1-1-2008

Valuations for Spike Train Prediction

Vladimir Itskov
Columbia University,, vladimir@neurotheory.columbia.edu

Carina Curto
University of Nebraska - Lincoln, ccurto2@math.unl.edu

Kenneth D. Harris
Rutgers, The State University of New Jersey, kdharris@rutgers.edu

LETTER Communicated by Jonathan Victor

Valuations for Spike Train Prediction

Vladimir Itskov
vladimir@neurotheory.columbia.edu
Carina Curto
ccurto@rutgers.edu
Kenneth D. Harris
kdharris@andromeda.rutgers.edu
Center for Molecular and Behavioral Neuroscience, Rutgers, The State University of New Jersey, Newark, NJ 07102, U.S.A.

The ultimate product of an electrophysiology experiment is often a decision on which biological hypothesis or model best explains the observed data. We outline a paradigm designed for comparison of different models, which we refer to as *spike train prediction*. A key ingredient of this paradigm is a prediction quality valuation that estimates how close a predicted conditional intensity function is to an actual observed spike train. Although a valuation based on log likelihood (L) is most natural, it has various complications in this context. We propose that a quadratic valuation (Q) can be used as an alternative to L. Q shares some important theoretical properties with L, including consistency, and the two valuations perform similarly on simulated and experimental data. Moreover, Q is more robust than L, and optimization with Q can dramatically improve computational efficiency. We illustrate the utility of Q for comparing models of peer prediction, where it can be computed directly from cross-correlograms. Although Q does not have a straightforward probabilistic interpretation, Q is essentially given by Euclidean distance.

1 Introduction

We consider a paradigm for analysis of neural spiking data that we refer to as *spike train prediction*. In this framework, a biological hypothesis is translated into a prescription for predicting a given cell's conditional intensity (i.e., firing rate) at each moment in time from a set of predictor variables such as sensory input, animal behavior, or spiking history. Different hypotheses are compared for suitability by determining which one better predicts the actual observed spike trains. This approach was previously used to study the organization of hippocampal cell assemblies during spatial behavior (Harris, Csicsvari, Hirase, Dragoi, & Buzsaki, 2003). Various related approaches have been used to study and compare statistical models of spiking neurons (Barbieri, Quirk, Frank, Wilson, & Brown, 2001; Brown,

Barbieri, Ventura, Kass, & Frank, 2002; Paninski, Pillow, & Simoncelli, 2004; Truccolo, Eden, Fellows, Donoghue, & Brown, 2005).

A traditional statistical analysis typically consists of model specification, parameter estimation, and assessing model goodness of fit. Our approach is different and is closer to the framework of model selection (Linhart & Zucchini, 1986). We start from the premise that for large and complex data sets, no model we choose will be exactly correct; our goal is hence to select the most appropriate from one or more approximating families of models. This requires constructing a prediction quality valuation, which assigns a real number to each model. Perhaps the most obvious choice of valuation is one based on log likelihood (L). However, while log likelihood has well-known optimality properties, it poses complications for spike train prediction. L may be highly sensitive to small changes in spike trains, which often occur as the result of spike sorting errors. In practice, this means that L loses resolution in its ability to pick out optimal parameters. Moreover, optimization with L requires iterative procedures that can make it unusable for large experimental data sets. These considerations lead us to propose a quadratic valuation (Q) as a viable alternative to L.

In the same way that maximum likelihood and least-squares analyses often yield similar results for statistical models of finite-dimensional random variables, we find that L and Q are similar valuations for models of point processes. Moreover, Q has many advantages, including robustness and computational efficiency. In particular, for a linear peer prediction model (defined in section 4), Q can be computed directly from spike train cross-correlograms, reducing optimization with Q to a linear problem and thus allowing its use in large-scale recordings. Q also shares important theoretical properties with L: both valuations are consistent (i.e., parameters estimated using L and Q converge to the true values in the limit of large data) and have the same global maxima. Q and L also perform similarly on simulated and experimental data. In contrast, a third valuation we consider, based on the Kolmogorov-Smirnov (KS) statistic, is quite different. Despite its many similarities with L, we show that Q is not a transformed version of L; rather, Q can be derived as a regularized Euclidean distance.

2 Spike Train Prediction

Of course, it is impossible to predict the timing of spikes exactly. Instead, our prediction of a spike train takes the form of a conditional intensity function:

$$\lambda(t; H_t) = \lim_{\Delta t \to 0} \frac{\text{Prob}\big(\text{One spike occurs in time interval } [t, t + \Delta t] \mid H_t\big)}{\Delta t},$$

where H_t denotes the spiking history preceding time t. A hypothesis can be formulated as a parameterized family of prescriptions for the conditional

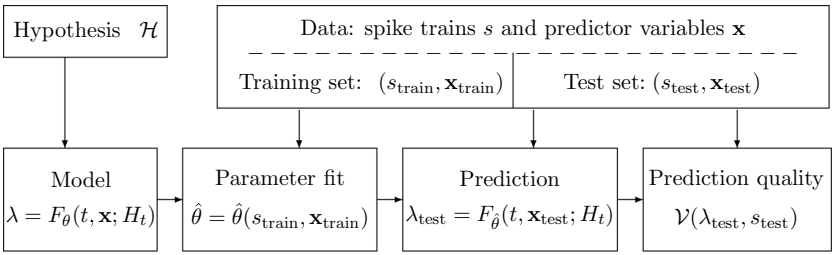

Figure 1: Spike train prediction paradigm. A hypothesis $\mathcal{H}$ is formalized as a family of prescriptions $F_\theta(t, \mathbf{x}; H_t)$ for predicting the intensity function λ from the predictor variables $\mathbf{x}(t)$ and spiking history H_t. The data are divided into training and test sets (for cross-validation). The parameters θ in each model are determined using a fitting procedure $\hat{\theta} = \hat{\theta}(s, \mathbf{x})$ on the training set. The predicted intensity λ_{test} is then computed from predictor variables and spiking history on the test set. A pair of hypotheses is compared using a prediction quality valuation $\mathcal{V}$, which compares the predicted intensity to the actual observed spike train s_{test}.

intensity function,

$$\lambda(t; H_t) = F_\theta(t, \mathbf{x}; H_t), \tag{2.1}$$

where $\mathbf{x}(t)$ is a set of predictor variables such as sensory inputs, animal behavior, or the activity of other simultaneously recorded neurons, and θ is a vector of model parameters. For a given spike train s, $H_t(s)$ denotes the history of that spike train up to time t.

One hypothesis may be compared to another based on how well the intensity function λ describes the actual spike train s for the optimal value of θ. This requires a method for fitting the parameters θ to a given data set and a prediction quality valuation $\mathcal{V}(\lambda, s)$ whose arguments are an intensity function λ and a spike train s.[1] The greater the value of $\mathcal{V}(\lambda, s)$, the greater the quality of the prediction λ, as compared to the actual observed spike train s.

As in any model selection procedure, we must ensure that the results are not biased toward more complex models merely because of their tendency to overfit statistical fluctuations as well as real structure. In a cross-validation paradigm (see Figure 1), the parameters for each model family are fit on the training set,[2] while model comparison is performed on the test

[1] A related but distinct concept is spike train metrics (Victor & Purpura, 1996; Aronov & Victor, 2004; van Rossum, 2001), which are used to compare one spike train to another.

[2] In simple cases, the fitting is done by maximizing the valuation on the training set: $\hat{\theta} = \arg\max_\theta \mathcal{V}\left(\lambda_\theta\left(t, \mathbf{x}_{\text{train}}; H_t\right), s_{\text{train}}\right)$. Frequently, regularization is also required (see note 5).

set. This ensures that overfitting will result in a worse prediction quality. In contrast, in penalty-based model comparison methods such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), the L valuation acquires a correction term that penalizes models with large numbers of parameters. However, these methods are guaranteed accurate only for correctly specified probabilistic models, a condition often not achieved in neurophysiology.

The procedure outlined in Figure 1 is used to compare models, not to evaluate a given model's goodness of fit.[3] When one model's parameters are a subset of another's, the more complex model can perform better under cross-validation only if at least one of the additional parameters is meaningful. The paradigm therefore provides a way of testing whether a larger model family will yield a better approximation to the data than a given simpler one.

