Convergence properties of three spike-triggered analysis techniques

Liam Paninski [a,1]

[a] *Center for Neural Science, New York University*

Abstract

We analyze the convergence properties of three spike-triggered data analysis techniques. All of our results are obtained in the setting of a (possibly multidimensional) linear-nonlinear (LN) model for stimulus-driven neural activity. We start by giving exact rate of convergence results for the common spike-triggered average (STA) technique. Next, we analyze a spike-triggered covariance method, variants of which have been recently exploited successfully by Bialek, Simoncelli, and colleagues. These first two methods suffer from extraneous conditions on their convergence; therefore, we introduce an estimator for the LN model parameters which is designed to be consistent under completely general conditions. We provide an algorithm for the computation of this estimator and derive its rate of convergence. We close with a brief analysis of the effects of refractoriness on the results presented here.

Key words: Spike-triggered average, cascade model, empirical process

1 Introduction

Systems-level neuroscientists have a few favorite 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 $P(spike|X = x)$, the probability that our cell emits a spike, given that some observable signal X in the world takes value x. Because data is expensive, neuroscientists 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 $P(spike|X = x)$

[1] We thank E. Simoncelli for many interesting discussions. LP is supported by a predoctoral fellowship from the Howard Hughes Medical Institute. Contact: liam@cns.nyu.edu; http://www.cns.nyu.edu/∼liam

for each possible x. In this paper, we analyze one such phenomenological model whose popularity seems to be on the rise:

$$p(spike|\vec{x}) = f(<\vec{k}_1, \vec{x}>, <\vec{k}_2, \vec{x}>, \ldots, <\vec{k}_m, \vec{x}>). \quad (1)$$

Here f is some arbitrary nonconstant, $\Re^m$-measurable, $[0,1]$-valued function, and $\{k_i\}$ are some linearly independent elements of the dual space, X', of some topological vector space, X — the space of possible "input signals." Interpret f as a regular conditional distribution. Roughly, then, the neuron projects the signal $\vec{x}$ onto some m-dimensional subspace spanned by $\{\vec{k}_i\}_{1 \leq i \leq m}$ (call this subspace K), then looks up its probability of firing based only on this projection.

This model has two important features. 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. Second, equation (1) implies:

$$p(spike|\vec{x}) = p(spike|\vec{x} + \vec{y}) \; \forall \; y \perp K. \quad (2)$$

In other words, the conditional probability of firing is constant along (hyper)planes in the input space. The natural generalization of this is a model for which these surfaces of constant firing probability are manifolds of low codimension. However, we will stick to the linear case here.

This model thus separates the (quite difficult) problem of learning $p(spike|\vec{x})$ into two pieces: learning K and learning f. If K is known, the problem of learning f reduces to a density estimation problem, about which much is known (see e.g. [4]). The problem of estimating K seems to be less well-understood; we focus primarily on this problem here. Our main goal will be to describe the convergence properties of K-estimators; as usual, rate-of-convergence results clarify exactly why a given estimator works well (in the sense that a only a small number of samples is needed for reliable estimates) in certain cases and poorly (sometimes not at all) in others.

2 Convergence rates

It turns out to be very easy to derive rates of convergence for both 1) the classical spike-triggered average estimate for K, and 2) the spike-triggered covariance techniques recently employed by Bialek, Simoncelli, et al. The proofs are fairly similar, involving only a change of basis and an application of the CLT and the delta method. For clarity, we assume K is one-dimensional for the next two paragraphs; the extension to the multidimensional case is straightforward and will be described elsewhere.

1) As pointed out in [3], radial symmetry of $p(\vec{x})$ is sufficient to make the STA an unbiased estimator of K; this radial symmetry condition turns out to be necessary for the general convergence of the STA, and we therefore assume for the moment that $p(\vec{x})$ is radially symmetric. We find that the spike-triggered average has variability

around K of asymptotic scale

$$\frac{1}{\sqrt{N}}\frac{\sigma(p)}{E(q)}\sqrt{\dim X - 1};$$

here $\sigma(p)$ is the square root of the second moment of $p(\vec{x})$, and q is a random variable with density given by

$$P(q) \equiv p(< \vec{x}, \vec{k}_1 > | spike) = \frac{f(< \vec{x}, \vec{k}_1 >)p(< \vec{x}, \vec{k}_1 >)}{\int_{\Re} f(< \vec{x}, \vec{k}_1 >)p(< \vec{x}, \vec{k}_1 >)}, \tag{3}$$

with f as defined in (1) and $p(< \vec{x}, \vec{k}_1 >)$ denoting the one-dimensional projection of $p(\vec{x})$.

