early work of Perkel et al. (1967a,b).

The new field of computational neuroscience uses detailed biophysical models and artificial neural networks to study emergent behavior of neural systems and the way neural systems represent and transmit information (e.g., Dayan and Abbott 2001). Statistical methods have become an essential complement: in addition to their ubiquitous role in summarizing experimental data, they provide estimates of biologically relevant parameters, assessments of uncertainty about them, and a formalism for evaluating the fit of theoretical predictions to observed data. The statistical paradigm begins with informal investigation of the data, a process that has been named *exploratory analysis* (Tukey 1977). Exploratory results, together with judgment based on experience, help guide construction of an initial probability model to represent variability in observed data. Typically, this model posits some underlying, and unobserved, regularity that is coupled with known or unknown sources of irregularity. Every such model, and every statistical method, makes some assumptions. These lead to a reduction of the data to some typically small number of interpretable quantities. The data may be used, again, to check the probabilistic assumptions, and to consider ramifications of departures from them. Should serious departures from the assumptions be found, a new model may be formed. Figure 1 attempts to highlight the iterative nature of probability modeling and model assessment, followed by statistical inference, all embedded into the production of scientific conclusions from experimental results (Box et al. 1978). We demonstrate this process by analyzing data from a hippocampal "place" cell, using a collection of techniques reviewed in subsequent sections.

Within the field of statistics there are standards for evaluating alternative data-reduction procedures. Sometimes methods that seem intuitive may be shown to work well. In particular, simple data summaries and graphical displays are often sufficient for demonstrating striking experimental findings, when noise is small relative to signal, for example, or when additional sources of variation (beyond those summarized) need not be made explicit. In more subtle situations, however, intuitive data summaries may be inconclusive, possibly because they may make inefficient use of the data. We will illustrate by considering first an elementary framework for spike count analysis, where a peculiar yet plausible method of estimating firing rate will be shown to make inefficient use of the data. We then work by analogy, showing how, in important problems, intuitive methods used by neurophysiologists may be equivalent to discarding most of the available data. Our overview focuses on 3 main points: *1*) maximum likelihood and Bayes-

Address for reprint requests and other correspondence: V. Ventura, Department of Statistics and Center for the Neural Basis of Cognition, 5000 Forbes Ave., Baker Hall 132 Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: vventura@stat.cmu.edu).

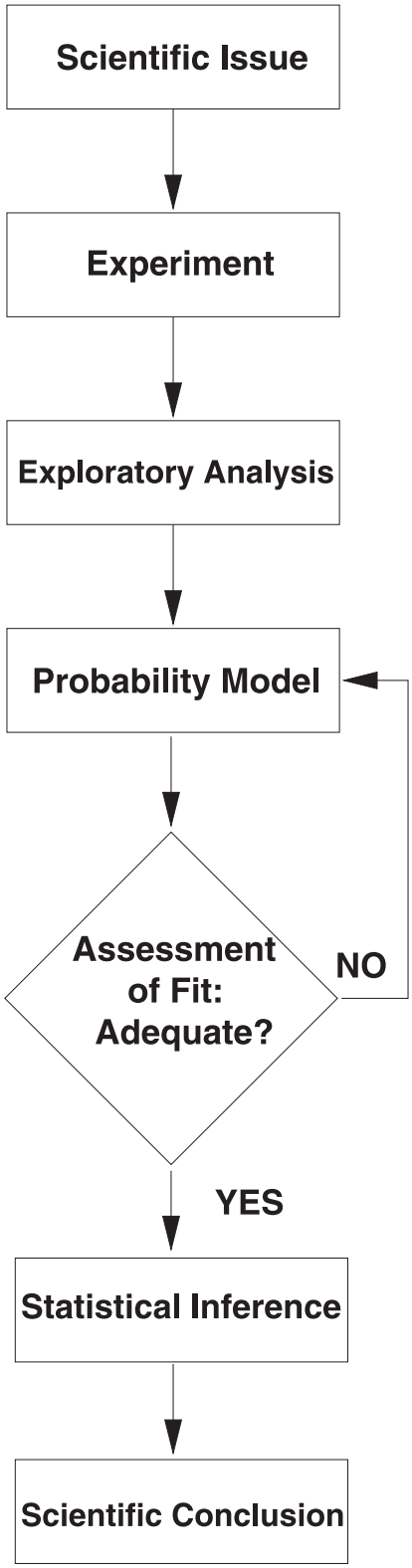

FIG. 1. Formal statistical inference within the process of drawing scientific conclusions. Probability model building is a prerequisite to formal inference procedures. Model building is iterative in the sense that tentative models must be assessed and, if necessary, improved or abandoned.

ian methods are able to use all of the available information to estimate an unknown parameter (such as a firing rate), 2) modern nonparametric methods, including the Bootstrap, are often applicable to neuronal data, and 3) many analyses based on a Poisson assumption can be extended to non-Poisson data.

INFORMATION AND STATISTICAL[1] EFFICIENCY

Spike trains recorded in vivo tend to be irregular both within and across trials. To capture the available information about stimulus or behavior it is helpful to formulate a reasonably accurate probability model for the noise. In probability theory stochastic sequences of event times are called *point processes*, and the simplest point process is the *Poisson process*. A basic property is that counts of Poisson process events follow Poisson distributions. Thus, the simplest assumption one might make about a spike train is that it may be described as a Poisson process, and then the resulting spike counts (the number of spikes in particular windows of time) would follow Poisson distributions. In this section we use Poisson spike counts for pedagogical purposes to discuss some fundamental statistical concepts.

We do not want to give the impression, however, that Poisson models are to be trusted without critical examination. Indeed, there are reasons to think it unlikely in principle that spike trains should follow Poisson processes. Assuming large numbers of excitatory and inhibitory inputs, in the time-homogeneous case (meaning that the inputs do not vary systematically across time, as they would with a time-varying stimulus), Gerstein and Mandlebrot (1964) showed that a simple model of voltage variation with a fixed spiking threshold leads to interspike intervals that follow an inverse Gaussian distribution. (See also Tuckwell 1988.) Spike trains that follow this model would be non-Poisson. There has been considerable documentation and discussion of the variability of spike trains and its sources (an early study is Smith and Smith 1965; see Shadlen and Newsome 1998, and the references therein). However, whether a particular set of spike train data should be considered Poisson is an empirical matter, subject to statistical examination. Later we discuss methods that either remove the Poisson assumption in treating counts or substitute a more general conception of an event–time process.

A parametric statistical model is a probability model with a fixed set of parameters that may be estimated from the data

For illustrative purposes we consider the analysis of count data that are assumed to follow a Poisson distribution. If μ is the mean number of times that a neuron fires during some specified time interval (a, b), then the probability of getting y spikes during that interval is

$$p(y|\mu) = e^{-\mu}\frac{\mu^y}{y!} \quad (1)$$

If we assume a particular value for μ we may use Eq. *1* to calculate the probability of observing any specified number of spikes during the interval (a, b). In this situation μ is called a *parameter*. Having observed n trials of data $y_1, y_2, \cdots, y_n$ the obvious estimate of μ would be the average number $\bar{y} = 1/n \sum_i y_i$ of spikes per trial during the interval (a, b). We call $\bar{y}$ an *estimator* of μ. It is not only intuitive, but is also an

[1] To those who want mathematical details on these topics we recommend the introductory text by Wasserman (2004).

example of a very general and generally very good method of estimation called *maximum likelihood* (ML), to which we will return shortly. We will, throughout, follow the standard statistical convention of using $\hat{\mu}$ to denote a generic estimator of μ, but often this estimator is obtained by ML.

Evaluation of an estimation procedure has four components

Although the sample mean is an intuitive estimator of the Poisson mean μ, one might dream up alternatives. For example, a property of the Poisson distribution is that its variance is also equal to μ; therefore, the sample variance $s^2 = 1/(n-1)\ \Sigma_i\ (y_i - \bar{y})^2$ could also be used to estimate the population variance μ. This may seem odd, and potentially inferior, on intuitive grounds because the whole point is to estimate the mean firing rate, not its variance. On the other hand, once we take the Poisson model seriously the population mean and variance become equal and, from a statistical perspective, it is reasonable to ask whether it is better to estimate one rather than the other from their sample analogues. Our purpose here is to present a simple analysis that demonstrates the inferiority of the sample variance to the sample mean as an estimator of the Poisson mean μ. We are going through this exercise so that we can draw an analogy to it later on.

A statistical estimator is itself subject to random variation: it is computed from a sample of data, and a new sample (a new set of n trials) would produce different data and therefore a different value of the estimator. To study the variation of an estimator we may calculate its expectation and variance, both obtained, theoretically, across repeated sets of n trials. The expectation of both $\bar{y}$ and s^2 is μ; on average, in repeated sets of trials, the positive errors tend to cancel the negative errors (they are both *unbiased*). But how accurate is s^2 compared to $\bar{y}$? Analytical calculation of the variance of each estimator gives[2]

$$V(\bar{y}) = \frac{\mu}{n}$$

$$V(s^2) = \frac{\mu}{n} + \frac{2\mu^2}{n-1}$$

where n is the number of repeated trials. Therefore, the variance of s^2 is always larger than that of $\bar{y}$ and it is, in this sense, less accurate: s^2 tends to be further from the correct value of μ than $\bar{y}$. For example, if we take $n = 100$ trials and $\mu = 10$, we find $V(\bar{y}) = 0.10$, whereas $V(s^2) = 2.12$. The estimator s^2 has about 21 times the variability of $\bar{y}$, so that estimating μ using s^2 would require about 2,100 trials of data to gain the same accuracy as using $\bar{y}$ with 100 trials. See Figure 2.

This simple analysis is compelling, as long as we are sure that the data follow a Poisson distribution. Because the distribution of real spike counts may well depart from Poisson, a realistic comparison of $\bar{y}$ versus s^2 should consider their behavior also under alternative assumptions. In this regard, the sample mean remains a reasonably good estimator of the population mean in large samples regardless of the probability distribution of the spike counts: the sample mean gets arbitrarily close to the population mean for sufficiently large samples (it is a *consistent* estimator). The sample variance, on the other hand, does so only if the population variance is truly equal to the population mean; otherwise, as the sample size increases it will converge to the wrong value (it is, in that case, an inconsistent estimator). We return to this issue below.

We now consider the theoretical framework that allows us to generalize these kinds of results to more complicated situations. We can list 4 components of formal statistical evaluation of an estimator: *1*) a quantity to be estimated, which is a numerical characteristic (in mathematical jargon, a functional) of a probability distribution; *2*) an estimator; *3*) a criterion according to which the accuracy of the estimator will be judged; and *4*) a set of assumptions about the (unknown) probability distribution, which allow us to carry out the evaluation of the estimator. In the Poisson case, *1*) μ, the average firing rate in our interval of time, was the quantity being estimated; *2*) $\bar{y}$ and s^2 were the 2 estimators being evaluated; *3*) we considered the variance as our criterion; and *4*) we assumed a Poisson distribution, with the trials being probabilistically identical and independent.

Concerning item *3*), in general terms, a simple and illuminating criterion is mean squared error (MSE), which is computed by squaring the error and then averaging across infinitely many hypothetical replications of the data. Formally, the MSE of an estimator $\hat{\mu}$ is written as

$$MSE(\hat{\mu}) = E[(\hat{\mu} - \mu)^2]$$

where $E(\cdot)$ represents the expectation (or expected value). By the linearity of the expectation we also obtain

$$MSE(\hat{\mu}) = \text{Bias}\ (\hat{\mu})^2 + V(\hat{\mu})$$

where Bias $(\hat{\mu}) = E(\hat{\mu}) - \mu$ represents the average amount by which the estimator differs from μ. This formula is the basis for what is called the "Bias–Variance trade-off" in comparing alternative methods. It often happens that one method has comparatively high bias, whereas the other has comparatively high variance; MSE takes both into consideration. In the case of $\bar{y}$ and s^2, the bias of both is zero, so the MSE is actually equal to the variance. Thus, the criterion we were applying was the MSE. MSE is commonly used in studies of alternative estimation methods.

Fisher information measures the optimal precision with which a parameter may be estimated from data

In a 1922 paper that laid the groundwork for much of the statistical theory developed in the 20th century, R. A. Fisher analyzed the general form of the estimation problem. Fisher observed that estimators often may be approximated by the sample mean of some suitably defined random variables so that, according to the central limit theorem, they will tend to be approximately normally distributed (Gaussian) for large sample sizes. Figure 2 shows a pair of histograms of $\bar{y}$ and s^2 values calculated from 1,000 randomly generated samples of size $n = 100$ when the true Poisson mean was $\mu = 10$. The *asymptotic normality* of the 2 estimators is indicated by the approximately normal shape of their histograms. (Much of statistical theory is asymptotic in the sense that it considers behavior when the sample size becomes arbitrarily large.) In the case of $\bar{y}$ and s^2,

[2] Our use of variance in this pair of equations may be confusing: here it refers to the variability of an estimator across repeated sets of n trials (i.e., across replications of the entire experiment); when s^2 is called the *sample variance,* the term "variance" instead refers to a computation of variability carried out across one set of n trials, within a particular experiment.

