INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **14** (2003) 437–464 PII: S0954-898X(03)53781-2

Convergence properties of three spike-triggered analysis techniques

Liam Paninski[1,2]

Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, USA

E-mail: liam@cns.nyu.edu

Received 6 September 2002, in final form 19 March 2003
Published 21 May 2003
Online at stacks.iop.org/Network/14/437

Abstract

We analyse the convergence properties of three spike-triggered data analysis techniques. Our results are obtained in the setting of a probabilistic linear–nonlinear (LN) cascade neural encoding model; this model has recently become popular in the study of the neural coding of natural signals. We start by giving exact rate-of-convergence results for the common spike-triggered average technique. Next, we analyse a spike-triggered covariance method, variants of which have been recently exploited successfully by Bialek, Simoncelli and colleagues. Unfortunately, the conditions that guarantee that these two estimators will converge to the correct parameters are typically not satisfied by natural signal data. Therefore, we introduce an estimator for the LN model parameters which is designed to converge under general conditions to the correct model. We derive the rate of convergence of this estimator, provide an algorithm for its computation and demonstrate its application to simulated data as well as physiological data from the primary motor cortex of awake behaving monkeys. We also give lower bounds on the convergence rate of any possible LN estimator. Our results should prove useful in the study of the neural coding of high-dimensional natural signals.

1. Introduction

Systems-level neuroscientists have a few favourite problems, the most prominent of which is the 'what' part of the neural coding problem: what makes a given neuron in a particular part of the brain fire? In more technical language, we want to know about the conditional probability distributions $P(\text{spike}|X = x)$, the probability that our cell emits a spike, given that some observable signal X in the world takes a value x. Because data are expensive, neuroscientists

[1] http://www.cns.nycu/edu/∼liam

[2] A brief summary of some of the results presented here is to appear in the *NIPS02 Conference Proceedings* (Paninski 2002).

0954-898X/03/030437+28$30.00

typically postulate a functional form for this collection of conditional distributions, and then fit experimental data to these functional models, in lieu of attempting to directly estimate $P(\text{spike}|X = x)$ for each possible x. Clearly, to interpret the results of this kind of statistical analysis, we must have a good understanding of the bias and variance properties of the estimation procedure in question. This is especially true in the case of high-dimensional data (e.g. natural sensory signals or complex motor behaviour), for which direct visualization is often impossible.

In this paper, we analyse the statistical properties of a phenomenological model whose popularity in the natural signal community seems to be on the rise (Theunissen *et al* 2001, Brenner *et al* 2001, Schwartz *et al* 2002, Ringach *et al* 2002):

$$p(\text{spike}|\vec{x}) = f(\langle \vec{k}_1, \vec{x} \rangle, \langle \vec{k}_2, \vec{x} \rangle, \ldots, \langle \vec{k}_m, \vec{x} \rangle). \tag{1}$$

Here f is some arbitrary [0, 1]-valued function and $\{k_i\}$ are some linearly independent elements of (the dual space of) some vector space, X—the space of possible 'input signals'. Interpret f as a regular conditional probability. This model says, then, that the neuron projects the signal $\vec{x}$ onto some m-dimensional subspace spanned by $\{\vec{k}_i\}_{1 \leqslant i \leqslant m}$ (call this subspace K) and then looks up its probability of firing based only on this projection. This model is often called a 'linear–nonlinear,' or 'LN,' cascade model; it is a probabilistic analogue of what are called 'Wiener cascade' models (Hunter and Korenberg 1986) in the system identification literature. (Note that this model is not the same as a Volterra series model (Marmarelis and Marmarelis 1978): these two classes of systems have very different approximation properties.)

The LN model has two important features which recommend it for complex natural signal data. First, the spike trains of the cell are given by a conditionally (inhomogeneous) Poisson process given $\vec{x}$; that is, there are no dynamics in this model beyond those induced by $\vec{x}$ and K. This makes the LN cell a simple starting point for more detailed modelling. Second, equation (1) implies

$$p(\text{spike}|\vec{x}) = p(\text{spike}|\vec{x} + \vec{y}) \qquad \forall y \perp K. \tag{2}$$

In other words, the conditional probability of firing is constant along (hyper)planes in the input space. This model thus separates the quite difficult nonparametric problem of learning $p(\text{spike}|\vec{x})$ into two much simpler pieces: learning K and learning f. For example, if f is known, the problem of learning K reduces to a fairly standard parametric estimation problem (for which, say, maximum likelihood methods are generally efficient); conversely, if K is known, learning f entails the nonparametric estimation of a density, about which, again, much is known (see e.g. Devroye and Lugosi 2001). The semiparametric problem of estimating K without *a priori* knowledge of f seems to be much less well understood; we focus primarily on this problem here.

The main goal of this paper is to describe the convergence properties of three different types of K estimator: (1) techniques based on the spike-triggered average (STA) (Theunissen *et al* 2001); (2) techniques based on the spike-triggered covariance (Brenner *et al* 2001); and (3) a new, more general technique based on a probabilistic distance measure between the spike-triggered and 'no-spike'-triggered distributions. We were motivated by two basic questions. First, when do these estimators work (in the sense of 'consistency', that is, given enough data, do they provide us with an accurate estimate of K (Schervish 1995))? Second, when the estimator is consistent, what is the statistical rate of convergence (that is, how much data do we need to be close to the correct K)? Our first main result is that the first two K estimators often do not work, even given infinite data. More precisely, the conditions for consistency of these estimators turn out to be surprisingly stringent; for example, these conditions are typically not satisfied by natural signal stimulation paradigms. Our second main result provides an antidote of sorts: the novel estimator we introduce here converges under very general conditions to the

correct K. Finally, we provide various results on the rate of convergence of these three classes of K estimator. Together, these results serve to put the growing subfield of LN statistical modelling on a more solid theoretical foundation.

2. Notation; outline

The basic semiparametric model space we will work in is defined as follows. An LN model is completely specified by knowledge of K and f, plus the stimulus distribution $p(\vec{x})$; thus, LN models take values in the space

$$(p, K, f) \in \mu(X) \times \mathcal{G}_m(X) \times L_{[0,1]}(\Re^m),$$

where $\mu(X)$ denotes the space of all probability measures on X, $\mathcal{G}_m(X)$ is the space of all m-dimensional subspaces of X, and $L_{[0,1]}(\Re^m)$ is the space of measurable functions on $\Re^m$, taking values in $[0, 1]$. In most cases, p is held fixed and/or presumed known and we discuss only (K, f) instead of (p, K, f); this should be clear from context. Note that K, the main parameter of interest, is finite dimensional, while f, the 'nuisance' parameter in statistical jargon, is infinite-dimensional.

Let N denote the number of available samples, drawn i.i.d. from $p(\vec{x})$. We assume throughout this paper that $p(\vec{x})$ has zero mean and finite second moments; the first assumption obviously entails no loss of generality and the second seems entirely reasonable on physical grounds. Then our basic results will take the following form:

$$E(\text{Error}(\hat{K})) \approx \alpha N^{-\gamma} + \beta, \tag{3}$$

as N becomes large. The estimator $\hat{K}$ is a map taking N observations of stimulus and spike data (where spikes are binary random variables, conditionally independent given the stimulus) into an estimate of the true underlying K:

$$\hat{K} : (X \times \{0, 1\})^N \to \mathcal{G}_m(X)$$
$$(\vec{x}_N, s_N) \to \hat{K}(\vec{x}_N, s_N),$$

where $(\vec{x}_N, s_N)$ denotes the N-sample data; the natural error metric, then, is the geodesic distance on $\mathcal{G}_m(X)$ (the 'canonical angle') between the true subspace K and the estimated subspace $\hat{K}$:

$$\text{Error}(\hat{K}) \equiv \cos^{-1}(s(P_K^t P_{\hat{K}})),$$

where P_V denotes the projection operator corresponding to the subspace V and $s(A)$ denotes the smallest singular value of the operator A. For notational ease, we will mostly work in the $m = 1$ case; here the metric takes the explicit form

$$\text{Error}(\hat{K}) \equiv \cos^{-1} \frac{\langle \hat{K}, \vec{k}_1 \rangle}{\|\hat{K}\|_2 \|\vec{k}_1\|_2}.$$

The scalar terms γ, α and β in (3) each depend on f, K and $p(\vec{x})$; γ is a constant giving the order of magnitude of convergence (usually, but not always, equal to $1/2$), α gives the precise convergence rate and β gives the asymptotic error. We will mostly be concerned with giving exact values for α, and simply indicating when β is zero or positive (i.e. when $\hat{K}$ is consistent in probability or not, respectively).

Most of the remainder of the paper will be devoted to deriving representation (3), including the constants α, β and γ, for the three classes of K estimator mentioned in the introduction. We carry out this programme for the spike-triggered average and the spike-triggered covariance technique in sections 3 and 4, respectively. Section 5 contains perhaps the central results of this

paper; here we give details of the analysis and computation of a new, universally consistent estimator. In section 6, we present some simulation results comparing the performance of the three estimators and applications of the new estimator to real physiological data. We provide a few lower bounds on the convergence rate of any possible LN estimator in section 7; these bounds provide rigorous measures of the difficulty of the K-estimation problem. Finally (section 8), we close with a brief discussion of a few important areas for future research. Proofs appear in an appendix.

3. Spike-triggered averaging

The first estimator, the STA, is classical and very intuitive. We define

$$\hat{K}_{STA} \equiv \frac{1}{N_s} \sum_{i=1}^{N_s} \vec{x}_i, \tag{4}$$

where $\vec{x}_i$ is the ith stimulus for which a spike occurred and N_s denotes the total number of spikes observed (N and N_s are, of course, roughly proportional, with constant $p(\text{spike}) = \int_X p(\vec{x}) f(K\vec{x})$). As is well known, $\hat{K}_{STA}$ is simply the sample mean of the spike-conditional stimulus distribution $p(\vec{x}|\text{spike})$; since the spike signal is binary-valued, this is the same as the cross-correlation between the spike and the stimulus signal. We will also consider the following linear regression-like modification (Theunissen *et al* 2001):

$$\hat{K}_{RSTA} \equiv A\hat{K}_{STA},$$

where A is an operator chosen to 'rotate out' correlations in the stimulus distribution $p(\vec{x})$ (A is typically the (pseudo-) inverse of the stimulus correlation matrix, which we will denote as $\sigma^2(p(\vec{x}))$. In this section and the next, we assume that $\sigma^2(p(\vec{x}))$ is known; this assumption seems fair because either: (1) $p(\vec{x})$ is chosen by the experimenter, or (2), in the natural signal paradigm, a sufficient number of samples from the natural distribution are available that $\sigma^2(p(\vec{x}))$ can be estimated to arbitrary accuracy, i.e. the experimenter has access to many more examples than N, the number of samples seen by the neuron. At any rate, even if $\sigma^2(p(\vec{x}))$ is unknown, the basic analysis presented here still works, although slightly worse constants are obtained.

We begin with necessary and sufficient conditions for consistency. As usual, we say $p(\vec{x})$ is radially symmetric if $p(B) = p(UB)$ for all measurable sets B and all unitary transformations (rotations) U; examples include the standard multivariate Gaussian density or the uniform density on the sphere. (Note that, if $p(\vec{x})$ has this radial symmetry property, then $\hat{K}_{STA} = \hat{K}_{RSTA}$.) Finally, since $\hat{K}_{RSTA}$ clearly returns a single vector, that is, a one-dimensional subspace of X, assume for the moment that $K = \vec{k}_1$ (i.e. K is a one-dimensional subspace). Then we have the following:

Theorem 1 (Consistency: $\beta(\hat{K}_{STA})$). *If $p(\vec{x})$ (resp. $p(A^{1/2}\vec{x})$) is radially symmetric and $E(\langle\vec{x}, \vec{k}_1\rangle|\text{spike}) \neq 0$, then $\beta(\hat{K}_{STA}) = 0$ (resp. $\beta(\hat{K}_{RSTA}) = 0$), that is, the spike-triggered average estimator is consistent.*

Conversely, if $p(\vec{x})$ is radially symmetric and $E(\langle\vec{x}, \vec{k}_1\rangle|\text{spike}) = 0$, then $\beta > 0$, and if $p(\vec{x})$ is not radially symmetric, then there exists an f for which $\beta > 0$.

In other words, spike-triggered averaging techniques always work (given enough data) if the input distribution p is radially symmetric and if the neuron's tuning f is sufficiently asymmetric, in the sense that $|E(\langle\vec{x}, \vec{k}_1\rangle|\text{spike})| > 0$; conversely, it is not hard to find examples for which these conditions are not met and the STA fails to recover $\vec{k}_1$. The above sufficiency conditions are fairly well known; for example, most of the sufficiency statement appeared

(albeit in somewhat less precise form) in Chichilnisky (2001) (see also Ringach *et al* (1997) and references therein for related results; (Bussgang 1952) seems to be the earliest). The condition on $E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})$ is discussed in more depth below. Note the lack of restrictions on f; this function is not required to be smooth, or even continuous.

