Methods for studying the neural code in high dimensions

Alexandro Diaz Ramirez

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

©2012

Alexandro D. Ramirez

All Rights Reserved

ABSTRACT

Methods for studying the neural code in high dimensions

Alexandro D. Ramirez

Over the last two decades technological developments in multi-electrode arrays and fluorescence microscopy have made it possible to simultaneously record from hundreds to thousands of neurons. Developing methods for analyzing these data in order to learn how networks of neurons respond to external stimuli and process information is an outstanding challenge for neuroscience. In this dissertation, I address the challenge of developing and testing models that are both flexible and computationally tractable when used with high dimensional data. In chapter 2 I will discuss an approximation to the generalized linear model (GLM) log-likelihood that I developed in collaboration with my thesis advisor. This approximation is designed to ease the computational burden of evaluating GLMs. I will show that our method reduces the computational cost of evaluating the GLM log-likelihood by a factor proportional to the number of parameters in the model times the number of observations. Therefore it is most beneficial in typical neuroscience applications where the number of parameters is large. I then detail a variety of applications where our method can be of use, including Maximum Likelihood estimation of GLM parameters, marginal likelihood calculations for model selection and Markov chain Monte Carlo methods for sampling from posterior parameter distributions. I go on to show that our model does not necessarily sacrifice accuracy for speed. Using both analytic calculations and multi-unit, primate retinal responses, I show that parameter estimates and predictions using our model can have the same accuracy as that of generalized linear models.

In chapter 3 I study the neural decoding problem of predicting stimuli from neuronal responses. The focus is on reconstructing zebra finch song spectrograms, which are high-dimensional, by combining the spike trains of zebra finch auditory midbrain neurons with information about the correlations present in all zebra finch

song. I use a GLM to model neuronal responses and a series of prior distributions, each carrying different amounts of statistical information about zebra finch song. For song reconstruction I make use of recent connections made between the applied mathematics literature on solving linear systems of equations involving matrices with special structure and neural decoding. This allowed me to calculate *maximum a posteriori* (MAP) estimates of song spectrograms in a time that only grows linearly, and is therefore quite tractable, with the number of time-bins in the song spectrogram. This speed was beneficial for answering questions which required the reconstruction of a variety of song spectrograms each corresponding to different priors made on the distribution of zebra finch song. My collaborators and I found that spike trains from a population of MLd neurons combined with an uncorrelated Gaussian prior can estimate the amplitude envelope of song spectrograms. The same set of responses can be combined with Gaussian priors that have correlations matched to those found across multiple zebra finch songs to yield song spectrograms similar to those presented to the animal. The fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations than temporal correlations. However the best reconstructions combine MLd responses with both spectral and temporal correlations.

Contents

1 Introduction **1**

2 Fast inference in Generalized Linear Models via expected log-likelihoods **5**

2.1 Introduction 6

2.2 Results 7

2.2.1 Generalized linear models 8

2.2.2 The computational advantage of using ELs over log-likelihoods in a GLM 11

2.2.3 Computational efficiency of maximum expected log-likelihood estimation for the LNP and Gaussian model 14

2.2.4 Analytic comparison of the accuracy of EL estimators with the accuracy of maximum-likelihood estimators 20

2.2.5 Fast methods for refining maximum expected log-likelhood estimators to obtain MAP accuracy 23

2.2.6 Marginal likelihood calculations 27

2.3 Discussion 32

2.4 Methods 36

2.4.1 Computing the mean-squared error for the MPELE and MAP 36

2.4.2 Computing the MPELE for an inhomogeneous Poisson process 41

2.4.3 Calculating $\hat{R}$ for the LNP model 43

2.4.4 Calculation of Bayesian credible regions 45
2.4.5 Simulated and real neuronal data applications 46

3 Incorporating Naturalistic Correlation Structure Improves Spectrogram Reconstruction From Neuronal Activity in the Songbird Auditory Midbrain 55

3.1 Introduction 56
3.2 Methods 58
3.2.1 Electrophysiology 58
3.2.2 Auditory Stimuli 59
3.2.3 Bayesian Decoding 60
3.2.4 Encoding Model 60
3.2.5 Birdsong Priors 62
3.2.6 Song Reconstructions 70
3.2.7 Simulating STRFs 72
3.2.8 Optimal Linear Estimator 73
3.2.9 Measuring Reconstruction Accuracy 74
3.3 Results 76
3.3.1 Two-alternative forced choice discrimination 76
3.3.2 Decoding Song Spectrograms Using Single Neurons 77
3.3.3 Population Decoding of Song Spectrograms 79
3.3.4 Effect of Prior Information on Song Estimation 80
3.3.5 Reconstruction Dependence On STRF Frequency Tuning . . . 85
3.4 Discussion 87
3.5 Acknowledgements 102
3.6 Appendix 102
3.6.1 Generating Samples from $p(\mathbf{x}, q|\mathbf{r}, \theta)$ 102

4 Concluding remarks 107

List of Figures

2.1 Numerical examples showing the accuracy of approximating the projection $x^T\theta$ as a normaly distributed random variable. We plot the output of $\mathbf{E}\Big[G\Big]$ (equation 2.2.18 in the text) with (vertical axis) and without (horizontal axis) approximating $p(q)$ as normal. G is taken to be exponential corresponding to Poisson responses. x has a binary white noise distribution (A) or Weilbull distribution (B). Without the normal approximation, $\mathbf{E}\Big[G\Big]$ was found by Monte-Carlo sampling using 12000 simulated samples of x for each projection. x is a 600 dimensional vector with mean 0.36 in A or drawn from a Weilbull distribution (B) with scale and shape parameters 0.15, 0.5. θ was restricted to have a Gaussian profile, in order to approximate filter shapes often found in applications, with mean 0, and with random variance and random amplitude. 48

2.2 The MELE comes with a minimal cost in accuracy (and sometimes improved accuracy) compared to maximum-likelihood estimators (MLE). A.) (left) The mean squared error (MSE) for the MELE (solid lines) and the MLE (dotted line) is shown as a function of the ratio of the number of parameters to number of samples. We plot results for the asymptotic case where the number of samples and dimensions goes to infinity but the ratio remains finite. Different colors denote different values for the true filter norm. The MLE mean squared error is larger than the MELE MSE when the number of samples is close to the number of dimensions. B.) MSE for both estimators when L2 regularization is added. MSE is similar for both estimators for a large range of ridge parameters and values of $\frac{p}{N}$. A.) (right) For each value of $\frac{p}{N}$, separate ridge parameters are chosen for the MPELE and MAP estimators to minimize their respective mean squared errors. Even at the optimal choice of regularization, the difference in performance between the two estimators remains small for a wide range of values of SNR and $\frac{p}{N}$. In figure 2.3, we show that when the MELE and MPELE is used on real data small differences in performance can be eliminated using fast gradient based optimization methods. 49

2.3 The MELE is a good initializer for finding maximum a posteriori (MAP) estimators using fast gradient based optimization methods. A.) The spatial-temporal receptive field of a randomly chosen Retinal Ganglion cell (RGC) responding to binary white-noise stimuli and fit with a linear-nonlinear Poisson (LNP) model with exponential non-linearity is shown (labeled MLE). The receptive field of the same cell fit using theMELE under an LNP model is also plotted (see label marg). The goodness-of-fit of the MLE (measured in terms of information rate, see methods) is slightly higher than that of the MELE (12 versus 11 bits/s). However this difference disappears after a couple pre-conditioned conjugate gradient (PCG) iterations using the true likelihood initialized at the MELE (see label +2PCG). Note that we are only showing spatial and temporal slices of the 9x9x10 dimensional receptive field for illustrative purposes. B.) Similar results hold when the same cell responds to 1/f correlated Gaussian noise stimuli (for correlated Gaussian responses, the MLE and MELE are both fit using an L2 regularizer). However, 9 PCG iterations instead of 2 were used to find an estimator with a goodness-of-fit equal to the MLE. C.) These conclusions hold for many cells. (top) Boxplot of information rates across a population of 91 cells, responding to binary white-noise stimuli, each fit independently using the MLE, MELE or a couple PCG iterations using the true likelihood and initialized at the MELE. (bottom) Information rates for the same population, responding to 1/f Gaussian noise, fit independently using an L2 regularized MLE, MPELE or 9 PCG iterations using the true regularized likelihood initialized at the MPELE. 50

2.4 The MELE of an LNP model combined with a coordinate descent algorithm quickly determines post-spike L1 penalized coupling parameters. A.) Example stimulus, self-history, and coupling filters for two different RGC cells (top and bottom rows). MAP parameters are found using coordinate descent to optimize the true GLM log-likelihood with L1 penalized coupling filters (labeled MAP). Fast estimates of the self-history and coupling filters are found by running the same coordinate descent algorithm with the stimulus filter (SF) fixed, up to a gain, to the MELE of an LNP model(labeled 'w fixed SF'; see text for full procedure used in determining history parameters). By fixing the stimulus current using the MELE, the correct non-zero coupling parameters are found 16 times faster than the MLE. B.)Boxplots comparing the cross-validated information rate of 100 different RGC cells show that the small differences in parameters between the two estimators do not affect goodness-of-fit. 51

2.5 The EL can be used to quickly find Type II Maximum Likelihood filters. Thirty simulated neural responses were drawn from a linear-nonlinear Poisson (LNP) model with stimuli drawn from an independent white-noise Gaussian. The true filter (shown in black in B) has 250 parameters. A.) Optimal hyper-parameters which maximize the evidence function using the EL (top left column, vertical axis) are similar to those which maximize the GLM evidence function (top left column, horizontal axis). After a single iteration of the fixed point algorithm used to maximize the GLM evidence (see text, section Marginal likelihood calculations), the two sets of hyper-parameters match to sufficient accuracy (bottom left column). B.) Median filter estimates (blue), and absolute median deviation (light blue), using either set of hyper-parameters match for a wide range of values of $\frac{p}{N}$. Below the filter estimates we plot the mean squared error (MSE) of both filters which also matches for a wide range of values of $\frac{p}{N}$. 52