3 Prediction Quality Valuations

3.1 Definitions of the Valuations. Here we define the three valuations we consider in this letter. Our main focus is on introducing the quadratic valuation Q as a reasonable alternative to the log likelihood valuation L. For purposes of comparison, we consider a third valuation based on the KS statistic; subsequent sections will show that KS behaves differently from L and Q.

Note that in comparing models with predictions λ_1 and λ_2, only the relative difference $\mathcal{V}(\lambda_1, s) - \mathcal{V}(\lambda_2, s)$ is meaningful. The actual value of $\mathcal{V}(\lambda, s)$ should not be interpreted as an absolute goodness-of-fit assessment. We also require that valuations be well defined in the sense that $\mathcal{V}(\lambda, s)$ should not depend on arbitrary parameters such as sampling rate or bin size.

The following definitions are all intended for general conditional intensity functions $\lambda(t; H_t)$. For history-dependent intensity functions, the value $\mathcal{V}(\lambda, s)$ is computed using the spiking history $H_t(s)$. For a fixed spike train s, $\lambda(t; H_t(s))$ is a function of t only; to avoid cumbersome notation, we write $\lambda(t)$ instead of $\lambda(t; H_t(s))$ throughout, since s is fixed in the valuation formulas.

3.1.1 The Log Likelihood Valuation L. We define the log likelihood valuation L as

$$\mathrm{L}(\lambda, s) \stackrel{\text{def}}{=} \frac{1}{T}\left(-\int_0^T \lambda(t)\,\mathrm{d}t + \sum_{k=1}^{N_s} \log \lambda(s^k)\right), \tag{3.1}$$

[3]Once a model is selected, a goodness of fit analysis using standard methods such as KS plots can be performed.

where s^k is the time of the kth spike and N_s is the total number of spikes in the spike train. This is the log likelihood per unit time of observing the spike train s given the conditional intensity function λ (Daley & Vere-Jones, 2003). A priori, L is the most natural valuation to use due to its simple probabilistic interpretation and optimality properties. In the limit of large data, log likelihood is consistent and saturates the Cramér-Rao lower bound. L can also be interpreted in units of bits; previously log likelihood has been used together with cross-validation in order to estimate the information carried by a cell about a stimulus (Kjaer, Hertz, & Richmond, 1994; Harris et al., 2003).

3.1.2 The Quadratic Valuation Q. We define the quadratic valuation Q as

$$\mathrm{Q}(\lambda, s) \stackrel{\text{def}}{=} \frac{1}{T}\left(-\int_0^T \lambda^2(t)\,\mathrm{d}t + 2\sum_{k=1}^{N_s} \lambda(s^k)\right). \tag{3.2}$$

To gain insight into the meaning of Q, we present here a derivation from Euclidean distances between discretized rate functions and spike trains. For a given spike train s, denote by n_i the number of spikes in the ith bin (here we split the time interval $[0, T]$ into N_{bins} bins of length Δt). The mean squared distance between the prediction $\lambda(t)$ and the binned spike train is

$$q_{\Delta t}(\lambda, s) \stackrel{\text{def}}{=} \frac{1}{N_{\text{bins}}} \sum_i \left(\lambda(t_i) - \frac{n_i}{\Delta t}\right)^2 = \frac{\Delta t}{T} \sum_i \left(\lambda^2(t_i) - 2\lambda(t_i)\frac{n_i}{\Delta t} + \frac{n_i^2}{\Delta t^2}\right),$$

where t_i denotes the time of the ith bin. For small enough bin size, we can approximate this by $q_{\Delta t}(\lambda, s) \approx \frac{1}{T}\int_0^T \lambda^2(t)\,\mathrm{d}t - \frac{2}{T}\sum_{k=1}^{N_s} \lambda(s^k) + \frac{N_s}{T\Delta t}$, where N_s is the total number of spikes in the spike train s. We can thus define Q as

$$\mathrm{Q}(\lambda, s) \stackrel{\text{def}}{=} \lim_{\Delta t \to 0} (q_{\Delta t}(0, s) - q_{\Delta t}(\lambda, s)).$$

Note that the sign of Q has been chosen such that smaller distances correspond to higher-quality predictions. The $q_{\Delta t}(0, s)$ term prevents the limit from diverging but does not depend on λ and hence will not affect comparisons between different λ's. Using Q is therefore equivalent to using the Euclidean distance $q_{\Delta t}$, for small enough Δt.

3.1.3 The KS Valuation. We use the time-rescaling theorem and other ideas from Barbieri et al. (2001) and Brown et al. (2002) to define the KS valuation. The time-rescaling theorem states that if $\lambda(t) > 0$ is a conditional intensity function and the spike train s is a realization of the associated point process, then the time-rescaled spike times $\Lambda(s^k) = \int_0^{s^k} \lambda(t)\,\mathrm{d}t$ are a Poisson process with unit rate. This suggests another way to evaluate the quality of

prediction λ: given the actual observed spike train s, use λ to compute the time-rescaled interspike intervals (ISIs) $\tau_k = \Lambda(s^k) - \Lambda(s^{k-1})$, and then see how well their distribution approximates the expected distribution $F(x) = 1 - e^{-x}$. From the rescaled ISIs $\{\tau_k\}$, we construct an empirical cumulative density function (cdf)

$$G(x) = \frac{1}{N_s - 1} \int_0^x \sum_k \delta(t - \tau_k)\, \mathrm{d}t.$$

We define the KS valuation as

$$\mathrm{KS}(\lambda, s) \stackrel{\text{def}}{=} 1 - \sup_x \{|G(x) - F(x)|\} = 1 - \sup_x \{|G(x) + e^{-x} - 1|\}.$$

The KS valuation provides a measure of the discrepancy between the distribution of ISIs observed in the data and the distribution predicted by the model. Of course, the KS statistic is not generally used in this way; rather, it is usually used as a test for goodness of fit. If a model does not pass the KS test, then it is presumed to be incorrect (although the converse is not true). Nevertheless, we find that the KS valuation yields a reasonable candidate for an objective function to be optimized (see section 3.2).

3.2 Simulated and Experimental Data. In this section we compare the performance of the three valuations on simulated and experimental data. In an example where the true underlying model is known, we find that L, Q, and KS all select the correct model from a one-parameter family. However, when the simulated spike train data are "contaminated" by random bursts,[4] L's performance falters while Q and KS continue to be optimized by the best model in the family. On experimental place field data, where the true underlying family of models is unknown, we find that L and Q perform very similarly but show little correlation with KS.

3.2.1 Simulation. To compare the performance of the three valuations in a case where the true underlying model is known, we simulated an inhomogeneous Poisson spike train using a place field model with time offset. In this model, $\lambda(t)$ is obtained from the place field $F(\mathbf{x})$ and evaluated on the time-offset trajectory $\mathbf{x}(t + \tau_0)$, where τ_0 is constant and $\mathbf{x}(t)$ is generated by Brownian motion in two dimensions. The conditional intensity function is thus $\lambda(t) = F(\mathbf{x}(t + \tau_0))$. We wanted to see how well the valuations work by using them to recover the correct value of τ_0.

[4]In practice, contamination of spike train data is common due to the difficulties of spike sorting.

We used the valuations L, Q, and KS to compare time-offset place field models for different values of τ using threefold cross-validation. For each value of τ, we computed the place field[5] on the training set and evaluated the prediction quality on the test set. The results were then averaged across the three training-test set pairs for each valuation. This procedure was repeated for a range of τ values that included the "true" τ_0. In this noise-free situation, all three valuations performed well (see Figure 2, pure traces), as they were all optimized by the correct value $\tau = \tau_0$.

3.2.2 Robustness. A potential difficulty with L is that it is not robust. If just one spike in the spike train s occurs at a time when the predicted intensity function λ is very small (in practice, this may happen due to spike sorting errors), the value of $\mathrm{L}(\lambda, s)$ will jump to a very large negative number due to the $\log \lambda(s^k)$ terms, which may render L useless for model comparison. In contrast, it is clear from the definitions of Q and KS that addition or deletion of a single spike will perturb these valuations only slightly.

To test the robustness of L, Q, and KS, we repeated the procedure in section 3.2.1 for "contaminated" spike train data. We simulated a Poisson train of bursts (10 spikes per burst with 2 ms spacing) and chose a number of bursts at random to provide different levels of contamination (2% and 5%). These bursts were then inserted into the original spike train, and equal numbers of spikes were randomly deleted from the contaminated spike train in order to preserve the average firing rate of 2 Hz. While the Q and KS valuations were essentially unaffected by this contamination procedure (both continued to select the correct optimal value for τ), the performance of L faltered (see Figure 2, contaminated traces).