On the other hand, the converse is novel, to our knowledge, and is perhaps surprisingly stringent: without a highly restrictive symmetry condition on $p(\vec{x})$, spike-triggered averaging methods often remain biased, even given infinite data; thus, these estimators will typically converge, but not necessarily to the correct $\vec{k}$. As is well known, distributions of natural signals tend to lack this symmetry property (Simoncelli 1999, Ruderman and Bialek 1994); thus, STA analyses of natural signal data must be interpreted with caution. The first part of the necessity statement will be obvious from the following discussion of $\alpha(\hat{K}_{RSTA})$ (and, in fact, appears implicitly in Chichilnisky (2001)). The second part, while perhaps unsurprising given the analysis of Chichilnisky (2001), is a little harder and seems to require characteristic function (Fourier transform) techniques. The proof proceeds by showing that a distribution is symmetric iff it has the property that the conditional mean of $\vec{x}$ is zero on all planar 'slices' (i.e. $E(\langle \vec{u}, \vec{x} \rangle | \langle \vec{v}, \vec{x} \rangle \in B) = 0$ for all $\vec{u} \perp \vec{v} \in X'$ and real measurable sets B).

Next we have the rate of convergence, to give a rough idea of how many samples is 'enough':

Theorem 2 (Convergence rate: $\alpha(\hat{K}_{STA})$). *Assume $p(\vec{x})$ is symmetric normal, with standard deviation $\sigma(p)$. If $\beta(\hat{K}_{STA}) = 0$, then $N_s^{1/2}(\hat{K}_{STA} - K)$ is asymptotically symmetric normal with mean zero (considered as a distribution on the tangent plane of $\mathcal{G}_m(X)$ at the true underlying value K), and scale*

$$\frac{\sigma(p)}{E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})}.$$

Thus,

$$\alpha(\hat{K}_{STA}) = \frac{\sigma(p)}{|E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})|} \sqrt{\dim X - 1}.$$

Thus the asymptotic error of the STA scales directly with the dimension of the ambient space and inversely with $|E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})|$, a measure of the asymmetry of the spike-triggered distribution along k_1. The standard example of a neuron for which $|E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})|$ is small is a complex cell in V1, whose responses are roughly symmetric with respect to sign inversion. The theorem serves to quantify the well-known result that spike-triggered averaging works poorly, if at all, for neurons with this kind of response symmetry.

The proof follows by applying the multivariate central limit theorem to the sample mean of N_s random vectors drawn i.i.d. from the spike-conditional stimulus distribution, $p(\vec{x}|\text{spike})$. The proof also supplies the asymptotic distribution of $\text{Error}(\hat{K}_{STA})$ (a noncentral F), which might be useful for hypothesis testing. The details are easy once the mean of this distribution is identified (as in Chichilnisky (2001), under the above sufficiency conditions).

Note that we stated the result under stronger-than-necessary conditions (i.e. $p(\vec{x})$ is Gaussian instead of just symmetric) in order to simplify the statement. (In this case, the form of α becomes quite simple under these stronger assumptions; α depends on the nonlinearity f only through $E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})$. The general case is proven by identical methods but results in a slightly more complicated, f-dependent, term in place of $\sigma(p)$.) This pattern of stating non-optimal results in the text, then giving the stronger, more general results in the appendix will reappear without comment below.

One final note: in stating the above two results, we have assumed that K is one-dimensional. Nevertheless, the two theorems extend easily to the more general case, after

Error($\hat{K}_{STA}$) is redefined to measure angles between m- and one-dimensional subspaces. (Of course, now $E(\hat{K}_{STA})$ depends strongly on the input distribution $p(\vec{x})$, even for radially symmetric $p(\vec{x})$; see, e.g., Schwartz *et al* (2002) for an analysis of a special case of this effect.)

4. Covariance-based methods

The next estimator was introduced in an effort to extend spike-triggered analysis to the $m > 1$ case (see, e.g., de Ruyter and Bialek 1988, Brenner *et al* 2001, Schwartz *et al* 2002). Where $\hat{K}_{STA}$ was based on the first moment of the spike-conditional stimulus distribution $p(\vec{x}|\text{spike})$, $\hat{K}_{CORR}$ is based on the second moment. We define

$$\hat{K}_{CORR} \equiv (\sigma^2(p))^{-1}\text{eig}(\widehat{\Delta\sigma^2}),$$

where eig(A) denotes the significantly non-zero eigenspace of the operator A, and $\widehat{\Delta\sigma^2}$ is some estimate (typically the usual sample covariance estimate) of the 'difference-covariance' matrix $\Delta\sigma^2$, defined by

$$\Delta\sigma^2 \equiv \sigma^2(p(\vec{x})) - \sigma^2(p(\vec{x}|\text{spike})).$$

Again, we start with β:

Theorem 3 ($\beta(\hat{K}_{CORR})$). *If $p(\vec{x})$ is Gaussian and*

$$\text{var}_{p(\vec{x}|\text{spike})}(\langle\vec{k},\vec{x}\rangle) \neq \text{var}_{p(\vec{x})}(\langle\vec{k},\vec{x}\rangle) \qquad \forall\vec{k} \in E_K,$$

for some orthogonal basis E_K of K, then $\beta(\hat{K}_{CORR}) = 0$. Conversely, if $p(\vec{x})$ is Gaussian and the variance condition is not satisfied for f, then $\beta > 0$, and if $p(\vec{x})$ is non-Gaussian, then there exists an f for which $\beta > 0$.

As before, the sufficiency is fairly well known (see the thesis of Odelia Schwartz for a proof, or Brenner *et al* (2001) for a sketch), while the necessity appears to be novel and relies on characteristic function arguments. It is perhaps surprising that the conditions on p for the consistency of this estimator are even stricter than for the STA. The essential fact here turns out to be that a distribution is normal iff, after a suitable change of basis, the conditional variance on all planar 'slices' of the distribution is constant.

We have, with Odelia Schwartz, developed a striking inconsistency example which is worth mentioning here:

Example. (Inconsistency of $\hat{K}_{CORR}$). There is a nonempty open set of nonconstant f and radially symmetric $p(\vec{x})$ such that $\hat{K}_{CORR}$ is asymptotically orthogonal to K almost surely as $N \to \infty$. (In fact, the f and p in this set can be taken to be infinitely differentiable.)

The basic idea is that, for nonnormal p, the spike-triggered variance of $\langle\vec{v},\vec{x}\rangle$ depends on f even for $\vec{v} \perp \vec{k}$; thus, one can find f for which

$$|\text{var}_{p(\vec{x}|\text{spike})}(\langle\vec{k},\vec{x}\rangle) - \text{var}_{p(\vec{x})}(\langle\vec{k},\vec{x}\rangle)|$$

is small but

$$|\text{var}_{p(\vec{x}|\text{spike})}(\langle\vec{v},\vec{x}\rangle) - \text{var}_{p(\vec{x})}(\langle\vec{v},\vec{x}\rangle)|, \qquad \vec{v} \perp \vec{k}$$

is large. We leave the details to the reader.

We can derive a similar rate of convergence for these covariance-based methods. To reduce the notational load, we state the result for $m = 1$ only; in this case, we can define $\lambda_{\Delta\sigma^2}$ to be the (unique and nonzero by assumption) eigenvalue of $\Delta\sigma^2$.

Theorem 4 ($\alpha(\hat{K}_{CORR})$)**.** *Assume* $p(\vec{x})$ *is symmetric normal. If* $\beta(\hat{K}_{CORR}) = 0$*, then* $N_s^{1/2}(\hat{K}_{CORR} - K)$ *is asymptotically symmetric normal with mean zero and*

$$\alpha = \frac{\sigma(p)\sqrt{\sigma^2(p) - \lambda_{\Delta\sigma^2}}}{|\lambda_{\Delta\sigma^2}|}\sqrt{\dim X - 1}.$$

(Again, while $\lambda_{\Delta\sigma^2}$ will not be exactly zero in practice, it can often be small enough that the asymptotic error remains prohibitively large for physiologically reasonable values of N_s.) The proof proceeds by applying the multivariate central limit theorem to the covariance matrix estimator, then examining the first-order Taylor expansion of the eigenspace map at $\Delta\sigma^2$. It is also worth emphasizing that the asymptotics in the above theorem (and indeed, in all of the results in this paper) are in N only; the theorem is not valid if $\dim X$ grows as well. (See, e.g., Everson and Roberts (1999), Johnstone (2000) and references therein for some useful asymptotic results on eigenspace analysis in the case that $\dim X$ is of order N.)

5. ϕ-divergence techniques

We have seen that the two most common K estimators are not consistent in general; that is, the asymptotic error β is bounded away from zero for many (non-pathological) combinations of $p(\vec{x})$, f, and K. In particular, we have to place very strong conditions on p to guarantee that $\hat{K}_{RSTA}$ and $\hat{K}_{CORR}$ will converge to the correct K. We now introduce a new class of estimator which is consistent ($\beta = 0$) in general.

The basic idea is that $K\vec{x}$ is, in a sense, a sufficient statistic for $\vec{x}$: $\vec{x} - K\vec{x} -$ spike forms a Markov chain. Let us give a few definitions. Given a continuous, strictly convex real function ϕ on $[0, \infty]$, with $\phi(1) = 0$, define the ϕ divergence (following Csiszar (1967)) between two measures μ and ν as

$$D_\phi(\mu; \nu) \equiv \int \mathrm{d}\nu\, \phi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) = \int \mathrm{d}\mu\, \tilde{\phi}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right),$$

where $\tilde{\phi}(t) = t\phi(t^{-1})$ and the densities $\mathrm{d}\mu$ and $\mathrm{d}\nu$ are interpreted as likelihood ratios. The best-known ϕ divergence is the Kullback–Leibler divergence ($\phi(t) = t\log t$). The main property of ϕ divergences we need is the so-called data-processing inequality (Cover and Thomas 1991): for any Markov morphisms S and T

$$D_\phi(S(\mu); T(\nu)) \leqslant D_\phi(\mu; \nu),$$

with equality only if S and T are sufficient. The above inequality is named for the following special case: μ is $p(x, y)$, the joint distribution of some r.v.'s X and Y, and ν is the product $p(x)p(y)$ of their marginals. Then, for any Markov chain X–Y–Z, $D_\phi(p(x, y);\, p(x)p(y)) \geqslant D_\phi(p(x, z);\, p(x)p(z))$, with equality iff X—Z—Y (i.e. iff $Y(Z)$ is sufficient for X.

Thus, if we identify the random variable 'spike' with X in the above Markov chain, $\vec{x}$ with Y, and $\langle K, \vec{x}\rangle$ with Z, it is clear from (1) that $\langle K, \vec{x}\rangle$ is sufficient for $\vec{x}$ with respect to 'spike', and the data processing inequality states that

$$M_\phi(V) \equiv D_\phi(p(\langle V, \vec{x}\rangle, \text{spike});\, p(\langle V, \vec{x}\rangle)p(\text{spike})),$$

considered as a function of vector spaces V of dimension $\dim K$, reaches a maximum on K, and this maximum is unique under certain weak conditions. (When $\dim V > \dim K$, the maximum will no longer be unique, but it is easy to show that the maximizers still contain K.)

The basic idea is that $\langle V, \vec{x}\rangle$ is equivalent to $\langle K, \vec{x}\rangle$ plus some noise term that does not affect the spike process (more precisely, this noise term is conditionally independent of 'spike' given $\langle K, \vec{x}\rangle$); this noise term is obviously 0 for $V = K$, and the larger this noise, the smaller $M_\phi(V)$.

Another way to put it is that $M_\phi(V)$ measures how strongly $\langle V, \vec{x}\rangle$ modulates the firing rate of the cell: for V near K, the conditional measures $p(\text{spike}|\langle V, \vec{x}\rangle)$ are on average very different from the prior measure $p(\text{spike})$ and $M_\phi(V)$ is designed to detect exactly these differences. Conversely, for V orthogonal to K, the conditional measures $p(\text{spike}|\langle V, \vec{x}\rangle)$ will appear relatively 'unmodulated' (that is, $p(\text{spike}|\langle V, \vec{x}\rangle)$ will tend to be much nearer the average $p(\text{spike})$) and $M_\phi(V)$ will be comparatively small.