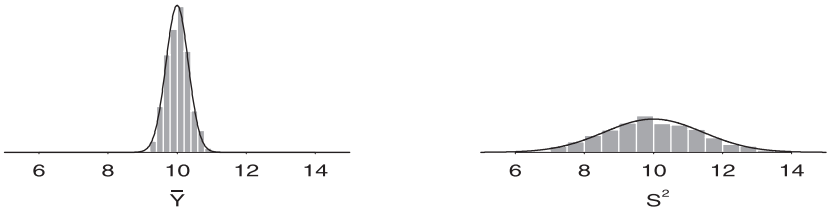

FIG. 2. Histograms displaying distributions of $\bar{y}$ and s^2 based on 1,000 randomly generated samples of size $n = 100$ from a Poisson distribution with mean parameter $\mu = 10$. In these repeated samples both $\bar{y}$ and s^2 have distributions that are approximately Normal (represented by the overlaid curves). Both distributions are centered at 10 but the values of s^2 fluctuate much more than do the values of $\bar{y}$.

both estimators are also centered at the true value of μ. In general, for large samples, the bias of any reasonable estimator will be small, so that it is at least centered *close* to the true value of the quantity it is estimating. Fisher considered the class of all estimators that were asymptotically normal, with the correct value as the asymptotic mean and asked what was the smallest possible variance. This is equivalent to asking what is the minimal MSE in large samples.

The answer is that the smallest possible variance is the reciprocal of what is now called *Fisher information.* Put differently, the Fisher information gives the best possible precision of any asymptotically "good" estimator (where precision is the reciprocal of variance). The statistical efficiency of an estimator is judged by its variance, or asymptotic variance, relative to the bound determined by Fisher information. If an estimator attains the bound it is said to be *efficient.* Fisher described efficient estimators by saying they contain the maximal amount of information supplied by the data about the value of a parameter. That is, the information in the data pertaining to the parameter value may be used well (or poorly) to make an estimator more (or less) accurate; in using as much information about the parameter as is possible, an efficient estimator uses the data most efficiently and reduces to a minimum the uncertainty attached to it.

For the Poisson model, the Fisher information about μ in n observations is

$$I(\mu) = n/\mu \tag{2}$$

and the large-sample variance bound μ/n is obtained by $\bar{y}$ so it is an efficient estimator. On the other hand, use of s^2 rather than $\bar{y}$ to estimate μ effectively discards 95% of the information in the data. In general, the Fisher information can be computed from the probability model, or more precisely, the likelihood, which we introduce in the next subsection.

Maximum likelihood estimators are efficient

In addition to deriving the information bound, Fisher showed that the method of maximum likelihood (ML) produces efficient estimators, i.e., as explained above, this means that an ML estimator (MLE) automatically has the smallest possible uncertainty, for large samples.

The MLE is the value of the parameter that maximizes the likelihood function. To explain what the likelihood function is, and why one might want to maximize it, let us consider the Poisson distribution. For any particular value of μ, *Eq. 1* gives the probability of observing y spikes during the interval (a, b) for a single trial. Table 1 displays the value of $p(y|\mu)$ for several values of y and μ. The argument of the probability density, that is, the quantity that is varying in *Eq. 1*, is the spike count y. Because *Eq. 1* is a probability density, the sum over all possible values of y is 1. In Table 1 this corresponds to summing across rows. The likelihood function reverses what is held fixed and what is varying: it applies $p(y|\mu)$ with the data y held fixed and the unknown parameter μ allowed to vary. Reasoning based on the likelihood function continues along this reversed, or "inverse," path: having observed a spike count y, we would find it implausible to think that the correct value of μ was one that would make y an improbable event. Instead, the MLE maximizes the likelihood function, producing the

TABLE 1. *Equation 1 evaluated at $y = 0, 1, \cdots$ for several values of μ, where μ is the theoretical mean spike count in the interval (a, b)*

	y								
μ	0	1	2	3	4	5	6	≥7	Products of Columns at $y = 4$ and at $y = 5$
3.5	3.0	10.6	18.5	21.6	18.9	13.2	7.7	6.5	2.5
3.75	2.4	8.8	16.5	20.7	19.4	14.5	9.1	8.6	2.8
4	1.8	7.3	14.7	19.5	19.5	15.6	10.4	11.2	3.0
4.25	1.4	6.1	12.9	18.3	19.4	16.5	11.7	13.7	3.2
4.5	1.1	5.0	11.2	16.9	19.0	17.1	12.8	16.9	3.3
4.75	0.9	4.1	9.8	15.5	18.4	17.4	13.8	20.1	3.2
5	0.7	3.4	8.4	14.0	17.5	17.5	14.6	23.9	3.1
5.25	0.5	2.8	7.2	12.7	16.6	17.4	15.3	27.5	2.9
5.5	0.4	2.2	6.2	11.3	15.6	17.1	15.7	31.5	2.7

All values are given as percentages (rounded to the nearest 0.1%). If $\mu = 4.5$, the probability of observing $y = 1$ spikes is 5% and the probability of observing $y = 2$ spikes is 11.2%. The greatest probability across the row $\mu = 4.5$ occurs for $y = 4$, where the probability of observing $y = 4$ spikes is 19%. Similarly, if $\mu = 5.25$, the most likely number of spikes is $y = 5$. The table also allows us to reverse the process, by reading down the columns rather than across the rows; we can thereby find the value of μ that is most likely to have given rise to a particular y. For example, the column $y = 4$ has its greatest value at $\mu = 4$. Thus if we observed $y = 4$, and if the only values of μ were those in the table, we would guess the unknown value of μ to be 4. If we observed $y = 5$, then we would guess that μ was 5. The values of μ thus guessed are in fact maximum likelihood estimates of μ based on a single observation y. The rows contain values of the probability density $p(y|\mu)$ for various potential values of y, which must add up to 1. The columns are values of the likelihood function for various values of μ, based on the observation y. The columns typically do not add up to 1. In the case of $n = 2$ trials, *Eq. 3* must be applied. For example, if we observed $y_1 = 5$ spikes on trial 1 and $y_2 = 4$ spikes on trial 2, and assuming that the trials are independent, so that $L(\mu) = p(y|\mu) = \prod_{i=1}^{n} p(y_i|\mu)$, the likelihood function for μ would be equal to the product of the columns at $y = 5$ and at $y = 4$. In this case the maximum would be achieved at $\mu = 4.5$, assuming once again that the only possible values of μ are those in the table.

parameter value μ that makes y the most probable to have occurred.

Suppose, now, that we have spike counts $y_1, \cdots, y_n$ for n trials. Let us denote the entire set of spike counts, now taken to be a vector, as $\mathbf{y} = (y_1, \cdots, y_n)$. The *likelihood function* is then based on the joint distribution of $\mathbf{y}$

$$L(\mu) = p(\mathbf{y}|\mu) \tag{3}$$

and the MLE of μ is the value that maximizes $L(\mu)$, that is the value of μ that gives the observed data its maximal probability of occurring under the assumed probability model. Here, as is usually the case, the parameter μ is allowed to vary continuously, and the MLE is usually obtained numerically by maximizing the *loglikelihood function*

$$\log \mathrm{L}(\mu) = \log L(\mu)$$

The MLE is the value of the parameter that provides the best fit of the model to the data, where "fit" is defined in terms of the probability assigned to the data by the model. There are other possible ways to define fit, but in his 1922 paper, Fisher pointed out that the MLE is efficient, in the sense stated earlier. Generally, the great virtues of ML estimation are *1*) for large samples it uses the maximal information (in Fisher's sense) available in the data, *2*) there are explicit algorithms to compute it even in complicated settings, and *3*) there is also an explicit method of assessing the uncertainty of MLEs.

In the Poisson case, an easy application of calculus shows that the MLE of μ is $\bar{y}$. Earlier we showed that, for estimating the Poisson parameter μ, the mean is much better than the variance, and then we stated that the mean is efficient in the sense of having the smallest possible MSE in large samples. We see here that the efficiency of $\bar{y}$ may be considered a special case of the general result that the MLE is efficient.

Bayes estimators are also optimal in large samples

Closely related to ML estimation, and more powerful in some circumstances, is Bayesian inference. Bayesian inference is based on Bayes' theorem, which is an elementary formula for computing conditional probabilities of the form $P(A|B)$ from probabilities of the form $P(B|A)$. The profound implication of Bayes' theorem for statistical inference becomes apparent when A signifies an unknown parameter and B signifies the available data: having observed some data y, Bayes' theorem allows us to express our uncertain knowledge about the parameter μ quantitatively, in the form of a probability distribution $p(\mu|y)$. The inputs to Bayes' theorem are the likelihood function $p(y|\mu)$ and a *prior* distribution $p(\mu)$ that represents a priori knowledge about μ. To contrast with the prior distribution, which represents knowledge that logically precedes the data, the distribution $p(\mu|y)$ produced using the data is called the *posterior*. The usual *Bayes estimator* is the mean of this posterior distribution.

In the Poisson case, Bayes' theorem provides $p(\mu|y)$ in terms of the formula for $p(y|\mu)$ given by *Eq. 1*. Specifically, according to Bayes' theorem, the probability of any set of values of μ is determined by the equation

$$p(\mu|\mathbf{y}) = \frac{p(\mathbf{y}|\mu)p(\mu)}{\int p(\mathbf{y}|\mu)p(\mu)\mathrm{d}\mu} \tag{4}$$

In words, *Eq. 4* says that after the data have been observed, the posterior probability density of μ is the normalized product of the likelihood function and the prior. The prior density of μ is often taken to be slowly varying across relevant parameter values, indicating that little is to be assumed a priori about the potential value of the parameter. For example, in estimating a Poisson mean firing rate, the prior probability density might be taken to be slowly varying across all conceivably realistic values of the firing rate. In such cases the posterior is determined almost entirely by the likelihood function and Bayes estimates become very close to MLEs. For this reason, Bayes estimates share with MLEs the 3 desirable properties listed above. Furthermore, when there is good a priori information, as in many decoding problems where one may assume the stimulus to be varying smoothly over time (see section on Bayesian decoding methods), Bayesian methods can incorporate this information to produce better procedures. In addition, Bayesian methods sometimes offer computational advantages because Monte Carlo simulation methods can be used to implement them.

It is easy to obtain standard errors for maximum likelihood and Bayesian estimation

Earlier we noted that Fisher information is reciprocal of the minimal possible variance of any "good" estimator, and that ML and Bayesian estimators achieve this bound. It follows that Fisher information may be used to obtain SEs for ML and Bayesian estimators. In practice, it is convenient to substitute the *observed information* instead. In the Poisson case we replace μ with $\hat{\mu} = \bar{y}$ in *Eq. 2* to obtain

$$SE(\hat{\mu}) = \sqrt{\bar{y}/n} \tag{5}$$

If the Poisson assumption is correct, an approximate 95% confidence interval can be obtained by inserting *Eq. 5* into the general formula

$$\hat{\mu} \pm 2SE(\hat{\mu}) \tag{6}$$

Note that the *SE* formula in *Eq. 5* is not the same as the usual formula for the SE of the sample mean

$$SE(\bar{y}) = \sqrt{s^2/n} \tag{7}$$

Equation 7 does not require the Poisson assumption, but it applies only to the sample mean, whereas *Eq. 5* is an instance of a widely applicable formula for the SE of a ML or Bayesian estimator.[3] This greatly enhances their practical utility. A method that may be used with virtually any estimator, and does not require a specific distributional assumption, is discussed in NONPARAMETRIC METHODS.

Likelihood ratio tests have good statistical power

To illustrate statistical considerations concerning estimation procedures, earlier we discussed estimation of the Poisson mean μ based on n repeated trials. Now suppose, under the same circumstances, we wish to test the null hypothesis H_0 : $\mu = \mu_0$. For example, μ_0 could be the baseline firing rate of a neuron and we may wish to demonstrate that some stimulus increases (or decreases) this rate, so that the alternative hy-

[3] In making this statement we are blurring the distinction between SEs in the "frequentist" and Bayesian senses (see Wasserman 2004).

pothesis $H_A : \mu \neq \mu_0$ would hold. In analyzing this situation we will follow a standard statistical convention, which we have ignored up to this point: capital and lowercase versions of a letter are used to distinguish a random variable ($\bar{Y}$) from a particular value ($\bar{y}$) taken by the random variable.

The obvious procedure to assess whether H_0 holds would be to examine $\bar{y}$ and if it is sufficiently far from μ_0, reject H_0. That is, we would define a quantity c and reject H_0 whenever $|\bar{y} - \mu_0| > c$. To determine c we usually require the probability of falsely rejecting H_0 to be small: under H_0 (that is, assuming H_0 were true), we would obtain c such that Prob (reject $H_0|H_0$ true) = Prob ($|\bar{Y} = \mu_0| > c$) = α, for a suitably small α, assuming each trial's spike count Y has a Poisson distribution with $\mu = \mu_0$. The usual criterion is to take $\alpha = 0.05$. This is called the *size* of the hypothesis test, or the probability of a *type I error.* When using data to assess H_0 a good practice is to report the *P*-value, which is the smallest value of α according to which the observed data would reject H_0. For example, one might report $P = 0.02$ rather than simply $P < 0.05$.