Of course, there are many regularization schemes by which log likelihood can be made more robust; however, they always involve choosing certain smoothing or cutoff parameters. Without a deep understanding of both the model and the optimization problem, these choices can be arbitrary and the results misleading. Q and KS are naturally robust; they do not need to be regularized. It is interesting to note that in our simulated example, we did regularize, as the computation of place fields involved smoothing that prevented the predicted λ from being very close to zero. Nevertheless,

[5] For fixed τ, the τ-offset place field is computed as

$$F_\tau(\mathbf{y}) = \frac{\sum_{\text{spikes } s^k} \varphi(\mathbf{x}(s^k + \tau) - \mathbf{y})}{\frac{1}{T}\int_0^T \varphi(\mathbf{x}(t + \tau) - \mathbf{y})dt},$$

where s^k is the time of the kth spike, and $\varphi(\mathbf{z}) = \exp(-|\mathbf{z}|^2/2\sigma^2)$, with $\mathbf{z} \in \mathbb{R}^2$ and $\sigma = .03$ (for real data, the smoothing parameter σ is scaled by the size of the real box L; here we used $L = 1$).

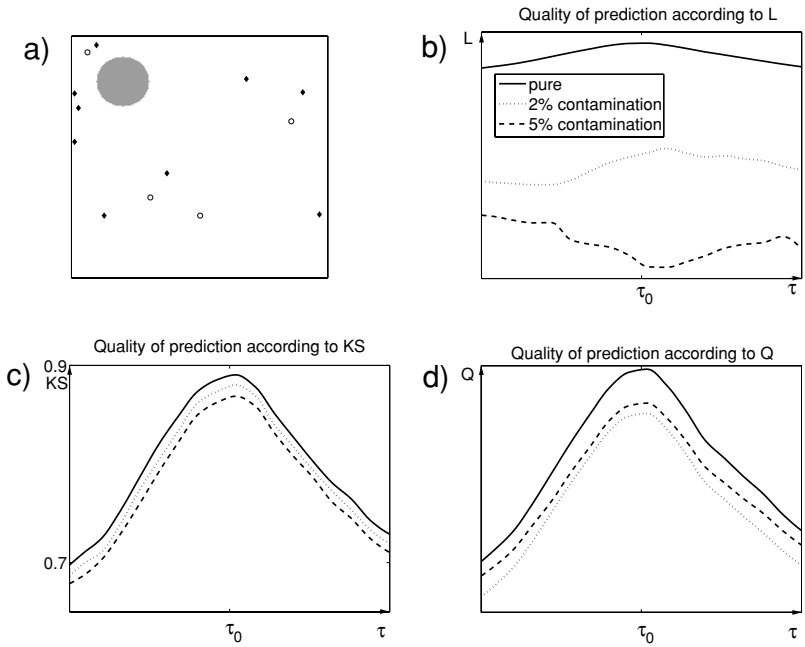

Figure 2: Simulation. A 20-minute spike train was generated from the time-offset place field model $\lambda(t) = F(\mathbf{x}(t + \tau_0))$. (a) Place field and positions of contamination bursts. $F(\mathbf{x})$ is constant inside the shaded region. Open circles represent locations of bursts in 2% contamination spike train; diamonds are bursts in the 5% contamination case. The average firing rate was 2 Hz for both pure and contaminated spike trains. (b–d) L, Q, and KS were computed on time-offset place field models for different values of τ. Only relative values within each curve should be compared. All three valuations recovered the correct value of τ_0 on uncontaminated data. For 2% contamination, L's performance was degraded; for 5% contamination, L completely failed to recover the correct value of τ_0. Q and KS performed equally well in all cases. Horizontal τ axes are all the same.

L's ability to pick out the correct value of τ was still compromised by the contamination.

3.2.3 Experimental Data. To compare the performance of the three valuations on real data, we used a recording of 56 hippocampal place cells in a rat performing a spatial task. These data were kindly provided by G. Buzsaki and previously analyzed in Harris et al. (2003). Place fields without time offsets ($\tau = 0$ only) were computed as in the simulation (see section 3.2.1) on a training set, and then the quality of prediction was evaluated on a separate test set.

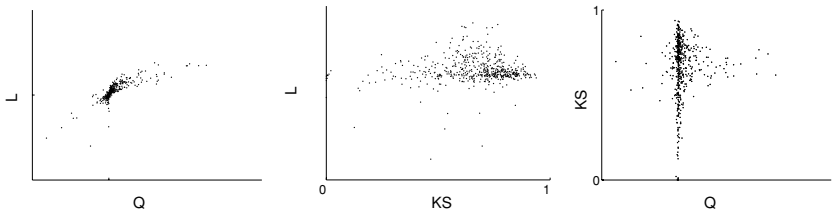

Figure 3: Valuation comparison on hippocampal place field data. The scatter plots show a comparison of the three considered valuations against each other. Place fields $F(\mathbf{x})$ were computed on the training set for each of 56 place cells, yielding predicted conditional intensity functions $\lambda(t) = F(\mathbf{x}(t))$, where $\mathbf{x}(t)$ is the instantaneous position of the rat. The coordinates of each dot represent the values $\mathcal{V}(\lambda, s)$ for a given place cell, where $\mathcal{V} =$ L, Q, or KS. While the correlation between L and Q is strong (Spearman's $\rho = 0.919$), the correlations between KS and L ($\rho = -0.042$) and between KS and Q ($\rho = -0.134$) are weakly negative.

As shown in Figure 3, the log likelihood and quadratic valuations L and Q exhibit a noticeably larger degree of correlation with each other than with KS. This reflects the fact that L and Q evaluate a different aspect of the fit of a spike train to the predicted intensity than does KS: namely, L and Q evaluate how well λ captures the observed instantaneous firing rate, while KS measures how well λ predicts the observed distribution of interspike intervals.

3.3 Relationship Between L and Q. We have seen that L and Q perform similarly on experimental data and that Q is more robust to contamination on simulated data. In section 4, it will become clear that Q also possesses great computational advantages over L. Nevertheless, L is the traditional favorite due to its theoretical properties, such as consistency and a straightforward probabilistic interpretation. In this section we show that for models of interest, Q is also consistent. Furthermore, we show that over the set of all possible conditional intensity functions λ, L and Q have the same global maxima. These findings may increase suspicions that Q is perhaps just a transformed version of log likelihood. We end this section by showing that this is not the case. In section 3.4, however, we show that Q has an interpretation in terms of Euclidean distance.

3.3.1 Consistency of L and Q. One of the reasons for using maximum likelihood estimation is that for data generated from a model that is included in the considered family, the parameter estimates given by L are consistent—they converge to the true values in the limit of large data (van de Geer, 2000). Here we show that Q is also consistent for models of the form

$$\lambda(t) = F(x(t); H_t), \tag{3.3}$$

where $x(t) \in \mathcal{X}$ is the value of stimulus, network, or behavioral variables at time t and $H_t = H_t(s) \in \mathcal{H}$ represents the observed spiking history of the spike train s at time t. We shall call the functions $F : \mathcal{X} \times \mathcal{H} \to \mathbb{R}_{\geq 0}$ *generalized tuning curves* (GTC), and the models (see equation 3.3) *GTC models*. Note that the only difference between this and the full set of models we consider in equation 2.1 is that here we do not allow explicit time dependence.

We formulate the large data limit as a single trial whose length tends to infinity; by ergodicity, this is equivalent to fixing the length of each trial and letting the number of trials go to infinity (though this requires some change of notations). We further assume that at any time $t \in [0, T]$, the relevant spiking history goes back some fixed and finite length of time that does not change in the large data limit $T \to \infty$. Finally, we discretize the time interval $[0, T]$ as well as the stimulus and history spaces $\mathcal{X}$ and $\mathcal{H}$.[6] The reasons for binning time and stimulus space are twofold: (1) it allows a simpler proof by circumventing certain analytical difficulties,[7] and (2) it ensures that we can cover the entire relevant domain in $\mathcal{X} \times \mathcal{H}$ with a large enough data set.

