Finding Likely Models that Describe Population Responses

Don H. Johnson and Jyotirmai Uppuluri
*Department of Electrical & Computer Engineering, MS 366
Rice University
6100 Main Street
Houston, Texas 77251–1892
dhj@rice.edu, juppu@rice.edu

Abstract

Analyzing the spike response of a neural population will be plagued by a lack of data unless some knowledge of the interrelationships among the discharge patterns of individual neurons places the available data in context. Unfortunately, determining that context —a model —is precisely what the data analysis must produce. We describe here a new approach to population data analysis that turns the usual data analysis procedure around: Instead of measuring response statistics, such as discharge rates and correlation coefficients, and then finding a single model from them, we find *all* statistical models consistent with the data. We use an objective information theoretic criterion for determining appropriate models. Our approach generalizes maximum likelihood techniques for estimating model parameters.

1 Introduction

A fundamental result from information theory casts a shadow on the ability to glean the structure of a population's response from data. For example, suppose we had a population of N neurons, and that our first concern was to measure inter-neuron correlations as well as determining the individual discharge rates. Using the typical data analysis procedure of binning the data (no more than one spike/bin) and ignoring temporal correlations, Weinberger [4] showed that the amount of data needed for accurate data analysis must be proportional to 2^N [2]. It is not what the data analysis measures that creates this "data hungry" result; rather any procedure based on binned data will face the problem that as another neuron is added to the population, twice as much data will be needed to sustain measurement accuracy. Worse yet, in order to analyze B-bin intra-neuron dependence as well as inter-neuron relationships, the amount of data needed is proportional to 2^{BN}. These statistical realities can be ignored by considering only pairwise interactions, but an encompassing model of the population response cannot be obtained from studying pairwise interactions alone. To obtain a complete model, we have little hope for analyzing anything other than very small populations over small time windows.

The fundamental reason for this depressing situation is that typically we have no model for the population response. *If* a model were available, the amount of data needed would not necessarily follow Weinberger's result. Essentially, a model puts the data in context and allows us to focus the data analysis. When we have a parametric model, we could find the maximum likelihood estimate of the parameters, which would give us a single model most consistent with the data *within those models having the assumed parametric form*. How would we know, however, that any of these models actually describes the data well? When we don't have a specific model in mind, we ideally would want the data to reveal the best model. However, Weinberger's result shows that using non-parametric statistical approaches (*i.e.*, no model) to analyze population responses won't work because we will lack enough data.

In this paper, we suggest a different approach to population data analysis. We try *all* models and find those most likely to have produced the data. Essentially, instead of demanding we have massive data resources (*i.e.*, the Weinberger result), we treat the data as scarce and precious and demand massive computational resources. The issues become how to determine the best models, how to evaluate their fits, and how to search efficiently through them.

2 Data Analysis Procedure

In the usual procedure for statistically analyzing a population's spiking response, each neuron's discharge pattern is binned in time, and the presence or absence of a spike in each time bin is represented by a binary symbol. At each time bin, we concatenate the individual symbols from each neuron to form the response measurement R to describe the entire population's response in a given bin [2]. We envision performing our analysis at each bin separately, which means we can consider one bin at a time. We assume we have M measured responses produced by repeating the stimulus M times; these form the response vector $\mathbf{R} = \{R_1, \ldots, R_M\}$ which represents a dataset for the population. We further assume these responses were produced by carefully repeating the stimulus to minimize sequential response effects. Consequently, the N responses are statistically independent of each other. If $P(R)$ denotes the probability of producing an individual population response, the probability of the entire set of measurements $P(\mathbf{R})$ thus equals $\prod_{m=1}^{M} P(R_m)$.