This all suggests that we could estimate K by maximizing $M_{\phi,N}(V)$, some estimator of the function $M_\phi(V)$. The rest of this section is devoted to describing the mathematical and computational properties of this type of estimator for several different forms of $M_{\phi,N}(V)$. The precise choice of ϕ here seems not to matter much for the asymptotic analysis, as long as ϕ is smooth enough; for mathematical and computational convenience, we choose $\phi(t) = t^2 - 1$. For this ϕ, a little algebra shows that

$$M_\phi(V) = \frac{\text{var}(p(\text{spike} = 1|\langle V, \vec{x}\rangle))}{p(\text{spike} = 1)\,p(\text{spike} = 0)}$$

(where, in a slight abuse of notation, $p(\text{spike} = 1)$ serves as a random variable—a function of $\langle V, \vec{x}\rangle$—in the numerator, and as a fixed probability in the denominator); this is a reasonably intuitive measure of firing rate modulation in the LN model. Originally, we chose this function because of the nice properties of ϕ and $t\tilde{\phi}(t)$ near zero, as previous work on the estimation of mutual information (Paninski 2003a) indicated that the smoothness of these functions plays a critical role in the estimability of D_ϕ; other advantages of this choice will become clear as we progress.

Before we move on to our main results, it is worth noting the recent work of Sharpee *et al* (2003), who independently presented an estimator based on maximizing the mutual information between $\langle V, \vec{x}\rangle$ and spike; this corresponds, in our notation, to maximizing $M_\phi(V)$, with $\phi(t) = t \log t$. While the methods and analysis presented below are somewhat more detailed than, and differ in several important respects from, those described in Sharpee *et al* (2003), it is worthwhile consulting their work for another illustration of the improved performance of this kind of estimator. We hope to undertake a more thorough comparison of the statistical and computational efficiency of the two estimators in the future.

5.1. Asymptotics

We will start by defining $M_{\phi,N}(V)$ more precisely. The simplest idea would be to let $M_{\phi,N}(V)$ be a 'plug-in' kernel or histogram estimator, that is,

$$M_{\phi,N}(V) \equiv D_\phi(\hat{p}_N(\langle V, \vec{x}\rangle, \text{spike});\ \hat{p}_N(\langle V, \vec{x}\rangle)\hat{p}_N(\text{spike})),$$

where $\hat{p}_N$, in turn, is an estimate of the underlying measure, either by kernel (that is, $\hat{p}_N$ is obtained by filtering the empirical measure

$$p_N \equiv \frac{1}{N}\sum_{i=1}^{N} \delta_i$$

according to some linear, shift-invariant kernel), or histogram (that is, X is partitioned into a countable number of bins and $\hat{p}_N$ is simply the discrete measure induced by p_N). We denote such an estimator of K by

$$\hat{K}_\phi \equiv \underset{V}{\text{argmax}}\, M_{\phi,N}(V).$$

We assume that the chosen kernel or histogram partition is roughly isotropic, and that the data has been pre-whitened, so that the global scale of the data is roughly the same for every V;

this helps to reduce the bias induced by the somewhat arbitrary scale imposed by the kernel width or average bin size. Fancier versions of these estimators adjust to the local scale as well (e.g. adaptive kernels or histograms), but for computational simplicity we will stick to nonadaptive estimators of the density for now. Obviously, for either type of estimator, we will have to let the kernel or bin width decrease with N; it is easy to come up with examples for which fixed bin width estimators fail (basically because, if the bin width is bounded from below, there exist f which are 'averaged over' by the kernel or histogram). Thus much of the labour in the analysis of these estimators is in dealing with shrinking bin sizes.

Our first result is a general consistency result for the kernel estimator. A nearly identical result holds for the histogram estimator.

Theorem 5 ($\beta(\hat{K}_\phi)$). *If p has a nonzero density with respect to Lebesgue measure, f is not constant a.e. and the kernel width goes to zero more slowly than N^{r-1}, for some $r > 0$, then $\beta = 0$ for the kernel estimator.*

In other words, this new estimator $\hat{K}_\phi$ works for very general neurons f and stimulus distributions p; in particular, $\hat{K}_\phi$ is suitable for application to natural signal data. Clearly, the condition on f is minimal; we ask only that the neuron be tuned. The condition on p is quite weak (and can be relaxed further); we are simply ensuring that we are sampling from all of X, and in particular, the part of X on which the cell is tuned.

Things get more complicated when it comes to computing the rate of convergence. The rough picture is as follows: for each V, $M_{\phi,N}(V)$ converges to $M_\phi(V)$, with an error that depends on N, V, the kernel width a_N and the parameters of the LN model (K, f, p). We have to choose a_N in such a way as to minimize the effect of these errors on $\hat{K}_\phi$. The error can be split up into a bias term and a variance term. It turns out that the variance term does not depend very strongly on a_N, so we ignore this for now. The bias term can be split up further into an approximation bias and a sample bias: the approximation bias measures the difference between $M_\phi(V)$ and its kernel- (or histogram-) smoothed version, defined in the obvious way, while the sample bias is the average difference between $M_{\phi,N}(V)$ and this smoothed version of $M_\phi(V)$. It is intuitively clear that these two types of bias behave differently as a function of a_N; if a_N goes to zero too slowly, the approximation bias will go to zero slowly but the sample bias will die quickly (roughly, because larger kernels or histogram bins average over more data), and vice versa. Thus, if we can compute the asymptotic approximation bias and sample bias as a function of a_N, we have a well-defined optimization problem: choose a_N to minimize their sum, the total bias.

We carry out this program in the appendix; the final result, for $m = 1$, for example, is that the sample bias behaves roughly like $(Na_N)^{-1}$, implying that the naive estimator $\hat{K}_\phi$ converges somewhat slowly. The following theorem follows from some simple algebra to obtain the optimal kernel width for minimizing the bias in $M_{\phi,N}(V)$, then a second-order expansion of $E(M_{\phi,N}(V))$ around K to obtain the corresponding behaviour for the bias of $\hat{K}_\phi$.

Theorem 6 (Bias of $\hat{K}_\phi$). *If the approximation error is of the order of a_N^r, then the optimal kernel width is of the order of $N^{-1/(r+1)}$, corresponding to an optimal bias in the kernel or histogram estimators which can be of the order of $N^{-r/[2(r+1)]}$.*

Again, a similar conclusion holds for the histogram version of $\hat{K}_\phi$. To understand what this result means for a given set of parameters (f, K, p), note that it is straightforward to show, using a Taylor expansion, that the approximation error behaves like a_N^2 if p is well behaved and f is, say, uniformly twice differentiable; this corresponds to a convergence rate of $N^{-1/3}$ for $\hat{K}_\phi$. As another example, step functions have an approximation error that behaves like a_N; this leads to an even slower convergence rate, $N^{-1/4}$.

This slow bias behaviour can be corrected using a standard statistical trick: we replace the naive 'plug-in' estimators for $M_{\phi,N}$ with their jackknifed versions, where for any function of the data $T(x_N)$, we define the jackknifed version of T to be

$$T_{JK} = NT - \frac{N-1}{N}\sum_{i=1}^{N} T_{-i},$$

where T_{-i} is T computed using all but the ith data sample. Simple computations prove that this procedure solves the bias problem (for simplicity, the next three results in this section are stated under some weak smoothness assumptions on f and p; see the appendix for details):

Proposition 7 (Jackknife bias). *If the kernel (or bin) width goes to zero more slowly than N^{r-1}, $r > 0$, then the sample bias of the jackknifed version of $M_{\phi,N}$ decays exponentially.*

This is almost enough to establish an $N^{-1/2}$ convergence rate for the estimator $\hat{K}_\phi$ given by maximizing the jackknifed kernel or histogram version of $M_{\phi,N}$, under suitable conditions on the smoothness of f. The last step is to show that $M_{\phi,N}(V)$ is asymptotically linear in N and smooth enough in V, that is,

$$M_{\phi,N}(V) = M_\phi(V) + p_N m_V + o_p(N^{-1/2}), \tag{5}$$

where $o_p(N^{-1/2})$ is a random variable which is negligible on an $N^{1/2}$ scale and

$$p_N m_V \equiv \frac{1}{N}\sum_{i=1}^{N} m_V(\vec{x}_i, \text{spike}_i)$$

denotes the 'empirical process' associated with some function $m_V(\vec{x}, \text{spike})$, uniformly differentiable in V and with $p(\vec{x}, \text{spike})$-mean zero. We leave the details behind representation (5) for the appendix; basically, m_V is computed as a derivative of $M_{\phi,N}(V)$. Now, the theory of empirical processes (van der Vaart and Wellner 1996) states that $p_N m_V$ converges in a suitable sense to a Gaussian stochastic process (this makes intuitive sense, given that, for fixed V, $p_N m_V$ is just a sample mean of N i.i.d. random variables with finite variance) and this leads, finally, to the asymptotic representation for $\hat{K}_\phi$:

Theorem 8 (γ and α for $(\hat{K}_\phi)$). *If the approximation error is of the order of a_N^r, $r > 1$, then the jackknifed kernel or histogram versions of $\hat{K}_\phi$, with bandwidth N^s, $-1 < s < -1/r$, converge at an $N^{-1/2}$ rate.*

Moreover, $N^{1/2}(\hat{K}_\phi - K)$ is asymptotically normal, with mean zero and

$$\alpha(\hat{K}_\phi) = (\text{trace}\, H^{-1} J H^{-1})^{1/2},$$

where

$$H \equiv \left.\frac{\partial^2 M(V)}{\partial V^2}\right|_K$$

and

$$J \equiv E_{p(\vec{x},\text{spike})}\left(\left.\frac{\partial m_V}{\partial V}\right|_K^2\right).$$

The methods follow, e.g., example 3.2.12 of van der Vaart and Wellner (1996)—basically, a generalization of the classical theorem on the asymptotic distribution of the maximum likelihood estimator in regular parametric families.

Numerical evidence indicates that $\alpha(\hat{K}_\phi)$ is often smaller than $\alpha(\hat{K}_{STA})$ or $\alpha(\hat{K}_{CORR})$ (that is, the ϕ-divergence estimator often converges faster than the STA or covariance methods, even in the cases when the latter two methods are known to converge to the correct K), but we have so far been unable to obtain any general bounds on these quantities. Section 6 details a few of these numerical experiments, using both simulated and real data; see also Sharpee *et al* (2003) for some simulations of a similar estimator using natural image data.

5.2. Computation

We still have not mentioned how to actually compute $\hat{K}_\phi$. Histogram methods for the evaluation of $M_{\phi,N}(V)$ suffer from several problems: it is difficult to non-adaptively place histogram partitions well for all V simultaneously, for example, and attempts to place the histogram adaptively greatly complicate hill-climbing algorithms for the maximization of $M_N(V)$. Kernel methods are more attractive, but require numerical integration of effectively unconstrained nonlinear functions over m-dimensional spaces. A more efficient approach is a 'resubstitution' estimator: we replace numerical integration with a kind of Monte Carlo integration, using the observed samples as our integration points. Thus, sticking with the example of $\phi(t) = t^2 - 1$, instead of computing the integral

$$\int \frac{\hat{p}(V\vec{x}, \text{spike})^2}{\hat{p}(V\vec{x})\hat{p}(\text{spike})} = \int \hat{p}(\text{spike}|V\vec{x})\hat{p}(V\vec{x}|\text{spike}),$$

we compute the sum

$$\frac{1}{N_s}\sum_{i\in S}\hat{p}(\text{spike} = 1|V\vec{x}_i) + \frac{1}{N - N_s}\sum_{i\in S^c}\hat{p}(\text{spike} = 0|V\vec{x}_i), \tag{6}$$

where S is the set of stimuli which induced a spike and ^{c} denotes the set complement. The conditional measures $\hat{p}(\text{spike}|V\vec{x})$ are estimated via kernel, as discussed above; again, the jackknife trick can be used to remove the sample bias and the asymptotic theory developed in the last section goes through.

To compute $\hat{K}_\phi$, now we have to maximize $M_{\phi,N}(V)$; unfortunately, this function is non-convex in general and no direct solution seems to exist. General iterative algorithms such as simulated annealing or gradient ascent with repeated restarts may, of course, be applied to this problem, but their convergence is extremely slow. We have developed a specialized ascent algorithm for maximizing expression (6) that is much more efficient. This algorithm makes use of several tricks which might be useful more generally for maximizing empirical functionals on spaces of vector spaces; we describe these ideas in turn below. We plan to make the algorithm publicly available at http://www.cns.nyu.edu/∼liam, in order to facilitate quantitative evaluation on as large a variety of neural and synthetic data as possible.