Proof of Consistency of Q. Suppose we have data generated from a GTC model as in equation 3.3. Intuitively, the best estimate for the true $F_0(y; h)$ is given by the expected number of spikes per unit time for a "stimulus" $y \in \mathcal{X}$ and spiking history $h \in \mathcal{H}$:

$$\hat{F}(y; h) = \frac{\text{number of spikes at } (y, h)}{\text{total time spent at } (y, h)} = \frac{S(y, h)}{N(y, h)\Delta t}, \tag{3.4}$$

where $S(y, h)$ denotes the total number of spikes that occurred at stimulus y with history h and $N(y, h)$ is the total number of (discretized) times in which the observed trajectory and spiking history $(x(t), H_t)$ passed through the point (y, h). Note that $N(y, h)$ will be strictly nonzero in time bins for which $S(y, h)$ is nonzero. For each stimulus history bin (y, h), the estimate $\hat{F}(y; h)$ follows a normalized binomial distribution with mean $F_0(y; h)$. By the law of large numbers, the expected value of $\hat{F}(y; h)$ converges to $F_0(y; h)$ as $T \to \infty$. In other words, in the limit of large data the estimate 3.4 converges to the true function F_0 that was used to generate the data.

For any finite data set, both L and Q are maximized by $\hat{F}$. A simple computation yields that for any spike train s with spike times $\{t_s\}$, and time

[6]In particular, we can think of the history space $\mathcal{H}$ as $\mathbb{R}^n$, where n is the number of prior time bins relevant to spiking history.

[7]In practice, there is always some limit to measurement precision, so collected data are naturally discretized, and binned spaces are the only case of interest.

bin size Δt,

$$\begin{aligned}
\mathrm{L}(\lambda, s) &= -\sum_t F(x(t); H_t)\Delta t + \sum_{t_s} \log F(x(t_s); H_{t_s}) \\
&= -\sum_{y,h} F(y; h)N(y, h)\Delta t + \sum_{y,h} \log F(y; h)S(y, h) \\
\mathrm{Q}(\lambda, s) &= -\sum_t F^2(x(t); H_t)\Delta t + 2\sum_{t_s} F(x(t_s); H_{t_s}) \\
&= -\sum_{y,h} F^2(y; h)N(y, h)\Delta t + 2\sum_{y,h} F(y; h)S(y, h).
\end{aligned}$$

(Here we have dropped the overall $1/T$ factors, as they are irrelevant for this computation.) Given a valuation $\mathcal{V}$, the maximum of $\mathcal{V}(\lambda, s)$ can be achieved only when the appropriate variational derivative vanishes:

$$\frac{\delta \mathrm{L}(\lambda, s)}{\delta F(y; h)} = -N(y, h)\Delta t + \frac{S(y, h)}{F(y; h)} = 0 \tag{3.5}$$

$$\frac{\delta \mathrm{Q}(\lambda, s)}{\delta F(y; h)} = -2F(y; h)N(y, h)\Delta t + 2S(y, h) = 0. \tag{3.6}$$

Both equations hold if and only if $F(y; h) = S(y, h)/N(y, h)\Delta t = \hat{F}(y; h)$. As can be seen by taking second derivatives with respect to $F(y; h)$, this solution provides the maximum for each valuation. Thus, for the class of GTC models, the maximum L estimate and the maximum Q estimate coincide. Since $\hat{F} \to F_0$ in the limit of large data, it follows that Q (and L) is consistent on this class of models. □

Is Q still consistent when we optimize over a parameterized subclass of GTC models? In general, for models of the form

$$\lambda_\theta(t) = F_\theta(x(t); H_t),$$

parameterized by θ, Q and L will not have the same optimum for finite data.[8] However, as long as the data were generated by a model from within a subclass, say, with parameter values θ_0, the optimal L and Q estimates $\hat{\theta}^L$ and $\hat{\theta}^Q$ will converge to the true values θ_0. Here we give a sketch of the proof:

Consider data generated by a true model with parameters θ_0 in a subclass of GTC models parameterized by θ. Assume the set of all models in the subclass F_θ forms a submanifold in the space of all possible GTC functions, and

[8] As an example, a short calculation shows that over the class of separable tuning curves $\lambda(t) = F_1(x(t))F_2(H_t)$, L and Q do not have the same maximum.

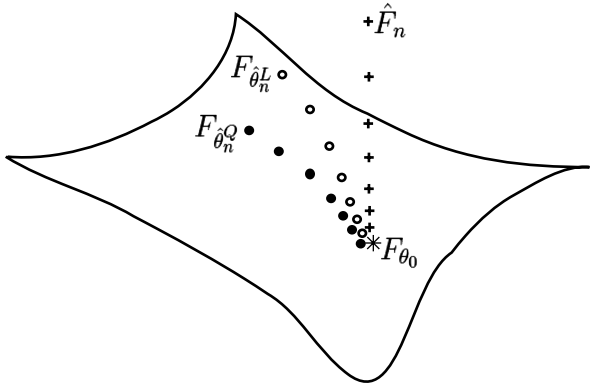

Figure 4: Consistency over a parameterized subclass of GTC models. Consider data generated by a model $\lambda_{\theta_0}(t) = F_{\theta_0}(x(t); H_t)$ belonging to a subclass of GTC models parameterized by θ. We may view the function F_{θ_0} as a point on a submanifold in the space of all GTC functions. For each data set $(x_n(t), s_n)$, the optimal L and Q estimates over the entire class of GTC functions coincide and are given by $\hat{F}_n$. For the constrained optimization over the submanifold, the optimal L estimate $F_{\hat{\theta}_n^L}$ generally differs from the optimal Q estimate $F_{\hat{\theta}_n^Q}$. In the limit of large data, $n \to \infty$, each sequence of estimates $\hat{F}_n$ (crosses), $F_{\hat{\theta}_n^L}$ (open circles), and $F_{\hat{\theta}_n^Q}$ (closed circles) converges to the true F_{θ_0} (star). See text for a sketch of the proof.

let F_{θ_0} denote the point on the submanifold corresponding to the true model (see Figure 4). Let $\{(x_n(t), s_n)\}$ be a sequence of trajectories and spike trains of increasing length, with the limit as $n \to \infty$ being the large data limit. For each n, L and Q yield continuous and convex functions on the entire space of GTC functions: $\mathrm{L}_n(F) = \mathrm{L}(F(x_n(t); H_t), s_n)$ and $\mathrm{Q}_n(F) = \mathrm{L}(F(x_n(t); H_t), s_n)$.

For each n, let $\hat{F}_n$ be defined by equation 3.4. We have just shown that $\hat{F}_n$ is the maximizer of both L and Q over the entire space of GTC functions and that as $n \to \infty$, $\hat{F}_n \to F_{\theta_0}$ on the submanifold, even though $\hat{F}_n$ need not lie in the model subclass for any n. Let $\hat{\theta}_n^L$ and $\hat{\theta}_n^Q$ denote the parameter values that optimize L and Q within the subclass of models; they correspond to points $F_{\hat{\theta}_n^L}$ and $F_{\hat{\theta}_n^Q}$ on the model submanifold (see Figure 4). As the global maximum $\hat{F}_n$ approaches the submanifold, the continuity and convexity of L and Q imply that the constrained maxima $F_{\hat{\theta}_n^L}$ and $F_{\hat{\theta}_n^Q}$ must approach $\hat{F}_n$. In other words, as $n \to \infty$, $\hat{F}_n \to F_{\theta_0}$ and $F_{\hat{\theta}_n^L}, F_{\hat{\theta}_n^Q} \to \hat{F}_n$. Thus, $F_{\hat{\theta}_n^L}, F_{\hat{\theta}_n^Q} \to F_{\theta_0}$, and hence $\hat{\theta}_n^L, \hat{\theta}_n^Q \to \theta_0$, showing that Q (and L) is consistent.