2.6 Statistics from the posterior parameter distribution approximated use the EL can be quickly computed and are in close agreement with those from the true posterior. 4000 responses were simulated from an 100 dimensional linear-nonlinear Poisson model using independent white-noise Gaussian stimuli ($\frac{p}{N} = 0.025$). 10^6 Markov Chain Monte Carlo samples, computed using Hybrid Monte Carlo, were then drawn from the posterior assuming a flat prior and using either a Poisson likelihood or the EL approximation to the Poisson. Both columns show the median vector along with 95% credible region (defined in the text) of the approximate (blue) and true (red) posterior distribution. In the middle and right column we have zoomed in around different elements for visual clarity. Statistics from both distributions are in close agreement for many elements, however some directions show discrepancies. It took 34 seconds to sample using the EL and 83 seconds using the true likelihood. Replacing the likelihood with the profile likelihood of the EL results yielded similar results to the approximate distribution (green). The Laplace approximation also provided a good approximation to the posterior (black). 53

2.7 Finite sample approximations to the limit $p, N \to \infty$, $\frac{p}{N} \to \rho$. We plot the true mean-squared error (MSE) (colored lines) as a function of finite sample size for different values of $\frac{p}{N}$ for both the MLE (left column) and MELE (right column). Black dashed lines show MSE when $p, N \to \infty$, $\frac{p}{N} \to \rho$. The quality of the approximation does not seem to depend on $\frac{p}{N}$ for the MELE and is within 1% accuracy after about 100 samples. For the MLE, this approximation depends on $\frac{p}{N}$. However, while the quality of the approximation depends on the estimator used, for values often used in real data sets ($\frac{p}{N} \sim 0.01$, $N \sim 10000$) this is an excellent approximation. 54

3.1 Encoding model and parameters. In the encoding model, each neuron is modeled with a spectrogram filter (STRF) and post-spike filter that captures stimulus-independent spiking properties. The stimulus is temporally convolved and frequency multiplied with the STRF and then exponentiated to obtain the instantaneous firing rate used for generating spikes. The spikes are convolved with the post-spike filter and used in the model as a feedback signal that affects future spike generation. 92

3.2 Least informative prior: uncorrelated Gaussian distribution. **A**, An example spectrogram with power spectral density indicated by color. **B**, Normalized histogram of power spectral density values across all songs and spectrogram bins (blue dots). The mean and variance of these power values is used to construct a Gaussian prior (black line) that confines estimated values of power spectral density to regions found in actual song spectrograms. **C**, To visualize the information provided by the prior, a sample spectrogram drawn from this prior is plotted. This prior does not provide information on spectrotemporal correlations in spectrograms, as demonstrated by this sample. 93

3.3 Spectrotemporaly Correlated Gaussian prior. **A**, The spectrotemporal covariance matrix is modeled as separable in frequency and time. The frequency component is the spectral covariance matrix (upper panel). The temporal component is fully described by the temporal autocorrelation function in song spectrogram power density (bottom panel, red line). The prior uses an approximation to this function using an Autoregressive model (blue line). **B**, The full spectrotemporal covariance matrix is a concatenation of several spectral covariance matrices, like those shown in the upper panel, each corresponding to the covariance at a different temporal lag. The bottom panel labeled 'Approximate Covariance' plots the separable covariance matrix and the middle panel labeled 'True Covariance' plots the non-separable covariance matrix. **C**, (Top) An example spectrogram used in determining song statistics for constructing the Gaussian prior. (Bottom) Sample spectrogram drawn from the Correlated Gaussian prior. **D**, Two-dimensional power spectra, also called the modulation power spectra (MPS), for song spectrograms (top) and for the prior (bottom); the prior does a good job of capturing information about spectrotemporal modulations except at joint regions of high spectral modulations and temporal modulations near zero. 94

3.4 Most informative prior: hierarchical model with a two-state hidden variable that infers whether the spectrogram is in a vocalization or silent period. These periods have different statistical properties not captured by a single Gaussian prior. The state variable determines which spectral covariance matrix and mean the prior uses to inform reconstructions. **A**, Example spectrogram overlaid with vocalization and silent states (black line). **B**, (Top left) Spectral covariance matrix used during vocal periods. (Top right) Spectral covariance matrix used for silent periods. (Bottom) Prior information of transition rates between silent and vocal periods determined from song spectrograms. **C**, Sample spectrogram drawn from this prior; the sharp transitions in song statistics during vocal and silent periods better matches song spectrograms. 95

3.5 Conspecific song discrimination based on likelihood of spike-trains from multiple neurons. **A**, Spike trains from multiple neurons in response to presentation of song segment 1. Under a Two-alternative forced choice (2AFC) test, song discrimination is performed by choosing the song which leads to a greater likelihood of observing the given spikes. spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches maximal value. Neurons with the same BF are plotted on the same row. **B**, 2AFC results as a function of response duration and the number of neurons used for discrimination. 2AFC was performed multiple times for each possible pairing of the twenty songs in the data set. Each panel shows the frequency of correct trials across all possible song pairings. Above each panel is reported the average of the histogram. On average, neurons performed at chance level when stimulus segments were only 3 ms in duration. Near perfect song discrimination can be achieved using 189 responses and response durations at least around 30 ms, or 104 neurons and durations of about 100 ms. 96

3.6 Single cell decoding of song spectrogram. **A**, (Top) Spectrogram of birdsong that elicited the two spikes shown immediately below. spikes are plotted at the frequency at which this neuron's receptive field reaches maximal value. **B**, (Top, left) The most probable spectrogram from the posterior distribution (MAP Estimate) given the two spikes shown in A and using an uncorrelated prior. When a single spike occurs the MAP is determined by the neuron's spectrotemporal receptive field (STRF, shown to the right). In the absence of spikes, the MAP is determined by the prior mean. (Bottom, left) MAP estimate using the correlated Gaussian prior; when a spike occurs the MAP is determined by the neuron's STRF multiplied by the prior covariance matrix (shown to the right). Immediately after a spike the MAP infers spectrogram values using prior knowledge of stimulus correlations. . 97

3.7 Population decoding of song spectrogram with varying degrees of prior information of song statistics. **A**, (Top) Spectrogram of birdsong played to 189 different neurons leading to the spike responses shown immediately below. spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches its maximal value. Neurons with the same BF are plotted on the same row. MAP estimate given the responses in A, using an uncorrelated prior **B**, a prior with temporal correlations and no spectral correlations **C**, a prior with spectral correlations and no temporal correlations **D**, and a prior with spectral and temporal correlations **E**. Combining the spike train with spectral information is more important for reconstructing the original spectrogram than combining the spike train with temporal information. However, combining spikes with joint spectrotemporal information leads to the best reconstructions. 98

3.8 Decoding performance given different amounts of prior information and numbers of neurons. **A**, (upper row) Spectrogram reconstructions for an example song (Figure 3.7A) using a Gaussian prior with spectrotemporal correlations and using varying numbers of neuronal responses (plotted in the lower row). Above each reconstruction is the signal-to-noise (SNR) used to measure similarity between the reconstructed song and the song presented to the bird. **B**, Solid lines show the signal-to-noise (SNR) ratio averaged across all decoded songs while dashed lines show one standard error. The prior used for decoding is denoted by color. Spectral prior information leads to faster growth in the SNR than temporal information. For reference, the magenta line shows the growth in SNR for the commonly used optimal linear estimator (OLE). The OLE has access to both spectral and temporal correlations. **C**, Coherence between spectrograms and reconstructions under the four different priors. The horizontal axis reports temporal modulations and the vertical axis reports spectral modulations. All plots display the highest coherence at low spectral and temporal modulations. The primary effect of adding spectrotemporal prior information is to improve reconstructions at lower spectral and temporal modulations. 99

3.9 Single neuron and population decoding using a hierarchical prior. **A**, Song spectrogram along with a single cell's response to this song (left column) and the response of this cell plus forty nine other cells with nearby characteristic frequencies (right column). **B**, MAP estimates using a single, correlated Gaussian prior (top row) are compared with estimates using the posterior mean and the hierarchical prior (bottom row); in both the single neuron and population decoding case, the estimate using a hierarchical prior looks similar to the MAP with a Gaussian prior. **C**, The expected value for vocalization state given responses; single cell responses do not yield enough information to accurately infer the spectrogram's vocalization state, however as the number of neurons used for inference increases the vocalization state becomes more pronounced. 100

3.10 Spectral blur of STRFs causes a small loss of information for reconstructions. **A**, (upper, left panel) Example STRF and localized point STRF (upper, right panel) with equivalent peak frequency. (lower, left panel) Frequency vectors at latency where STRF obtains maximal value for the population of neurons used in this study. The equivalent plot for point STRFs (lower, right panel). Point STRF peak locations were randomly drawn from a distribution constructed using the peak locations of real STRFs. **B**, (first two rows) Song spectrogram and evoked responses of 189 real neurons. (middle rows) Reconstructed song spectrogram given simulated responses using a point STRF model. Simulated responses are shown immediately below the reconstruction. (bottom two rows) Reconstructed song spectrogram given simulated responses using full STRFs. Responses are shown immediately below the reconstruction. Reconstructions with full STRFs show slightly different spectral details but otherwise look *very* similar to reconstructions using point STRFs. **C**, SNR growth (plus and minus one standard error) as a function of the number of neurons used in decoding for point STRFs and full STRFs; on average the point STRFs have higher SNR than full STRFs. 101

Acknowledgments

I have been truly privileged to learn from so many smart, generous people. I would like to first acknowledge my advisor Liam Paninksi. He has been incredibly generous with his time and knowledge and is always willing to listen to new ideas. He has an uncanny ability of integrating seemingly disparate ideas and concepts from both science and mathematics. I have learned a great deal from our collaboration and discussions.

I have had several teachers who, by lending me countless hours of their time, have had a tremendous impact on my understanding of science; Larry Abbott, Bill Bialek, Stefano Fusi, John Krakauer, Ken Miller, Chris Wiggins and Ernst-Ludwig Florin at the University of Texas Austin. I am grateful for all of them.

My scientific interests were also greatly influenced by my time as a student in the Neural Systems and Behavior course at the Marine Biological Laboratory (MBL). A special thanks to Michael H. Dickenson and Sarah W. Bottjer for creating a wonderful course and exposing me to the exciting and beautiful ideas in neuroethology. Thank you to the MBL for generously providing funding during my stay.

Thank you to Wilson Truccollo for generously agreeing to serve on my thesis defense.

Thank you to my collaborator and thesis committee member Sarah Woolley. Thank you for opening up your lab to me, serving on my committee, lending me your time and fostering a fabulous environment for collaboration. Thank you to my other collaborators; Joe Schumacher, David Schneider and Yashar Ahmadian. A special

thanks to Joe for opening my ears to the joys of Sonic Youth and agreeing to spend many evenings together either learning Bayesian methods of analysis or playing music.