To compute Prob ($|\bar{Y} - \mu_0| > c$) one could rely on the approximate normality of the sample mean, which is likely to be adequately accurate unless μ_0 and n are both small. Specifically, under H_0, because each Y has mean μ_0 and variance μ_0, it follows that $\bar{Y}$ has mean μ_0 and variance μ_0/n; therefore the calculation of Prob ($|\bar{Y} - \mu_0| > c$) may be based on a normal distribution with mean 0 and variance μ_0/n, so that $c = 2\sqrt{\mu_0/n}$ when $\alpha = 0.05$.

Alternatively, if there are doubts about the adequacy of this approximation, Prob ($|\bar{Y} - \mu_0| > c$) may be obtained (even for small n and μ_0) by computer simulation. We simulate samples $U_1, U_2, \cdots, U_n$, where each U_1 has a Poisson distribution with mean μ_0 and for each sample compute the sample mean $\bar{U}$. If we repeat this procedure a large number of times to obtain a large number of samples (say, 10,000) we can then approximate Prob ($|\bar{Y} - \mu_0| > c$) accurately by the proportion of samples for which $|\bar{U} - u_0| > c$. This is an example of a *parametric Bootstrap* test. We discuss the nonparametric Bootstrap in the next section. Simulation-based calculation of *P*-values is an important technique because it applies to situations where normal approximations (or other, similar approximations) are either unavailable or of dubious validity.

In complicated settings there may not be an obvious statistical test, or there may be several (as in the hypothetical example of alternative estimators for the Poisson mean μ used above), or one may want some reassurance that, as with ML estimation, the method is a good one. The general approach based on the likelihood function is to define the ratio of the likelihood under H_0 to that under H_A and to reject H_0 whenever this ratio is sufficiently small. This procedure is called the *likelihood ratio test.* Thus, to test $H_0 : \mu = \mu_0$ versus $H_A : \mu = \mu_1$ we would use $LR = L(\mu_0)/L(\mu_1)$ and reject H_0 when $LR < c$, where c is chosen so that Prob ($LR < c$) = α when H_0 is true. Here, μ_1 would be some prespecified alternative value of the Poisson mean. Usually, however, no such specific alternative to μ_0 is apparent and the more generic alternative $\mu \neq \mu_0$ is used. In this case the likelihood ratio $L(\mu_0)/L(\hat{\mu})$ is used, where $\hat{\mu}$ is the MLE. For the Poisson case $H_0 : \mu = \mu_0$, it turns out that the likelihood ratio test yields the "obvious" procedure based on $\bar{y}$ discussed above. The likelihood ratio test produces, similarly, many familiar statistical tests (such as *t*-tests and *F*-tests).

When 2 distinct procedures are available for testing the same null hypothesis they may be compared by first aligning them to have the same size (i.e., the same value of α) and then computing their *power,* which is the probability of correctly rejecting H_0 for particular values of the parameter that satisfy H_A. Extensive theory and simulation studies have shown that the likelihood ratio test tends to have good power, often nearly or exactly the best possible power, in a wide variety of situations, at least when sample sizes are reasonably large (e.g., van der Vaart 1998). It is therefore the most commonly applied statistical testing procedure when there are explicit parametric probability models under both hypotheses. Bayesian methods of hypothesis testing are beyond the scope of our review here, but they have received considerable attention in recent years and the interested reader should consult Kass and Raftery (1995); an application to neuronal data analysis is in Behseta and Kass (2005).

Applicability of a probability model to a data set should be assessed

Previous subsections have been concerned with fundamental principles of statistical inference. In our introductory discussion, however, we emphasized the inductive process diagrammed in Fig. 1 and, according to that figure, statistical inferences should occur only after determination that the probability model on which the inferences will be based is adequate. A variety of procedures have been developed to assess the adequacy of a probability model in representing the regularity and variability in a particular set of data. One method is to consider a more elaborate model and conduct a likelihood ratio test to choose between the simpler model and the more elaborate model. A second approach is to compare suitable features of the data to predictions made by the model, and a general procedure for doing so is to examine a set of functions of the data that would, if the probability model were correct, constitute a sample from a particular probability distribution, say $F(y)$; when the function values are ordered and plotted against theoretical quantiles of the putative probability distribution $F(y)$, the result, called a Q–Q (for quantile–quantile) plot, should be roughly linear (e.g., Hogg and Tanis 2001). Both of these methods are discussed later in POISSON AND NON-POISSON MODELS. In addition, chi-squared goodness-of-fit statistics may be used to evaluate fit when applicable (e.g., Sokal and Rohlf 1995).

NONPARAMETRIC METHODS

In the previous section we discussed parametric statistical methods, first considering estimation and then describing hypothesis testing based on the likelihood ratio test. In the following subsection we review very briefly the purpose of nonparametric methods, and a bit of the terminology, so that readers will be equipped to tackle the literature. We next describe a very general approach to assessing uncertainty, called the *Bootstrap,* illustrating it by revisiting the simple hypothesis testing problem posed earlier.

Modern nonparametric methods allow greater flexibility in probability models at the cost of some loss of efficiency

The Poisson distribution uses a single parameter μ. Other distributions or models, such as linear regression models, typically depend on a small number of parameters that must be estimated from (or "fitted to") the data at hand. It is important to consider carefully what might happen if the probability model were to be incorrect in some plausible way. For example, as we already mentioned, in the case of estimating μ, the sample mean $\bar{y}$ remains consistent (and the MSE becomes arbitrarily small with sufficiently much data) for non-Poisson data, whereas s^2 will estimate the population variance, which may differ from the population mean. Often a restrictive model, such as the Poisson, is replaced with a more flexible model. Highly flexible models are often called "nonparametric."

Classical nonparametric methods substitute transformed versions of the data (such as ranks), to obtain analogues of standard elementary methods (such as ANOVA). Modern nonparametric methods use models with large numbers of parameters, and their theoretical analysis often assumes there are infinitely many parameters. In some contexts, the parameters are used to characterize a probability distribution, as in the Bootstrap, which we discuss next. Sometimes, large numbers of parameters are used to replace linear relationships with nonlinear relationships whose form is not picked a priori but rather is determined from the data. For example, curve fitting is often accomplished with very flexible models (having large numbers of adjustable parameters), a process statisticians usually call *nonparametric regression.* We discuss the application of generalized nonparametric regression to neuronal data in the next section, where we view the PSTH as a nonparametric estimator of the firing rate function and contrast it with more efficient alternatives.

Nonparametric methods apply more generally than parametric methods. This generality comes at a cost: when the assumptions of a parametric model hold (or hold to a close approximation), they will be more efficient than nonparametric competitors. In deciding whether to apply a parametric or a nonparametric method, therefore, one must consider the likelihood of a substantial departure from the parametric assumptions. The notion of "substantial departure" will depend on the context, and some parametric methods are more likely than others to perform poorly. Numerical simulations may be used to determine performance of nonparametric methods in various situations (compared, say, to ML, which would be fully efficient). Some modern nonparametric methods can perform very well, with relatively little loss of efficiency (e.g., van der Vaart 1998). An example concerning the fitting of directional tuning curves is discussed in the next section.

The Bootstrap is a general nonparametric method of assessing uncertainty

Earlier we pointed to the use of simulation in computing *P*-values. There we were discussing the parametric case in which, under H_0, the probability distribution of each observation Y_i was Poisson with parameter $\mu = \mu_0$. However, because neuronal spike counts are known to exhibit non-Poisson behavior, it may be desirable to use a procedure that does not assume Poisson counts. For this purpose, the *translated* observed spike counts $Y_1 - (\bar{Y} - \mu_0)$, $Y_2 - (\bar{Y} - \mu_0)$, $\cdots$, $Y_n - (\bar{Y} - \mu_0)$, with Y_i being the count for the ith trial, may be repeatedly *resampled*. That is, we draw observations U_1, U_2, $\cdots$, U_n where each U_i takes on the value of one of the observed values of $Y_i - (\bar{Y} - \mu_0)$, and all such values occur with probability $1/n$. This is known as sampling *with replacement.* The reason for the translation is explained below. Having thereby obtained a large number of samples we then (as with the Poisson U_i values in the previous section) compute the proportion of samples for which $|\bar{U} - \mu_0| > c$, which, again, estimates the desired Prob $(|\bar{Y} - \mu_0| > c)$. This method is known as the *Bootstrap.* It is nonparametric because it does not require the assumption of a specific parametric family of distributions (here, Poisson). Instead, it requires the data values Y_i to be *independent replications from the same probability distribution.* It works because it implicitly uses the data Y_1, Y_2, $\cdots$, Y_n to estimate the probability distribution from which they are assumed to be drawn, and then computes the desired probability from this estimate. With a large simulation, the Poisson-based parametric boot strap method can compute Prob $(|\bar{Y} - \mu_0| > c)$ with arbitrarily good accuracy. The nonparametric Bootstrap *P*-value, however, is accurate only for large sample sizes n. The great value of the Bootstrap is that it may be applied in many complicated situations. We describe its use in testing for correlated activity between 2 neurons in CORRELATED PAIRS OF NEURONS. Furthermore, a substantial literature has demonstrated both theoretically and in numerical studies that it is widely effective (Davison and Hinkley 1997; Efron and Tibshirani 1993). In addition to treating the testing problem, the Bootstrap may be used to obtain SEs and confidence intervals.

The simplicity of the Bootstrap conceals an important point: its properties depend on the precise manner in which the resampling is conducted; arbitrary shuffles of the data do not necessarily accomplish desired statistical goals. In hypothesis testing, the *P*-value must be obtained under the hypothetical reality imposed by H_0. For example, in the problem discussed above, we wished to compute the hypothetical probability Prob $(|\bar{Y} - \mu_0| > c)$ under the supposition that the observations had mean $\mu = \mu_0$. We cannot merely sample the Y_i values because their mean μ may not equal μ_0. (Indeed, this is precisely the possibility we are examining when we perform the statistical hypothesis test.) Instead, the calculation may be based on new data that resemble the Y_i values except that their mean is shifted to equal μ_0. To produce a version of Y_i that has mean μ_0 rather than μ we would, in principle, want to subtract from Y_i the quantity $\mu - \mu_0$. Because we do not know the value of μ we instead subtract $(\bar{Y} - \mu_0)$. Thus, in the procedure outlined above, the observed values of $Y_i - (\bar{Y} - \mu_0)$ were sampled.

STATISTICAL SUMMARIES OF FIRING RATE

The PSTH communicates firing rate and its evolution across time

Within an experiment, each trial's set of recordings from a single neuron produces a different spike train. By recording many repeated trials the regularity in the neuron's response to a stimulus, or the production of some behavior, may be obtained. A raster plot displays the complete set of spike times

for all trials for a single neuron in a particular experimental condition, whereas the peristimulus time histogram (PSTH) accumulates these to show both the overall activity of the neuron and the way its firing rate varies across time (Gerstein and Kiang 1960).

One reason the PSTH works well as a visual summary of the data is that our eye is able to pick up trends, smoothing the PSTH so that we see the temporal evolution of the firing rate. In Fig. 3 we have overlayed a smooth curve on a PSTH, based on data recorded from a locust antennal lobe neuron. The PSTH jumps somewhat erratically from bin to bin because of noise and does not track the kind of continously varying firing rate one would expect. For this reason, the curve is closer to what we understand from looking at the PSTH than is the PSTH itself.

We will later discuss the method we used to generate the fitted curve in Fig. 3. We should emphasize our use of the modifier "fitted." A fundamental conceptualization in statistical theory distinguishes the unknown "true" firing rate curve (which would result from the hypothetical possibility of running infinitely many trials) from an *estimate* of it obtained from actual data, in the sense described in *Information and statistical efficiency.* Furthermore, when we speak of estimating the firing rate we mean that we will use the data to produce an estimate of the *instantaneous* firing rate at each time *t,* where t varies across the whole range of experimentally interesting values. We write the instantaneous firing rate as $\lambda(t)$ and are interested in estimating the entire curve described by $\lambda(t)$ for t in some interval, which we write as $[A, B]$. In statistics, one usually writes an estimate obtained from data with a hat, so an estimate of the firing-rate function would be written $\hat{\lambda}(t)$.