2) Covariance-based methods [5,6] have appeared specifically to circumvent the problems faced by the STA, namely, the strong symmetry condition on $p(\vec{x})$ and the positivity condition on $|E(q)|$. The basic idea is to find the eigenvectors of some estimate of the "difference-covariance" matrix $\Delta\sigma^2$, defined by

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

a little linear algebra shows that the nonzero eigenspace of $\Delta\sigma^2$ is contained in K. We can derive a similar rate of convergence for these methods:

$$\frac{1}{\sqrt{N}}\frac{\kappa(\Delta\sigma^2)}{\lambda_{\Delta\sigma^2}}\sqrt{\dim X - 1};$$

here, $\kappa(\Delta\sigma^2)$ is a term that measures the variability of the estimate of $\Delta\sigma^2$ ($\kappa(\Delta\sigma^2)$ contains kurtotic terms in general but can be thought of as scaling, roughly, with $\sigma^2(p(\vec{x}))$), and $\lambda_{\Delta\sigma^2}$ is the (unique by assumption) nonzero eigenvalue of $\Delta\sigma^2$.

In general, then, the two most commonly used techniques for estimating K extract only a subspace of K; while covariance-based approaches suffer from fewer symmetry assumptions than do STA methods, both techniques require the positivity (preferably the large positivity) of a first or second moment of $p(\vec{x})$ and $p(\vec{x}|spike)$. It is easy to think of non-pathological examples for which neither of these moment positivity conditions is satisfied (just as it is easy to think of probability distributions which are not completely determined by their means and variances). Therefore, we will develop an estimator for K which is universally consistent — i.e., convergence of the estimator does not require conditions either on f (in equation 1) or on $p(\vec{x})$.

The basic idea is that $K\vec{x}$ is in a sense a sufficient statistic for $\vec{x}$ (that is, $\vec{x}$ — $K\vec{x}$ — $spike$ forms a Markov chain). The data processing inequality states that

$$M(V) \equiv D_{KL}(p(< V, \vec{x} >, spike); p(< V, \vec{x} >)p(spike))$$

(where D_{KL} denotes Kullback-Leibler divergence), considered as a function of vector spaces V of dimension $\dim K$, reaches a unique maximum on K, and is 0 precisely on all "irrelevant" subspaces — that is, those subspaces completely orthogonal to K. (When $\dim V \neq \dim K$, the maximum is no longer unique, but it is easy to show that

maximizers lie in K or contain K, according as $\dim V < \dim K$ or $\dim V > \dim K$, respectively.) This suggests that we could estimate K by maximizing

$$M_N(V) \equiv D_f(q_N(< V, \vec{x} >, spike); q_N(< V, \vec{x} >)q_N(spike)),$$

where D_f is some functional for which something like the data processing inequality is true, and q_N is some estimate of p. For example, we could let q_N be some kernel estimate, that is, a filtered version of the empirical measure

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

This doesn't quite work, however, because the kernel induces an arbitrary scale; if this scale is larger than the natural scale of $p(< V, \vec{x} >)$ for some V but not others, our estimate will be biased away from K. Therefore, D_f and p_N have to be scale-free in some sense.

One computationally efficient solution to our problem is as follows. Set

$$M(V) \equiv \sup_{g,h \in \mathcal{G}\times\mathcal{H}} |p(g(< V, \vec{x} >), h(spike)) - p(g(< V, \vec{x} >))p(h(spike))|$$

and

$$M_N(V) \equiv \sup_{g,h \in \mathcal{G}\times\mathcal{H}} |p_N(g(< V, \vec{x} >), h(spike)) - p_N(g(< V, \vec{x} >))p_N(h(spike))|,$$

where $\mathcal{G}$ and $\mathcal{H}$ are totally bounded function classes and Pg denotes the expectation of the function g with respect to the probability measure P. When $\mathcal{G}$ and $\mathcal{H}$ are classes of set indicator functions, M can be written in the form of one of Csiszar's f-divergences, a class of distance-like functionals for which the data processing inequality is true. The theory of empirical processes [7] shows that, when $\mathcal{G}$ and $\mathcal{H}$ are not too large, a functional (infinite-dimensional) CLT holds, in that $N^{1/2}(M_N - M)$ converges uniformly to a centered Gaussian process whose covariance function can be computed easily. We choose $\mathcal{G}$ to be the quantile indicators; this makes M and M_N scale-free, as desired ($\mathcal{H}$ is discrete, hence scale-free automatically). All this means that deriving consistency and convergence rates for $\hat{K} \equiv \text{argmax}_V M_N(V)$ proceeds as in the corresponding classical proofs for maximum likelihood estimators (see [7] for details). When $M(V)$ is smooth to second order at K, for example, we have that $\hat{K}$ converges to K at a $N^{-1/3}$ rate; here, the critical prefactor is inversely proportional to $|M''(K)|$, and when $M(V)$ has a nondifferentiable peak at K a $N^{-1/2}$ rate can be recovered. We have developed an algorithm for the computation of $\text{argmax}_V M_N(V)$, and numerical results show that $\hat{K}$ can be competitive with spike-triggered average or covariance techniques even in cases where $E(q)$ or λ are positive; that is, for reasonable neural models, $M''(K)$ is often sufficiently large compared to $E(q)$ or λ that the difference between $N^{-1/3}$ and $N^{-1/2}$ is irrelevant. It is unclear at present whether a universally $N^{-1/2}$-consistent estimator for K exists.