I would like to acknowledge many of the special friendships I have made in graduate school;"taco crew" members: Evan Schaffer, Ben Matthews, and Dave Malito; "pastry/pancake club" members: Annegret Falkner, Daniela Hernandez, David Schneider, Dave Malito, Joe Schumacher, Ana Calabrese and special guest stars; Kelley Remole, Dara Sosulski, Megan Corty, Barney Bryson, Heather McKellar, Amitabha Gupta, Sean Escola, Michael Vidne, Dan Rubin, Eftychios Pnevmatikakis, Carl Smith, Yuriy Mishchenko, Maxim Nikitchenko, David Pfau, Ari Pakman, Tim Machado, Jake Hoffman, Andrew Mugler, Johnathan Bronson, Alexandra Walczak, Stephanie Palmer, Taro Toyoizumi, Xaq Pitkow, Tim Vogels, David Sussillo, Kanaka Rajan, Baktash Babadi, Greg Wayne, Saul Kato, Ann Kennedy, Brian DePasquale, Omri Barak, Mattia Rigotti, Yashar Ahmadian, Pablo Jercog, Avi Ziskind, Srdjan Ostojic, Tatjana Tchumatchenko, and George Zhao. We've had oh so many wonderful conversations.

Thank you to directors past and present of the doctoral program in neurobiology and behavior, Darcy Kelley, John Koester, Ken Miller, and Carol Mason. Thank you to Cecil Oberbeck, Alla Kerzhner, Liz Ryan, Dood Kalicharan, Richard McFarland, and Jessica Keyer for making things run smoothly.

I was generously supported by a National Science Foundation predoctoral fellowship during my early years here.

My family for being a continual source of love, care and support. Thank you to the Barkan family for your hospitality and some really great meals.

Finally, thank you to Charlotte for all that we share.

–To Nellie Ramirez, Carlos G. Ramirez, Soccorro Medina, and Heliodoro S. Medina.

Chapter 1

Introduction

The motivation for this work began, roughly, a hundred years ago when physiologists were first able to detect electrical signals from individual cells. A series of experiments started by Keith Lucas and completed by Edgar Adrian provided convincing evidence that sensory neurons communicate with the brain in a 'all-or-nothing' manner, much like what had been found before from isolated motor nerve fibers (Adrian & Zotterman 1926, Adrian 1964). The implication of this result was quickly realized; it is only the timing of these 'all-or-nothing' events, now known to be action potentials, that can provide information to the brain. Since then neuroscientists have attempted to decipher the function of the nervous system by analyzing sequences of these impulses, referred to as spike trains, and their relationship to external stimuli. The general problem (often called the neural coding problem) is to understand what stimulus information is encoded by neurons and, even more broadly, understanding how this information can be used to make behavioral and perceptual decisions (Perkel & Bullock 1969, Markowitsch 1988, Rieke, et al. 1997, Eggermont 2001).

There is a long-standing history in neuroscience of addressing these questions using mathematical models. Mathematical formulations of the neural coding problem have been well described in the literature (for excellent introductions see (Rieke et al. 1997, Abbott & Dayan 2001). Therefore, I will not go into much detail here and

simply outline the framework to be used throughout this thesis. Neuronal responses (impulse times) will be denoted by the vector $\mathbf{r}$ and external stimuli denoted by $\mathbf{x}$. The question of understanding what information is encoded by neurons is then transformed into one of predicting $\mathbf{r}$ when neurons are excited by $\mathbf{x}$. Since responses to stimuli are known to be variable, it is possible to observe many different $\mathbf{r}$'s to a given $\mathbf{x}$, the problem of interest is actually in understanding and characterizing this variability by estimating the probability of observing $\mathbf{r}$ to a given $\mathbf{x}$, $p(\mathbf{r}|\mathbf{x})$.

Much of the interest in modern neuroscience is in understanding how networks of neurons act in concert to define the function of particular brain regions. In a statistical context this is equivalent to understanding $p(\mathbf{r}|\mathbf{x})$ when $\mathbf{r}$ contains the responses of large groups of neurons. This interest has been stimulated in large part by technological developments in fluorescence microscopy (Cossart, et al. 2003, Ohki, et al. 2005, Nikolenko, et al. 2008), and multi-electrode arrays (Nicolelis, et al. 2003, Litke, et al. 2004) that make it possible to simultaneously record from large neuronal populations. Developing methods for analyzing these data remains a key challenge in neuroscience (Brown, et al. 2004, Paninski, et al. 2009, Stevenson & Kording 2011).

In large part, the difficulty is due to the fact that $\mathbf{r}$ and $\mathbf{x}$ are high-dimensional. For example, these technologies allow $\mathbf{r}$ to contain temporal responses from thousands, and perhaps someday hundreds of thousands, of neurons (Alivisatos, et al. 2012) and $\mathbf{x}$ could be chosen to imitate natural stimuli such as a movie or a spectrogram of natural sounds (Woolley, et al. 2005). Furthermore, when $\mathbf{x}$ is a movie there is evidence that the resolution and thus the dimension of $\mathbf{x}$ matters. Field, et al. (2010), showed that high resolution, and thus high-dimensional, movies can be an effective way to probe network connectivity even between cell pairs where one cell is not being directly recorded. Classical tools for analyzing responses, such as histogram-based methods, require more samples than can possibly be recorded during the time of a typical experiment to be of use (Brown et al. 2004). The problem is made even more difficult for 'online' applications' where one must determine $p(\mathbf{r}|\mathbf{x})$, and often

$p(\mathbf{x}|\mathbf{r})$, in real-time. For example in the field of neural prosthetics $\mathbf{x}$ could correspond to the desired position of the subject's arm or hand that must be determined from responses $\mathbf{r}$ in motor or parietal cortex (Bansal, et al. 2012, Malik, et al. 2011, Wu, et al. 2009, Yu, et al. 2007, Donoghue 2002a). In this case, the method of analysis must be able to process high-dimensional data while being computationally efficient.

Several models for $p(\mathbf{r}|\mathbf{x})$ have been proposed in the literature for responses recorded from a variety of brain areas (Brillinger 1988, Chornoboy, et al. 1988, Brillinger 1992, Utikal 1997, Martignon, et al. 2000, Iyengar 2001, Schnitzer & Meister 2003, Paninski, et al. 2004a, Truccolo, et al. 2005, Nykamp 2005, Schneidman, et al. 2006, Pillow, et al. 2008). In this thesis, I concentrate on the use of parametric generalized linear models (GLM) of $p(\mathbf{r}|\mathbf{x})$ (Brillinger 1988, Chornoboy et al. 1988, Plesser & Gerstner 2000, Paninski et al. 2004a, Truccolo et al. 2005, Rigat, et al. 2006, Pillow et al. 2008, Vidne, et al. 2011a). These models are a good starting point for dealing with high-dimensional $\mathbf{r}$ and $\mathbf{x}$ since they approximate $p(\mathbf{r}|\mathbf{x})$ with $p(\mathbf{r}|\mathbf{x}, \theta)$, a function parameterized by θ that has a smaller dimensionality than the dimensionality of the original problem (the size of $\mathbf{r}$ times the size of $\mathbf{x}$). Additionally, efficient statistical methods for inferring θ from data have been worked out (McCullagh & Nelder 1989, Lehmann & Casella 1998, Paninski 2004). Furthermore, this model fairly accurately describes neural responses in a variety of brain areas including the retina (Vidne et al. 2011a, Pillow et al. 2008), motor cortex (Saleh, et al. 2012a), hippocampus (Truccolo et al. 2005) and auditory midbrain (Calabrese, et al. 2011).

Unfortunately, when the number of GLM parameters is large, inferring their values from data can be computationally difficult. In general, the number of parameters grows with the stimulus dimension and therefore will be large for high-dimensional stimuli. Stevenson & Kording (2011) give evidence that the GLM's prediction power grows when covariates to account for the, pairwise, interactions of neighboring cells on the observed cell are added to the model, at least for neurons in the primary visual cortex and primary/pre-motor cortex. If we define the stimulus as including these

interaction terms then recording from more neurons increases the dimensionality of the stimulus. Since much research with the GLM includes these interaction terms (Stevenson & Kording 2011, Vidne et al. 2011a, Pillow et al. 2008), analyzing data with the GLM may become too sluggish for practical use in the future, as the number of simultaneously observed cells grows and external stimuli increase in size. This will certainly be true for 'online' applications of the model.

In this thesis I study neural coding problems with a particular interest in developing and applying statistical models that have good predictive power and a tractable computational cost. That is, models that one can evaluate and fit to data in a reasonable amount of computation time even if stimuli and responses are high-dimensional. I describe a method, developed by my advisor and I, for predicting neuronal responses to a given stimulus and argue that this method fits the above criteria. I then turn to decoding and apply recently developed methods for fast inference of the mode of $p(\mathbf{x}|\mathbf{r})$ to the auditory system of zebra finch. I show how exploiting the speed of these methods allows one to ask interesting questions related to the interplay of prior information and the quality of reconstructed zebra finch songs from neural data.

In chapter 2 I will introduce the GLM framework that will be used for the rest of the thesis. I then discuss an approximation to the GLM log-likelihood that we have generalized from a technique known as the 'Expected' likelihood (Park & Pillow 2011). I will show that this approximation reduces the computational cost of evaluating the GLM log-likelihood without a large sacrifice in model accuracy. In chapter 3 I focus on decoding stimuli given neural responses from the zebra finch auditory midbrain. I will show how decoded song spectrograms depend on prior knowledge of the second-order statistics of zebra finch song. In particular, I will show that the feature detectors in the auditory midbrain carry enough information to reconstruct coarse qualities of spectrograms from individual conspecific songs if they are combined with knowledge of the spectro-temporal correlations averaged across all zebra finch songs.

Chapter 2

Fast inference in Generalized Linear Models via expected log-likelihoods

Abstract

Recent work has shown how a quickly computable estimator of the parameters in an inhomogeneous Poisson model, known as the spike-triggered average (STA) in the neuroscience literature, can be derived from the 'expected log-likelihood' (EL) of that model. We generalize these results to the broad class of models known as generalized linear models (GLM) and clearly elucidate the computational advantage of the EL approximation. We show how Maximum Likelihood estimation of the GLM parameters can be sped up by orders of magnitude under some simple conditions on the priors and likelihoods involved. We perform a risk analysis, using both analytic and numerical methods, and show that maximum EL estimators come with little cost in accuracy (and in some cases even improved accuracy) compared to standard Maximum Likelihood estimates. Finally we find that these methods can significantly decrease the computation time of marginal likelihood calculations for model selection

and of Markov chain Monte Carlo methods for sampling from the posterior parameter distribution. We use multi-unit, primate retinal responses to validate our findings.