Conceptually, $\lambda(t)$ is the *trial-averaged* firing-rate function, as opposed to the within-trial firing-rate function. As we note later, the latter may include effects resulting from "membrane memory" (the refractory period) or erratic bursting (bursting that is not time-locked to experimental stimulus or behavior), so that at time t it may depend on the precise timing of spikes that occur before time t. The function $\lambda(t)$ represents the tendency to fire at time t that a neuron would have if such memory or erratic bursting effects were removed, as they would be by averaging (hypothetically) over infinitely many trials. It is often, implicitly, considered to be a representation of the firing of large numbers of independently acting neurons similar to the one being recorded (each trial effectively representing the activity of one such similar neuron). Our focus, in the remainder of this section, on estimation of $\lambda(t)$ is not meant to imply that every analysis should begin with a time-domain representation of firing rate. In many situations, especially when oscillatory stimuli are used, frequency-domain analyses can produce valuable interpretations of the data (Brillinger 1992; Mechler et al. 1998; Pesaran et al. 2002).

Some neuroscience questions involve instantaneous firing rate

Sometimes, questions of scientific interest are posed naturally in terms of instantaneous firing rate. For example, Olson et al. (2000) examined neurons in the supplementary eye field (SEF) when a monkey moved his eyes in response to either an explicit external cue (the point to which the eyes were to move was illuminated) or an internally generated translation of a complex cue (a particular pattern at a fixation point determined the location to which the monkey was to move his eyes). In one part of their study, they were interested in the time at which maximal firing rate was achieved, and the delay of this maximum for the internally generated cue compared to the external cue. Formally, the problem is to determine the value of t that maximizes $\lambda(t)$. Similarly, when the maximal firing rate, or the difference between the maximal rate and a baseline rate, is of interest the problem involves estimation of $\lambda(t)$.

Probability models in terms of instantaneous firing rate make efficient use of the data

A second reason for estimating the instantaneous firing rate $\lambda(t)$ is that it appears in probability models and, as we have already noted, when coupled with maximum likelihood or Bayesian methods of estimation, probability models reduce the data efficiently. For example, if we assume a set of n spike times $s_1, \cdots, s_n$ follow a time-varying Poisson process then their probability density is

$$p(s_1, \cdots, s_n) = e^{-\int_A^B \lambda(t)\mathrm{d}t} \prod_{k=1}^{n} \lambda(s_k) \tag{8}$$

We discuss Poisson processes in POISSON AND NON-POISSON MODELS. Here, our point is that to apply the formula in *Eq. 8*, which is used in many theoretical and data-analytic calculations, we must be able to obtain values of $\lambda(t)$ for various values of t. For this purpose the function $\lambda(t)$ would have to be estimated from the data.

As an estimate of firing rate the PSTH can be improved by smoothing

It would be possible to use the PSTH as an estimate of $\lambda(t)$. However, the PSTH is relatively noisy and, under the assump-

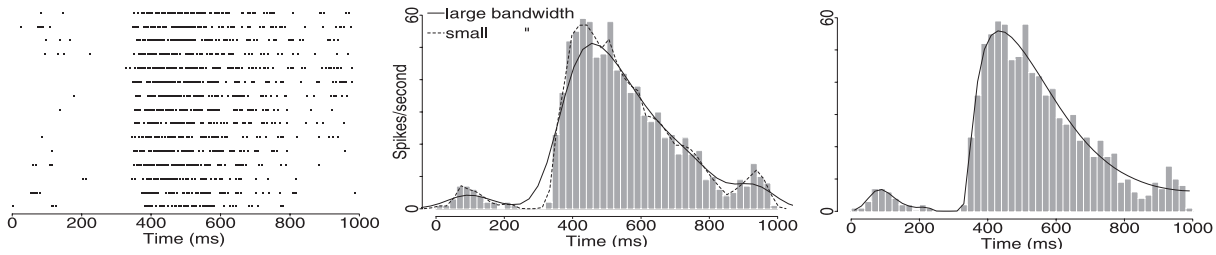

FIG. 3. Estimates of $\lambda(t)$ based on smoothed versions of the peristimulus time histogram (PSTH), for data from a locust antennal lobe neuron in response to a stimulating odor. *Left:* raster plot of the 15 trials. *Center:* PSTH with 2 Gaussian filter fits, one having small bandwidth and one having larger bandwidth, the former of which undersmooths where the firing rate varies slowly, whereas the latter oversmooths where the firing rate varies rapidly. *Right:* PSTH and Bayesian adaptive regression splines (BARS) fit.

tion that $\lambda(t)$ is itself smooth, it is possible to produce much better estimates $\hat{\lambda}(t)$ by smoothing ("filtering"). Kass et al. (2003) provided an illustration in which smoothing increased the efficiency of estimating $\lambda(t)$, compared to the PSTH, by a factor of 14. This situation is analogous to the estimation of the Poisson mean μ we considered earlier. Using the PSTH to estimate $\lambda(t)$ is like using s^2 to estimate μ: it would take about 140 trials to get the same accuracy with the PSTH that may be achieved by only 10 trials with a smoothed version of the PSTH. Not all cases are as dramatic as a 14-fold improvement, but in our experience smoothing is likely to provide gains of at least severalfold.

There are many ways to produce a smooth firing-rate function; spline-based methods work very well

One simple way to reduce the noisiness of the PSTH is to use a moving average: one may pick a suitable time window δ (such as 30 ms) and at each time t, average the values of the PSTH between $t - \delta$ and $t + \delta$. A variant of this, often called a Gaussian filter (or, in the statistics literature, a Gaussian kernel density estimator), uses a weighted average, putting weights (defined by a Gaussian probability density) on each PSTH value that decrease as the values to be averaged get further away from time t.

When $\lambda(t)$ varies slowly, Gaussian filters do a good job of estimating it. When the firing rate varies quickly, however, Gaussian filters are unable to capture the variation without introducing artificial high-frequency fluctuations. An illustration is given in Fig. 3, which uses data from an experiment on olfactory coding in the locust antennal lobe (Stopfer et al. 2003). Gaussian filters can capture a rapid jump in firing rate only by allowing noise in time periods where the firing rate is varying slowly. When the bandwidth is increased so that the noise is filtered out, the rapid jump in firing rate is underestimated. In other words, to filter high-frequency noise, the Gaussian filter must remove the high-frequency signal. The problem here cannot be solved by a fixed-bandwidth filter. A better estimate of $\lambda(t)$, produced by a method called BARS, is given in the *third panel* of Fig. 3. It has the desirable characteristic of strongly smoothing the firing rate function, while allowing sudden increases or decreases—BARS effectively filters high-frequency noise while retaining high-frequency signal. BARS stands for Bayesian adaptive regression splines (DiMatteo et al. 2001). A spline is a collection of polynomial curves that are joined at selected points (here, time points) called "knots." BARS uses a Bayesian Monte Carlo method to allocate knots optimally, to adapt to sharp variations in the intensity function. An application of BARS to an analysis of many single-unit firing rate functions is provided by Behseta et al. (2005).

An alternative approach to estimating $\lambda(t)$ is to define a plausible parametric form for it, with a small number of parameters, and then estimate these parameters (by ML). For example, Olson et al. (2000) used 6 parameters to characterize firing rate intensity functions in 84 single units from the supplementary eye field. However, nonparametric approaches using BARS or related methods (cf. Hansen and Kooperberg 2002; Loader 1999) are often easier to implement, remain accurate when the parametric form is incorrect, and suffer relatively little loss of efficiency even when the parametric form holds. For example, in a closely related context, Kaufman et al. (2005) considered the problem of fitting tuning curves to spike count data collected during wrist movement in 8 2-dimensional directions, with the goal of relaxing the usual assumption of cosine tuning. Kaufman et al. modified BARS to make it applicable to functions defined on a circle, and they showed in simulation studies that this nonparametric method was almost as efficient as cosine tuning when the true tuning function was exactly of cosine form, while having the advantage of being able to fit departures from cosine tuning. It is also possible to use partially parametric fitting methods. Barbieri et al. (2001) showed how Zernike polynomials may be used to characterize hippocampal place fields. These involve both Gaussian and non-Gaussian components, and are able to capture departures from Gaussian place field tuning.

An important subtlety is that all smoothing methods require a choice of degree of smoothness. For example, a histogram requires a choice of bin width and a Gaussian filter requires a choice of bandwidth. These can be determined by statistical methods, but they are often selected based on visual appearance of results, possibly with some past experience in mind. To make a sensible choice, the data analyst must consider limits on the rate at which $\lambda(t)$ can change, based on physiology and properties of the stimulus or behavior. For example, one might ask whether the firing rate is likely to jump substantially within 10 ms, throughout the time interval during which the recording is made. If so (if, for instance, the stimulus is itself rapidly fluctuating), then a less smooth estimate of $\lambda(t)$ is desirable than if slow variations of $\lambda(t)$ are to be expected. More recently developed methods, including BARS, are able to use the data to determine smoothness across the time domain, but they always involve assumptions about how rapid the fluctuation of $\lambda(t)$ might be.

POPULATION CODING AND DECODING

Population coding refers to the information contained in the combined activity of multiple neurons. The general challenge of decoding, meaning to extract this information, reverses the process: it is to determine from a set of spike trains (obtained from multiple neurons) the experimental condition or stimulus that produced them. In the simplest case, some small set of conditions is used, and the problem is to infer which condition led to the observed response. Decoding problems arise in several contexts. Examples include position representation by ensembles of rat hippocampal neurons (Brown et al. 1998; Zhang et al. 1998), velocity encoding by fly H1 neurons (Bialek et al. 1991), velocity and position encoding by M1 neurons (Georgopoulos et al. 1986; Moran and Schwartz 1999; Serruya et al. 2002), and natural scene representations by cat lateral geniculate nucleus neurons (Stanley et al. 1999). Aside from their use to study how populations of neurons represent information, decoding algorithms are being studied for their use in the brain-controlled neural prosthetic devices (Schwartz et al. 2004; Serruya et al. 2002, Wessberg et al. 2000).

Statistical analysis proceeds in 2 stages. First, the neural firing that results from the relevant stimulus or behavioral condition must be estimated from some history of data. Second, this relationship must somehow be "inverted" to provide a prediction of stimulus or behavior based on neuronal activity. These 2 stages may be intertwined, but are conceptually and,

usually, analytically separable. The first is sometimes called "learning," "estimation," or "encoding." The second is often called "prediction" or "decoding." In many situations the encoding step takes place in a steady-state environment. This is the most common framework for "supervised learning." With brain-controlled ("closed loop") movement, however, the evolution of the system's state is important because the subject will continue to adapt over time (Taylor et al. 2002; Wessberg et al. 2000) and part of the goal is to enable the subject to learn.

After reviewing some simple, linear decoding methods we discuss Bayesian decoding. The Bayesian decoding framework is based on "state-space" modeling, where data (spike counts) are assumed to depend on an underlying internal state (such as planned velocity of hand movement), which in turn is assumed to vary according to some specified dynamic model. We find the framework appealing because it facilitates incorporation of useful particulars, such as spike count distributions and movement smoothness constraints. In addition it is widely applicable because it can accommodate internal states that are multidimensional and relationships of data to state that are highly nonlinear.

The stimulus may be reverse-predicted from spiking activity using linear regression

Reverse regression (also called reverse correlation) is a very simple and widely used decoding method (e.g., Stanley et al. 1999; Warland et al. 1997). The "reverse" part of the terminology comes from the reversal of roles played by the stimulus and spike-activity response: the spike count data are treated as if they were the inputs, i.e., the fixed explanatory variables (the x values in the usual regression notation), whereas the stimulus is considered the output, i.e., the response variable (the Y variable, which in the regression formulation is assumed to be subject to random error).

Let us first describe the procedure as it might be applied to data from a single neuron. The explanatory variables are defined by forming a series of successive bins of spike counts at some suitable resolution, such as 30 ms. Thus, $(x_1, x_2, \cdots, x_T)$ would represent the vector of spike counts in T successive bins after the stimulus. Given a training set of many stimulus and firing-rate combinations one computes the usual least-squares coefficients

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{9}$$

where Y is the vector of observed stimulus values and the ith row of the matrix X is the spike count vector corresponding to the ith stimulus value. The predictor of a new, unobserved stimulus y_* given a spike count vector x_* is then

$$\hat{y} = x_* \hat{\beta} \tag{10}$$

In the case of N neurons all NT spike counts are used as explanatory variables and then *Eqs. 9* and *10* are applied. In some applications nonparametric regression methods are used in place of linear regression (Warland et al. 1997).

The population vector algorithm is simple and often effective

One widely successful decoding method is the population vector algorithm (PVA; Georgopoulos et al. 1982, 1986) and its modifications (Taylor et al. 2002). The PVA has enabled investigation of cortical control of arm movement (e.g., Moran and Schwartz 1999) including phenomena such as illusion perception (Schwartz et al. 2004) and has also been used in quite different contexts, such as the representation of moving tactile stimuli in sensory cortex (Ruiz et al. 1995). The method originated from the observation that motor cortex neurons are directionally tuned, with broad tuning curves that may be characterized reasonably well by 2 parameters, average firing rate and preferred direction. The preferred direction $\vec{D}$ is the direction in which the neuron's firing rate is highest. The preferred directions are obtained by fitting tuning curves to observed training data. Once one knows a neuron's preferred direction and its average firing rate, its firing rate for an arbitrarily specified direction may be determined, subject to some error. The PVA reverses the process, predicting direction from firing rate, by combining the observed activity from a large number of neurons. Specifically, the movement velocity $\vec{v}$ at a suitable time lag τ after spiking activity is predicted by the "population vector" $\vec{P}$ according to the equation

$$\vec{P}(t + \tau) = \sum w_i(t) \vec{D}_i \tag{11}$$

where $\vec{D}_i$ is the ith neuron's preferred direction and the "weight" $w_i(t)$ is the neuron's firing rate at time t (after being normalized in some fashion). The intuitive interpretation of the population vector is that it represents the population when each neuron sends in its "vote" for its preferred direction, which is then weighted by its activity.