3.3.2 L and Q Have Same Global Maxima. Even when we optimize over all possible GTC models $\lambda(t) = F(x(t); H_t)$, λ is not allowed explicit time dependence. No matter what the actual spiking was, the conditional

intensity function is forced to satisfy the constraint $\lambda(t_1) = \lambda(t_2)$ whenever $(x(t_1), H_{t_1}) = (x(t_2), H_{t_2})$. What if we optimize over all possible λ? A hint is provided by observing that for fixed spike train s, the variational derivatives of L and Q with respect to λ satisfy the following simple relationship:

$$\frac{\delta \mathrm{Q}}{\delta \lambda} = 2\lambda \frac{\delta \mathrm{L}}{\delta \lambda}.$$

For discretized time, the above equation becomes a relation between the gradient vectors of L and Q in $\mathbb{R}^{N_{\text{bins}}}$ (note that multiplication by $\lambda(t)$ scales each coordinate and time bin differently). This implies that on the set of all possible (discretized) nonvanishing λ, Q and L have the same extrema.

We will now show that for a given spike train s, L and Q have the same global maxima when optimized over the entire space of conditional intensity functions λ. Informally, what we show is that

$$\arg\max_{\{\lambda\}} \mathrm{L}(\lambda, s) \text{ “}=\text{” } \arg\max_{\{\lambda\}} \mathrm{Q}(\lambda, s).$$

Unfortunately, L and Q are unbounded from above on the space of all possible λ, so we must be more careful in defining what we mean by the statement that “L and Q have the same global maxima.” We can avoid this complication by approximating λ's with discretized functions.

Given any bin size Δt and corresponding binning $0 = t_0 < t_1 < \cdots < t_{N_b} = T$, we can define the discretization of an integrable function $f(t)$ as the step function

$$f_{\Delta t}(t) = \frac{1}{\Delta t} \int_{t_{i-1}}^{t_i} f(\tau) d\tau \qquad \text{for } t \in (t_{i-1}, t_i].$$

Similarly, for a fixed spike train s, the conditional intensity function λ can also be discretized. We denote $\lambda^s(t) \stackrel{\text{def}}{=} \lambda(t; H_t(s))$, and the discretization $\lambda^s_{\Delta t}$. We can also define the discretization of a spike train s as the step function

$$s_{\Delta t}(t) = \frac{n_i}{\Delta t}, \qquad \text{for } t \in (t_{i-1}, t_i],$$

where n_i is the number of spikes in the ith bin.

Definition. *We say that two valuations $\mathcal{V}_1$ and $\mathcal{V}_2$ have the same global maxima if for every spike train s there exists an ε_s such that*

$$\arg\max_{\{\lambda\}} \mathcal{V}_1\big(\lambda^s_{\Delta t}, s\big) = \arg\max_{\{\lambda\}} \mathcal{V}_2\big(\lambda^s_{\Delta t}, s\big)$$

for every $\Delta t < \varepsilon_s$.

Equivalently, we can say that $\mathcal{V}_1$ and $\mathcal{V}_2$ have the same global maxima if for every spike train s and $\Delta t < \varepsilon_s$ they have the same global maxima among all possible step functions $f_{\Delta t}$:

$$\arg\max_{\{f_{\Delta t}\}} \mathcal{V}_1(f_{\Delta t}, s) = \arg\max_{\{f_{\Delta t}\}} \mathcal{V}_2(f_{\Delta t}, s).$$

In practice, this second definition is simpler to work with when checking whether a pair of valuations satisfies this property.[9]

The log likelihood valuation L and the quadratic valuation Q have the same global maxima. Indeed, for any step function $f_{\Delta t}$,

$$\mathrm{L}(f_{\Delta t}, s) = \frac{1}{T}\sum_{i=1}^{N_b}\left(n_i \log f^i_{\Delta t} - \Delta t f^i_{\Delta t}\right)$$

$$\mathrm{Q}(f_{\Delta t}, s) = \frac{1}{T}\sum_{i=1}^{N_b}\left(2n_i f^i_{\Delta t} - \Delta t\left(f^i_{\Delta t}\right)^2\right),$$

where $f^i_{\Delta t}$ denotes the value of $f_{\Delta t}$ in the ith bin. As can be seen by partial differentiation with respect to $f^i_{\Delta t}$, both valuations have a unique global maximum at $f_{\Delta t} = s_{\Delta t}$.

The KS valuation, however, has a different global maximum. For any spike train s and any $\Delta t < \frac{1}{2}\min\{s^k - s^{k-1}\}$, the time-rescaled spike times $\Lambda(s^k) = \int_0^{s^k} s_{\Delta t}(t)\,\mathrm{d}t = k$ correspond to a regular spike train with ISIs $\tau_k = 1$ for all k. We find $\mathrm{KS}(s_{\Delta t}, s) = e^{-1}$, much less than the maximal value of 1, demonstrating that the global maxima of L and Q do not optimize the KS valuation. Intuitively, this is because $s_{\Delta t}$ minimizes statistical fluctuations, predicting every spike time with perfect accuracy and yielding a distribution of ISIs very different from expected.

The fact that when optimizing over the space of all possible λ's, both L and Q are maximized by the conditional intensity function that predicts each spike time perfectly presents an obvious problem for generalizability. For this reason, we again stress that when using these valuations, a method such as cross-validation should be employed to avoid overfitting. On the other hand, sometimes spike times that were originally assumed stochastic are revealed on further investigation to be more reliable than previously expected (Bair and Koch, 1996; Ishikane, Gangi, Honda, & Tachibana, 2005). In these cases, fluctuations (although randomly distributed) may reflect not noise but additional structure in the data. A predicted intensity function

[9]The proof of the equivalence of these two definitions follows easily from considering the equivalence relation $\sim_s$ defined by $\lambda_1 \sim_s \lambda_2 \iff \lambda_1(t, H_t(s)) = \lambda_2(t, H_t(s))$, and observing that every $\lambda(t; H_t) \sim_s \lambda^s(t)$.

that matches the observed spike trains more closely may be desirable as long as it holds up under cross-validation.

3.3.3 Q Is Not L in Disguise. Because L and Q are correlated on real data, both are consistent over the class of GTC models, and both share the same global maxima over the space of all λ's, one might wonder whether Q is really a log likelihood in disguise. Clearly, the valuation $\mathrm{Q}(\lambda, s)$ is not the log likelihood of a point process for which λ is the conditional intensity function. However, there might be another conditional intensity function for which Q can be interpreted as a log likelihood. This can be rephrased into a somewhat more general question: Can λ be considered as a parameter for a continuous family of point processes whose log likelihood is

$$\ell_\lambda(s) = \mathrm{Q}(\lambda, s) + H(s)$$

for some function $H(s)$ that depends only on spike trains? In appendix A we show that the answer to this question is negative, even for the restricted class of history-independent $\lambda(t)$. This implies there is no way we can reinterpret Q as a log likelihood, and hence there is no (simple) probabilistic interpretation for Q. This is a fundamental theoretical difference between L and Q.

3.4 Interpretation of Q. We have seen that Q does not have a simple probabilistic interpretation. Does Q have any sensible interpretation? In this section, we show that Q has a straightforward interpretation in terms of Euclidean distance for history-independent intensity functions $\lambda(t)$. What follows is rather technical and independent from section 4 on peer prediction.

We would like to think of $\mathrm{Q}(\lambda, s)$ as a measure of the "distance" between the function $\lambda(t)$ and the spike train s. Ideally, we would write $\mathrm{Q}(\lambda, s) = -\|\lambda - s\|^2 + \|s\|^2$, using the usual L^2 norm $\|f\|^2 = \int_{-\infty}^{\infty} |f(t)|^2 \, \mathrm{d}t$. The comparison between two predictions would thus be given as a difference of Euclidean distances:

$$\|\lambda_2 - s\|^2 - \|\lambda_1 - s\|^2.$$

Unfortunately, since δ-functions are not in the function space $L^2(\mathbb{R})$, none of these norms are defined.