2.1 Introduction

Continual progress of multi-electrode (Field et al. 2010, Brown et al. 2004) and imaging techniques (Lütcke, et al. 2010, Bloodgood & Sabatini 2007) is resulting in the activity of more cells being recorded simultaneously than was previously possible. These developments have the promise of offering new insights into classic neuroscience questions such as what is the code used by the nervous system to transmit information? To do so requires the development of models and analyses to characterize network activity patterns using these data. However the computational time required for learning these patterns grows, as the number of cells being recorded, and hence the number of potential patterns, increases (Stevenson & Kording 2011). This cost limits the types of models and analyses that can be performed in experiments which require real-time calculations of model parameters, e.g. Brain-Machine Interfaces (BMI) (Santhanam, et al. 2006a, Donoghue 2002b) and adaptive closed-loop experiments (Lewi, et al. 2009a).

We address this issue by presenting novel methods for fitting the broad class of models known as generalized linear models (GLM) which are orders of magnitude faster than standard techniques. GLMs are useful for quantifying the relationship between neural responses and external stimuli or behavior (Brillinger 1988, Paninski 2004, Truccolo et al. 2005). This model's flexibility has been welll described in the statistics literature (McCullagh & Nelder 1989) and underlies its successful application to a variety of brain areas including the retina (Vidne, et al. 2011b, Pillow et al. 2008), motor cortex (Saleh, et al. 2012b), sensorimotor cortex (Truccolo et al. 2005) and auditory midbrain (Calabrese et al. 2011). Accurate estimation of model parameters is critical for predicting responses to novel stimuli and for neu-

ronal decoding, but determining model parameters for large data sets often becomes computationally intractable.

We build upon recent results (Park & Pillow 2011, Paninski 2004) which derive a quickly computable estimator of model parameters, popular in the neuroscience literature and referred to as the spike-triggered average (STA), from the 'expected log-likelihood' (EL) of an inhomogeneous Poisson model. First, we generalize this result to a broad family of GLMs and detail the computational advantage of estimating parameters using the EL of a GLM. Even though parameter estimates can be obtained more quickly than standard methods using the EL they may come with a large cost in accuracy. We assess the error associated with this estimator using linear Gaussian models and compare it with the error of standard *maximum a posteriori* (MAP) estimators. We find that the two estimators have comparable accuracy depending on the ratio of model parameters to sample size. We then present a novel method for using these results to fit GLMs whose covariates include spike-history and interneuronal effects. Finally we discuss how to use the EL to increase the computational efficiency of other statistical inference methods, such as Markov Chain Monte Carlo sampling and Type II maximum likelihood.

2.2 Results

The results are organized as follows. We first introduce the EL approximation to the GLM log-likelihood. We discuss the computational savings gained by computing this approximate likelihood compared to computing the GLM log-likelihood. We then discuss the computational advantage of utilizing ELs in various applications including parameter estimation, marginal likelihood calculations, and sampling from the posterior parameter distribution. For each application, we then compare results using both likelihoods. In the case of parameter estimation we validate results using real data and we present analytic calculations showing that estimators that maximize

the EL can come with a small cost in accuracy compared to standard estimators.

2.2.1 Generalized linear models

Consider a vector of observations, $r = (r_1, ..., r_N)$, resulting from N presentations of a p dimensional stimulus vector, x_i (for $i = 1, ..., N$). Under a GLM, with model parameters θ, the likelihood for r is chosen from the exponential family of distributions. From the definition of the exponential family the log-likelihood for r, modeling the observations as conditionally independent given x (an assumption we will later relax), can be written as

$$L(\theta) = \sum_{n=1}^{N} \frac{1}{c(\phi)} \Big(a(x_n^T\theta) r_n - G\Big(a(x_n^T\theta)\Big)\Big) + \text{const}(\theta), \tag{2.2.1}$$

for some functions $a(), G(), c()$ and where we have written terms that are constant with respect to θ as const(θ)(McCullagh & Nelder 1989). For the rest of the paper we will consider the scale factor $c(\phi)$ to be known and for convenience we will set it to one. This family encompasses many well known distributions, we will give some examples later in the text, such as the Gaussian, Gamma and Poisson distribution (McCullagh & Nelder 1989). The GLM also specifies the conditional mean of r given the stimulus and parameters, which we denote by $\mathbf{E}\Big[r_n | x_n, \theta\Big]$, as being related to a linear mapping of the stimulus vector, $x^T\theta$, through what is known as the link function, $g()$ (McCullagh & Nelder 1989)

$$g\Big(\mathbf{E}\Big[r_n | x_n, \theta\Big]\Big) = x_n^T\theta. \tag{2.2.2}$$

For each choice of the likelihood there exists a particular link function, known as the canonical link, which endows the model with convenient mathematical relationships. The canonical link is defined as the g such that the following is true $a(x_n^T\theta) = x_n^T\theta$.

For the rest of text we assume that g is chosen to be the canonical link function.

With this choice, it follows from equation 2.2.1 that the GLM log-likelihood is the sum of a linear and non-linear function of θ

$$L(\theta) = \sum_{n=1}^{N} (x_n^T\theta) r_n - G(x_n^T\theta) + \text{const}(\theta). \tag{2.2.3}$$

As we show later in the text, this property will be useful for motivating the definition of the EL. To demonstrate this theory with more well-known distributions, consider the case where observations are normally distributed. The canonical link is the identity function which means the conditional mean of r is linearly related to the parameters

$$\mathbf{E}\Big[r_n | x_n, \theta\Big] = x_n^T\theta. \tag{2.2.4}$$

The log-likelihood for r, up to a scale factor set by the variance, is then

$$L(\theta) = \sum_{n=1}^{N} -\frac{(r_n - x_n^T\theta)^2}{2} + \text{const}(\theta) \tag{2.2.5}$$

$$= \sum_{n=1}^{N} (x_n^T\theta) r_n - \frac{1}{2}(x_n^T\theta)^2 + \text{const}(\theta). \tag{2.2.6}$$

The non-linear function of the parameters in this case is seen to be $\frac{1}{2}(x_n^T\theta)^2$. Note that the non-linear function is not equal to the canonical link. As another example, if responses are distributed by an inhomogeneous Poisson process, where time is discretized so that r_n stores the number of events (in a neuroscience setting this could be spikes) emitted in time bin n, the canonical link function is the logarithm

$$\log\Big(\mathbf{E}\Big[r_n | x_n, \theta\Big]\Big) = x_n^T\theta. \tag{2.2.7}$$

The log-likelihood for r is then

$$L(\theta) = \sum_{n=1}^{N} \log\Big(\exp\Big(-\exp(x_n^T\theta)\Big)\frac{\Big(\exp(x_n^T\theta)\Big)^{r_n}}{r_n!}\Big) \tag{2.2.8}$$

$$= \sum_{n=1}^{N}(x_n^T\theta)r_n - \exp(x_n^T\theta) + \text{const}(\theta). \tag{2.2.9}$$

From the above equation, the non-linear function of the parameters is seen to be $\exp\Big(x^T\theta\Big)$. In the neuroscience literature this model is referred to as a linear-nonlinear-Poisson (LNP) model (Simoncelli, et al. 2004). As a final example, consider the case where responses are distributed according to a binary logistic regression model so that r_n only takes two values, say 0 or 1. The canonical link for this model is the logit function

$$p_n \equiv \mathbf{E}\Big[r_n|x_n,\theta\Big], \tag{2.2.10}$$

$$\log\Big(\frac{p_n}{1-p_n}\Big) = x_n^T\theta. \tag{2.2.11}$$

Using the above function for p_n, the log-likelihood for r is

$$L(\theta) = \sum_{n=1}^{N}\log\Big(p_n^{r_n}(1-p_n)^{1-r_n}\Big) \tag{2.2.12}$$

$$= \sum_{n=1}^{N}(x_n^T\theta)r_n + \log(1-p_n) \tag{2.2.13}$$

$$= \sum_{n=1}^{N}(x_n^T\theta)r_n - \log\Big(1+\exp\Big(x_n^T\theta\Big)\Big). \tag{2.2.14}$$

The corresponding non-linear function of the parameters is $\log\Big(1+\exp\Big(x_n^T\theta\Big)\Big)$.

2.2.2 The computational advantage of using ELs over log-likelihoods in a GLM

For large values of N and p there is a large difference in the computational cost between the linear and non-linear terms. Because we can trivially rearrange the linear term as $\sum_{n=1}^{N}(x_n^T\theta)r_n = (\sum_{n=1}^{N} x_n^T r_n)\theta$, its computation only requires a single evaluation of the sum $\sum_{n=1}^{N}(x_n^T r_n)$ no matter how many times the log-likelihood is evaluated. More precisely, if we evaluate the log-likelihood K times, the number of operations to compute the linear term is $O(Np + Kp)$. This will not be true in general for the non-linear sum and its computation will require $O(KNp)$ operations. The EL, denoted by $\tilde{L}(\theta)$, is an approximation to the log-likelihood that can alleviate this difference in computational cost. It is found by invoking the law of large numbers to approximate the sum over the non-linearity in equation 2.2.3 by its expectation (Sadeghi & Paninski 2012, Park & Pillow 2011, Field et al. 2010, Paninski 2004)

$$
\begin{aligned}
L(\theta) &= \sum_{n=1}^{N}(x_n^T\theta)r_n - G(x_n^T\theta) + \text{const}(\theta) && (2.2.15)\\
&\approx \sum_{n=1}^{N}(x_n^T\theta)r_n - N\mathbf{E}\Big[G(x^T\theta)\Big] && (2.2.16)\\
&\equiv \tilde{L}(\theta), && (2.2.17)
\end{aligned}
$$

where the expectation is with respect to the distribution of x. The EL trades in the $O(KNp)$ cost of computing the nonlinear GLM term for the cost of computing $\mathbf{E}\Big[G(x^T\theta)\Big]$ at K different values of θ, $O(Kp + Kz)$ where z is the cost of computing the expectation. In general then, ignoring terms that don't scale with K, the EL (and its gradient and hessian) can be be computed $\frac{p+Np}{2p+z} = \frac{1+N}{2+\frac{z}{p}} \approx \frac{Np}{z}$ times faster, than the true GLM log-likelihood. Note that because G only depends on the projection of x onto θ, calculating the expectation only requires the computation of a unidimensional

integral

$$\mathbf{E}\Big[G(x^T\theta)\Big] \;=\; \int G(x^T\theta)p(x)dx = \int G(q)\zeta_\theta(q)dq, \tag{2.2.18}$$

where ζ_θ is the, θ dependent, distribution of the one-dimensional variable $q = x^T\theta$.