The population vector algorithm is a special case of linear regression

We may connect the PVA to least-squares regression by thinking of the preferred direction $\vec{D}_i$ as an explanatory variable, which, for a given velocity $\vec{v}$, predicts spiking activity w_i according to the linear regression model

$$w_i = b_{i0} + \vec{D}_i \cdot \vec{v} + \varepsilon_i$$

where the coefficients b_{i0} are intercepts representing baseline firing rate and $\vec{D}_i \cdot \vec{v}$ is the dot product. If we now collect the firing rates together into a vector $Y = (w_1, \cdots, w_n)$ and the preferred directions (along with the intercepts) into a corresponding matrix X, once we take $\beta = \vec{v}$ we obtain the usual matrix form of the linear regression model

$$Y = X\beta + \varepsilon \tag{12}$$

as a description of the dependency of spiking activity on preferred direction.[4] Furthermore, we also have

$$X^T Y = \sum w_i(t) \vec{D}_i \tag{13}$$

which is the PVA prediction in *Eq. 11*. Now, if we assume that the preferred directions are uniformly distributed, then it may be shown that the matrix $X^T X$ reduces to the identity matrix. In this case, the least-squares fit for the velocity vector $\beta = \vec{v}$ given by *Eq. 9* reduces to the PVA in *Eq. 13*. Having made this observation, one may immediately generalize to the case in

[4] The PVA may be considered a special case of linear regression, but it is not a special case of reverse regression. In both *Eq. 12* and reverse regression, the role of stimulus is being played by velocity. In *Eq. 12*, however, velocity becomes the parameter β rather than the response variable y, which it is in reverse regression.

which the preferred directions are non-uniform by replacing the PVA estimator X^TY in *Eq. 13* with the least-squares estimator $(X^TX)^{-1}X^TY$ in *Eq. 9*. In addition, it should be noted that ordinary least squares assumes the responses, in this case the firing rates of the various neurons, have equal variances and are independent. Otherwise, classical statistical analysis shows that *weighted* least squares (WLS) should be used

$$\hat{\beta}_{\Sigma} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

where Σ is the variance matrix of Y (Kutner et al. 2004). Salinas and Abbott (1994) refer to the WLS estimator as the optimal linear estimator.

Bayesian decoding methods are efficient and flexible

We indicated earlier that Bayesian methods are efficient and we also mentioned that they can effectively incorporate a priori information. To implement Bayesian decoding for velocity, as an alternative to the PVA outlined above, we must apply Bayes' theorem (analogously to *Eq. 4*). This requires *1*) a probability model for the data and *2*) a prior probability assumption for the unknown velocities (which are to be predicted).

A simple version of Bayesian decoding obtains the predicted velocity vector at time $t + \tau$, which we now write as $v_{t+\tau}$, from a pair of equations. (One simplification in this exposition is that we are assuming the time lag τ between neural firing and movement remains fixed throughout the experiment and is the same for all neurons.) The first equation specifies the posterior distribution given by Bayes' theorem (analogously to *Eq. 4*)

$$p(v_{t+\tau}|y_1, \cdots, y_t) \propto p(y_t|v_{t+\tau})p(v_{t+\tau}|y_1, \cdots, y_{t-1}) \qquad (14)$$

In *Eq. 14* the likelihood function is the probability $p(y_t|v_{t+\tau})$ of a firing rate y_t when a movement $v_{t+\tau}$ will be made at time $t + \tau$. The second probability density on the right-hand side of *Eq. 14* plays the role of the prior for velocity at time $t + \tau$ based on the spiking activity that has preceded time t. This density is determined, recursively, from the equation

$$p(v_{t+\tau}|y_1, \cdots, y_{t-1}) = \int p(v_{t+\tau}|v_{t-1+\tau})p(v_{t-1+\tau}|y_1, \cdots, y_{t-1})\mathrm{d}v_{t-1+\tau} \qquad (15)$$

In *Eq. 15* we obtain the integrand $p(v_{t-1+\tau}|y_1, \cdots,| y_{t-1})$ from the previous step of *Eq. 14*. The factor $p(v_{t+\tau}|v_{t-1+\tau})$ is where the prior information (information separate from, or "prior" to, any spiking activity) is inserted: we assume the velocity $v_{t+\tau}$ will tend to resemble $v_{t-1+\tau}$, but will deviate from it by a small amount (governed by a smoothing parameter, which may also be obtained from the data); as a result the velocity will be smoothed across time. The details we are skipping may be found in Brockwell et al. (2004) and the references therein. *Equation 15* is sometimes called the Chapman–Kolmogorov equation (Brown et al. 1998).

Equations 14 and *15* are very general. Alternative evolving behavioral parameters may take the place of velocity. If the likelihood function [the probability $p(y_t|v_{t+\tau})$ in *Eq. 14*] describes reasonably well the dependency of spiking activity on these parameters, there are sufficient data, and the prior smoothness condition [the probability $p(v_{t+\tau}|v_{t-1+\tau})$ in *Eq. 14*] is appropriate, then Bayesian decoding will predict their evolution accurately.

Brockwell et al. (2004) compared the PVA to optimal linear estimation (OLE) and Bayesian decoding for a computer-simulated hand movement using 200 neurons with tuning that was similar to directional tuning observed in motor cortex data. In their simulation study PVA was less efficient than OLE by a factor of 2 and less efficient than Bayesian decoding by a factor of 10. Thus, for example, Bayesian decoding of 25 neurons would be roughly as accurate as PVA decoding of 250 neurons. Brockwell et al. (2004) also obtained roughly 7-fold improvement in MSE for a set of motor cortical data. Large gains in accuracy with Bayesian decoding have also been reported by Gao et al. (2002).

To illustrate the flexibility of the approach, we provide a second example of Bayesian decoding in the context of reconstructing the path of a foraging rat based on the activity of pyramidal neurons in the CA1 region of the hippocampus. It is well known that when a rat moves through its environment these neurons fire only in certain regions of space (O'Keefe and Dostrovsky 1971; Wilson and McNaughton 1993) and, as a result, they are called place cells and regions of space in which they fire are termed place receptive fields. Brown et al. (1998) analyzed place cell spike trains and position data from a rat freely foraging in a circular environment. Here, we contrast Bayesian decoding with reverse correlation, as in *Eq. 9* and *10*. We analyze similar data in *An example of iterative probability modeling*.

To display results we show the X and Y components of the animal's position separately. Figure 4 displays the fit of the reverse correlation model (green line) and the Bayesian decoding algorithm (red line) to the X and Y components of position (blue line) for the decoding stage. Reverse correlation, which makes no assumption of smoothness in the animal's path, provides a very noisy prediction, and strays substantially from the correct path during several epochs. The proportion of variability explained by reverse correlation is $R^2 = 0.23$. Compared with the reverse correlation fit, Bayesian decoding predicts the path of the animal in the separate directions very closely. Its deviations from the model fit, which are much smaller than those of the reverse correlation technique, occur at the extremes of the individual position components. The proportion of variability in the path data explained by Bayesian decoding is $R^2 = 0.87$.

POISSON AND NON-POISSON MODELS

At the beginning of INFORMATION AND STATISTICAL EFFICIENCY we noted that non-Poisson variation in spike trains is to be expected, and has been documented, under particular conditions (see also Barbieri et al. 2001; Kass and Ventura 2001; Reich et al. 1998; and the references therein). One phenomenon leading to non-Poisson behavior is the refractory period: immediately after a spike there is a short interval of time during which another spike is impossible and a longer interval of time during which the probability of a spike is reduced. For high firing rates refractory effects become detectable and may be important for some analyses. Another point, mentioned earlier, is that theoretical integrate-and-fire–type models produce non-Poisson behavior. In addition, either bursting or excess trial-to-trial variation (arising from changes in stimuli or internal state of the subject) may be present. These may require spe-

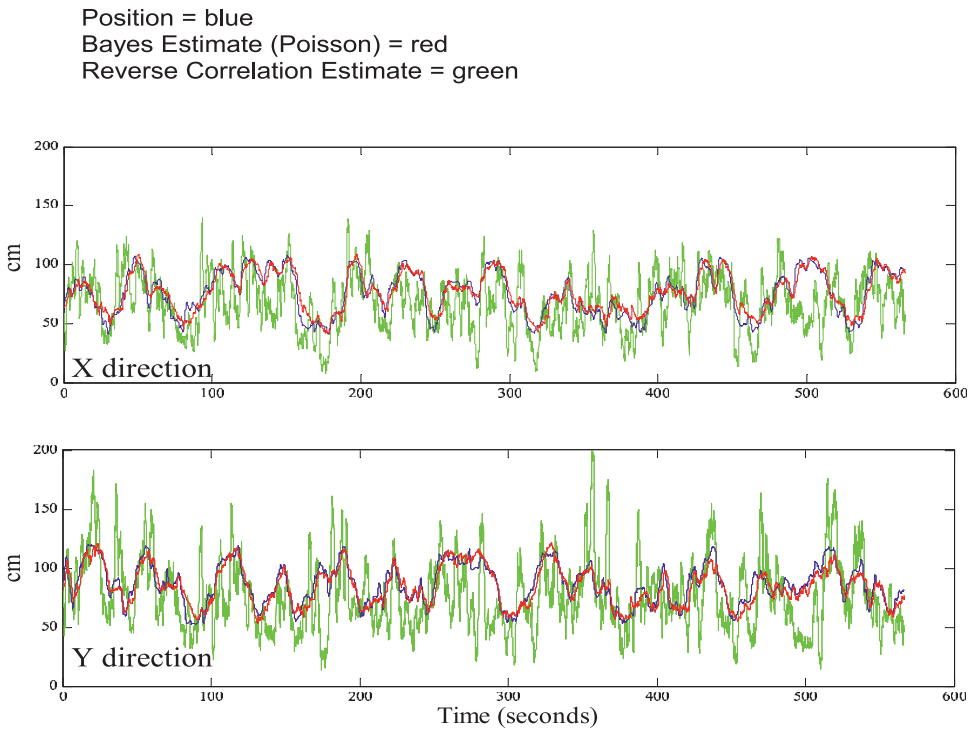

FIG. 4. Decoding of the position of a rat. Simultaneous activity of 34 place cells was recorded while the animal foraged for chocolate pellets for 23 min. Electroencephalogram (EEG) data were recorded from the same electrodes and bandpass filtered from 6 to 14 Hz to extract the theta rhythm, which is known to be an important determinant of place cell firing. Position of each animal was measured at 30 Hz by a camera that tracked the location of 2 infrared diodes mounted on the animal's headstage. Brown et al. (1998) defined the encoding stage as the first 13 min of spike train, path, and theta rhythm data and estimated the parameters of an inhomogeneous Poisson process model for each place cell, together with those in a spatial random-walk model for the prior probabilities. They defined the decoding stage as the last 10 min of the experiment for each animal. Position estimates for the decoding analysis were updated every 33 ms, the frame rate of the tracking camera. In total 18,000 (30 estimates/s × 60 s/min × 10 min) decoded position estimates were computed. *X* and *Y* directions are plotted separately. Bayesian decoding provides an accurate reconstruction of the path, whereas reverse regression is noisy and subject to sustained, substantial error in some parts of the experiment (e.g., around 500 s).

cialized techniques that either complement or extend Poisson-based statistical methods.

In the first subsection we briefly describe the Poisson process and emphasize its usefulness in analyzing data pooled across trials. In the next subsection we indicate ways in which probability models may be extended so that they may be applied to non-Poisson data to produce within-trial analyses. Finally, we illustrate assessment of fit for Poisson and non-Poisson models. These methods are then used to demonstrate the presence of non-Poisson bursting in a hippocampal place cell in the next section on iterative probability modeling.

Time-varying Poisson processes are suitable for analyzing data pooled across many trials

The simplest conception of neuronal firing is that it follows a Poisson process with a time-invariant firing rate λ. In this case, the spike-generating process is said to be *stationary* and *Eq. 8* specializes to

$$p(s_1, \cdots, s_n) = e^{-\lambda T}\lambda^n \tag{16}$$

where $T = B - A$ is the total observation time. The process is also called *homogeneous* (as opposed to time-varying or *inhomogeneous*). The simplicity of the homogeneous case is that the function $\lambda(t)$ is replaced by a single number λ. When *Eq. 16* holds, the probability that a spike will occur in the infinitesimal interval $(t, t + dt)$ is $\lambda \cdot dt$, regardless of the time t (i.e., there is a constant firing rate over time). Because spike trains are very rarely stationary, *Eq. 8* which allows for a time-varying firing rate, is an important generalization.