This joint probability of the measured response can be written in a simple way using the theory of types [1].

$$P(\mathbf{R}) = \prod_{m=1}^{M} P(R_m) = \exp\left\{-M\left[\mathcal{H}(\widehat{P}) + \mathcal{D}\left(\widehat{P}\|P\right)\right]\right\} \tag{1}$$

$\widehat{P}$ denotes the *type* of the data. A type is simply a histogram formed from the measured responses. For one neuron, the type computed at a bin amounts to the PST histogram. For two neurons, the type estimates four probabilities: that neither neuron produced spikes, that neuron "one" fired and neuron "two" did not, that neuron "two" fired and neuron "one" did not, and that both neurons fired. Result (1) means that the measurements are completely summarized by the type, and that the type is *all* that is needed in *any* subsequent analysis, information theoretic or not. $\mathcal{H}(\widehat{P})$ is the entropy of the type [1] and $\mathcal{D}\left(\widehat{P}\|P\right)$ is the Kullback-Leibler distance [2, 3] between the type and the probability distribution that describes how the data were produced.

$$\mathcal{H}(\widehat{P}) = -\sum_{r} \widehat{P}(r) \log \widehat{P}(r) \tag{2}$$

$$\mathcal{D}\left(\widehat{P}\|P\right) = \sum_{r} \widehat{P}(r) \log \frac{\widehat{P}(r)}{P(r)} \tag{3}$$

Here, r ranges over all of the 2^N possible values for the population's binary-coded response. In this paper, the base of the logarithm and the exponential are arbitrary; most frequently, the base equals two, which means that entropy and Kullback-Leibler distance have units of bits, and that $\exp\{\cdot\}$ means $2^{\{\cdot\}}$. The entropy and the Kullback-Leibler distance terms are each greater than or equal to zero.

The key idea rests on the fact that *any* statistical description of how the data were produced, whether it is right or wrong, has a probability of having produced the data. Assuming we have many possible models, we augment the notation for the response probability by a parameter vector $\boldsymbol{\theta}$ to indicate that possible probability models within a parametric family result from specific parameter choices.

$$P(\mathbf{R};\boldsymbol{\theta}) = \exp\left\{-M\left[\mathcal{H}(\widehat{P}) + \mathcal{D}\left(\widehat{P}\|P(\boldsymbol{\theta})\right)\right]\right\} \tag{4}$$

Within the context of a parametrically controlled model, the maximum likelihood estimate of the parameter vector is found by finding that $\boldsymbol{\theta}$ which maximizes the response probability $P(\mathbf{R};\boldsymbol{\theta})$. Because the Kullback-Leibler distance term in (4) is the only one containing the parameters, maximizing the likelihood function on the left side is equivalent to minimizing the Kullback-Leibler distance term on the right. Consequently,

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathcal{D}\left(\widehat{P}\|P(\boldsymbol{\theta})\right)$$

Here, $\mathrm{argmin}_{\boldsymbol{\theta}}$ means the value of $\boldsymbol{\theta}$ that minimizes the Kullback-Leibler distance. It is important to note that the minimum value of the Kullback-Leibler distance is usually not zero. For example, one possible model one might use would describe the responses of two neurons as being statistically independent. This model would have two parameters: the two single-neuron response probabilities. Even if the statistically independent model were correct, the measured joint response probabilities would only rarely be *exactly* consistent with that model. As the number of measured responses increases, the fit would be better and better, which means that the minimum value of the Kullback-Leibler distance between the measured type and the model would be smaller and smaller. The smallest the distance can be, which corresponds to an exact fit between data and model, is zero.

Therefore, the *theoretical maximal* value of the likelihood function (which occurs when there is an exact fit between the data nad the model) corresponds to the entropy term. We define the operator tmax to be this theoretical maximum.

$$\underset{\boldsymbol{\theta}}{\mathrm{tmax}}\, P(\mathbf{R};\boldsymbol{\theta}) = \exp\left\{-M\mathcal{H}(\widehat{P})\right\}$$

The ratio of the model's probability function evaluated at the population response measurements and the theoretical maximum can be interpreted as the relative probability that the measurements are consistent with the model. The more the model disagrees with the data, the smaller this relative probability. By accepting models for the data having a relative probability greater than some threshold value, we can determine *all* models reasonably consistent with the data. Because this ratio is related only to the Kullback-Leibler term in (4), this computation amounts to

$$-\log\frac{P(\mathbf{R};\boldsymbol{\theta})}{\mathrm{tmax}_{\boldsymbol{\theta}}\, P(R;\boldsymbol{\theta})} = M\mathcal{D}\left(\widehat{P}\|P(\boldsymbol{\theta})\right) \leq -\log P_0\ , \tag{5}$$