3 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. What happens to the above results when spikes which are close to each other in time are dependent, as is (of course) the case in any reasonable neural model? Does the above analysis fail in this case? If so, why? Can we correct for these non-Poisson effects? We can give fairly complete answers to all of these questions, at least in the following special case:

$$p(spike|\vec{x}, s_-) = f(<\vec{k}_1, \vec{x}>, <\vec{k}_2, \vec{x}>, \ldots, <\vec{k}_m, \vec{x}>)g(T(s_-)). \tag{4}$$

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 $p(spike|\vec{x}, s_-)$ a regular conditional distribution (aside from measurability issues, it is sufficient that $f, g \geq 0$, $fg \leq 1 \; \forall <\vec{k}, \vec{x}>$).

To see why the memory effects displayed by (4) complicate the analysis presented in the previous sections, recall the basic idea behind the proof of the fact that whenever $p(\vec{x})$ is radially symmetric and $E(\vec{x})$ exists, $E(STA)$ lies in K. We will write $E(STA)$ out and show the essential point of the proof; then we will show why the memory effects seen in (4) cause problems, and how these problems can be "fixed," in some suitable sense.

$$\begin{aligned} E(STA) &= \int p(\vec{x}|spike)\vec{x}d\vec{x} \\ &= \int p(spike|\vec{x})\frac{p(\vec{x})}{p(spike)}\vec{x}d\vec{x} \\ &= \int f(<\vec{K}, \vec{x}>)\frac{p(\vec{x})}{p(spike)}\vec{x}d\vec{x}. \end{aligned} \tag{5}$$

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 $<\vec{K}, \vec{x}>$ - the proof that $E(STA) \in K$ follows immediately (after a suitable change of basis). This key equality does not hold in general for (4):

$$\begin{aligned} p(spike|\vec{x}) &= \int p(spike|\vec{x}, s_-)p(s_-|\vec{x})ds_- \\ &= \int f(<\vec{K}, \vec{x}>)g(T(s_-))p(s_-|\vec{x})ds_- \\ &= f(<\vec{K}, \vec{x}>)\int g(T(s_-))p(s_-|\vec{x})ds_- \\ &= f(<\vec{K}, \vec{x}>)h(\vec{x}). \end{aligned} \tag{6}$$

The first equality is (4); 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 $E(STA) \in K$. 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 $E(STA) \notin K$.

So we need to modify 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., [2]). [1] suggest ignoring all spikes for which $g(T(s_-)) \neq 1$, i.e., form

$$S\vec{T}A^* \equiv \frac{1}{N} \sum_{i \in S} \delta(g(T(s_{i-})) - 1)\vec{x}_i,$$

where S indicates the set of stimuli corresponding to spikes, and δ 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.

We suggest the following: form

$$S\vec{T}A_* \equiv \frac{1}{N} \sum_{i \in S} g(T(s_{i-}))^{-1}\vec{x}_i.$$

It is easy to see, from the above discussion, that $E(S\vec{T}A_*) \in K$.

References

[1] Aguera y Arcas, B., Fairhall, A. & Bialek, W. NIPS 13: 75-81 (2001).

[2] Berry, M. & Meister, M. J. Neurosci. 18: 2200-2211 (1998).

[3] Chichilnisky, E. Network 12: 199-213 (2001).

[4] Devroye, L. & Lugosi, G. Combinatorial Methods in Density Estimation. Springer Verlag, New York (2001).

[5] de Ruyter van Steveninck, R. & Bialek, W. Proc. R. Soc. Lond. B, 234: 379-414 (1988).

[6] Schwartz, O., Chichilnisky, E. & Simoncelli, E. NIPS 14 (2002).

[7] van der Vaart, A. & Wellner, J. Weak convergence and empirical processes. Springer-Verlag, New York (1996).