Equation 8 still involves a strong simplifying assumption: the probability of a spike at time t [i.e., in the infinitesimal interval $(t, t + dt)$] is given by

$$\text{Prob}\,[\text{spike in } (t, t + dt)] = \lambda(t)dt \tag{17}$$

that is, the firing rate depends only on time. For both homogeneous and time-varying Poisson processes the probability of a spike at time t is independent of the number and timing of the spikes that have occurred before time t. Thus, effects stemming from a refractory period or bursting are being ignored.

One of the reasons the general Poisson model *Eq. (8)* is important is that it holds, at least approximately, when data are pooled across trials. Specifically, a general theorem (Daley and Vere-Jones 1988, Theorem 9.2.V) shows that, as the number of replications (trials) of a non-Poisson process increases, the pooled observations will approximately follow a Poisson process. It may be verified statistically that spikes times pooled across trials are nearly Poisson (e.g., Olson et al. 2000; Ventura et al. 2001). This is illustrated in the section on assessment of goodness-of-fit models. Whenever the general Poisson model holds, the count of the number of spikes found in any given interval of time (a, b) follows a Poisson distribution, as specified in *Eq. 1*.

Non-Poisson data may be analyzed using probability models

Although data pooled across trials may be safely analyzed under the Poisson assumption, it is frequently important to examine data without such pooling. Within trials, effects such as the refractory period and bursting may create worrisome departures from a Poisson process.

To generalize *Eq. 17* we replace the firing-rate function $\lambda(t)$, also called the *intensity* function, with the *conditional intensity* that involves not only t but also the spiking history. Writing the history as $H_t = [s_1, \cdots, s_{n(t)}]$, where the series $[s_1, \cdots, s_{n(t)}]$ represents the spike times before t, we then write the conditional intensity as $\lambda(t|H_t)$ and obtain the generalization of *Eq. 17* as

$$\text{Prob}\,[\text{spike in } (t, t + dt)] = \lambda(t|H_t)dt \tag{18}$$

This conditional intensity may, in principle, depend on the

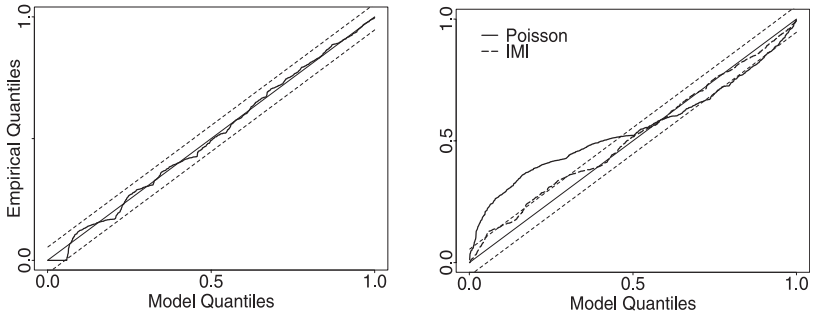

FIG. 5. Goodness-of-fit Q–Q plots, with 95% probability bands. *Left:* plot of empirical quantiles vs. theoretical model quantiles, for the locust antennal lobe neuronal data after pooling spike times across trials [thereby destroying the original interstimulus interval (ISI) structure of the data]. The plot indicates very close agreement with the Poisson process predictions. *Right:* plots for within-trial data based on Poisson and non-Poisson (IMI, for Inhomogeneous Markov Interval) models, the latter following *Eq. 19*. This plot uses all 15 trials, without pooling. The Poisson process model clearly does not fit these data, whereas the IMI model fits quite well. Both plots are based on the time-rescaling method discussed by Brown et al. (2002).

entire history of spikes up to time t. For example, the probability that a neuron will have a spike 500 ms after the animal has been presented with a cue could, in principle, depend on precisely when and how many times the neuron fired during the 500 ms after the cue. That conception, however, is too complicated to be practical: it involves more parameters than can be estimated from limited data. One useful simplification is to assume that the probability of a spike at time t depends on both t and on the time delay since the previous spike. Letting s_* be the previous spike time, we write

$$\text{Prob}\,[\text{spike in } (t,\, t + \mathrm{d}t)] = \lambda(t,\, t - s_*)\mathrm{d}t \qquad (19)$$

Kass and Ventura (2001) showed how generalized regression methods may be used to fit models of the form of *Eq. 19* to neuronal data, and how the likelihood ratio test may be used both to examine the suitability of the Poisson model and to examine the suitability of *Eq. 19*. Additional references to models for which *Eq. 19* holds are contained in Kass and Ventura (2001) and Johnson (1996).

Instead of a sequence of spike times, spike trains may be represented, equivalently, as a sequence of interspike intervals (ISIs). In the case of homogeneous Poisson processes, the ISIs are probabilistically independent and follow an exponential distribution. When the ISIs are independent and from some particular (possibly non-exponential) distribution, the spike train, which is again homogeneous in time, is said to follow a *renewal* process. Barbieri et al. (2001) showed how inhomogeneous versions of renewal processes may be constructed and fit to neuronal data by rescaling time (see also Brown et al. 2002). Such time-rescaled probability models may be motivated from biophysical considerations, such as integrate-and-fire (IF) or other models of the neuron's spiking mechanism. They are similar in spirit to the models specified by *Eq. 19* and are likely to produce similar data-analytical results, but as yet no detailed comparison of these 2 alternative statistical models has been carried out.

Goodness-of-fit may be assessed for both Poisson and non-Poisson models

In *Information and statistical efficiency* we mentioned Q–Q plots as a general methodology for assessing goodness-of-fit. Here we illustrate with the locust antennal lobe data considered earlier. We consider the fit of Poisson and non-Poisson processes. The methodology was previously discussed by Brown et al. (2002). In the Poisson case, the idea is that it is possible to remove the effect of $\lambda(t)$, suitably transforming the spike train, and thereby obtain a homogeneous Poisson process, with exponentially distributed ISIs. We then order these ISIs and plot them against quantiles from an exponential distribution. When the Poisson fits well, the plotted points will fall roughly on a line. Furthermore, 95% probability bands may be superimposed so that it is possible to judge substantial departures from linearity in the plot. Specifically, 95% of spike trains artificially generated from an assumed probability model will have Q–Q plots lying within the 95% probability bands. Thus, spike trains having Q–Q plots within the 95% probability bands would be judged to have ISI distributions that are consistent with the hypothesized model. This is illustrated in Fig. 5 first when the data are pooled across trials. According to our earlier remarks, pooling across trials should make the resulting point process agree closely with what is expected for an inhomogeneous Poisson process. This is what we observe in the *left panel* of Fig. 5.

The transformation of the spike train used in producing Fig. 5 is a rescaling of time and, as discussed by Brown et al. (2002), the same method is applicable to other point processes. The *right side* of Fig. 5 shows the within-trial Q–Q plots for both the inhomogeneous Poisson process and the model of *Eq. 19*. The non-Poisson process from *Eq. 19* fits these data quite well, and much better than the Poisson.

AN EXAMPLE OF ITERATIVE PROBABILITY MODELING

To illustrate the iterative model-building process displayed in Fig. 1 we will present results based on 3 probability models for the activity of a single neuron recorded by Frank et al. (2001) from the CA1 region of the hippocampus in a behaving rat. Bayesian decoding was based on position and theta rhythm. Associated with the theta rhythm, place cells often exhibit bursting. A simple question is: What is the relative importance of position compared with the theta rhythm and bursting in describing the spiking activity of the place cell neurons? We will display results from 3 models: the first based only on the rat's position, the second including theta phase, and the third including both theta phase and bursting.

In the experiment described by Frank et al. (2001), the spiking activity of approximately 30 CA1 hippocampal neurons was recorded from a rat during 20 min of running on U-shaped track. We begin with some exploratory analysis in Fig. 6. As is evident from Fig. 6*A*, along the 150-cm U-shaped track the spiking activity of this neuron is high only in positions 40 to 90 cm. From Fig. 6*B* there is a suggestion of increased activity between theta phases 90 to 180°. The ISI histogram, in Fig. 6*C*, shows a large number of short ISIs (<20 ms), an observation that is consistent with the presence of bursting. The second, much smaller peak in the ISI histogram

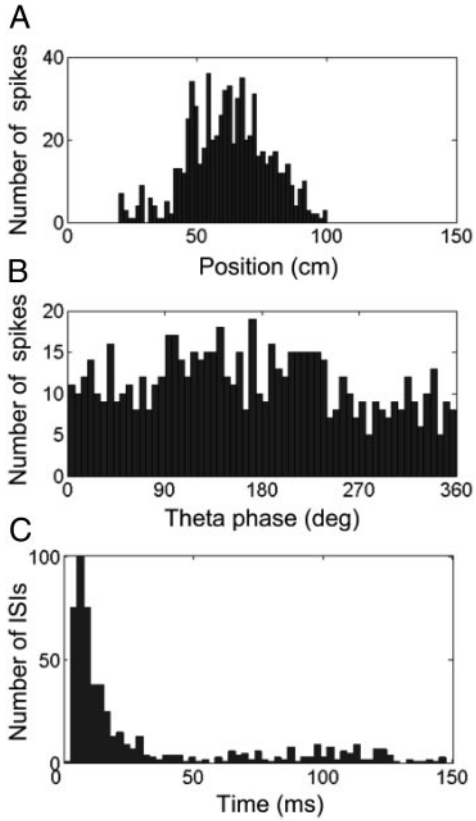

FIG. 6. Histograms of CA1 neuron's spiking activity. *A*: spike counts plotted against position. *B*: spike counts plotted against theta rhythm phase. *C*: ISI histogram: the large early peak and smaller and broader late peak indicate the presence of bursting and theta rhythm, respectively.

between 100 and 150 ms is consistent with the theta rhythm modulation.

As an initial probability model for the neuronal spiking activity we used an inhomogeneous Poisson process with firing rate being a function of the animal's position in the environment—without consideration of theta rhythm. Barbieri et al. (2004) found Zernike polynomials effective in representing place fields in a foraging environment enclosed by a circular boundary. To adapt their approach to the example we are presenting here, the domain of the polynomials was restricted to the *U*-shaped track. Dependency of firing rate on position was thus represented as

$$\log \lambda(t) = \theta_0 + \sum_{j=1}^{5} \theta_j z_j[x(t)] \qquad (20)$$

where $x(t)$ is the position of the animal at time t.

The model was fitted to the data using ML. The resulting estimate of the spatial extent of the place field is shown in Fig. 7*A*. It matches well the histogram in Fig. 6*B*. However, the Q–Q plot (Fig. 7*C*, black curve), based on the method described earlier, clearly shows the simple spatial model to be inaccurate: the plot ranges far outside the 95% probability bands. To determine whether adding a theta phase component could improve the fit, the model was reformulated by adding a term of the form $\theta_6 \cos[\phi(t)] + \theta_7 \sin[\phi(t)]$, where $\phi(t)$ represents the theta phase. The resulting Q–Q plot (Fig. 7*C*, red curve) indicates no improvement in fit.