where P_0 is the threshold relative probability value. We can thus search for *all* parametric models reasonably consistent with the data by computing the Kullback-Leibler distance between the type derived from the measurements and the all models in the assumed parametric family. The region of parameter space for which the Kullback-Leibler distance is sufficiently small describes likely population models. We call this region the *model feasibility region*: all models within this region had a relative probability of producing the measurements which is greater than the chosen threshold value P_0. The smaller this probability threshold, the greater the number of models that we want to consider. Note that even the maximum likelihood model may not yield a sufficiently small Kullback-Leibler distance; in this case, *all* models in a particular parametric class cannot describe the data well. It is this aspect of determining models that seems the most promising about this data processing technique: We can accept or reject models within a given (large) class.

3 Example

Figure 1 illustrates our approach. We simulated a two-neuron population and analyzed the models that describe the response in a single bin. The simulated neural responses were heavily correlated, and two models were used to analyze the simulated data. One model assumed the neurons produced statistically independent responses. This model had two parameters, the two neuron's response probabilities. The second model allowed the responses to be correlated, giving the model three parameters: the two response probabilities and the correlation coefficient. Two datasets differing only in the number of stimulus presentations M were analyzed. For each dataset, the type was accumulated, and the Kullback-Leibler distance between the type and all models computed. The parameter values of models likely to have produced the data are enclosed within the indicated boundaries.

This simulation shows that this approach of finding all the likely models has several interesting properties. First of all, as the number of responses increases, the boundary shrinks, which restricts the models capable of producing the data. In the top row, which used an incorrect model, no boundary existed when enough responses were present because the Kullback-Leibler distance exceeded the threshold in all cases: no