The basic algorithm alternates between a local step and a global step until a convergence criterion is satisfied. The local step is straightforward: given the current V_0, we compute the gradient of (6), using a smooth (Gaussian, say) kernel; call the gradient $\vec{e}_0 \perp V_0$. The global step consists of finding the constrained maximum of (6), where V is allowed to vary only over the circle

$$(1 + t^2)^{-1/2}(V_0 + t\vec{e}_0), \qquad t \in \Re. \tag{7}$$

The first, and most important, trick now is to compute $M_{\phi,N}(V)$ using, not a smooth kernel, but rather a simple boxcar function. This allows us to compute our function for all t very quickly (and therefore to find its global maximum over all t very quickly (noniteratively) as well. The idea is simple: for a boxcar kernel, $M_{\phi,N}(t)$ changes value a finite number of times, namely at the points t_i at which kernels centred on different points intersect. Since precomputing these 'crossing times' t_i is simple trigonometry, we only need to sort the times and keep track of the value of each change (this turns out to be very simple as well, since (6) is a sum over the same index i) to compute the full function. This mix of global and local maximizations greatly increases the efficiency of the algorithm. This also obviates the need for conjugate gradient ascent techniques (Press *et al* 1992), as the boxcar kernels make $M_{\phi,N}(V)$ highly unsmooth (i.e. we do not become trapped in any long, smooth valleys).

Our other tricks do not have quite the same impact, but are helpful nevertheless. The next two ideas are about choosing the search direction $\vec{e}_0$ intelligently when local maxima are

encountered (i.e. when the circle search described above returns V_0 as the global maximum over t). First, if we have kept a list of circles we have already searched over, we can use a self-avoiding procedure to choose our next search direction: basically, we choose the next search to be in the direction $\vec{e}_0$ such that

$$\vec{e}_0 = \operatorname*{argmax}_{\vec{e} \perp V_0} \max_z D(V_z \oplus \vec{e}_z, V_0 \oplus \vec{e}_0),$$

where $D(\ )$ denotes geodesic distance and z indexes all past searches. This prevents us from searching over ground we have already covered.

The second trick along these lines is a little more interesting, but requires that at least dim X searches have already been made (roughly, we will need the set of old search circles to span X before this method becomes useful). The idea is that, with each search, we gain some information on the global structure of $M_{\phi,N}(V)$ beyond the simpler local structure we use to choose gradients, do circle maximizations, and so on. If we can use this global information to guide our choice of the next search direction, we should gain in efficiency. The simplest way to do this is a variant of the principal component analysis-style trick used by the spike-triggered covariance estimator. We form two 'covariance' matrices, U and V, as follows: U is the covariance of a set of points $\vec{y}_i$ sampled randomly from the set of all previous search circles (this is a set of dim X-dimensional points, all of length unity) and V is the covariance of the same set of points, with norm scaled now by the value of $M_{\phi,N}(V)$ at each point, i.e. $M_{\phi,N}(\vec{y}_i)\vec{y}_i$. By the unitary symmetry of $M_{\phi,N}(V)$, we can hope that the 'variance' of the data $M_{\phi,N}(\vec{y}_i)\vec{y}_i$ should be largest near K, even if we have not searched (i.e. collected points $\vec{y}_i$) near K yet. Now our best guess at a good search direction $\vec{e}_0$ solves the usual eigenvector problem associated with the Rayleigh quotient corresponding to U and V.

Finally, we can use a few not-so-specialized tricks to help speed up convergence. Most of these are some version of the coarse-to-fine idea. Since the speed of the algorithm scales inversely with N, but the accuracy scales proportionally with N, we can run the algorithm for a few iterations on a subsampled data set (artificially reducing N) to get a rough estimate, then gradually scale up N to refine our original coarse estimate. Similar tricks can be played with the kernel width a and dim X, assuming f and K, respectively, vary slowly enough that coarsening makes sense.

6. Application to simulated and real data

In this section we give examples of data sets, both simulated and real, for which the novel estimator introduced in section 5 reveals structure that is either undetected or contaminated by the usual estimators $\hat{K}_{STA}$ and $\hat{K}_{CORR}$ (see, e.g., Paninski 2003b, Sharpee *et al* 2003 for further numerical comparisons).

6.1. Numerical comparisons

Figures 1–3 present simple comparisons of the performance of $\hat{K}_\phi$ to that of the standard estimators $\hat{K}_{STA}$ and $\hat{K}_{CORR}$ on simulated data. Simulations here have the advantage, as usual, that we know the 'right answer'; this allows us to rigorously quantify the distribution of error of these estimators in simple, easy-to-understand situations, and to illustrate, in a less technical way, some of the ideas presented in more mathematical language in the preceding sections. Each point in each of these first three figures corresponds to the error of the two estimators ($\hat{K}_\phi$ versus $\hat{K}_{STA}$ in 1 and 2, and versus $\hat{K}_{CORR}$ in figure 3), given N samples drawn i.i.d. from a fixed distribution $p(\vec{x})$ and presented to an LN model whose parameters were chosen randomly on each set of N trials. In each case, the LN model is one-dimensional ($m = 1$), for simplicity.

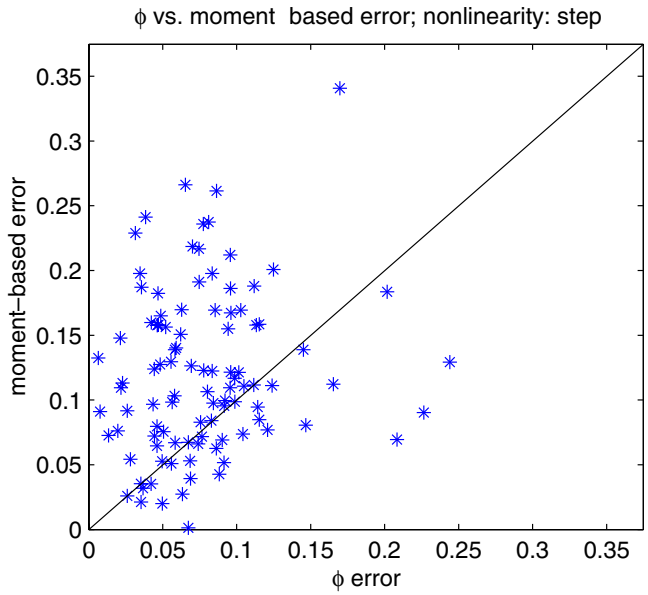

Figure 1. Plot of the error for $\hat{K}_\phi$ versus that of $\hat{K}_{STA}$. $p(\vec{x})$ = Gaussian white noise; f is a step function, where the step position is chosen randomly. Axes index error in radian units. $N = 80$ and dim $X = 3$ here; these small values were chosen for computational efficiency, but similar results are seen with larger values (see figure 3, for example). The error of $\hat{K}_\phi$ is slightly (but significantly) smaller than that of $\hat{K}_{STA}$ for these parameter settings.

In figure 1, we chose the input distribution $p(\vec{x})$ and the parameters of the LN model to be entirely favourable to the performance of $\hat{K}_{STA}$: $p(\vec{x})$ was chosen to be a standard Gaussian to satisfy the conditions of theorem 1 (implying that the STA does not suffer from an asymptotic bias), while the nonlinearity f was chosen to be a simple Heaviside step function (taking values zero and one), where the step position was chosen randomly according to a standard normal as well (by theorem 2, this form of f implies that $\alpha(\hat{K}_{STA})$ is always finite, and indeed fairly small with high probability; the value of the linear filter $\vec{k}$ is irrelevant, by the symmetry of p). Nevertheless, somewhat surprisingly, $\hat{K}_\phi$ significantly outperforms $\hat{K}_{STA}$ on average ($p < 0.05$, rank test).

We chose the LN model parameters randomly in figure 1, partly in an effort to emulate physical reality, where we have no control over the parameters, and partly to avoid picking an LN model that happened to confound either estimator to an abnormal degree. However, it is worth showing an example of the estimators' performance on a single, fixed model and input distribution, both because single models are perhaps easier to think about than a family of random models and in order to give a sense of the variability involved in the above numerical experiment. Thus, in figure 2, we present an identical simulation, except with the position of the step in the nonlinearity f (the only random parameter) fixed at 0. The results are essentially identical, if anything favouring the new estimator even more.

In figure 3, we use a nonlinearity which is more suited to $\hat{K}_{CORR}$: f is quadratic, of the form

$$f(t) = a(t - b)^2,$$

with a, b chosen randomly ($a > 0$; we have in mind an energy-type model for visual cortex cells (see, e.g., Simoncelli and Heeger 1998 and references therein). For Gaussian input data,

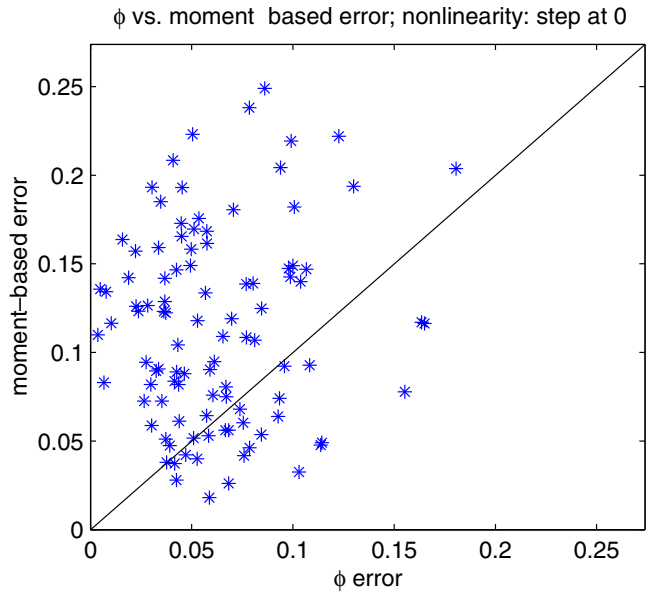

Figure 2. Plot of the error for $\hat{K}_\phi$ versus that of $\hat{K}_{STA}$; parameters as in figure 1, except the step is always at zero. Conventions are as in figure 1.

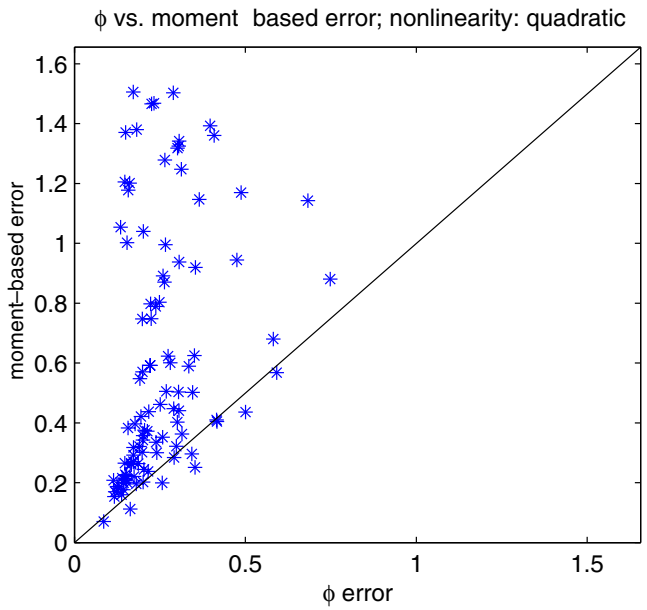

Figure 3. Plot of the error for $\hat{K}_\phi$ versus that of $\hat{K}_{CORR}$ $p(\vec{x}) =$ uniform on a hypercube; $\vec{k}$ is chosen randomly; f is quadratic, with the centre and scale chosen randomly. $N = 200$ and dim $X = 10$ here; conventions are as in figure 1.

$\hat{K}_{CORR}$ is competitive with $\hat{K}_\phi$ (data not shown), as expected given theorem 3. To provide a physiologically plausible example for which this is not the case, we took the input distribution

$p(\vec{x})$ to be uniform on a hypercube; this corresponds, for example, to a temporal signal whose value is chosen independently and identically distributed at each time step. The physical examples we have in mind here are the full-field white visual flicker stimulus employed, for example, in Berry and Meister (1998), Chander and Chichilnisky (2001), or the random 'checkerboard' spatial stimulus used in cortical and thalamic studies (e.g. Reid and Shapley 2002 and references therein). Finally, we chose the linear projection $\vec{k}$ randomly on the sphere for each new set of data; the results did not depend strongly on the identity of the chosen filter (for instance, the ratio $\text{Error}(\hat{K}_{CORR})/\text{Error}(\hat{K}_{\phi})$ was uncorrelated with the smoothness of $\vec{k}$; data not shown). Figure 3 shows that the new estimator outperforms the covariance-based estimator by a wide margin, essentially because of the asymptotic bias effects caused by the non-Gaussian nature of the data, as discussed in theorem 3.

6.2. Motor cortical data

In the preceding, we provided some encouraging numerical comparisons between $\hat{K}_{STA}$, $\hat{K}_{CORR}$ and the new estimator $\hat{K}_{\phi}$. This last subsection presents some preliminary results which are of interest more for their physiological relevance than for methodological reasons.