To capture the effect of both bursting and the theta rhythm in the description of the neuron's spiking activity a representation of spiking history, following the general form of *Eq. 18*, was included. The complete specification of within-trial firing rate was

$$\log \lambda(t|H_t) = \theta_0 + \sum_{j=1}^{5} \theta_j z_j[x(t)] + \sum_{j=1}^{20} \beta_j n_{t-j} \qquad (21)$$

where n_{t-j} is the number of spikes in the 10 ms interval $10j$ ms before time t. In this model, current spiking activity is related not only to current position and to theta phase, but also to spiking activity going back 200 ms before the current time point. In particular, *Eq. 21* represents the theta phase in terms of the tendency for increased firing rate whenever the spike counts n_{t-j} are relatively large, with j being roughly 125 ms, assuming the corresponding β coefficients are positive; the positivity of the β coefficients must be verified from the fit of the model to the data. The Q–Q plot based on the ML fit of model *Eq. 21* is also displayed in Fig. 7*C* (blue curve). This model provides a far better fit than the previous 2 models, with the Q–Q plot lying within the 95% bounds. The spatial component of the new model was found to be essentially identical

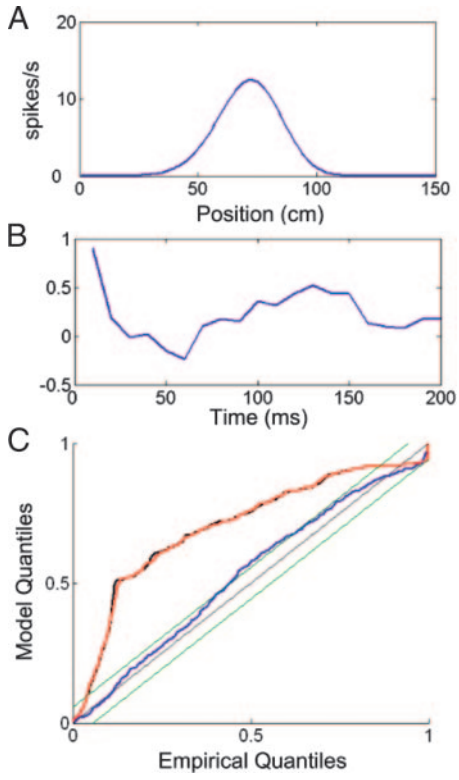

FIG. 7. Maximum likelihood fits from models incorporating *1*) position; *2*) position and theta rhythm; and *3*) position, theta rhythm, and bursting. *A*: position component, which remains the same for all 3 models. *B*: coefficients of the temporal component in *Eq. 21*. First 2 coefficients correspond to times 0–10 and 11–20 ms before the current spike time and their positive values represent the effect of bursting. Positive values of the coefficients from 100 to 110, 111 to 120, 121 to 130, 131 to 140, and 141 to 150 ms represent the effect of the theta rhythm. *C*: goodness-of-fit Q–Q plots for the 3 models: the position model Q–Q plot is in black; the position and theta rhythm model Q–Q plot is in red; and the position, theta rhythm, and bursting model Q–Q plot is in blue; 45° line represents a perfect fit, which would not account for statistical variation; green lines are the 95% probability bands. Q–Q plots show that the ISIs are consistent with the 3rd model but not the 1st and 2nd models.

to that shown in Fig. 7*A*. The coefficients β_j, plotted in Fig. 7*B*, are positive between roughly 5 and 20 ms, consistent with bursting, and also between 100 and 150 ms, consistent with the theta rhythm, as mentioned above. To check whether all the β coefficients improve the fit, the likelihood ratio test was applied sequentially, first to the model using only the coefficients β_1, β_2, next to the model including coefficients $\beta_1, \cdots, \beta_9$, and finally to the model with all 20 β_j coefficients. In each case the results were highly statistically significant. Considering these findings together with the goodness-of-fit assessment, we conclude that this neuron responds to position, with evident effects of theta phase and bursting.

These results, and similar analyses based on many more neurons, suggest that spatial and temporal structure are both important for understanding how CA1 represents information (Frank et al. 2004, *J Neurosci*). Note that, unlike the general situation described in *Statistical summaries of firing rate,* there has been no pooling across repeated trials. This is important because the precise paths taken by the animal across experimental replications may be different. Similar modeling efforts would be useful whenever trials involve behavioral changes or neuronal adaptation. We return to such trial-to-trial variation in the next section.

CORRELATED PAIRS OF NEURONS

Discussions of correlation may refer to any of several distinct phenomena. Neurons may be similar physiologically in the sense that they may respond to the same set of stimuli with the same temporal firing-rate profile, yet they may be statistically independent in the sense that the spiking activity of one neuron may be described adequately by its conditional intensity function, whereas the activity of the other neuron provides no additional predictive information (i.e., no additional ability to predict when the spikes of the first neuron will occur). In particular, 2 neurons may tend to fire in close temporal proximity to each other even if the neurons are statistically independent: if they both respond to a stimulus with nearly the same time-dependent firing-rate function, and if that function moves suddenly from a low rate to a much higher rate, both neurons will spike frequently at roughly the same experimentally defined response time. As a matter of definition, then, when we speak of correlation, we will be referring to an *additional* tendency of 2 neurons to fire together after their firing-rate functions (their conditional intensities) have been accounted for.

A second point is that the "tendency to fire together" may refer either to the tendency of the neurons to fire in close temporal proximity or to the tendency for one neuron to have larger or smaller spike counts on trials for which the other neuron has correspondingly larger or smaller spike counts. The former implies the latter: if 2 neurons tend to fire in close temporal proximity (beyond what is predicted by their individual firing-rate functions), they will also tend to have larger or smaller spike counts together, across trials. On the other hand, the latter tendency for 2 neurons to vary together across trials may occur over and above what is caused by proximal firing within trials and may occur in the absence of any correlation within trials. The variation in a neuron's spiking activity across trials beyond that predicted by the neuron's firing-rate function is called "trial-to-trial variation." Shared trial-to-trial variation can be an important source of correlation in spike counts.

The tendency of two neurons to act together may be displayed by the normalized JPSTH and cross-correlogram

Because the firing of each neuron is highly variable, care may be required in judging the importance of any apparent tendency for 2 neurons to fire together. Of course, 2 neurons could fire together as a result of chance fluctuations, and the general statistical problem is to compare observed tendencies to those that would be predicted by chance alone, under the assumption that the neurons are actually operating independently. If it could be assumed, in addition, that each neuron had a time-invariant firing rate the problem would not be very difficult. However, rigorous assessment of association in the presence of neuronal firing rates that vary across time is somewhat involved, particularly when the association may itself vary across time.

A very useful graphical display for this purpose is the joint poststimulus time histogram (JPSTH; Aertsen et al. 1989). This extends the PSTH by counting the number of trials on which spikes occur for both neuron 1 at time u and neuron 2 at time v, with the aim of indicating the times (and relative time lags $v - u$) at which large numbers of joint spikes occur. To judge what numbers are "large," the raw JPSTH values may be standardized in units that reflect the variation that would be expected if the 2 neurons were operating independently. This is what is done in the "normalized JPSTH." To judge the tendency of 2 neurons to fire together at a time lag of $\delta = v - u$ ms the values along the diagonals (defined by $v - u = \delta$) of the normalized JPSTH are summed, producing the *cross-correlogram,* which is a function of δ.

The cross-correlogram provides an indication of correlated firing at various lags δ. Assuming it is based on the normalized JPSTH (in the literature, definitions vary), the cross-correlogram adjusts for the time-varying firing rate. This is important: if the 2 neurons both had little activity over most of the range of time being examined but very substantial firing rates over some small subinterval of time, a sizeable fraction of spikes would occur in fairly close proximity to one another. By some measures of association the 2 neurons would therefore appear to be acting together even when, in fact, they were operating independently. Thus, the ability of the normalized JPSTH and cross-correlogram to adjust for firing rate is essential. There remain several statistical concerns, however.

A more powerful assessment of time-varying correlated spiking activity may be obtained by smoothing the JPSTH and applying Bootstrap significance tests

First, it is important to note that there is no uniquely compelling method of normalizing the JPSTH: the normalization of Aertsen et al. (1989) produces at each set of coordinates (u, v) the usual correlation (Pearson correlation), across trials, of the spiking activity for neuron 1 at time u with the spiking activity for neuron 2 at time v. Because this is a widely used measure of association it is appealing, but there are many other equally good measures and they can yield different results (Ito and Tsuji 2000). Second, there are alternative methods of defining statistical significance tests. These may use *1*) alter-

native association measures and test procedures, *2*) smoothing of measures across neighboring time values, and *3*) alternative methods of computing *p*-values. Ventura et al. (2005a) defined a statistical test based on the smoothed ratio $\xi_\delta(t)$ of the joint spiking activity (the diagonal of the JPSTH at time lag δ) to that expected under independence: the test uses the Bootstrap to define a threshold c and compute a *P*-value for the magnitude of the peak of the function $\xi_\delta(t) > c$. The resulting procedure can be more accurate and sensitive than tests based solely on the normalized JPSTH.

It is important to distinguish trial-to-trial variation from correlated spiking activity within trials

A third consideration is that the normalized JPSTH, and cross-correlogram, assume that all trials are statistically identical, in the sense that the variation across trials arises solely from the variation within trials. In other words, they assume that the additional trial-to-trial variation is zero. Nonnegligible variation between trials could arise from experimental effects, such as nonidentical stimuli, or changes in the internal state of the subject. The importance of distinguishing trial-to-trial variation from within-trial variation has been emphasized by Brody (1999a,b), Ben-Shaul et al. (2001), and Ventura et al. (2005b), among others. To consider trial-to-trial variability formally let us write the conditional intensity for trial r as $\lambda_r(t|H_t)$. The statistical problem of estimating $\lambda_r(t|H_t)$ is difficult because each trial contributes only a modest amount of information, in the form of a handful of spike times. Progress requires fairly strong assumptions about the form of $\lambda_r(t|H_t)$. The approach used by Ventura et al. (2005b) was to write

$$\lambda_r(t|H_t) = g_r(t)\lambda(t|H_t) \tag{22}$$

and to represent the trial-dependent gain factor $g_r(t)$ in terms of a small number of freely varying parameters. In this way, the relatively small amount of information available from a small number of spikes could be used efficiently. Ventura et al. (2005b) then showed how the smoothing and Bootstrap methods of Ventura et al. (2005a), for analyzing correlated activity of neurons, could be extended to accommodate trial-to-trial variation and, furthermore, how graphical displays such as the cross-correlogram could then be corrected for trial-to-trial variation.

Brody (1999a,b) distinguished latency effects, which refer to the tendency for a neuron to respond with differing delays on different trials, from excitability effects, which refer to changes in the magnitude of the firing rate across trials. Although these may in some cases be difficult to disambiguate, it is useful to consider them separately. Ventura (2004) proposed a simple testing and estimation procedure for trial-dependent shift effects as a way of accommodating latency. (See also Baker and Gerstein 2001.)

Decoding may be either adversely affected or improved by correlation

An important observation about the potential deleterious effects of correlation was made by Zohary et al. (1994; see also Shadlen and Newsome 1998). They examined spike counts in large time intervals across trials among pairs of neurons recorded simultaneously from MT in response to moving-dot stimuli and observed small but nonnegligible correlations, averaging around $r = 0.12$. Zohary et al. (1994) went on to point out that when a sample mean is computed from observations having similar correlations its SE no longer declines as $1/\sqrt{n}$ but rather reaches an asymptote. As a consequence, the sample mean based on many thousands of correlated observations would provide the same information as that based on only a few independent observations. Thus, a downstream neuron that integrates correlated spike counts by summing (equivalently, averaging) effectively collects only the amount of information provided by a small number of independent neurons. If this is the way neuronal information is transmitted, the system is grossly inefficient because of seemingly excessive redundancy.

A response to this observation is that populations of neurons may use more complicated integration mechanisms than simple summation. If, for example, some form of population coding is used, the effect disappears: as Abbott and Dayan (1999) have shown, in a simple but intuitively reasonable framework, correlation does not destroy the effectiveness of efficient decoding schemes. More specifically, they showed that the Fisher information increases with the number of neurons when the correlation takes one of several reasonable forms, assuming that not all neurons have identical tuning functions. Thus, if neuronal systems use more complicated computations than summation their capacity to transmit information need not be degraded dramatically by small trial-to-trial correlations.

A natural follow-up question is whether efficient decoding schemes such as those discussed earlier benefit from using correlation. A complete answer is not yet available. However, there is some evidence that the correlation will enhance decoding if it itself contains relevant information about the encoded parameter, such as movement velocity. For static scenarios like those considered by Abbott and Dayan (1999), Wu et al. (2001) showed that decoding schemes that ignore correlation will remain efficient if the correlation does not depend on the encoded parameter. Meanwhile, the M1 decoding procedures of Gao et al. (2002) benefited from additional information about movement parameters contained in the correlation structure of multiple neurons. [See Hatsopoulos et al. (2003), Nirenberg and Latham (2003), Reich et al. (2002), Seriés et al. (2004), and references therein for related results and discussion.]

DISCUSSION

The field of statistics is devoted to the creation and elucidation of a large body of tools for learning from data. We began this overview by presenting several foundational ideas: that probability models are used to describe data, that estimation procedures may be formalized and evaluated using well-established criteria such as mean squared error, and that methods based on the likelihood function are optimal when the probability model is correct. These ideas are important partly because, with effort and care, it is often possible to develop good probability models for experimental data. The evaluation framework (e.g., the use of mean squared error) is also valuable because it continues to apply to nonparametric methods, where modeling assumptions are relaxed. We illustrated the general points with concrete applications to smoothing the PSTH and JPSTH, and to decoding. Our intention has been, in

particular, to explain the benefits of Bayesian methods and the Bootstrap because both produce powerful and effective applications of simple concepts over a wide range of problems.

The highest-level message from our review is 3-fold. First, simple statistical methods remain important, and we would emphasize the great desirability of keeping things simple and intuitive, and using easily comprehended figures whenever possible. Second, when complicated situations present themselves and alternative statistical methods are available, it is possible to invoke well-established statistical principles to determine which of the methods is better and under what circumstances. Finally, as we have illustrated, the differences among alternative methods can be very substantial and the issue of choosing among them can therefore be of great practical importance.