In section 3.1.2 we derived Q using the squared Euclidean distance between a discretized intensity function and a spike train, in the limit as bin size $\Delta t \to 0$. Instead of discretizing, we could have chosen to compute the L^2 distance between λ and a smoothed spike train $s * \Omega_\rho$, where Ω_ρ is a smoothing function with smoothing parameter ρ. This is equivalent

to using the Fourier transform $\hat{\Omega}_\rho$ as a "high-frequency cutoff function" to define a new scalar product $\langle \cdot, \cdot \rangle_{\Omega,\rho}$, which is also defined on spike trains.[10]

For any smoothing function Ω with bounded and continuous Fourier transform $\hat{\Omega} \in L^1(\mathbb{R})$ and $\hat{\Omega}(0) = 1$ (i.e., $\int_{-\infty}^{\infty} \Omega(t)\,\mathrm{d}t = \sqrt{2\pi}$), we can define a family of functions $\hat{\Omega}_\rho(\xi) \stackrel{\text{def}}{=} \hat{\Omega}(\rho\xi)$. Here ξ has units of frequency, and ρ is a dimensionless parameter. For each ρ, we define a scalar product[11]

$$\langle f, g \rangle_{\Omega,\rho} \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \hat{\Omega}_\rho(\xi) \hat{f}(\xi) \hat{g}(\xi)^* \mathrm{d}\xi.$$

This yields a metric $\|f - g\|^2_{\Omega,\rho} = \langle f - g, f - g \rangle_{\Omega,\rho}$, which is defined for a variety of functions, including spike trains. In terms of the ordinary scalar product, we can write $\langle f, g \rangle_{\Omega,\rho} = \langle f, \frac{1}{\sqrt{2\pi}} \Omega_\rho * g \rangle$. This approach is also sometimes used to define spike train metrics.[12]

For each choice of smoothing function Ω, a "Euclidean distance" valuation can be defined:

$$\mathcal{V}_\Omega(\lambda, s) \stackrel{\text{def}}{=} \frac{1}{T} \lim_{\rho \to 0} \left(-\|\lambda - s\|^2_{\Omega,\rho} + \|s\|^2_{\Omega,\rho} \right).$$

This is the limit in which the smoothing function Ω_ρ tends to a δ-function (no smoothing). At this point, one might worry that different choices of Ω could yield different valuations. We show in appendix B that for any bounded and continuous $\hat{\Omega} \in L^1(\mathbb{R})$, for any $\hat{\lambda} \in L^1(\mathbb{R})$, and for any spike train s,

$$\mathcal{V}_\Omega(\lambda, s) = \mathrm{Q}(\lambda, s).$$

In short, although one must regularize or discretize (as in section 3.1.2) the spike train in order to compute a true Euclidean distance between s and $\lambda(t)$, we have shown that Q can be defined as the limit of such "Euclidean distance" valuations regardless of the regularization (or discretization) procedure.

[10] Hats denote the Fourier transform $\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-it\xi}\,\mathrm{d}t$.

[11] This is a generalization of Sobolev norms, which use cutoffs of the form $\hat{\Omega}(\xi) = \frac{1}{(1+|\xi|)^p}$.

[12] For example, in van Rossum (2001), a spike train metric is defined by convolving the spike trains with a decaying exponential. There are also spike train metrics that cannot be derived from Euclidean distance (see Victor & Purpura, 1996).

4 Peer Prediction

Here we present an example of spike train prediction, which will help to illustrate practical differences between L and Q. Assume we have a simultaneous recording of n cells on an interval $[0, T]$, with spike trains $s_\alpha(t) = \sum_k \delta(t - s_\alpha^k)$, where $\alpha = 1, \ldots, n$ runs over cells, and $k = 1, \ldots, N_\alpha$ is the number of spikes for neuron α. We can analyze the structure of neuronal coordination by computing the effective weights[13] $w_{\alpha\beta}$ in the model

$$\lambda_\alpha(t) = g\left(\sum_{\beta\neq\alpha} w_{\alpha\beta}\, \phi * s_\beta + \bar{\lambda}_\alpha\right), \tag{4.1}$$

where g is a link function. Here $\bar{\lambda}_\alpha$ is constant on the interval $[0, T]$ and zero elsewhere,[14] $(\phi * s)(t) = \int_{-\infty}^{+\infty} \phi(t-\tau)s(\tau)d\tau$ is convolution, and $\phi(t)$ is the function that determines the timing of synchronization,[15] typically a gaussian

$$\phi_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}},$$

where the parameter σ might be interpreted as the timescale on which spikes are synchronized.

The computation of model parameters is greatly simplified if the link function can be chosen to be identity. Note, however, that in this linear model, the weights $w_{\alpha\beta}$ are allowed to be negative, which would result in negative intensities $\lambda_\alpha(t)$. Such a λ has no probabilistic interpretation; the quality of this prediction cannot be evaluated using L because this would involve taking the log of negative numbers. Nevertheless, Q is still defined for $\lambda < 0$, but such λ's always give worse predictions than nonnegative ones. We will use Q in what follows.

For a given function $\phi(t)$, we can find the effective weights by maximizing the quadratic function of weights,

$$\mathcal{Q}(w, \bar{\lambda}) = \sum_\alpha \left(\mathrm{Q}(\lambda_\alpha, s_\alpha) - k^2 \sum_{\beta\neq\alpha} w_{\alpha\beta}^2\right),$$

where Q is the quadratic valuation and the second term is a penalty added to reduce overfitting (as in ridge regression).

[13] This represents an effective connectivity in the network and should not be confused with actual synaptic connections.

[14] With abuse of notation, we shall denote both the function and the constant by $\bar{\lambda}_\alpha$.

[15] For example, having $\phi(t) = \delta(t)$ would mean that all the cells in a given cell assembly tend to fire at exactly the same time.

A direct calculation (see appendix C) shows that under some mild conditions on ϕ, $\mathcal{Q}(w, \bar{\lambda})$ may be expressed as a function of cross-correlograms:[16]

$$\mathcal{Q}(w, \bar{\lambda}, \{s_\alpha\}) = \mathcal{F}(w, \bar{\lambda}, \{c_{\alpha\beta}\}, \{N_\alpha\}),$$

where $c_{\alpha\beta}$ denotes the cross-correlogram

$$c_{\alpha\beta}(\tau) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} s_\alpha(t) s_\beta(t+\tau)\, dt = \sum_{i,j} \delta\big(\tau + s_\alpha^i - s_\beta^j\big), \tag{4.2}$$

and s_α^i is the time of the ith spike in the spike train s_α. In other words, the optimal parameters for linear peer prediction using Q depend on only correlations between the spike trains, not on individual spike times.

Because Q is quadratic in $(w_{\alpha\beta}, \bar{\lambda}_\alpha)$ the problem of finding the optimal weights becomes a linear problem. The optimal weights $w_{\alpha\beta}$ and baseline rates $\bar{\lambda}_\alpha$ can be found as a solution of the linear system

$$\sum_{\gamma \neq \alpha} w_{\alpha\gamma} A_{\gamma\beta} = B_{\alpha\beta} \qquad \forall \alpha \neq \beta \tag{4.3}$$

$$\bar{\lambda}_\alpha = \frac{N_\alpha}{T} - \sum_{\beta \neq \alpha} w_{\alpha\beta} \frac{N_\beta}{T}, \tag{4.4}$$

where

$$A_{\gamma\beta} = -k^2 \delta_{\gamma\beta} + \frac{N_\gamma N_\beta}{T^2} - \frac{1}{T} \int_{-\infty}^{+\infty} \phi_2(t) c_{\beta\gamma}(t) dt \tag{4.5}$$

$$B_{\alpha\beta} = \frac{N_\alpha N_\beta}{T^2} - \frac{1}{T} \int_{-\infty}^{+\infty} \phi(t) c_{\beta\alpha}(t) dt, \tag{4.6}$$

$\phi_2(t) = \int_{-\infty}^{+\infty} \phi(\tau)\phi(\tau - t) d\tau$, and N_α is the total number of spikes of cell α on the training set. To see this, observe that since $\mathcal{Q}(w, \bar{\lambda})$ is quadratic in the variables $(w_{\alpha\beta}, \bar{\lambda}_\alpha)$, it achieves its maximum at the solution to the system $\nabla_{(w,\bar{\lambda})} \mathcal{Q}(w, \bar{\lambda}) = 0$. Computing this gradient yields equations 4.3 to 4.6.

Harris et al. (2003) used a nonlinear link function, together with the valuation L. The resulting nonlinear optimization was computationally intensive. In contrast, optimization with Q using linear peer prediction was up to two orders of magnitude faster, as would be expected from solving a linear problem. Nevertheless, the two valuations yielded similar results. As shown in Figure 5, both Q and L select models of peer prediction with the same timescale σ.

[16]See appendix C for an explicit formula.