We have begun to apply these new spike-triggered analysis techniques to data collected in the primary motor cortex (MI) of awake, behaving monkeys, in an effort to elucidate the neural encoding of time-varying hand position signals in MI. This analysis has led to several interesting findings on the encoding properties of these neurons, with immediate applications to the design of neural prosthetic devices (Paninski *et al* 2002, Shoham *et al* 2003). The monkeys are performing a random drawing task, designed roughly to mimic everyday (for humans, but perhaps not monkeys in the wild) manual movement (for methodological details, see Paninski *et al* 1999, Fellows 2001, Paninski *et al* 2003a, Serruya *et al* 2002); the 'stimulus' space X in this context is the fairly high-dimensional space of time-varying hand position signals.

One novel and surprising result of this analysis is that the relevant K for MI cells appear to be one-dimensional. In other words, the conditional firing rate of these neurons, given a specific time-varying hand path, is well captured by the following model (figure 4): $p(\text{spike}|\vec{x}) = f(\langle\vec{k}_1, \vec{x}\rangle)$, where $\vec{x}$ represents the two-dimensional hand position signal in a temporal neighbourhood of the current time, $\vec{k}_1$ (in a slight abuse of notation) is a cell-specific affine functional and $f(t)$ is a scalar nonlinearity which turns out to be relatively cell-independent. There is no reason to have assumed MI cells would have this kind of one-dimensional tuning—for example, it is easy to find V1 cells which are notably multidimensional (e.g. Touryan *et al* 2002, Rust *et al* 2003)—but it is not hard to see that our observations are consistent with and extend the classical 'cosine' model of MI tuning (Georgopoulos *et al* 1986, Moran and Schwartz 1999).

We support the qualitative one-dimensional picture in figure 4 with two somewhat more quantitative results. First, we could find no two-dimensional parametric model which fits the nonlinearity (in a likelihood sense) better than a simple one-dimensional model in any of the cells we examined (after suitable correction for differences in dimensionality (Schwarz 1978)). Second, the mutual information in the second most modulatory axis $I(\text{spike}; \langle\vec{k}_2, \vec{x}\rangle)$ is not significantly different from zero (according to a Monte Carlo test constructed by simulating spike trains from a one-dimensional model whose parameters were matched and whose inputs were identical to those of the real cell, then estimating K and $I(\text{spike}; \langle\vec{k}_2, \vec{x}\rangle)$ for this model, repeating the procedure often enough to construct a nonparametric estimate of the null distribution to test against). Further details of $\vec{k}_1$, f and this information analysis will be presented elsewhere (Paninski *et al* 2002, 2003b, Shoham *et al* 2003).

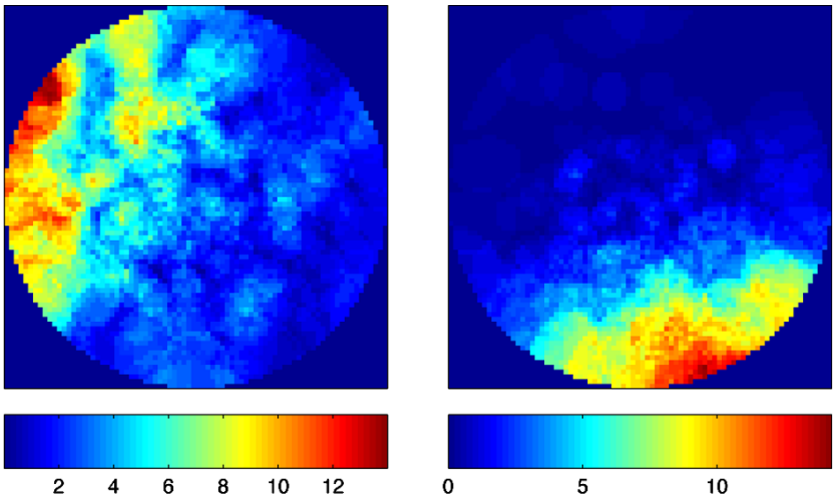

Figure 4. Example of $\hat{f}(\hat{K}\vec{x})$ functions, computed from two different MI cells, with rank $\hat{K} = 2$; the x- and y-axes index $\langle \hat{k}_1, \vec{x} \rangle$ and $\langle \hat{k}_2, \vec{x} \rangle$, respectively, while the colour axis indicates the value of $\hat{f}$ (the conditional firing rate given $\hat{K}\vec{x}$), in Hz. The scale on the x and y axes is arbitrary and has been omitted. $\hat{K}$ was computed using the ϕ-divergence estimator and $\hat{f}$ was estimated using an adaptive kernel within the circular region shown (where sufficient data were available for reliable estimates). Note that the contours of these functions are approximately linear; that is, $\hat{f}(\hat{K}\vec{x}) \approx f_0(\langle \vec{k}_1, \vec{x} \rangle)$, where $\vec{k}_1$ is the vector orthogonal to the contour lines and f_0 is a suitably chosen scalar function on the line.

(This figure is in colour only in the electronic version)

7. Lower bounds

Our final mathematical results are lower bounds on the convergence rates of any possible K estimator; these kinds of bounds provide rigorous measures of the difficulty of a given estimation problem, or of the efficiency of a given estimator. The first lower bound is local, in the sense that we assume that the true parameter is known *a priori* to be in some small neighbourhood of parameter space. Recall that the Hellinger metric between any two densities is defined as (half of) the L_2 distance between the square roots of the densities.

Theorem 9 (Local (Hellinger) lower bound). *For simplicity, let p be standard normal. For any fixed differentiable f, uniformly bounded away from 0 and 1 and with a uniformly bounded derivative f', and any Hellinger ball $\mathcal{F}$ around the true parameter (f, K),*

$$\liminf_{N\to\infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(\text{Error}(\hat{K})) \geqslant \left(\sigma(p)\left(E_p\left(\frac{|f'|^2}{f(1-f)}\right)\right)^{1/2}\right)^{-1} \sqrt{\dim X - 1}.$$

The second infimum above is taken over all possible estimators $\hat{K}$. The right-hand side plays the role of the inverse Fisher information in the Cramer–Rao bound and is derived using a similarly local analysis; see Jongbloed (2000) for details on the Hellinger technique or Gill and Levit (1995) on the Bayesian Cramer–Rao technique.

Global bounds are more subtle. We want to prove something like

$$\liminf_{N\to\infty} a_N \inf_{\hat{K}} \sup_{\mathcal{F}(\epsilon)} E(\text{Error}(\hat{K})) \geqslant C(\epsilon),$$

where $\mathcal{F}(\epsilon)$ is some large parameter set containing, say, all K and all f for which some relevant measure of tuning is greater than ϵ, a_N is the corresponding convergence rate and $C(\epsilon)$ plays

the role of $\alpha(\hat{K})$ from the previous sections. So far, our most interesting results in this direction are negative:

Theorem 10 (ϕ-divergences are poor indices of K-difficulty). *Let $\mathcal{F}(\epsilon)$ be the set of all (K, f) for which the ϕ-divergence 'information' between $\vec{x}$ and* spike *is greater than ϵ, that is,*

$$D_\phi(p(K\vec{x}, \text{spike});\ p(\text{spike})\,p(K\vec{x})) > \epsilon.$$

Then, for $\epsilon > 0$ small enough, for any putative convergence rate a_N,

$$\liminf_{N\to\infty} a_N \inf_{\hat{K}} \sup_{\mathcal{F}(\epsilon)} E(\text{Error}(\hat{K})) = \infty.$$

In other words, strictly information-theoretic measures of tuning do not provide a useful index of the difficulty of the K-learning problem; the intuitive explanation of this result is that purely measure-theoretic distance functions, like ϕ divergences, ignore the topological and vector space structure of the underlying probability measures, and it is exactly this structure that determines the convergence rates of any efficient K estimator. To put it more simply, the learnability of K depends on the smoothness of f, just as we saw in the last section (cf theorem 6), a common theme in nonparametric statistics.

8. Conclusion and directions for future work

We have presented here a fairly detailed analysis of the statistical properties of the LN model (1). In particular, we have attempted to elucidate when and why the common estimators for the LN model parameters work well, or not. More importantly, we have provided a new estimator which is guaranteed to recover the true parameters in much greater generality than was previously possible. We hope that our results will find application in understanding the neural processing of naturalistic stimuli; as mentioned briefly in section 6.2, these methods have already led to a better understanding of the neural coding of dynamic hand movement signals in primary motor cortex.

We take this opportunity to outline one obvious avenue for future work: how do we extend the basic LN model (1) in a way that allows us to capture more of the details of the neural code, while at the same time retaining some of the simplicity that allows us to estimate the model?

8.1. Non-Poisson effects

As noted in the introduction, model (1) generates spike trains which are (conditionally inhomogeneous) Poisson processes (note that, even if the stimulus ensemble is time-translation-invariant, the spike train is not necessarily a marginally homogeneous Poisson process); given the input signal $\vec{x}$, the spikes in one time bin do not depend on those in any other nonoverlapping bin. We can extend this model by allowing spikes which are close to each other in time to be dependent (the importance of such an extension has been noted in several contexts; see, e.g., Berry and Meister (1998), Brown *et al* (2002) and Pillow and Simoncelli (2003)). Some natural questions immediately arise. Does the standard spike-triggered analysis fail in this case? If so, why? Can we correct for these non-Poisson effects? We can give at least preliminary answers to all of these questions, at least in the following special case:

$$p(\text{spike}|\vec{x}, s_-) = f(\langle\vec{k}_1, \vec{x}\rangle, \langle\vec{k}_2, \vec{x}\rangle, \ldots, \langle\vec{k}_m, \vec{x}\rangle)g(T(s_-)). \tag{8}$$

Here T is some arbitrary statistic of s_-, the spike train up to the present time (e.g. T could encode the time since the last spike); the 'modulation function' g maps the range of T into the half-interval $[0, \infty)$. The only conditions on f and g are those necessary to make $p(\text{spike}|\vec{x}, s_-)$

a regular conditional distribution (aside from measurability issues, it is sufficient that $f, g \geqslant 0$, $fg \leqslant 1\,\forall(K\vec{x}, s_-)$).

To see why the memory effects displayed by (8) complicate the analysis presented in the previous sections, recall the basic idea behind Chichilnisky's proof of the fact that, for model (1), whenever $p(\vec{x})$ is radially symmetric, $E(\hat{K}_{STA})$ lies in K (we are abusing notation slightly; K here denotes the subspace generated by K, which is assumed to be one-dimensional, as in section 3). We will write $E(\hat{K}_{STA})$ out and show the essential point of the proof; then we will show why the memory effects seen in (8) cause problems and how these problems can be 'fixed', in some suitable sense. We have

$$\begin{aligned} E(\hat{K}_{STA}) &= \int p(\vec{x}|\text{spike})\vec{x}\,\mathrm{d}\vec{x} \\ &= \int p(\text{spike}|\vec{x})\frac{p(\vec{x})}{p(\text{spike})}\vec{x}\,\mathrm{d}\vec{x} \\ &= \int f(\langle\vec{K},\vec{x}\rangle)\frac{p(\vec{x})}{p(\text{spike})}\vec{x}\,\mathrm{d}\vec{x}. \end{aligned}$$

The first equality is Bayes, the second (1). The essential point is that the conditional probability of a spike given $\vec{x}$ depends only on $\langle\vec{K},\vec{x}\rangle$—the proof that $E(\hat{K}_{STA}) \in K$ follows immediately (after a suitable change of basis). This key equality does not hold in general for (8):

$$\begin{aligned} p(\text{spike}|\vec{x}) &= \int p(\text{spike}|\vec{x}, s_-)\,p(s_-|\vec{x})\,\mathrm{d}s_- \\ &= \int f(\langle\vec{K},\vec{x}\rangle)g(T(s_-))\,p(s_-|\vec{x})\,\mathrm{d}s_- \\ &= f(\langle\vec{K},\vec{x}\rangle)\int g(T(s_-))\,p(s_-|\vec{x})\,\mathrm{d}s_- \\ &= f(\langle\vec{K},\vec{x}\rangle)h(\vec{x}). \end{aligned}$$

The first equality is (8), the second linearity; the last is by way of definition: h is an abbreviation for the conditional expectation of $g(T(s_-))$ given $\vec{x}$. If $g \equiv 1$ (as in (1)), then $h(\vec{x}) \equiv 1$, and we recover $E(\hat{K}_{STA}) \in K$. However, in general, h is nonconstant in $\vec{x}$: h depends on $\vec{x}$ not only through its projection onto $\vec{K}$ but also through its projection on all time-translates of K to the left (i.e. all functions $k_{-\tau}$ such that $k_{-\tau}(t) = k(t+\tau)$, for some $k \in K$ and $\tau > 0$). Most K, of course, are not time-translation-invariant. This breaks the proof and the result; indeed, it is easy to think of simple (non-pathological) examples of f, g, and radially symmetric $p(\vec{x})$ for which $E(\hat{K}_{STA}) \notin K$.