We have chosen the topics of summarizing firing rate, population coding and decoding, Poisson and non-Poisson modeling, and assessing correlation in pairs of neurons partly because they are important problems but also because we could use them to illustrate the general statistical ideas we began with. The general parametric and nonparametric probability-modeling framework emphasized here is widely accepted in the field of statistics. For an interesting dissenting opinion see Breiman [2001; the discussions to that article by Cox (2001) and by Efron (2001) are consistent with the point of view expressed here]. We have not attempted here a comprehensive survey of statistical methods in neurophysiology, and there are clearly important gaps in our coverage. In particular, we would like to single out the deep and important statistical challenges associated with the analysis of multiple simultaneously recorded spike trains (Brown et al. 2004). Two conspicuous difficulties are the complexity of spike sorting (e.g., Bar-Gad et al. 2001; Harris et al. 2001) and the absence of well-developed statistical methods for multiple point process data analysis. In addition, there are algorithmic challenges associated with construction of effective brain–machine interfaces, which is one of the exciting applications of multiple spike-train signal processing. We expect, however, that statistical thinking of a kind that begins with the concepts we have presented in this review will play a key role in meeting these challenges in the future.

ACKNOWLEDGMENT

We thank J. Victor for comments and R. Barbieri for help preparing the figures.

GRANTS

R. E. Kass and V. Ventura were supported by National Institutes of Health Grant MH-64537 and E. N. Brown was supported by Grants MH-59733, MH-61637, DA-015664, and R25 MH-66410.

REFERENCES

Abbott LF and Dayan P. The effect of correlated variability on the accuracy of a population code. *Neural Comput* 111: 91–101, 1999.

Aertsen AMHJ, Gerstein MK, Habib MK, and Palm G. Dynamics of neuronal firing correlation: modulation of "effective connectivity." *J Neurophysiol* 61: 900–918, 1989.

Baker SN and Gerstein GL. Determination of response latency and its application to normalization of cross-correlation measures. *Neural Comput* 13: 1251–1357, 2001.

Barbieri R, Frank LM, Nguyen DP, Quirk MC, Solo V, Wilson MA, and Brown EN. Dynamic analyses of information encoding by neural ensembles. *Neural Comput* 16: 277–307, 2004.

Barbieri R, Quirk MC, Frank LM, Wilson MA, and Brown EN. Construction and analysis of non-Poisson stimulus response models of neural spike train activity. *J Neurosci Methods* 105: 25–37, 2001.

Bar-Gad I, Ritov Y, Vaadia E, and Bergman H. Failure in identification of multiple neuron activity causes artificial correlations. *J Neurosci Methods* 107: 1–13, 2001.

Behseta S and Kass RE. Testing equality of two functions using BARS. *Stat Med* 2005 In press.

Behseta S, Kass RE, and Wallstrom G. Hierarchical models for assessing variability among functions. *Biometrika* In press.

Ben-Shaul Y, Bergman H, Ritov Y, and Abeles M. Trial to trial variability in either stimulus or action causes apparent correlation and synchrony in neuronal activity. *J Neurosci Methods* 111: 99–110, 2001.

Bialek W, Rieke F, de Ruyter van Steveninck RR, and Warland D. Reading a neural code. *Science* 252: 1854–1857, 1991.

Box GEP, Hunter WG, Hunter JS. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* New York: Wiley, 1978.

Breiman L. Statistical modeling: the two cultures (with discussion). *Stat Sci* 16: 199–231, 2001.

Brillinger DR. Nerve cell spike train data analysis: a progression of technique. *J Am Stat Assoc* 87: 260–271, 1992.

Brockwell AE, Rojas AL, and Kass RE. Recursive Bayesian decoding of motor cortical signals by particle filtering. *J Neurophysiol* 91: 1899–1907, 2004.

Brody CD. Correlations without synchrony. *Neural Comput* 11: 1537–1551, 1999a.

Brody CD. Disambiguating different covariation. *Neural Comput* 11: 1527–1535, 1999b.

Brown EN, Barbieri R, Ventura V, Kass RE, and Frank LM. The time-rescaling theorem and its application to neural spike data analysis. *Neural Comput* 14: 325–346, 2002.

Brown EN, Frank LM, Tang D, Quirk MC, and Wilson MA. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J Neurosci* 18: 7411–7425, 1998.

Brown EN, Kass RE, and Mitra PP. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nat Neurosci* 7: 456–461, 2004.

Cox DR. Comment on article by Breiman. *Stat Sci* 16: 216–218, 2001.

Curran-Everett D and Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. *J Neurophysiol* 92: 669–671, 2004.

Daley DJ and Vere-Jones D. *An Introduction to the Theory of Point Processes.* New York: Springer-Verlag, 1988.

Davison AC and Hinkley DV. *Bootstrap Methods and Their Applications.* Cambridge, UK: Cambridge Univ. Press, 1997.

Dayan P and Abbott LF. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Cambridge, MA: MIT Press, 2001.

DiMatteo I, Genovese CR, and Kass RE. Bayesian curve-fitting with free-knot splines. *Biometrika* 88: 1055–1071, 2001.

Efron B. Comment on article by Breiman. *Stat Sci* 16: 218–219, 2001.

Efron B and Tibshirani R. *An Introduction to the Bootstrap.* New York: Chapman & Hall, 1993.

Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans R Soc Lond A Phys Sci* 222: 309–322, 1922.

Frank LM, Brown EN, and Wilson MA. A comparison of the firing properties of putative excitatory and inhibitory neurons from CA1 and the entorhinal cortex of the awake behaving rat. *J Neurophysiol* 86: 2029–2040, 2001.

Frank LM, Stanley GB, and Brown EN. Hippocampal plasticity across multiple days of exposure to novel environments. *J Neurosci* 24: 7681–7689, 2004.

Gabbiani F and Koch C. Principles of spike train analysis. In: *Methods of Neuronal Modeling,* edited by Koch C and Segev I. Cambridge, MA: MIT Press, 1998.

Gao Y, Black MJ, Bienenstock E, Shoham S, and Donoghue JP. Probabilistic inference of hand motion from neural activity in motor cortex. In: *Advances in Neural Information Processing Systems,* vol. 14. Cambridge, MA: MIT Press, 2002.

Georgopoulos AP, Kalaska JF, Caminiti R, and Massey JT. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci* 2: 1527–1537, 1982.

Georgopoulos AP, Schwartz AB, and Kettner RE. Neuronal population coding of movement direction. *Science* 233: 1416–1419, 1986.

Gerstein GL and Kiang NY-S. An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophys J* 1: 15–28, 1960.

Gerstein GL and Mandlebrot B. Random walk models for the spike activity of a single neuron. *Biophys J* 4, 41–68, 1964.

Hansen MH and Kooperberg C. Spline adaptation in extended linear models (with discussion). *Stat Sci* 17: 2–51, 2002.

Harris KD, Hirase H, Leinekugel X, Henze DA, and Buzsaki G. Temporal interaction between single spikes and complex spike bursts in hippocampal pyramidal cells *Neuron* 32: 141–149, 2001.

Hatsopoulos NG, Paninski L, and Donoghue JP. Sequential movement representations based on correlated neuronal activity. *Exp Brain Res* 149: 478–486, 2003.

Hogg RV and Tanis EA. *Probability and Statistical Inference* (6th ed.). Upper Saddle River, NJ: Prentice-Hall, 2001.

Ito H and Tsuji S. Model dependence in quantification of spike interdependence by joint peri-stimulus time histogram. *Neural Comput* 12: 195–21, 2000.

Johnson D. Point process models of single-neuron discharges. *J Comput Neurosci* 3: 275–299, 1996.

Kass RE and Raftery AE. Bayes factors. *J Am Stat Assoc* 90: 773–795, 1995.

Kass RE and Ventura V. A spike-train probability model. *Neural Comput* 13: 1713–1720, 2001.

Kass RE, Ventura V, and Cai C. Statistical smoothing of neuronal data. *Network-Comp Neural* 14: 5–15, 2003.

Kaufman CG, Ventura V, and Kass RE. Spline-based nonparametric regression for periodic functions and its application to directional tuning of neurons. *Stat Med* In press.

Kutner MH, Nachtsheim CJ, and Neter J. *Applied Linear Regression Methods* (4th ed.). Chicago, IL: McGraw-Hill/Irwin, 2004.

Loader C. *Local Regression and Likelihood.* New York: Springer-Verlag, 1999.

Mechler F, Victor JD, Purpura KP, and Shapley RM. Robust temporal coding of contrast by V1 neurons for transient but not steady-state stimuli, *J. Neurosci* 18: 6583–6598, 1998.

Moran DW and Schwartz AB. Motor cortical activity during drawing movements: population representation during spiral tracing. *J Neurophysiol* 82: 2693–2704, 1999.

Nirenberg S and Latham PE. Decoding neuronal spike trains: how important are correlations? *Proc Natl Acad Sci USA* 100: 7348–7353, 2003.

O'Keefe J and Dostrovsky J. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain Res* 34: 171–175, 1971.

Olson CR, Gettner SN, Ventura V, Carta R, and Kass RE. Neuronal activity in macaque supplementary eye field during planning of saccades in response to pattern and spatial cues. *J Neurophysiol* 84: 1369–1384, 2000.

Perkel DH, Gerstein GL, and Moore GP. Neuronal spike trains and stochastic point processes. I. The single spike train. *Biophys J* 7: 391–418, 1967a.

Perkel DH, Gerstein GL, and Moore GP. Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys J* 8: 419–440, 1967b.

Pesaran B, Pezaris JS, Sahani M, Mitra PP, and Andersen RA. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat Neurosci* 5: 805–811, 2002.

Reich DS, Mechler F, and Victor JD. Independent and redundant information in nearby cortical neurons. *Science* 294: 2566–2568, 2002.

Reich DS, Victor JD, and Knight BW. The power ratio and the interval map: spiking models and extracellular data. *J Neurosci* 18: 10090–10104, 1998.

Ruiz S, Crespo P, and Romo R. Representation of moving tactile stimuli in the somatic sensory cortex of awake monkeys. *J Neurophysiol* 73: 525–537, 1995.

Salinas E and Abbott LF. Vector reconstruction from firing rates. *J Comput Neurosci* 1: 89–107, 1994.

Schwartz AB, Moran DW, and Reina GA. Differential representation of perception and action in the frontal cortex. *Science* 303: 380–383, 2004.

Seriés P, Latham PE, and Pouget A. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat Neurosci* 7: 1129–1135, 2004.

Serruya MD, Hatsopoulos NG, Paninski L, Fellows MR, and Donoghue JP. Instant neural control of a movement signal. *Nature* 416: 141–142, 2002.

Shadlen MN and Newsome WT. The variable discharge of cortical neurons: implications for connectivity, computation and information coding. *J Neurosci* 18: 3870–3896, 1998.

Smith DR and Smith GK. A statistical analysis of the continual activity of single cortical neurones. *Biophys J* 5: 47–74, 1965.

Sokal RR and Rohlf FJ. *Biometry* (3rd ed.). New York: Freeman, 1995.

Stanley GB, Li FF, and Dan Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J Neurosci* 19: 8036–8042, 1999.

Stopfer M, Jayaraman V, and Laurent G. Intensity versus identity coding in an olfactory system. *Neuron* 39: 991–1004, 2003.

Taylor DM, Helms Tillery SI, and Schwartz AB. Direct control of 3D neuroprosthetic devices. *Science* 296: 1829–1832, 2002.

Tuckwell HC. Introduction to theoretical neurobiology. In: *Nonlinear and Stochastic Theories.* Cambridge, UK: Cambridge UP, 1988, vol 2.

Tukey JW. *Exploratory Data Analysis.* Reading, MA: Addison–Wesley, 1977.

van der Vaart AW. *Asymptotic Statistics.* Cambridge, UK: Cambridge Univ. Press, 1998.

Ventura V. Testing for, and estimating latency effects for Poisson and non-Poisson spike trains. *Neural Comput* 16: 2323–2350, 2004.

Ventura V, Carta R, Kass RE, Gettner SN, and Olson CR. Statistical analysis of temporal evolution in single-neuron firing rates. *Biostatistics* 3: 1–20, 2001.

Ventura V, Cai C, and Kass RE. Statistical assessment of time-varying dependence between two neurons. *J Neurophysiol* In press (a).

Ventura V, Cai C, and Kass RE. Trial-to-trial variability and its effect on time-varying dependence between two neurons. *J Neurophysiol* In press (b).

Warland DK, Reinagel P, and Meister M. Decoding visual information from a population of retinal ganglion cells. *J Neurophysiol* 78: 2336–2350, 1997.

Wasserman L. *All of Statistics.* New York: Springer-Verlag, 2004.

Wessberg J, Stambaugh CR, Kralik JD, Beck PD, Laubach M, Chapin JK, Kim J, Biggs SJ, Srinivasan MA, and Nicolelis AL. Real-time prediciton of hand trajectory by ensembles of cortical neurons in primates. *Nature* 408: 361–365, 2000.

Wilson MA and McNaughton BL. Dynamics of the hippocampal ensemble code for space. *Science* 261: 1055–1058, 1993.

Wu S, Nakahara H, and Amari S-I. Population coding with correlation and an unfaithful model. *Neural Comput* 13: 775–797, 2001.

Zhang K, Ginzburg I, McNaughton BL, and Sejnowski TJ. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J Neurophysiol* 79: 1017–1044, 1998.

Zohary E, Shadlen MN, and Newsome WT. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* 370: 140–143, 1994.