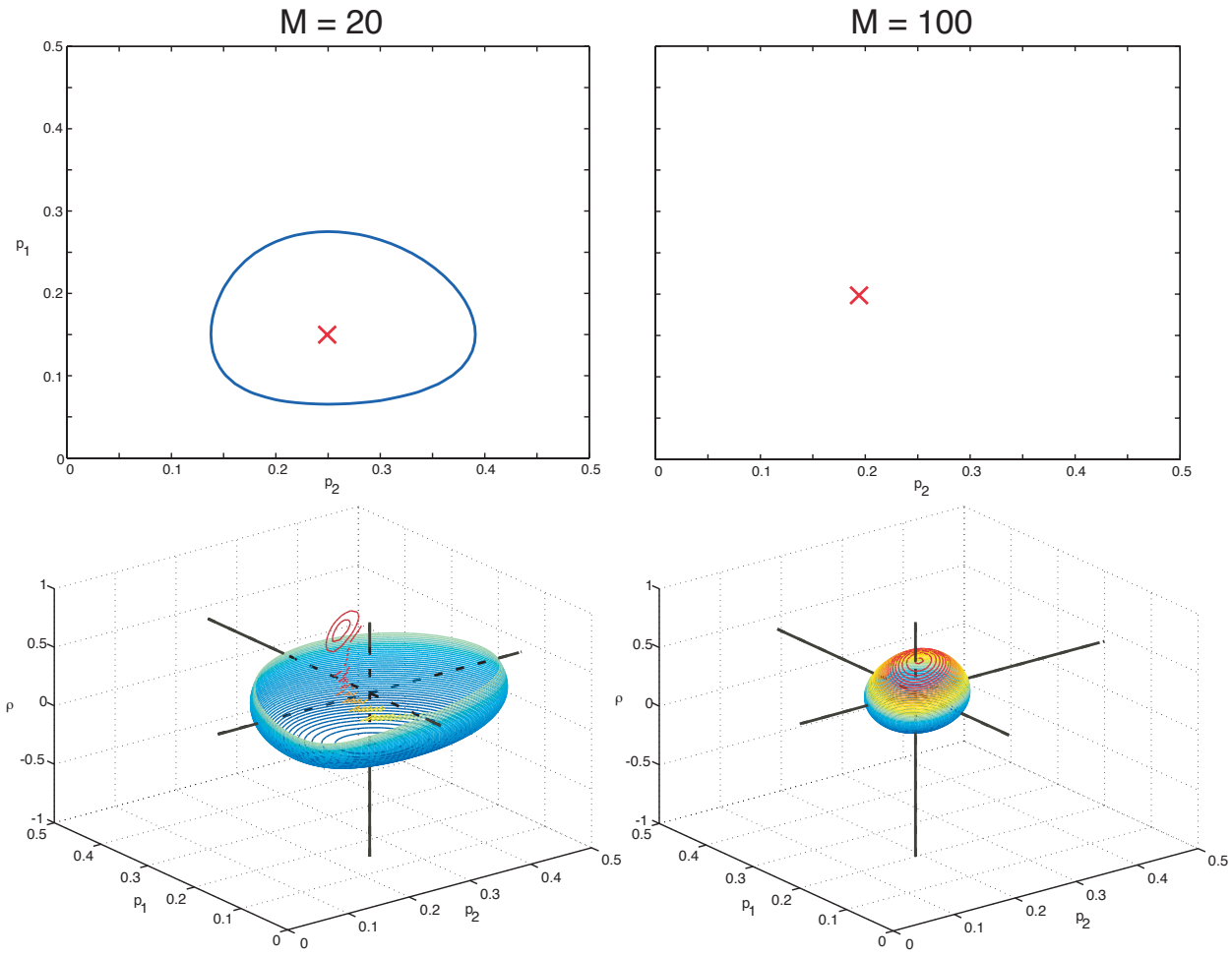

Figure 1: The plots show model feasibility regions for two models that can describe the single-bin response of a two-neuron population. In the simulation, the response probabilities were both 0.2 and the correlation coefficient was 0.5. In the left column, 20 responses were used in the analysis; in the right, one hundred. The data producing the results in the left column are a subset of the dataset used in the right. In the top row, the neural responses were considered independent, and the model parameters consisted of the two neuron's response probabilities p_1 and p_2. The cross indicates the maximum likelihood estimate and the blue contour encloses the parameter values for those models consistent with the data at a threshold probability $P_0 = 0.1$. No contour appears in the right column because no model was consistent with the larger dataset at the required probability level.

In the bottom row, the model allowed the responses to be correlated. Here, $\boldsymbol{\theta} = \{p_1, p_2, \rho\}$, where ρ is the correlation coefficient. Here, surfaces enclose parameter values for models likely to have produced the simulated data at the same relative probability level as used in the top row. As the left panel shows, the surface may not be entirely closed. In the right panel, the parameter surface shrinks because more data are available, but it does not vanish. The maximum likelihood parameter estimates are indicated by the intersecting lines. For the left panel, they were $\widehat{p}_1 = 0.15$, $\widehat{p}_2 = 0.25$, and $\widehat{\rho} = 0.40$; on the right, $\widehat{p}_1 = 0.20$, $\widehat{p}_2 = 0.19$, and $\widehat{\rho} = 0.40$.

model that assumed independent neural responses could have produced the data that the specified probability level. Despite the model misfit, note that the maximum likelihood estimate of the response probabilities in this case is quite accurate. Consequently, just computing the maximum likelihood estimate and judging parameter estimate precision with the Cramér-Rao bound would not have revealed the model's inadequacies. Secondly, note that the boundary is not ellipsoidal, which would occur if a Gaussian-like conception of modeling errors were forced on the analysis. Furthermore, the maximum likelihood parameter estimate, which always corresponds to the most likely model, is not necessarily located in the center of the boundary. Thus, our approach allows an exact specification of the range of models that can describe the data reasonably well rather than relying on confidence intervals for individual parameters.

4 Conclusions

We have not yet analyzed a large population response, but our approach looks to be promising. The approach consists of explicitly finding parametric probabilistic descriptions that are consistent with the data to a specified level of probability. No approximations are used in our approach. The calculations amount to finding a Kullback-Leibler distance between the measurements and all possible models, then focusing on those models for which the distance is sufficiently small (equation 5). The best model fitting the data presumably has the smallest minimum value for the Kullback-Leibler distance. We do not need to find the variance of our parameter estimates; the parameter feasibility surface does that for us.

References

[1] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[2] D.H. Johnson, C.M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of neural coding. *J. Comp. Neuroscience*, 10:47–69, 2001.

[3] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[4] M.J. Wienberger, J.J. Rissanen, and M. Feder. A universal finite memory source. *IEEE Trans. Info. Th.*, 41:643–652, 1995.