So we need to modify $\hat{K}_{STA}$ somehow to bring its expectation back into the desired subspace. Assume for simplicity that g is bounded below away from zero and that g and $T(s_-)$ are known (the simultaneous estimation of f, g, and K appears to be more difficult; no consistent estimator for (f, g, K) seems to be known, although attempts have appeared, e.g., Berry and Meister (1998). Aguera y Arcas *et al* (2001) suggest ignoring all spikes for which $g(T(s_-)) \neq 1$: i.e. form

$$\hat{K}_{STA^*} \equiv \frac{1}{N_s}\sum_{i\in S}\delta(g(T(s_{i-})) - 1)\vec{x}_i,$$

where S, again, indicates the set of stimuli corresponding to spikes and δ is the usual Dirac functional. However, the above string of equations shows that this procedure can actually make the situation worse: this effectively sets g equal to zero at all of these points where $g \neq 1$, which, in many cases, makes h more strongly $\vec{x}$-dependent, not less. In addition, of course,

ignoring these 'bad' spikes is expensive from a data collection point of view. An obvious alternative would be to form

$$\hat{K}_{STA^*} \equiv \frac{1}{N_s} \sum_{i \in S} g(T(s_{i-}))^{-1} \vec{x}_i.$$

It is easy to see, from the above discussion, that $E(\hat{K}_{STA^*}) \in K$.

More complete analysis of this kind of model and estimator would clearly be useful.

Acknowledgments

We thank Eero Simoncelli, Jonathan Pillow and Odelia Schwartz for extensive useful discussions, and Nicole Rust and Tatyana Sharpee for preliminary discussions of Sharpee *et al* (2003). This work was supported by a predoctoral fellowship from the Howard Hughes Medical Institute.

Appendix A. Proofs

A.1. $\hat{K}_{STA}$

Consistency of $\hat{K}_{RSTA}$*: sufficiency.* By the strong law of large numbers, the proof comes down to a bias calculation, and Chichilnisky's proof (Chichilnisky 2001) for $\hat{K}_{STA}$ illustrates the source of this bias very nicely. First, the conditional expectation $E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})$ exists by the finite-variance assumption on $p(\vec{x})$. Then, for $\hat{K}_{RSTA}$, we have the following string of equalities:

$$\begin{aligned}
E(\hat{K}_{RSTA}) &= E(A\hat{K}_{STA}) \\
&= \int A\vec{x}\, p(\vec{x}|\text{spike})\, \mathrm{d}\vec{x} \\
&= \int A\vec{x} \frac{p(\text{spike}|\vec{x})}{p(\text{spike})} p(\vec{x})\, \mathrm{d}\vec{x} \\
&= \int A\vec{x} \frac{f(\langle \vec{K}, \vec{x} \rangle)}{p(\text{spike})} p(\vec{x})\, \mathrm{d}\vec{x} \\
&= \int A^{1/2}\vec{y} \frac{f(\langle \vec{K}, A^{-1/2}\vec{y} \rangle)}{p(\text{spike})} p(A^{-1/2}\vec{y}) |A|^{1/2}\, \mathrm{d}\vec{y} \\
&= \int A^{1/2}\vec{y} \frac{f(\langle A^{-1/2}\vec{K}, \vec{y} \rangle)}{p(\text{spike})} p(A^{-1/2}\vec{y}) |A|^{1/2}\, \mathrm{d}\vec{y} \\
&= A^{1/2} \int \vec{y} \frac{f(\langle A^{-1/2}\vec{K}, \vec{y} \rangle)}{p(\text{spike})} p(A^{-1/2}\vec{y}) |A|^{1/2}\, \mathrm{d}\vec{y}.
\end{aligned}$$

The first two equalities are by definition, the third Bayes, the fourth (1), the fifth a linear change of coordinates $y = A^{1/2}x$, the sixth by the symmetry of $A^{-1/2}$, and the seventh by linearity. The rest of the proof follows Chichilnisky (2001) (see also section 8.1).

Consistency of $\hat{K}_{STA}$*: necessity.* The claim is that, if p is asymmetric, then there exists some f and $\vec{v}$ for which

$$\int \vec{x} \frac{f(\langle \vec{v}, \vec{x} \rangle)}{p(\text{spike})} p(\vec{x})\, \mathrm{d}\vec{x} \neq C_{f,\vec{v}}\vec{v},$$

for some scalar $C_{f,\vec{v}}$. The claim is equivalent to the following: if

$$\int \vec{x} \frac{f(\langle \vec{v}, \vec{x} \rangle)}{p(\text{spike})} p(\vec{x})\, \mathrm{d}\vec{x} = C_{f,\vec{v}} \vec{v} \qquad \forall f, \vec{v}, \tag{9}$$

then p is symmetric. It suffices to prove (9) for simple functions, that is, $f = 1$ on some set B, and $f = 0$ everywhere else. Thus condition (9) reduces to

$$\int_{\langle \vec{v}, \vec{x} \rangle \in B} \langle \vec{u}, \vec{x} \rangle\, p(\vec{x})\, \mathrm{d}\vec{x} = 0 \qquad \forall B \in \mathcal{B}, \vec{u} \perp \vec{v},$$

with $\mathcal{B}$ the class of all measurable sets. This, in turn, implies that the characteristic function (Fourier transform) of $p(\vec{x})$, $\check{p}(\vec{s})$, satisfies the following differential equation:

$$\frac{\partial \check{p}(\vec{s})}{\partial \vec{t}} = 0 \qquad \forall \vec{s} \perp \vec{t}.$$

Since $\check{p}$ is everywhere differentiable, by the finite-power assumption on p, the above equation implies that $\check{p}(\vec{s})$ is radially symmetric in $\vec{s}$, which, finally, implies the symmetry of p. □

Convergence rate $\alpha(\hat{K}_{RSTA})$. Assume p is elliptically symmetric. By the multivariate CLT and the computations above, $\hat{K}_{RSTA}$ is asymptotically normally distributed with mean

$$E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike}) \vec{k}_1$$

and covariance matrix

$$\frac{1}{N_s} A^2 C,$$

where C denotes $\sigma^2(p(\vec{x}|\text{spike}))$.

The asymptotic error behaves like the norm of this distribution orthogonal to $\vec{k}_1$, normalized by the projection of the mean of the distribution onto $\vec{k}_1$. The final result is

$$\alpha = \frac{(\text{trace}\, E^t A^2 C E)^{1/2}}{|E(\langle \vec{x}, \vec{k}_1 \rangle | \text{spike})|},$$

where E is any matrix whose columns are an orthonormal basis for the subspace of X' orthogonal to K. This reduces to the quoted result when p is Gaussian (in which case

$$E^t A^2 C E = E^t A E$$

and white. □

A.2. $\hat{K}_{CORR}$

Consistency of $\hat{K}_{CORR}$*: necessity.* The argument for $\hat{K}_{CORR}$ is similar to that for $\hat{K}_{STA}$. We want to prove that, if p is non-Gaussian, then there exists some f and $\vec{u} \perp \vec{v}$ for which

$$\int \langle \vec{u}, (\vec{x} - E_{p(\vec{x}|\text{spike})} \vec{x}) \rangle^2 \frac{f(\langle \vec{v}, \vec{x} \rangle)}{p(\text{spike})}\, \mathrm{d}p(\vec{x}) \neq \int \langle \vec{u}, \vec{x} \rangle^2\, \mathrm{d}p(\vec{x})$$

(recall we assumed that $E_{p(\vec{x})} \vec{x} = 0$). Without loss of generality, we assume that p is white, that is,

$$\int \langle \vec{u}, \vec{x} \rangle^2\, \mathrm{d}p(\vec{x}) = 1 \qquad \forall \vec{u} : \|\vec{u}\|_2 = 1;$$

translating into the contrapositive again, we reformulate the claim as follows: if

$$\int \langle \vec{u}, (\vec{x} - E_{p(\vec{x}|\text{spike})} \vec{x}) \rangle^2 \frac{f(\langle \vec{v}, \vec{x} \rangle)}{p(\text{spike})}\, \mathrm{d}p(\vec{x}) = 1 \qquad \forall f, \vec{u} \perp \vec{v}, \tag{10}$$

then p is Gaussian. The proof proceeds in two stages: first, we prove that the above condition implies that p is symmetric (making use of the result for $\hat{K}_{STA}$); then we prove that any symmetric p satisfying (10) is Gaussian. Again, we may restrict our attention to simple functions f.

First the symmetry. Note that (10) can be written as a mixture of conditional variances, given $\langle \vec{v}, \vec{x} \rangle$. More formally, for simple f,

$$\int \langle \vec{u}, (\vec{x} - E_{p(\vec{x}|\text{spike})}\vec{x}) \rangle^2 \frac{f(\langle \vec{v}, \vec{x} \rangle)}{p(\text{spike})} \, \mathrm{d}p(\vec{x}) = \frac{1}{p(\text{spike})} \int_{\langle \vec{v}, \vec{x} \rangle \in B} \langle \vec{u}, (\vec{x} - E_{p(\vec{x}|\text{spike})}\vec{x}) \rangle^2 \, \mathrm{d}p(\vec{x});$$

in other words, (10) says that the conditional variance of $\langle \vec{u}, \vec{x} \rangle$, given $\langle \vec{v}, \vec{x} \rangle \in B$, is constant (for all vectors $\vec{u} \perp \vec{v}$ and all measurable sets B). Now consider disjoint subsets of B, B_1 and B_2, $B = B_1 \cup B_2$. It is clear that the following 'mixture' equation holds:

$$\begin{aligned} p(\langle \vec{u}, \vec{x} \rangle | \langle \vec{v}, \vec{x} \rangle \in B) = \frac{1}{p(\langle \vec{v}, \vec{x} \rangle \in B)} \big(& p(\langle \vec{v}, \vec{x} \rangle \in B_1) p(\langle \vec{u}, \vec{x} \rangle | \langle \vec{v}, \vec{x} \rangle \in B_1) \\ & + p(\langle \vec{v}, \vec{x} \rangle \in B_2) p(\langle \vec{u}, \vec{x} \rangle | \langle \vec{v}, \vec{x} \rangle \in B_2) \big). \end{aligned}$$

Now, since the mixture of two densities with the same positive finite variance but different means has strictly greater variance than either of the two original densities, and each component in the above equation has the same variance, each component must also have the same mean. That is,

$$\begin{aligned} \int_{\langle \vec{v}, \vec{x} \rangle \in B} \langle \vec{u}, \vec{x} \rangle p(\vec{x}) \, \mathrm{d}\vec{x} &= \int_{\langle \vec{v}, \vec{x} \rangle \in B_1} \langle \vec{u}, \vec{x} \rangle p(\vec{x}) \, \mathrm{d}\vec{x} \\ &= \int_{\langle \vec{v}, \vec{x} \rangle \in B_2} \langle \vec{u}, \vec{x} \rangle p(\vec{x}) \, \mathrm{d}\vec{x} \\ &= 0. \end{aligned}$$

The above equations hold for all such B, B_1, B_2, and are equivalent to condition (9) from the proof for $\hat{K}_{STA}$; thus p is symmetric.

Now, given that p is symmetric, we can write

$$p(\vec{x}) = g(\|\vec{x}\|_2^2)$$

for some scalar function g; it turns out that (10) provides us with a simple differential equation for p (and hence g) in Fourier space. For simple f and symmetric p, (10) reduces to

$$\int_{\langle \vec{v}, \vec{x} \rangle \in B} \langle \vec{u}, \vec{x} \rangle^2 \, \mathrm{d}p(\vec{x}) = \int_{\langle \vec{v}, \vec{x} \rangle \in B} \mathrm{d}p(\vec{x}) \qquad \forall B \in \mathcal{B}, \vec{u} \perp \vec{v}.$$

In the Fourier domain, this means that

$$\frac{\partial^2 \check{p}}{\partial \vec{t}^2} = -\check{p}(\vec{s}), \qquad \forall \vec{s} \perp \vec{t} : \|\vec{t}\|_2 = 1.$$

Applying this equation to g, we find that

$$\frac{\partial \check{g}(s)}{\partial s} = -\check{g}/2,$$

i.e.

$$\check{g}(s) = c\mathrm{e}^{-s/2},$$

for some constant c; the proof is complete upon applying the inverse Fourier transform and normalizing. □

Convergence rate $\alpha(\hat{K}_{CORR})$. Assume p is nondegenerate Gaussian. By the multivariate CLT, $\widehat{\Delta\sigma^2}$ is asymptotically normal with mean $\Delta\sigma^2$ and covariance

$$C \equiv E(\widehat{\Delta\sigma^2_{ij}} - \Delta\sigma^2_{ij})(\widehat{\Delta\sigma^2_{gh}} - \Delta\sigma^2_{gh}) = \frac{1}{N_s}(\sigma^2_{s,ih}\sigma^2_{s,jg} + \sigma^2_{s,ig}\sigma^2_{s,jh}),$$

where σ_s abbreviates the spike-triggered covariance matrix. Again, the proof relies on an analysis of the (normalized) behaviour of the estimate on the orthogonal complement of K. This comes down to the usual local analysis, as follows.