In certain cases, when one has information about the stimulus distribution, the integral can be performed analytically. The simplest example is when is given the first and second moments of the stimulus covariates, say $\mathbf{E}\Big[x\Big] = 0$, $\mathbf{E}\Big[xx^T\Big] = C$. If one considers a GLM with Gaussian noise,

$$G(x^T\theta) \;=\; \frac{\theta^T x^T x\theta}{2}, \tag{2.2.19}$$

the average required by the EL is

$$\mathbf{E}\Big[G(x^T\theta)\Big] \;=\; \frac{\theta^T C\theta}{2}. \tag{2.2.20}$$

Another example is the LNP model if $p(x)$ is normally distributed, in which case

$$\begin{aligned}
\mathbf{E}\Big[G(x_n^T\theta)\Big] &= \int \exp(x^T\theta)\frac{1}{(2\pi)^{\frac{p}{2}}|C|^{\frac{1}{2}}}\exp\Big(-x^TC^{-1}x/2\Big)dx && (2.2.21)\\
&= \exp\Big(\frac{\theta^TC\theta}{2}\Big)\frac{1}{(2\pi)^{\frac{p}{2}}|C|^{\frac{1}{2}}}\int \exp\Big(-(x-x')^TC^{-1}(x-x')/2\Big)dx\\
&= \exp\Big(\frac{\theta^TC\theta}{2}\Big). && (2.2.22)
\end{aligned}$$

The second line follows by combining the exponentials and completing the square, with x' being a vector that vanishes after integration. Note that in this case the integral only depends on the scalar quantity $\theta^TC\theta$. In fact, this result generalizes and will be true for all distributions $p(x)$ of the form

$$p(x) \;\propto\; h(x^TC^{-1}x), \tag{2.2.23}$$

for arbitrary h. This broad class of functions are known as Elliptical distributions (assuming zero mean) and include multivariate Student's-t, logistic and exponential power families (Fang, et al. 1990). Changing coordinates to the basis $y = C^{-1/2}x$, these distributions only depend on the norm $y^T y$ and are therefore symmetric with respect to rotations of y. Since G only depends on the projection $y^T C^{1/2}\theta$, rotations of y can alternatively be viewed as rotations of θ which means that the integral only depends on the scalar norm $\theta^T C^{1/2} C^{1/2}\theta = \theta^T C\theta$. In practice, this is beneficial since we can numerically evaluate $\mathbf{E}\Big[G(x^T\theta)\Big]$ for all possible θ no matter the dimension of θ. We can then pre-compute the integral before running routines which may require large numbers of calls to the likelihood, and do away with the need to compute the integral in equation 2.2.18 at each call.

In fact equation 2.2.22 also holds if we approximate the variable q in equation 2.2.18 as normally distributed, with mean $\mathbf{E}\Big[\theta^T x\Big] = \theta^T \mathbf{E}\Big[x\Big] = 0$ and variance $\mathrm{var}(\theta^T x) = \theta^T C\theta$. Classic results using central limit theorems for random projections, suggest that for high-dimensional $p(x)$, almost all 1-d projections lead to a Gaussian distribution, i.e. for almost all θ we can treat $p(q)$ as Gaussian (Diaconis & Freedman 1984). Numerically we find that this approximation can work quite well for different distributions of x. Figure 2.1 examples this for simulated stimuli drawn from two non-Elliptic distributions (binary white noise stimuli in A and Weilbull distributed stimuli in B). Therefore we just need to compute equation 2.2.18 for two scalars, all possible values of the mean and variance of this Gaussian.

There are many utilizations of likelihoods where the computational savings of using ELs would be beneficial. For example, a standard approach for estimating the model parameters θ from data is to compute the maximum likelihood estimator (MLE) (van der Vaart 1998), $\arg\max_\theta \; L(\theta)$. While there are too many methods available for numerical optimization to be discussed here, the majority require several evaluations of the log-likelihood. If we instead infer θ by maximizing the expected likelihood (MELE) $\arg\max_\theta \; \tilde{L}(\theta)$, the inference can be performed faster since each

EL call can be computed more quickly than the true log-likelihood. In fact, as discussed in the next section, the MELE can sometimes be computed analytically when the MLE cannot. In the Bayesian framework one is often interested in measuring the probability of the data, r, with the parameter dependence integrated out (this probability is sometimes referred to as the marginal likelihood) (Gelman, et al. 2003a, Kass & Raftery 1995a). This is done by assigning a prior distribution $f(\theta|R)$, with its own 'hyper-parameters' R, over θ and calculating

$$F(R) \equiv p(r|x_1, ..., x_N, R) \quad = \quad \int p(r, \theta|x_1, ..., x_N, R) d\theta. \tag{2.2.24}$$

In practice, a variety of numerical methods such as Markov Chain Monte Carlo (MCMC) sampling, Evidence Propagation, the Laplace Approximation (which we will say more about later), etc., are used to evaluate $F(R)$ (Bishop 2006). These methods all involve several calls to the log-likelihood and can be performed faster using ELs. MCMC can also be used to sample and calculate statistics from the posterior parameter distribution, $P(\theta|X, r)$. Thus sampling is another application of likelihoods where faster computation time would be beneficial. We will discuss each of these applications in turn in the following sections.

2.2.3 Computational efficiency of maximum expected log-likelihood estimation for the LNP and Gaussian model

To illustrate the results of the last section with specific examples, we consider the cost of computing the maximum likelihood estimator for an LNP and Gaussian model with and without using the expected likelihood. The EL under the Gaussian model is,

$$\tilde{L}(\theta) \quad = \quad \theta^T X^T r - N \frac{\theta^T C \theta}{2}, \tag{2.2.25}$$

where we have used equation 2.2.20 and defined $X = (x_1, ..., x_N)^T$. Optimizing $\tilde{L}$ to find the MELE is a convex problem, the solution of which can be found finding the value of θ where the gradient of equation 2.2.25 equals zero. The solution is

$$\hat{\theta}_{\text{MELE}} = C^{-1}\frac{X^T r}{N}. \tag{2.2.26}$$

The maximum likelihood solution for linear regression under a Gaussian model, without the EL approximation, has a well known solution given by (Johnson & Wichern 2007)

$$\hat{\theta}_{MLE} = (X^T X)^{-1} X^T r. \tag{2.2.27}$$

The computational cost of determining both estimators is the cost of solving p-dimensional linear systems of equations, which in general is $O(p^3)$. If C is chosen by an experimenter to have special structure, e.g. $C = I$ (white-noise) or C is Banded, Circulant, Toeplitz etc., this cost can be reduced to $O(p)$ or $O(p\log(p))$, a time that is much faster than the general case (Golub & van Van Loan 1996). This can be exploited in many applications, since real experiments are often done by choosing stimuli from distributions which have such structure. The MLE will not enjoy this decrease in speed because $X^T X$ typically will not have structure even when C does.

We can analytically solve for the MELE under an LNP model as well if we allow for an offset term, θ_0. If we approximate $x^T\theta$ as normally distributed, the MELE has an analytic solution (see (Park & Pillow 2011) for a proof also methods). Briefly, the main intuition why is because if one first optimizes the EL (equation 2.2.16) with respect to the offset θ_0 and then substitutes the optimal θ_0 back into the EL, the resulting cost function $\max_{\theta_0} \tilde{L}(\theta, \theta_0)$ also known as the profile likelihood (Murphy &

van der Vaart 2000), is a quadratic function of θ

$$\hat{\theta}_{\text{MELE}} = \arg\max_{\theta} \; \theta^T X^T r - \sum_{n=1}^{N} r_n \frac{\theta^T C \theta}{2}. \tag{2.2.28}$$

Note that this is essentially the same quadratic problem as in the Gaussian case (equation 2.2.25) with the total number of spikes $\sum_n r_n$ replacing the number of samples N in equation 2.2.25. Optimizing this quadratic function results in a regression-like estimator

$$\hat{\theta}_{\text{MELE}} = \frac{C^{-1} X^T r}{\sum_n r_n}. \tag{2.2.29}$$

In the neuroscience literature, the function $\frac{X^T r}{\sum_n r_n}$ is referred to as the spike-triggered average. This name refers to the fact that if time is discretized fine enough so that the entries of r are 0 or 1, the product $X^T r$ is simply an average of the stimulus conditioned on the occurrence of a 'spike' (r=1). The computational cost for computing $\hat{\theta}_{\text{MELE}}$ again reduces to the cost of solving a linear system of equations $C\hat{\theta}_{\text{MELE}} = \frac{X^T r}{\sum_n r_n}$. While there are too many optimization methods to discuss here, a standard method for optimizing the GLM log-likelihood under an LNP model to find the MLE is the Newton-Raphson (NR) method (Nesterov 2004). The NR method is an iterative procedure where each iteration optimizes a local, quadratic approximation to L. This quadratic problem is solved by computing the so-called Newton direction, $-H^{-1}g$, where H is the Hessian and g is the gradient of the GLM log-likelihood. Constructing the Hessian requires Np^2 operations. Solving for the Newton direction requires at most $O(p^3)$ operations. These operations must be performed at each iteration and represent the limiting cost of the NR method. Therefore if there are K total iterations the limiting cost of the NR method is $O(K(Np^2 + p^3))$ which is considerably larger than the cost of solving equation 2.2.29 for large values of N, p or K.