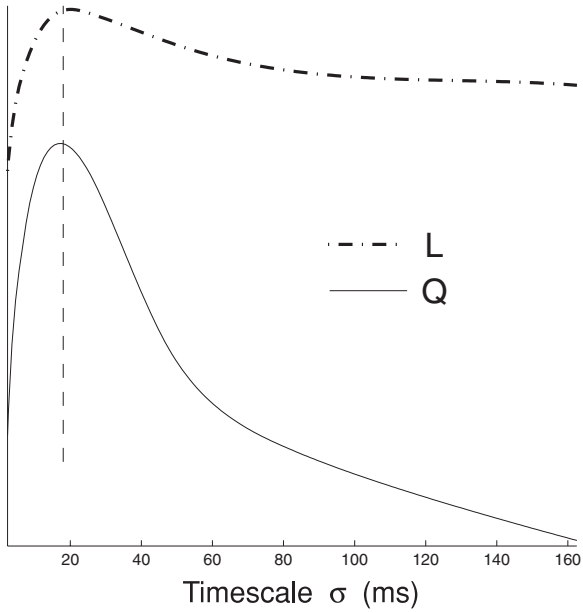

Figure 5: Peer prediction with L and Q. For each timescale σ, linear and nonlinear peer prediction models were fit on a training set using Q and L, respectively. The plotted curves represent the test set values of L and Q, for the corresponding predictions, averaged across the population of cells. L and Q peak together, selecting peer prediction models with the same timescale $\sigma \approx 20$ ms. Note that L and Q are on different scales; their numerical values should not be compared.

5 Conclusion

We have considered a paradigm for spike train analysis in which biological hypotheses are translated into prescriptions for predicting the conditional intensity λ. Models are then compared using a prediction quality valuation. We considered the L, Q, and KS valuations. We found that L and Q behaved similarly on experimental data, whereas KS was quite different. On simulated data, where the true model was known, all three valuations found the correct model on uncontaminated data. For contaminated spike trains, however, L lost resolution and could no longer be used to find the true value, while Q and KS were robust to contamination and continued to find the correct model parameter. Q's quadratic nature also yields important practical advantages in computational efficiency. As seen in peer prediction, optimization with Q may be orders of magnitude faster than with L.

We further showed that Q shares some important theoretical properties with L: both valuations are consistent on models of interest and share the

same global maxima. Unlike L, Q does not have a simple probabilistic interpretation but can be interpreted in terms of Euclidean distance. The similarities between L and Q, as well as Q's natural interpretation, lead us to propose that Q may serve as an alternative to L. In cases where considerations of robustness or computational efficiency are important, Q may be the more attractive choice.

Appendix A: Q Cannot Be Turned into Log Likelihood

Proposition 1. *There is no family of point processes, parameterized by $\lambda(t)$, whose log likelihood is given by*

$$\ell_\lambda(s) = Q(\lambda, s) + H(s), \tag{A.1}$$

where $H(s)$ is some function defined on spike trains.

Proof. We first consider the case of constant functions $\lambda(t) = f = \text{const}$. Without loss of generality we may assume that time is measured in units of the total time T. Thus, equation 3.2 becomes

$$\mathrm{Q}(f, s) = \left(-f^2 + 2fN_s\right).$$

If $\ell_\lambda(s)$ were the log likelihood of observing a spike train s, then computing the integral over the set $\mathcal{S}$ of all possible spike trains would yield

$$\int_{\mathcal{S}} e^{\ell_\lambda(s)} ds = \sum_{n=0}^{\infty} \int_{\mathcal{S}_n} e^{-f^2+2fn+H(s)} ds = e^{-f^2} \sum_{n=0}^{\infty} h_n e^{2fn} = 1, \tag{A.2}$$

where

$$h_n \stackrel{\text{def}}{=} \int_{\mathcal{S}_n} e^{-H(s)} ds,$$

and $\mathcal{S}_n$ denotes the set of all spike trains that have exactly n spikes.

Since $\mathrm{Q}(f, s)$ is continuous in f, equation A.1 should hold for constant $\lambda(t) = f$ in some nontrivial interval $[a, b]$, and the last equality in equation A.2 should hold for all $f \in [a, b]$. This contradicts lemma 1 below. Therefore regardless of the choice of $H(s)$, the function $\ell_\lambda(s)$, equation A.1, cannot be a log likelihood of observing the spike train s for constant λ.

This proof is easily generalized to the case when $\lambda(t)$ is a linear combination of step functions. Since step functions are dense in the space of all functions on a given time interval and the correspondence $\lambda \mapsto \ell_\lambda(s)$ is assumed to be continuous, the proof is valid for any function $\lambda(t)$.

Lemma 1. *For any sequence of positive numbers $\{h_n\}_{n\geq 0}$ and any non-trivial interval $[a, b] \subset \mathbb{R}$, the series $S(f) \stackrel{def}{=} e^{-f^2} \sum_{n=0}^{\infty} e^{2fn} h_n$ fails to converge to 1 in at least one point $f \in [a, b]$.*

Proof. Denote $z = e^{2f}$. If $S(f)$ converges to 1 for all $f \in [a, b]$, then the power series $\sum_{n=0}^{\infty} z^n h_n$ converges and is equal to $F(z) = e^{(\frac{\ln z}{2})^2}$ for $z \in [e^{2a}, e^{2b}]$. Thus, the power series must converge inside the circle $|z| < e^{2b}$ on the complex plane. This contradicts the observation that the function $F(z)$ blows up at the origin and thus does not have a holomorphic continuation inside any circle containing the origin (see, e.g., Rudin, 1987).

Appendix B: Q Is Independent of Regularization

Proposition 2. *Let $\hat{s} \in L^{\infty}(\mathbb{R})$ (so s can be a spike train) and $\hat{\lambda} \in L^1(\mathbb{R})$ (this is guaranteed for $\lambda \in C^2(\mathbb{R})$, with compact support), and let $\mathcal{V}_{\Omega}$ and $\|\cdot\|^2_{\Omega,\rho}$ be defined as in section 3.4. Then*

$$Q(\lambda, s) = V_{\Omega}(\lambda, s) = \lim_{\rho \to 0} \frac{1}{T}\left(-\|\lambda - s\|^2_{\Omega,\rho} + \|s\|^2_{\Omega,\rho}\right),$$

for every bounded and continuous $\hat{\Omega}$, with $\hat{\Omega}(0) = 1$.

Proof. Observing that $\mathcal{V}_{\Omega}(\lambda, s) = \frac{1}{T} \lim_{\rho\to 0}(-\langle \lambda, \lambda\rangle_{\Omega,\rho} + 2\langle \lambda, s\rangle_{\Omega,\rho})$, and $Q(\lambda, s) = \frac{1}{T}(-\langle \lambda, \lambda\rangle + 2\langle \lambda, s\rangle)$, the proof follows from the following lemma.

Lemma 2. *If $\hat{\Omega}$ is bounded and continuous with $\hat{\Omega}(0) = 1$, $\hat{f} \in L^1(\mathbb{R})$ and $\hat{g} \in L^{\infty}(\mathbb{R})$, then*

$$\lim_{\rho\to 0} \langle f, g\rangle_{\Omega,\rho} = \langle f, g\rangle.$$

Proof. By Lebesgue's dominated convergence theorem (Rudin, 1987),

$$\lim_{\rho\to 0} \langle f, g\rangle_{\Omega,\rho} = \lim_{\rho\to 0} \int_{-\infty}^{\infty} \hat{\Omega}_{\rho}(\xi) \hat{f}(\xi) \hat{g}(\xi)^* \, d\xi = \int_{-\infty}^{\infty} \lim_{\rho\to 0} \hat{\Omega}_{\rho}(\xi) \hat{f}(\xi) \hat{g}(\xi)^* \, d\xi$$

if there exists $h \in L^1(\mathbb{R})$ such that $\left|\hat{\Omega}_{\rho}(\xi) \hat{f}(\xi) \hat{g}(\xi)\right| \leq h(\xi)$, for all $\rho \geq 0$. Since $\hat{\Omega}$ is bounded, $|\hat{\Omega}_{\rho}(\xi)| \leq \|\hat{\Omega}\|_{\infty}$ for all ρ, where $\|\cdot\|_{\infty}$ denotes the supremum norm. Thus, we can take $h = \|\hat{\Omega}\|_{\infty} \|\hat{g}\|_{\infty} \hat{f}$. Moreover, since $\hat{\Omega}$ is continuous, $\lim_{\rho\to 0} \hat{\Omega}_{\rho}(\xi) = \lim_{\rho\to 0} \hat{\Omega}(\rho\xi) = \hat{\Omega}(0) = 1$, which completes the proof.