Let eig_1 denote a top eigenvector map, that is, a map

$$\mathrm{eig}_1 : \Re^{(\dim X)^2} \to \Re^{\dim X},$$

taking a matrix to its (normalized) top eigenvector (in the case we will be dealing with, this map is uniquely defined almost surely). We know that, for N_s large enough,

$$\mathrm{eig}_1 \widehat{\Delta\sigma^2} - \mathrm{eig}_1 \Delta\sigma^2 \approx D\,\mathrm{eig}_1(\Delta\sigma^2)(\widehat{\Delta\sigma^2} - \Delta\sigma^2),$$

where $D\,\mathrm{eig}_1(\Delta\sigma^2)$ denotes the Jacobian matrix of eig_1 at the point $\Delta\sigma^2$; $\widehat{\Delta\sigma^2} - \Delta\sigma^2$ is normally distributed with mean zero and covariance C, and so we know everything we need to know about the asymptotic behaviour of $\widehat{\Delta\sigma^2}$ if we can compute $D\,\mathrm{eig}_1(\Delta\sigma^2)$ on the (proper) subspace of $\Re^{(\dim X)^2}$ on which C is a positive definite operator (this subspace is clearly contained in the space of all possible symmetric matrices, for example). The final result is that

$$\alpha(\hat{K}_{CORR})N_s^{-1/2} = (\mathrm{trace}\, E(\sigma^2)^{-1} D\,\mathrm{eig}_1(\Delta\sigma^2) C D\,\mathrm{eig}_1(\Delta\sigma^2)^t (\sigma^2)^{-1} E^t)^{1/2}.$$

The computation of the derivative turns out to be fairly straightforward. We want to look at how much the symmetric perturbation ϵB, ϵ small, affects the ith component of the first eigenvector of the symmetric matrix $A = VDV^t$, with V orthonormal and D diagonal. This is not difficult if V is the identity matrix; in this case, if D is zero everywhere but the first element is λ, say, then a little direct computation shows that

$$\mathrm{eig}_1(D + \epsilon B) - \mathrm{eig}_1 D \approx \frac{\epsilon}{\lambda} Z_1 B,$$

where Z_1 is the operator mapping a matrix to its first column, after setting the first element to zero. The general result now follows after a change of basis or two:

$$\mathrm{eig}_1(A + \epsilon B) - \mathrm{eig}_1 A \approx \frac{\epsilon}{\lambda} V Z_1 V^t B V.$$

Plugging everything in, we get the stated result. □

A.3. $\hat{K}_\phi$

Consistency of $\hat{K}_\phi$. We want to prove that

$$\operatorname*{argmax}_V M_N(V) \to K$$

almost surely. According to arguments like those leading to corollary 3.2.3 of van der Vaart and Wellner (1996), it suffices to prove the following two statements:

(1) $M(V)$ has a well-separated, unique maximum at K;
(2) $\sup_V |M_N(V) - M(V)| \to 0$ almost surely.

When we say that $M(K)$ is a 'well-separated' maximum of $M(V)$, we mean that

$$M(K) > \sup_{V \in O^c} M(V),$$

where O is any open set containing K.

Part (1) is fairly straightforward. Under the conditions of the theorem, the sufficiency part of the data processing inequality ensures that K is a unique maximum. To see that this unique maximum is well-separated we need only note (van der Vaart and Wellner 1996) that $M(V)$ is continuous in V under the conditions of the theorem, with compact domain, and that continuous functions on compact domains attain their suprema; thus, since $M(V)$ attains its maximum on the compact set O^c, $\max_{V \in O^c} M(V)$ must be strictly less than the unique maximum $M(K)$.

Part (2) requires a little more effort. Letting $W_{a_N} * g$ denote the convolution of the kernel W_{a_N} with the function g, define the deterministic sequence of functions

$$M_N^*(V) \equiv \int_{\text{spike},X} \frac{(W_{a_N} * p(\text{spike}, V\vec{x}))^2}{p(\text{spike})(W_{a_N} * p(V\vec{x}))}.$$

Then the proof splits into two parts:

$$\text{(2a)} \ \sup_V |M_N(V) - M_N^*(V)| \to 0 \text{ a.s.};$$

$$\text{(2b)} \ \sup_V |M_N^*(V) - M(V)| \to 0.$$

We handle part (a) with probability inequalities on uniform deviations of sample means from expectations (the standard VC inequalities (van der Vaart and Wellner 1996, Devroye *et al* 1996) are sufficient); since $M_N(V)$ is continuous on compact subsets of $\mathcal{G}_m(X)$ in the topology generated by uniform convergence and p is tight, the almost sure convergence follows.

We prove (b) by noting that $M_N^*(V)$ is uniformly continuous in kernel width and V. Thus it is enough to prove pointwise convergence; this can be done under standard conditions on W (Devroye and Lugosi 2001), either using Fourier transforms or by direct argument. □

Bias of $\hat{K}_\phi$. We need to quantify the rate of decay of $M_N(V) - M(V)$. As indicated in the proof of the consistency theorem, this error has two parts: sample error and approximation error. The sample error, in addition, can be broken up into a bias term and a variance term. The bias term is what will cause us some problems, and it turns out that we can compute it explicitly.

We have

$$\begin{aligned} E(M_N(V) - M(V)) &= E\left(\int_{X,\text{spike}} \frac{\hat{p}(V\vec{x}, \text{spike})^2}{\hat{p}(V\vec{x})\hat{p}(\text{spike})} - \int_{X,\text{spike}} \frac{p(V\vec{x}, \text{spike})^2}{p(V\vec{x})p(\text{spike})}\right) \\ &= \int_X \left(E\left(\sum_{\text{spike}} \left(\frac{\hat{p}(V\vec{x}, \text{spike})^2}{\hat{p}(V\vec{x})\hat{p}(\text{spike})}\right)\right) - \sum_{\text{spike}} \left(\frac{p(V\vec{x}, \text{spike})^2}{p(V\vec{x})p(\text{spike})}\right)\right), \end{aligned}$$

by definition, linearity and Fubini.

Now we write out the expectation inside the integral. To simplify the computations, we assume either that we are dealing with the histogram estimator or that the kernel is a simple boxcar; this makes $\hat{p}$ a constant multiple of a binomial random variable. (The extension to more general kernels is not conceptually difficult (van der Vaart and Wellner 1996) but precludes the direct calculations presented below.) Assume that N is large enough to replace $\hat{p}(\text{spike})$ with $p(\text{spike})$; this can be made rigorous with the usual exponential (Chernoff) inequalities (Devroye *et al* 1996). Let $W_a(x)$ denote the m-dimensional cube of width a centred on x, p^*

the smoothed version of p, as above, s the event (spike $= 1$) and $B(i, N, p)$ the probability mass of a binomial with parameters N and p on count i; then

$$
\begin{aligned}
E\left(\sum_{\text{spike}}\left(\frac{\hat{p}(V\vec{x}, \text{spike})^2}{\hat{p}(V\vec{x})\hat{p}(\text{spike})}\right)\right) &\approx E\left(\sum_{\text{spike}}\left(\frac{\hat{p}(V\vec{x}, \text{spike})^2}{\hat{p}(V\vec{x})p(\text{spike})}\right)\right) \\
&= \sum_{i=0}^{N} B\left(i, N, \int_{W_a(V\vec{x})} p(V\vec{x})\right) \sum_{j=0}^{i} B(j, i, p^*(s|V\vec{x})) \\
&\quad \times \frac{\frac{1}{p(s)}\left(\frac{j}{a^m N}\right)^2 + \frac{1}{1-p(s)}\left(\frac{i-j}{a^m N}\right)^2}{\frac{i}{a^m N}} \\
&= \sum_{i=1}^{N} B\left(i, N, \int_{W_a(V\vec{x})} p(V\vec{x})\right) \frac{1}{i a^m N} \\
&\quad \times \left[\frac{1}{p(s)}((i p^*(s|V\vec{x}))^2 + i p^*(s|V\vec{x})(1 - p^*(s|V\vec{x}))) \right. \\
&\quad \left. + \frac{1}{1-p(s)}((i(1 - p^*(s|V\vec{x})))^2 + i p^*(s|V\vec{x})(1 - p^*(s|V\vec{x})))\right] \\
&= \frac{1}{a^m N}\left[\sum_{i=1}^{N} B\left(i, N, \int_{W_a(V\vec{x})} p(V\vec{x})\right) i\left(\frac{p^*(s|V\vec{x})^2}{p(s)} + \frac{(1 - p^*(s|V\vec{x}))^2}{1 - p(s)}\right)\right. \\
&\quad + p^*(s|V\vec{x})(1 - p^*(s|V\vec{x}))\left(\frac{1}{p(s)} + \frac{1}{1 - p(s)}\right) \\
&\quad \left. \times \left(1 - (1 - \int_{W_a(V\vec{x})} p(V\vec{x}))^N\right)\right] \\
&= \sum_{\text{spike}}\left(\frac{p^*(V\vec{x}, \text{spike})^2}{p^*(V\vec{x})p(\text{spike})}\right) + B_s(V),
\end{aligned}
$$

where we have abbreviated the sample bias in our estimate of $M(V)$ as

$$
B_s(V) \equiv \frac{1}{a^m N}\int_X \left(\frac{p^*(s|V\vec{x})(1 - p^*(s|V\vec{x}))}{p(s)(1 - p(s))}\left(1 - \left(1 - \int_{W_a(V\vec{x})} p(V\vec{x})\right)^N\right)\right).
$$

To get a sense of how this behaves, let p be bounded and continuous, say; then the sample bias is roughly

$$
\frac{1}{a^m N}\int_X \left(\frac{p^*(s|V\vec{x})(1 - p^*(s|V\vec{x}))}{p(s)(1 - p(s))}(1 - \mathrm{e}^{-a^m N p(V\vec{x})})\right).
$$

Now if p decays quickly enough (say it has compact support, to make things obvious), then the final term in the integral above tends to unity and we are left with a bias in our estimate of $M(V)$ of order $(a^m N)^{-1}$. In turn, since the maximum of the above integral with respect to V is clearly not K in general, we are left with a bias of size up to $(a^m N)^{-1/2}$ in our estimate of $\text{argmax}_V M(V)$, as is easy to see after expanding $E(M_{\phi,N})$ about K. We should note that most of the above can be generalized to other choices of ϕ, using a second-order Taylor expansion.

If the sample bias for estimating $M(V)$ is of the order of $(a^m N)^{-1}$, and the approximation bias is of the order of a^r, say, for $r > 0$, then if we equate the two rates to minimize their sum we get that the optimal rate of decay in kernel width is

$$
a \sim N^{-1/(r+m)},
$$

corresponding to an optimal bias rate for $M(V)$ of

$$
\text{bias} \sim N^{-r/(r+m)},
$$

which in turn means that the optimal bias for estimating $\operatorname{argmax}_V M(V)$ can be of the order of $N^{-r/[2(r+m)]}$. □

Bias of the jackknifed kernel estimator. We write out the bias of the jackknifed estimator, using the formula above:

$$\begin{aligned}
ET_{JK} &= NET - \frac{N-1}{N}\sum_{i=1}^{N} ET_{-i} \\
&= N\frac{1}{a^m N}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\left(1-\left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^N\right) \\
&\quad - \frac{N-1}{N}\frac{N}{a^m(N-1)}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))} \\
&\quad \times \left(1-\left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^{N-1}\right) \\
&= \frac{1}{a^m}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\left(\left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^{N-1}\right. \\
&\quad \left. - \left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^N\right) \\
&= \frac{1}{a^m}\int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^{N-1}\int_{W_a(V\vec{x})} p(V\vec{x}) \\
&= \int_X \frac{p^*(s|V\vec{x})(1-p^*(s|V\vec{x}))}{p(s)(1-p(s))}\left(1-\int_{W_a(V\vec{x})} p(V\vec{x})\right)^{N-1}(W_a * p(V\vec{x})).
\end{aligned}$$

Under the conditions stated above, this dies exponentially. □

Convergence rates γ and α for $\hat{K}_\phi$. Representation (5) follows from a fairly classical first-order expansion; see, e.g., Serfling (1980) for background. $p_N m_V$ is the Frechet differential (aka the 'functional derivative' to physicists) of $M_{\phi,N}(V)$ at $M_\phi(V)$ in the direction of $p_N - p$; the representation as a sum of i.i.d. random variables follows from the linearity of the differential. We obtain

$$m_V(\vec{x}, \text{spike}) = 2\frac{p(\text{spike}|V\vec{x})}{p(\text{spike})} - \sum_{\text{spike}} \frac{p(\text{spike}|V\vec{x})^2}{p(\text{spike})} - \int \mathrm{d}p(\vec{x})\left(\frac{p(\text{spike}|V\vec{x})}{p(\text{spike})}\right)^2.$$

The random variable $m_V(\vec{x}, \text{spike})$ is bounded, thus obviously has finite variance. That the remainder term in (5) is $o_p(N^{-1/2})$ follows by computing its variance explicitly, roughly following the computation of the bias terms above. We skip the details. The convergence rate and limit distribution is obtained by applying theorem 3.2.10 of van der Vaart and Wellner (1996) to $p_N m_V$. □

A.4. Lower bounds

Local (Cramer–Rao/Hellinger) lower bounds. The basic idea behind the proof is as follows. For any sufficiently regular finite-dimensional statistical model, the Cramer–Rao bound gives a lower bound on the convergence rate. The models we are dealing with are not finite-dimensional; nevertheless, we can apply the bounds to finite-dimensional submodels within the complete, infinite-dimensional family and then try to make the bound as large as possible

by choosing the most difficult submodel. By 'most difficult' we mean, roughly: as close as possible to the true f, K, p in some probabilistic sense, but as far away as possible in the sense of the error metric (on the manifold $\mathcal{G}_m(X)$). In other words, we want the models to be as wrong as possible but easily confusable with f, K, p.