The conjugate gradient (CG) algorithm is another commonly used iterative method used to find the MLE (Nesterov 2004, Golub & van Van Loan 1996). This method avoids the costly construction and inversion of the Hessian required to find the Newton direction by optimizing the log-likelihood along search directions that only require g to be computed. The dominant cost of computing the GLM gradient scales like $O(Np)$ and therefore if K_{CG} iterations are required the cost of CG scales like $O(K_{CG}Np)$. This method's disadvantage over the NR algorithm is that it typically requires many more iterations to converge. In practice, the CG search directions are often multiplied by matrices known as "pre-conditioners" which have the effect of lowering the number of iterations for the algorithm to converge (Nesterov 2004, Golub & van Van Loan 1996). When a pre-conditioner is used the algorithm is referred to as preconditioned conjugate gradients (PCG). In general the per iteration cost of PCG will be higher than that of CG due to preconditioning, $O(p^2 + Np)$ versus $O(Np)$. In any case, when N is large and/or when many iterations are required for convergence, optimizing the GLM log-likelihood via PCG or CG will be more expensive than solving equation 2.2.29.

We can naturally incorporate regularization to our estimator, by maximizing the EL with some type of penalty of the form $\log(f(\theta))$ resulting in a maximum posterior expected log-likelihood estimator (MPELE)

$$\hat{\theta}_{\text{MPELE}} = \arg\max_{\theta} \ \tilde{L}(\theta) + \log(f(\theta)). \tag{2.2.30}$$

We can exploit special structure in C when solving for the MPELE as well. For example, if we use an $L2$ penalty so that $\log(f(\theta))$ in equation 2.2.30 equals $-\frac{R}{2}\|\theta\|_2^2$ with R a scalar, the MPELE for the LNP model is again a regularized spike-triggered average (see methods, computing the MPELE for an inhomogeneous Poisson process)

$$\hat{\theta}_{\text{MPELE}} = \left(C + \frac{R}{\sum_n r_n}\right)^{-1} \frac{X^T r}{\sum_n r_n}. \tag{2.2.31}$$

For general matrices R and C, the dominant cost of computing $\hat{\theta}_{\text{MPELE}}$ will be $O(Np+p^3)$. This is still faster to compute than the MAP, which cost $O(KNp)$ or $O(K(Np^2+p^3))$, for large N or K. When $C+R$ share some special structure, e.g. C and R are both circulant or banded, computing $\hat{\theta}_{\text{MPELE}}$ can again be reduced to $O(Np+p)$ or $O(Np+p\log(p))$. The MPELE under a Gaussian model will be nearly identical because of the similarity between the profile likelihood of the LNP model with EL and the Gaussian likelihood noted earlier. In fact, the MPELE with an $L2$ penalty, has been used to reduce computation time in the more general context of Gaussian process regression (Rasmussen & Williams 2005, Sollich & Williams 2005). In this setting one allows the regression function, say $f(x)$, to have a general form, rather than maintaining the restriction $f(x) = x^T\theta$, and only assumes that this function is drawn from a Gaussian process.

If we use an L1 penalty so that $\log(f(\theta))$ in equation 2.2.30 equals $-\frac{\lambda}{2}\|\theta\|_1$ with λ a scalar, $\hat{\theta}_{\text{MPELE}}$ under a Gaussian model solves the equation

$$\hat{\theta}_{\text{MPELE}} = \arg\max_{\theta} \; \theta^T X^T r - N\frac{\theta^T C\theta}{2} - \frac{\lambda}{2}\|\theta\|_1. \tag{2.2.32}$$

If C is a diagonal matrix, classic results from subdifferential calculus (Nesterov 2004) show that $\hat{\theta}_{\text{MPELE}}$ is a solution to equation 2.2.32 if and only if $\hat{\theta}_{\text{MPELE}}$ satisfies the subgradient optimality conditions which in this case are

$$-NC_{jj}(\hat{\theta}_{\text{MPELE}})_j + (X^T r)_j = \lambda \text{sign}(\hat{\theta}_{\text{MPELE}})_j \text{ if } (\hat{\theta}_{\text{MPELE}})_j \neq 0 \tag{2.2.33}$$

$$\Big| -NC_{jj}(\hat{\theta}_{\text{MPELE}})_j + (X^T r)_j \Big| \leq \lambda \text{ otherwise}, \tag{2.2.34}$$

for $j = 1, ..., p$. The above equations imply that $\hat{\theta}_{\text{MPELE}}$ is a soft-thresholded function of $X^T r$: $(\hat{\theta}_{\text{MPELE}})_j = 0$ if $|(X^T r)_j| \leq \lambda$, otherwise

$$(\hat{\theta}_{\text{MPELE}})_j = \frac{1}{NC_{jj}}\Big((X^T r)_j - \lambda \text{sign}(\hat{\theta}_{\text{MPELE}})_j\Big), \tag{2.2.35}$$

for $j = 1, ..., p$. Note that equation 2.2.35 implies that we can independently solve for each element of $\hat{\theta}_{\text{MPELE}}$ along all values of λ (the so-called regularization path). Also, only a single matrix-vector multiply, $X^T r$, is required resulting in a complexity $O(Np)$. Finally we note that nearly identical results hold for the MPELE under an LNP model with L1 regularization since the profile likelihood (equation 2.2.28) is quadratic. The only difference is that under an LNP model with L1 regularization, N, in equation 2.2.35 is replaced by the total spike count $\sum_n r_n$.

Without our approximation, there are several methods in the literature to solve the linear regression problem with L1 regularization (often referred to as the Lasso) (Osborne, et al. 2000, Efron, et al. 2004, Friedman, et al. 2010). However, none can be solved as quickly as the MPELE. The main reason being that the MPELE can exploit structure present in C but the MAP has no access to C and must use $X^T X$ which typically is not structured. For example, one fast method is the homotopy algorithm (Osborne et al. 2000, Efron et al. 2004). The computational complexity of this algorithm is limited by several inversions of different subsets of the matrix $X^T X$. The size of the subsets is inversely related to the penalty parameter λ and in the worst case this algorithm has a cost $O(p^3)$. Another recent method for solving the Lasso problem is a cyclical coordinate descent method (Friedman et al. 2010) colloquially titled glmnet. Briefly, this is an iterative algorithm which treats the Lasso as if it were a problem where each element of the penalized maximum likelihood vector, $\hat{\theta}_{MAP}$, could be solved independently, as we can do in solving equation 2.2.35. Accordingly, glmnet maximizes the likelihood by cycling through each element of θ and maximizing the likelihood for that element while holding the others fixed. The advantage in doing so is that an analytic expression exists for the optimization at each coordinate step which aids in efficiently evaluating the entire regularization path. Each coordinate step requires a matrix-vector multiply, resulting in a computational cost of order $O(Np)$ if X and r are dense. This is in contrast to solving equation 2.2.35 which only requires a single matrix-vector multiply. This difference in the number of matrix-

vector multiplies between the two problems arises because the elements of the $\hat{\theta}_{MAP}$ are not independent whereas those of $\hat{\theta}_{\text{MPELE}}$ are due to the structure present in C.

When C is not a diagonal matrix we can no longer solve equation 2.2.32 analytically. However we can still solve this equation numerically using interior-point methods (Nesterov 2004). Briefly, these methods solve a sequence of auxiliary, convex problems whose final solution is the desired vector. Unlike problems with an $L1$ penalty, these auxiliary problems are constructed to be smooth, a gradient and Hessian exist, by introduction of a barrier function. We can solve these auxiliary problems in small number of iterations using the NR method. Solving for the Newton direction again requires solutions to equations of the form $C\theta = b$ which can be performed in a time $O(p)$ if C is banded. In general if C is sparse we can use CG to solve these auxiliary problems in a time faster than the $O(Np)$ required by multiplication with $X^T X$.

2.2.4 Analytic comparison of the accuracy of EL estimators with the accuracy of maximum-likelihood estimators

We have considered several applications where the computational cost of using the EL is lower than performing equivalent computations using the GLM log-likelihood. In particular when one is interested in estimating model parameters we have shown that the computational cost of estimating θ using the MELE can be orders of magnitude lower than estimates computed using the MLE. We now focus our attention on the accuracy of MEL estimators. When responses and stimuli are normally distributed we can analytically calculate the mean-squared error (MSE) of the MELE and MLE. These expressions are useful for determining how the difference in MSE between the two estimators depends on model parameters. For convenience we do not consider an offset, however the results are easily generalizable.

Assume

$$r|x \sim \mathcal{N}(x\theta, \sigma^2 I), \tag{2.2.36}$$

$$x \sim \mathcal{N}(0, I). \tag{2.2.37}$$

For convenience we set $\sigma^2 = 1$. We prove in the methods that the MSE for the MELE, $\hat{\theta}_{\text{MELE}}$, and the maximum-likelihood estimator, $\hat{\theta}_{MLE}$, is

$$\mathbf{E}\Big[\|\hat{\theta}_{\text{MELE}} - \theta\|_2^2\Big] = \frac{\theta^T\theta + p(\theta^T\theta + 1)}{N} \tag{2.2.38}$$

$$\mathbf{E}\Big[\|\hat{\theta}_{MLE} - \theta\|_2^2\Big] = \frac{p}{N - p - 1} \tag{2.2.39}$$

$$= \frac{p/N}{1 - p/N - 1/N}. \tag{2.2.40}$$

The first thing we note is that the MSE of both estimators approaches 0 as $N \to \infty$. This is to be expected since their is no difference between the GLM log-likelihood and EL in this limit and since the MLE is consistent (van der Vaart 1998). For many real-world applications, it is more appropriate to consider the limit where the number of samples and parameters are large, both $N \to \infty$ and $p \to \infty$, but their ratio is fixed $\frac{p}{N} \to \rho$ where ρ is finite. In this limit we find

$$\mathbf{E}\Big[\|\hat{\theta}_{\text{MELE}} - \theta\|_2^2\Big] \to \rho(\theta^T\theta + 1) \tag{2.2.41}$$

$$\mathbf{E}\Big[\|\hat{\theta}_{MLE} - \theta\|_2^2\Big] \to \frac{\rho}{1 - \rho}. \tag{2.2.42}$$

See figure 2.7 and the methods (Computing the mean-squared error for the MPELE and MAP) to see the validity of approximating real data, with finite N and p by this limit.