Appendix C: Peer Prediction and Cross-Correlograms

Let $Q(\lambda_{\alpha}, s_{\alpha})$ denote the valuation 3.2 where $\lambda_{\alpha}(t) = \sum_{\beta\neq\alpha} w_{\alpha\beta}\, \phi * s_{\beta} + \bar{\lambda}_{\alpha}$ is the prediction for the cell α, and $s_{\alpha}(t) = \sum_i \delta(t - s^i_{\alpha})$ is the corresponding

spike train. Recall from section 4 that

$$\mathcal{Q}(w, \bar{\lambda}) = \sum_{\alpha} \left(\mathrm{Q}(\lambda_\alpha, s_\alpha) - k^2 \sum_{\beta \neq \alpha} w_{\alpha\beta}^2 \right).$$

Let $\sigma^2 = \int_{-\infty}^{+\infty} t^2 \phi(t) dt$, and let d_α denote the average distance between the spikes of s_α and the boundary of $[0, T]$.

Lemma 3. *Assume that $\int_{-\infty}^{+\infty} \phi(t) dt = 1$, and $\sigma \ll \min_\alpha d_\alpha$. Then*

$$\begin{aligned}\mathcal{Q}(w, \bar{\lambda}) = &- \sum_{\alpha} \left(\bar{\lambda}_\alpha^2 - 2\bar{\lambda}_\alpha \frac{N_\alpha}{T} \right) + \sum_{\alpha} \sum_{\beta \neq \alpha} \sum_{\gamma \neq \alpha} w_{\alpha\beta} w_{\alpha\gamma} \\ &\times \left(-k^2 \delta_{\beta\gamma} - \frac{1}{T} \int_{-\infty}^{+\infty} \phi_2(t) c_{\beta\gamma}(t) dt \right) \\ &+ \frac{2}{T} \sum_{\alpha} \sum_{\beta \neq \alpha} w_{\alpha\beta} \left(\int_{-\infty}^{+\infty} \phi(t) c_{\beta\alpha}(t) dt - \bar{\lambda}_\alpha N_\beta \right),\end{aligned}$$

where $\phi_2(t) = \int_{-\infty}^{+\infty} \phi(\tau)\phi(\tau - t) d\tau$, and N_α is the total number of spikes of cell α on the training set.

Proof. Denote $\langle f, g \rangle = \int_{-\infty}^{+\infty} f(t) g(t) dt$. Then

$$\begin{aligned}\mathcal{Q}(w, \bar{\lambda}) = &\sum_{\alpha} \left(\mathrm{Q}(\lambda_\alpha, s_\alpha) - k^2 \sum_{\beta \neq \alpha} w_{\alpha\beta}^2 \right) \\ = &\sum_{\alpha} \left[-k^2 \sum_{\beta \neq \alpha} w_{\alpha\beta}^2 - \frac{1}{T} \left\langle \sum_{\beta \neq \alpha} w_{\alpha\beta}\, \phi * s_\beta + \bar{\lambda}_\alpha, \sum_{\gamma \neq \alpha} w_{\alpha\gamma} \phi * s_\gamma + \bar{\lambda}_\alpha \right\rangle \right. \\ &\left. + \frac{2}{T} \left\langle \sum_{\beta \neq \alpha} w_{\alpha\beta}\, \phi * s_\beta + \bar{\lambda}_\alpha, s_\alpha \right\rangle \right].\end{aligned}$$

Using the identity $\langle f * g, h \rangle = \langle f, g^- * h \rangle$, where $g^-(t) \stackrel{\text{def}}{=} g(-t)$, we obtain

$$\begin{aligned}\mathcal{Q}(w, \bar{\lambda}) = &\sum_{\alpha} \left[\sum_{\beta \neq \alpha} \sum_{\gamma \neq \alpha} w_{\alpha\beta} w_{\alpha\gamma} \left(-k^2 \delta_{\beta\gamma} - \frac{1}{T} \langle \phi * \phi^-, s_\beta^- * s_\gamma \rangle \right) - \frac{1}{T} \langle \bar{\lambda}_\alpha, \bar{\lambda}_\alpha \rangle \right. \\ &\left. + \frac{2}{T} \sum_{\beta \neq \alpha} w_{\alpha\beta} (\langle \phi, s_\beta^- * s_\alpha \rangle - \langle s_\beta * \phi, \bar{\lambda}_\alpha \rangle) + \frac{2}{T} \langle \bar{\lambda}_\alpha, s_\alpha \rangle \right].\end{aligned}$$

Noting that $\langle \bar{\lambda}_\alpha, \bar{\lambda}_\alpha \rangle = \bar{\lambda}_\alpha^2 T$ and $\langle \bar{\lambda}_\alpha, s_\alpha \rangle = \bar{\lambda}_\alpha N_\alpha$ (here $\bar{\lambda}_\alpha$ denotes the function on the left-hand side, and the constant value on the right-hand side of each equation), and using the approximation[17]

$$\langle s_\beta * \phi, \bar{\lambda}_\alpha \rangle = \bar{\lambda}_\alpha \sum_i \int_0^T \phi\big(t - s_\beta^i\big) dt \simeq \bar{\lambda}_\alpha N_\beta, \tag{C.1}$$

we can express $\mathcal{Q}(w, \bar{\lambda})$ in terms of cross-correlograms $c_{\beta\gamma} = s_\beta^- * s_\gamma$:

$$\mathcal{Q}(w, \bar{\lambda}) \simeq \sum_\alpha \Bigg[\sum_{\beta \neq \alpha} \sum_{\gamma \neq \alpha} w_{\alpha\beta} w_{\alpha\gamma} \left(-k^2 \delta_{\beta\gamma} - \frac{1}{T} \langle \phi * \phi^-, c_{\beta\gamma} \rangle \right) + \frac{2}{T} \sum_{\beta \neq \alpha} w_{\alpha\beta} (\langle \phi, c_{\beta\alpha} \rangle - \bar{\lambda}_\alpha N_\beta) - \bar{\lambda}_\alpha^2 + \frac{2}{T} \bar{\lambda}_\alpha N_\alpha \Bigg].$$

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was partly supported by NIH (R01MH073245) and an Alfred P. Sloan research fellowship to K.D.H.; V.I. was also supported by the Swartz Foundation.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Aronov, D., & Victor, J. D. (2004). Non-Euclidean properties of spike train metric spaces. *Physical Review E, 69*(6), 061905.

Bair, W., & Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation, 8*, 1185–1202.

Barbieri, R., Quirk, M. C., Frank, L. M., Wilson, M. A., & Brown, E. N. (2001). Construction and analysis of non-Poisson stimulus-response models of neural spiking activity. *Journal of Neuroscience Methods, 105*(1), 25–37.

Brown, E., Barbieri, B., Ventura, V., Kass, V., & Frank, L. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput., 14*(2), 325–346.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes.* New York: Springer-Verlag.

[17]This is where we need the assumption $\sigma \ll \min_\alpha d_\alpha$. In other words, there are few spikes "near" the boundary as compared to the timescale σ.

Harris, K., Csicsvari, J., Hirase, H., Dragoi, G., & Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature, 424*, 552–556.

Ishikane, H., Gangi, M., Honda, S., & Tachibana, M. (2005). Synchronized retinal oscillations encode essential information for escape behavior in frogs. *Nat. Neurosci., 8*(8), 1087–1095.

Kjaer, T., Hertz, J., & Richmond, B. (1994). Decoding cortical neuronal signals: Network models, information estimation and spatial tuning. *J. Comput. Neurosci., 1*, 109–139.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York: Wiley.

Paninski, L., Pillow, J., & Simoncelli, E. (2004). Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model. *Neural Comput., 16*(12), 2533–2561.

Rudin, W. (1987). *Real and complex analysis*. New York: McGraw-Hill.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist., 6*(2), 461–464.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol., 93*(2), 1074–1089.

van de Geer, S. A. (2000). *Applications of empirical process theory*. Cambridge: Cambridge University Press.

van Rossum, M. C. W. (2001). A novel spike distance. *Neural Computation, 13*(4), 751–763.

Victor, J. D., & Purpura, K. P. (1996). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J. Neurophysiol., 76*(2), 1310–1326.

Received April 18, 2005; accepted May 1, 2007.