The most obvious such submodel to try is obtained by keeping f fixed (we assume p is fixed), and simply rotating K around in $\mathcal{G}_1(X)$. More concretely, we define our family to be

$$\mathcal{F}_0 \equiv \{(q, g, V) : q = p, g = f, V \in \mathcal{G}_1(X)\}.$$

To apply Cramer–Rao to this family (under the stated conditions), we need to define an orthonormal basis of the tangent space to $\mathcal{G}_1(X)$ at K, $\{e_i\}_{1 \leqslant i < m}$; this induces a natural coordinate chart of $\mathcal{G}_1(X)$:

$$k_{\epsilon,i} \equiv (1 + \epsilon^2)^{-1/2}(k + \epsilon e_i).$$

The ith component of the score vector when a spike occurs is given by

$$\frac{\partial \log f_i}{\partial \epsilon} = \frac{f'(K\vec{x})\langle e_i, \vec{x}\rangle}{f(K\vec{x})};$$

plugging this into the asymptotic minimax form of the standard Cramer–Rao bound (Gill and Levit 1995), we have

$$\liminf_{N\to\infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(\text{Error}(\hat{K})) \geqslant (\text{trace}\, I_{\mathcal{F}_0}(p, f, K)^{-1})^{1/2},$$

where the Fisher information for $\mathcal{F}_0$ at the true model is given by

$$I_{\mathcal{F}_0}(p, f, K) = E_p\left(\frac{\langle e_i, \vec{x}\rangle\langle e_j, \vec{x}\rangle f'(\langle \vec{k}, \vec{x}\rangle)^2}{f(\langle \vec{k}, \vec{x}\rangle)(1 - f(\langle \vec{k}, \vec{x}\rangle))}\right).$$

This reduces to the quoted result when p is, e.g., standard normal.

A more systematic approach to the search for 'hard' subfamilies requires a more rigorous definition of the notion of 'confusability' between probability measures. While the detailed theory is beyond the scope of this paper, we mention that an appropriate measure of confusability is given by the Hellinger distance between two probability measures; recall that this distance is a kind of L_2 norm between (the square roots of) probability distributions, and can be written in our case as the square root of

$$H_p^2(f, K; h, V) \equiv \tfrac{1}{2}\int_X \left(f(K\vec{x})^{1/2} - h(V\vec{x})^{1/2}\right)^2 \mathrm{d}p(\vec{x}).$$

For our purposes, it suffices to note that, for sufficiently close models (K, f) and (V, h),

$$H_p^2(f, K; h, V) \sim \int_X \left(\frac{(f(K\vec{x}) - h(V\vec{x}))^2}{f(K\vec{x})}\right) \mathrm{d}p(\vec{x}).$$

Simple computations with this asymptotic form of Hellinger distance indicate a stronger subfamily:

$$\mathcal{F}_1 \equiv \{(q, g, V) : q = p, V \in \mathcal{G}_1(X), g(t) = E_{p(\vec{x}|\langle V,\vec{x}\rangle = t)} f(K\vec{x})\}.$$

The final result is

$$\liminf_{N\to\infty} N^{1/2} \inf_{\hat{K}} \sup_{\mathcal{F}} E(\text{Error}(\hat{K})) \geqslant (\text{trace}\, I_{\mathcal{F}_1}(p, f, K)^{-1})^{1/2},$$

with

$$I_{\mathcal{F}_1}(p, f, K) = E_p\left(\frac{\langle e_i, \vec{x} - E_{p(\vec{x}|\langle k,x\rangle)}\vec{x}\rangle\langle e_j, \vec{x} - E_{p(\vec{x}|\langle k,x\rangle)}\vec{x}\rangle f'(\langle \vec{k}, \vec{x}\rangle)^2}{f(\langle \vec{k}, \vec{x}\rangle)(1 - f(\langle \vec{k}, \vec{x}\rangle))}\right).$$

This inequality is in general stronger, but reduces to the first when p is, say, elliptically symmetric. □

Global minimax lower bound. We mimic Ritov and Bickel (1990) (theorem 2). Let K_0 and K_N be separated by a distance a_N. It suffices to put prior distributions π_N on the space of ϵ-tuned LN models supported on these two planes—that is, the conditional distributions given by model (1), with K given by K_0 or K_N, such that

$$D_\phi(p(K\vec{x}, \text{spike});\ p(K\vec{x})p(\text{spike})) > \epsilon$$

—such that the conditional error probability given N data samples of the best hypothesis test between K_0 and K_N converges to $1/2$ as $N \to \infty$. Since the best Bayesian test between two a_N-separated subspaces has error bounded away from zero, we have an order bound on the error of any minimax estimator and the claim is proven. The basic idea behind the construction of the prior is to let 'typical' functions (roughly, any function contained in the support of π_N) vary much more rapidly than the average distance between the projected samples $K\vec{x}_i$; this makes it impossible for any hypothesis test to discern the direction of the underlying conditional probability contour lines which run orthogonal to K. We skip the details, which are easy to verify given (Ritov and Bickel 1990). □

References

Aguera y Arcas B, Fairhall A and Bialek W 2001 What can a single neuron compute? *NIPS* **13** 75–81

Berry M and Meister M 1998 Refractoriness and neural precision *J. Neurosci.* **18** 2200–11

Brenner N, Bialek W and de Ruyter van Steveninck R 2001 Adaptive rescaling optimizes information transmission *Neuron* **26** 695–702

Brown E, Barbieri R, Ventura V, Kass R and Frank L 2002 The time-rescaling theorem and its application to neural spike train data analysis *Neural Comput.* **14** 325–46

Bussgang J 1952 Crosscorrelation functions of amplitude-distorted Gaussian signals *RLE Technical Reports* 216

Chander D and Chichilnisky E 2001 Adaptation to temporal contrast in primate and salamander retina *J. Neurosci.* **21** 9904–16

Chichilnisky E 2001 A simple white noise analysis of neuronal light responses *Network: Comput. Neural Syst.* **12** 199–213

Cover T and Thomas J 1991 *Elements of Information Theory* (New York: Wiley)

Csiszar I 1967 Information-type measures of difference of probability distributions and indirect observations *Stud. Sci. Math. Hungar.* **2** 299–317

de Ruyter R and Bialek W 1988 Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transmission in short spike sequences *Proc. R. Soc.* B **234** 379–414

Devroye L, Gyorfi L and Lugosi G 1996 *A Probabilistic Theory of Pattern Recognition* (New York: Springer)

Devroye L and Lugosi G 2001 *Combinatorial Methods in Density Estimation* (New York: Springer)

Everson R and Roberts S 1999 Inferring the eigenvalues of covariance matrices from limited, noisy data *IEEE Trans. Signal Process.* **48** 2083–91

Fellows M, Paninski L, Hatsopoulos N and Donoghue J 2001 Diverse spatial and temporal features of velocity and position tuning in mi neurons during continuous tracking *SFN Abstracts* p 940.1

Georgopoulos A, Caminiti R and Kalaska J 1986 Neuronal population coding of movement direction *Science* **233** 1416–19

Gill R and Levit B 1995 Applications of the van trees inequality: a Bayesian Cramer–Rao bound *Bernoulli* **1/2** 59–79

Hunter I and Korenberg M 1986 The identification of nonlinear biological systems: Wiener and hammerstein cascade models *Biol. Cybern.* **55** 135–44

Johnstone I 2000 On the distribution of the largest principal component *Technical Report* 2000-27 (Stanford)

Jongbloed G 2000 Minimax lower bounds and moduli of continuity *Stat. Probab. Lett.* **50** 279–84

Marmarelis P and Marmarelis V 1978 *Analysis of Physiological Systems: The White-noise Approach* (New York: Plenum)

Moran D and Schwartz A 1999 Motor cortical representation of speed and direction during reaching *J. Neurophysiol.* **82** 2676–92

Paninski L 2002 Convergence properties of some spike-triggered analysis techniques *NIPS* at press

Paninski L 2003a Estimation of entropy and mutual information *Neural Comput.* **15** 1191–254

Paninski L 2003b Some rigorous results on the neural coding problem *PhD Thesis* New York University

Paninski L, Fellows M, Hatsopoulos N and Donoghue J 1999 Coding dynamic variables in populations of motor cortex neurons *Soc. Neurosci. Abstr.* **25** 665.9

Paninski L, Fellows M, Hatsopoulos N and Donoghue J 2003a Spatiotemporal tuning properties for hand position and velocity in motor cortical neurons *J. Neurophysiol.* submitted

Paninski L, Fellows M, Shoham S and Donoghue J 2002 Nonlinear encoding and decoding in primary motor cortex (MI) *Soc. Neurosci. Abstr.* **28**

Paninski L, Fellows M, Shoham S, Hatsopoulos N and Donoghue J 2003b Nonlinear population models for the encoding of dynamic hand position signals in MI *Computation Neuroscience Meeting,* at press

Pillow J and Simoncelli E 2003 Biases in white noise analysis due to non-poisson spike generation *Neurocomputing* at press

Press W, Teukolsky S, Vetterling W and Flannery B 1992 *Numerical Recipes in C* (Cambridge: Cambridge University Press)

Reid R and Shapley R 2002 Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus *J. Neurosci.* **22** 6158–75

Ringach D, Hawken M and Shapley R 2002 Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences *J. Vis.* **2** 12–24

Ringach D, Sapiro G and Shapley R 1997 A subspace reverse correlation technique for the study of visual neurons *Vis. Res.* **37** 2455–64

Ritov Y and Bickel P 1990 Achieving information bounds in non- and semi-parametric models *Ann. Stat.* **18** 925–38

Ruderman D and Bialek W 1994 Statistics of natural images: scaling in the woods *Phys. Rev. Lett.* **73** 814–17

Rust N, Schwartz O, Movshon A and Simoncelli E 2003 Spike-triggered characterization of excitatory and suppressive stimulus dimensions in monkey v1 directionally selective neurons *CNS03 Meeting* at press

Schervish M 1995 *Theory of Statistics* (New York: Springer)

Schwartz O, Chichilnisky E and Simoncelli E 2002 Characterizing neural gain control using spike-triggered covariance *NIPS* **14**

Schwarz G 1978 Estimating the dimension of a model *Ann. Stat.* **7** 461–4

Serfling R 1980 *Approximation Theorems of Mathematical Statistics* (New York: Wiley)

Serruya M, Hatsopoulos N, Paninski L, Fellows M and Donoghue J 2002 Instant neural control of a movement signal *Nature* **416** 141–2

Sharpee T, Bialek W and Rust N 2003 Maximally informative dimensions: analyzing neural responses to natural signals *Neural Comput.* submitted

Shoham S, Fellows M, Hatsopoulos N, Paninski L, Donoghue J and Normann R 2003 Optimal decoding for a primary motor cortical brain–computer interface *IEEE Trans. Biomed. Eng.* submitted

Simoncelli E 1999 Modeling the joint statistics of images in the wavelet domain *Proc. SPIE, 44th Annual Mtg* pp 188–95

Simoncelli E P and Heeger D J 1998 A model of neuronal responses in visual area MT *Vis. Res.* **38** 743–61

Theunissen F, David S, Singh N, Hsu A, Vinje W and Gallant J 2001 Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli *Network: Comput. Neural Syst.* **12** 289–316

Touryan J, Lau B and Dan Y 2002 Isolation of relevant visual features from random stimuli for cortical complex cells *J. Neurosci.* **22** 10811–18

van der Vaart A and Wellner J 1996 *Weak Convergence and Empirical Processes* (New York: Springer)