Figure 2.2A (left panel) plots the MSE for the MLE (dotted line) and MELE (solid line) given by equations 2.2.41 and 2.2.42 as a function of ρ. Note that we

do not plot the MSE for values of $\rho > 1$ because the MLE is non-unique when p is greater than N. Different colors correspond to different values of $\theta^T\theta$. Note that we have set the noise variance to one, so $\theta^T\theta$ is the signal variance divided by the noise variance or signal-to-noise ratio (SNR)

$$SNR = \frac{\mathbf{E}\left[\theta^T x^T x\theta\right]}{\sigma^2} \tag{2.2.43}$$

$$= \frac{\theta^T\theta}{1}. \tag{2.2.44}$$

The second line follows from the fact that we choose stimuli with identity covariance. For each line there is a value of ρ where the MELE outperforms the MLE. It can be seen from equation 2.2.42 that the MLE MSE diverges when $\rho \to 1$. However, equation 2.2.41 shows that the MSE of the MELE remains finite in this limit. Therefore, as demonstrated in the plot, we expect the MELE to outperform the MLE as ρ approaches 1, i.e when the amount of data is the same as the number of parameters. More specifically, from equations 2.2.41 and 2.2.42 it can be seen that the MELE outperforms the MLE when

$$\rho > \frac{SNR}{1 + SNR}. \tag{2.2.45}$$

Conversely equation 2.2.45 shows that for a fixed value of ρ the MLE will outperform the MELE if the SNR is larger than $(\rho^{-1} - 1)^{-1}$. This general tradeoff in accuracy between SNR and ρ has been noted before and is an example of the general fact that for models in which the variance in r is constant and $\mathbf{E}\left[r|X\right] = X\theta$, no unique, unbiased (averaged over X and r) linear estimator exists which has minimum variance for all values of θ (Shaffer 1991).

Typically some form of regularization is added to the likelihood to bias the MLE, which alleviates the MLE's decline in performance when $\rho \approx 1$. We calculated (see methods) the MSE for the model specified by equations 2.2.36,2.2.37 with an

L2-norm penalty of the form, $\log(f(\theta)) = -\frac{R}{2}\|\theta\|_2^2$ with R a scalar. Figure 2.2B (top panel) plots the MSE for both estimators (see equations 2.4.86, 2.4.93 in the methods for the equations being plotted) as a function of R and ρ for an SNR value of 1. Note that we now plot MSE values for $\rho > 1$ since regularization makes the MAP solution unique. It can be seen that the two estimators have similar accuracy over a large region of parameters. For each value of ρ we also find the value of R that minimizes each estimator's MSE. This is plotted in Figure 2.2A (right panel). Even at each estimators respective peak accuracy the two solutions show similar performance. Furthermore as ρ grows the difference in MSE between θ_{MAP} and $\hat{\theta}_{\text{MPELE}}$ shrinks.

In conclusion, when $\rho \approx 1$ so that the number of parameters being fit is comparable to the number of samples, the MELE will outperform the MLE for a wide range of SNR values. When ρ is not close to 1, equation 2.2.45 shows that the MELE can still outperform the MLE if the SNR is low. When regularization is added both estimators will have similar accuracy, at least for low values of SNR.

2.2.5 Fast methods for refining maximum expected log-likelhood estimators to obtain MAP accuracy

When the MAP outperforms the MPELE, we can always use $\hat{\theta}_{\text{MPELE}}$ as a quickly computable initializer for optimization algorithms used to compute $\hat{\theta}_{MAP}$. Gradient based optimization methods might quickly converge if the MPELE is a close approximation to the MAP. Since, as previously discussed, each iteration of these methods requires $O(Np)$ operations this could provide a quick method for computing the MAP, or an estimator that is just as accurate. We first tested this idea on real data, by fitting an LNP model to a population of ON and OFF parasol ganglion cells (RGCs) recorded *in vitro*. Cells responded to either binary white-noise stimuli or naturalistic stimuli (spatial-temporally correlated Gaussian noise with spatial correlations having a 1/F power spectrum and changes in time following a first order autoregressive process). As described in the methods (section simulated and real neuronal data applications),

our receptive fields are fit with 810 parameters and $\frac{p}{N} = 0.021$. We chose the inverse Hessian of the EL evaluated at the MELE or MPELE as a pre-conditioner. In this case, using the same notation as in equations 2.2.29 and 2.2.31, we use $(C \sum_n r_n)^{-1}$ or $(C \sum_n r_n + R)^{-1}$ as a pre-conditioner. These pre-conditioners are advantageous for two reasons. First, if the MPELE is reasonably close to the MAP then the likelihood we must optimize can be approximated as quadratic with curvature given by the Hessian of the true likelihood evaluated at the MAP. Under this approximation, a good pre-conditioner can decrease the number of PCG iterations required for optimization if it approximates the inverse of this Hessian. Our pre-conditioner will have this property when the MPELE and MAP estimators are close. Second, when C and R have special structure, as in this problem, we can multiply vectors by our pre-conditioner in a time faster than the general $O(p^2)$.

For binary white-noise stimuli we find that the MELE (approximated by equation 2.2.29 with $C = \mathbf{I}$) and MLE yield similar filters and accuracy, with the MLE slightly outperforming the MELE (see figure 2.3A). Note that in this case, $\hat{\theta}_{\text{MELE}}$ can be computed quickly, $O(pN)$, since we only need to compute the matrix-vector multiplication $X^T r$. On average across a population of 126 cells, we find that terminating the PCG algorithm, initialized at the MELE, after 2 iterations yielded an estimator with the same accuracy as the MAP. To measure accuracy we use the information rate on a hold-out data set not used for training (see methods section simulated and real neuronal data applications). Our estimator retained its computational advantage after running PCG for a small number of iterations. On an Intel Core 2.8 GHz processor–all timings quoted in this paper use this processor–the MLE took on average 15 times longer to compute than PCG initialized at the MELE (88±2 vs 6±0.1 seconds). When fitting responses to correlated stimuli, we used a ridge regression prior so that $\hat{\theta}_{\text{MPELE}}$ is given by equation 2.2.31. R is chosen based on the log-likelihood of a novel data set not used for training (section simulated and real neuronal data applications). In this case, we find that 9 PCG iterations are required to reach MAP

accuracy (see figure 2.3B), however our estimator was still faster to compute than the MAP(33±1 vs 107±8 seconds).

So far we have only discussed models which assume conditional independence of responses given an external stimulus. In neuroscience, it is known that neurons are directly or indirectly coupled to each other and that the predictive performance of the GLM can be improved in a variety of brain areas by not assuming conditional independence (Saleh et al. 2012b, Vidne et al. 2011b, Pillow et al. 2008, Truccolo et al. 2005). This is done by including self-history and interaction terms with all observed neurons. Specifically, assume we have recordings from M neurons and let r_i be the vector of responses across time of the ith neuron. Each neuron is modeled with a GLM where the weighted covariates, $x_n^T\theta_i$, can be broken into an offset, an external stimulus component x^s and a history-dependent coupling component

$$x_n^T\theta_i \quad = \quad \theta_{i0} + (x^s)^T\theta_i^s + \sum_{j=1}^{M}\sum_{k=1}^{\tau} r_{j,n-k}\theta_{ijk}^H, \tag{2.2.46}$$

for $n = 1, ..., N$. Note this is the same model as before when $\theta_{ijk}^H = 0$ for $j = 1, 2, ..., M$. A central, open challenge for this model is developing methods to determine model parameters from data that scale well with the number of observed neurons. This is critical since experimental methods continue to improve resulting in rapid growth of the number of observable neurons (Stevenson & Kording 2011).

We address this by devising a simple method that uses the MELE of an LNP model to fit a GLM with history terms. The idea behind our algorithm is if estimates of θ^s fit with $\theta_{ijk}^H = 0$ are similar to estimates of θ^s without this hard constraint, we can leverage the simple formula for the marginalized estimate of θ^s by using equation 2.2.29 as an initializer and fit the full model with $(x^s)^T\hat{\theta}_{\text{MELE}}$ held fixed up to a multiplicative gain factor. More precisely, to infer θ_i for $i = 1, 2, ..., M$, we first estimate θ_i^s, assuming no coupling or self-history terms ($\theta_{ijk}^H = 0$ for $j = 1, 2, ..., M$) using the MELE (equation 2.2.29 or a regularized version). We then update the

history components and offset by optimizing the GLM likelihood with the stimulus filter held fixed

$$\begin{aligned}(\theta_{i0}, \alpha_i, \theta_i^H) &= \arg\max_{(\theta_0,\alpha,\theta^H)} L\Big((\theta_0, \alpha X\hat{\theta}_{\text{MPELE}}, \theta^H)\Big) \\ &\quad + \log\Big(f(\theta_0, \alpha X\hat{\theta}_{\text{MPELE}}, \theta^H)\Big). \end{aligned} \tag{2.2.47}$$

Holding the stimulus filter fixed can greatly reduce the number of parameters being optimized and hence the computational cost of finding the history components. Finally, if further accuracy is required, we can update the estimates of the full parameter vector by running a small number of PCG iterations down the log-likelihood (or posterior) without fixing any parameters to zero.

We fit the GLM model specified by equation 2.2.46 to a population of 100 RGC cells responding to binary white-noise using 250 stimulus parameters and 105 parameters related to neuronal history ($\frac{p}{N} = 0.01$). Previous work has shown that only a subset of simultaneously recorded neurons will be coupled (Pillow et al. 2008) and we would like our fitting procedure to accurately determine this subset. To do so we regularize the coupling history components using an $L1$ penalty of the form $\lambda \sum_j |\theta_{ij}^H|$. We compute our estimates over a large range of values of λ using glmnet. Figure 2.4A compares filter estimates for two example cells using out method and the MAP. All filters are plotted on the same scale. The filter estimates are similar but not identical, however both methods find the same coupled, nearest neighbor cells. To evaluate the extent the differences between the filters affect the model's predictive power, we again measure the cross-validated information rate. Figure 2.4B shows that the information rate across the population is the same for both algorithms. Our method took an average of 1 minute to find the entire regularization path while finding the full MAP path took 16 minutes on average. In the machine learning and signal processing communities (Zhang 2011, Barron, et al. 2008, Tropp 2004, Mallat & Zhang 1993), computationally efficient, greedy algorithms are also known to quickly

select the appropriate features in sparse signals. The basic idea behind so-called forward greedy methods is to iteratively determine the number of non-zero features by adding features at each iteration to lower some cost function which measures performance ($L(\theta) + \log f(\theta)$ in our case). In addition to $L1$ regularization, we also found that a forward greedy approach could be used to determine the proper coupling coefficients in a manner that is scalable with large M.

2.2.6 Marginal likelihood calculations

As previously mentioned, the marginal likelihood can be computed more quickly with the log-likelihood approximated by the EL. We have shown examples in point estimation where the MELE estimator can leverage information we may have about the moments of stimulus covariates to speed up the estimation procedure. We show that this is also true when calculating marginal likelihoods. To demonstrate this, we first consider the case where the joint distribution, $p(r, \theta|X, R)$, required to compute the marginal likelihood (equation 2.5) can be written as the product of two Gaussian distributions

$$p(r, \theta|X, R) = p(r|X, \theta) f(\theta|R) \tag{2.2.48}$$

$$= \mathcal{N}(X\theta, \sigma^2\mathbf{I})\mathcal{N}(0, R^{-1}), \tag{2.2.49}$$

where $\mathbf{I}$ is the identity matrix. In this special case, computing the marginal likelihood is analytically tractable

$$\log F(R) = \frac{1}{2}\log\Big(\det(\Sigma R)\Big) + \frac{r^T X \Sigma X^T r}{2\sigma^4} + \text{const}(R), \tag{2.2.50}$$

$$\Sigma = (X^T X \sigma^2 + R)^{-1}. \tag{2.2.51}$$

If we approximate $p(r|X, \theta)$ by the the expected likelihood, the integral is still tractable and has the same form as equation 2.2.50 however with $\Sigma = (NC\sigma^2 + R)^{-1}$ assum-

ing $\mathrm{Cov}[x] = C$. When C and R are diagonalized by a common, known basis with known eigenvalues, then computing $\log(\det(\Sigma R))$ can be computed in $O(p)$, the time required to sum the p diagonal components of ΣR. This will also be true when C and R are banded matrices; if the bandwidth of $C + R$ is q the determinant can be computed in $O(pq^2)$. This is much faster than the usual $O(p^3)$ time required to compute the determinant via matrix decomposition when C and R have no special structure. In both of these cases the product $r^T X \Sigma X^T r$ can also be computed faster when using the EL in place of the log-likelihood.

When the likelihood and prior are not Gaussian we can approximate the marginal likelihood using the Laplace approximation (Kass & Raftery 1995b)

$$\begin{aligned} \log F(R) &\approx \log p(\theta_{MAP}|R) - \frac{1}{2}\log(\det(-H(\theta_{MAP}))) && (2.2.52) \\ &= \theta_{0MAP}\sum_n r_n + (\theta_{MAP})^T X^T r - \sum_n G(\theta_{0MAP} + x_n^T \theta_{MAP}) \\ &\quad + \log\Big(f(\theta_{MAP}|R)\Big) - \frac{1}{2}\log(\det(-H(\theta_{MAP}))), && (2.2.53) \end{aligned}$$

again neglecting factors that are constant with respect to R. $H(\theta_{MAP})$ is the posterior Hessian

$$H_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j}\Big(-\sum_n G(\theta_0 + x_n^T\theta) + \log\Big(f(\theta|R)\Big)\Big), \qquad (2.2.54)$$

evaluated at the MAP. We note that there are other methods for approximating the integral in equation 2.5 such as Evidence propagation (Bishop 2006), however we have not explored these yet.

Here too we can leverage information about stimulus covariates to decrease computation time if we use the EL. To demonstrate this we consider the LNP model where we can approximate $\mathbf{E}\Big[G(\theta_0 + x^T\theta)\Big]$ as in equation 2.2.22 and consider a normal prior with $f(\theta^s|R) = \mathcal{N}(0, R^{-1})$. As mentioned in the lines preceding equation 2.2.29, the MELE is found by optimizing the quadratic function $\max_{\theta_0} \tilde{L}(\theta, \theta_0)$. The MPELE

is found by optimizing the quadratic function $\max_{\theta_0} \tilde{L}(\theta, \theta_0) - \frac{R}{2}\|\theta\|_2^2$, with solution given by equation 2.2.31. The negative Hessian of this function,

$$-H = C(\sum_n r_n) + R, \tag{2.2.55}$$

can be computed more efficiently than $\log(\det(-H(\theta_{MAP})))$ when C and R have special structure for the same reasons as in the Normal case.

Marginal likelihood calculations have several applications. For example, when estimating parameters in the Bayesian framework one is faced with the problem of choosing the appropriate prior hyper-parameters, R. A common solution (Type II maximum likelihood) is to choose hyper-parameters that maximize the marginal likelihood (Gelman, et al. 2003b). That is estimate parameters using $\hat{R}$ such that

$$\hat{R} = \arg\max_R \; F(R). \tag{2.2.56}$$

Continuing with the LNP example given above, we find (see methods, section calculating $\hat{R}$ for the LNP model) that $\hat{R}$ can be computed analytically if the EL is used in place of the log-likelihood and $C = \mathbf{I}, R = \beta\mathbf{I}$:

$$\hat{R} = \begin{cases} \frac{p}{\frac{q}{N_s^2} - \frac{p}{N_s}}\mathbf{I} & \text{if } p < \frac{q}{N_s} \\ \infty & \text{if } p \geq \frac{q}{N_s}, \end{cases} \tag{2.2.57}$$

with $q \equiv \|X^T r\|_2^2$ and $N_s = \sum_{n=1}^N r_n$. Since $\hat{\theta}_{\text{MPELE}} = \left(C + \frac{R}{\sum_n r_n}\right)^{-1} \frac{X^T r}{\sum_n r_n}$ (equation 2.2.31), an infinite value of R corresponds to $\hat{\theta}_{\text{MPELE}}$ equal to zero. The intuitive interpretation of equation 2.2.57 is that the MPELE should equal zero when there isn't enough data, quantified by $\frac{q}{N_s}$, compared to the dimensionality of the problem p. Similar results hold when there are correlations in the stimulus, $C \neq I$, as shown in the methods (Calculating $\hat{R}$ for the LNP model).

We tested whether optimizing the marginal likelihood with the EL made a large

difference on the values of $\hat{R}$ found compared with marginal likelihood optimization using the GLM log-likelihood. We simulated Poisson responses to white-noise Gaussian stimuli using stimulus filters that have 250 parameters. By setting the gradient with respect to $\hat{R}$ of the marginal likelihood (under the Laplace approximation and using the GLM log-likelihood) to zero it is known that the optimal $\hat{R}$ obeys the following equation (Bishop 2006)

$$\hat{R} = \frac{p - \hat{R}\sum_i H^{-1}(\hat{\theta}_{MAP})_{ii}}{\hat{\theta}_{MAP}(\hat{R})^T\hat{\theta}_{MAP}(\hat{R})}. \tag{2.2.58}$$

The above equation only implicitly solves for $\hat{R}$ since $\hat{R}$ is present on both sides of the equation. However, this observation has led to the following iterative solution used in the literature (Bishop 2006) for determining the optimal R. At the $i+1$ iteration the running estimate of the optimal R is updated by

$$R_{i+1} = \frac{p - R_i\sum_i H^{-1}(\hat{\theta}_{MAP})_{ii}}{\hat{\theta}_{MAP}(R_i)^T\hat{\theta}_{MAP}(R_i)}. \tag{2.2.59}$$

We use this algorithm to optimize the marginal likelihood with respect to R. We initialize R_0 to 0 and run the algorithm until the change in R is less than a tolerance of $1e-5$. The distance between the average values of $\frac{\hat{R}}{N}$ using the GLM likelihood and expected likelihood appears to increase with $\frac{p}{N}$, with the EL systematically choosing lower values of $\frac{\hat{R}}{N}$ (Figure 2.5A top). This difference shrinks after a single iteration of equation 2.2.59 initialized using equation 2.2.57 (Figure 2.5A bottom). The remaining differences in $\hat{R}$ between the two methods did not lead to differences in the filters chosen using Type II maximum likelihood (Figure 2.5B). In these simulations the MAP typically took 20 times longer to compute than a single iteration of equation 2.2.59 initialized using equation 2.2.57 (10 versus 0.5 seconds).

Decreasing the computation time of Markov Chain Monte Carlo methods

We have shown how the EL can be used to significantly decrease the computation time required for computing the mode of posterior parameter distributions, $P(\theta|X, r)$, when the likelihood is a GLM. If we want to compute other statistics of this distribution, such as the variance or confidence intervals about the mean, we can use Markov Chain Monte Carlo (MCMC) methods to sample from $P(\theta|X, r)$ and then use these samples to compute any desired statistic. Running MCMC with the EL and its gradient can greatly reduce the computational burden of MCMC methods.

Briefly, MCMC is an iterative method that works by building a sequence of vectors $\{\Theta\}_{i=1}^{\infty}$ that converge to a stationary distribution that approximates $P(\theta|X, r)$ (Bishop 2006, Robert & Casella 2005). At each iteration some method, there are many, is used to propose a vector and the Metropolis-Hastings algorithm is used to accept or reject this proposal. Each iteration requires at least one evaluation of the posterior parameter distribution and sampling from a representative volume of the posterior typically requires thousands of iterations, due to correlations between proposals (Robert & Casella 2005). Certain MCMC algorithms such as the Hybrid Monte-Carlo (HMC) algorithm, can reduce the total number of iterations required but at a cost. Each HMC iteration uses several evaluations of the gradient of the log-posterior to reduce the correlation between proposals (Duane, et al. 1987). Since the cost of evaluating the likelihood and its gradient scales with the number of dimensions times the number of likelihood samples, these methods can be time consuming.

As previously mentioned, the cost of evaluating the EL and its gradient does not scale with the number of samples from the likelihood and can therefore reduce the computation time of MCMC. Running the HMC algorithm with the GLM likelihood, and its gradient, replaced by the EL and its gradient allows us to rapidly sample from an approximate posterior parameter distribution. Simulation results suggest that statistics computed from this distribution are often in good approximation to analogous statistics from the true posterior (Sadeghi & Paninski 2012), but not al-

ways. In figure 2.6 we compare the median and 95% credible region of the true and approximate posterior corresponding to a model with a flat prior and LNP likelihood (see methods section-Calculation of Bayesian credible regions- for a description of these credible regions were computed). For many elements of the θ vector, statistics computed between the true and approximate distribution are in close agreement, however some directions show discrepancies. For comparison, we also show the same statistics approximating the posterior distribution as a multivariate Gaussian either under the Laplace approximation or by replacing the GLM likelihood, and its gradient, replaced by the profile likelihood, $\max_{\theta_0} \tilde{L}(\theta, \theta_0)$. Using the profile likelihood instead of the EL does not make a difference in the results. The Laplace approximation also appears to be in close agreement with the true posterior.

We can sample from the exact posterior by evaluating proposals using the true posterior in the Metropolis-Hastings algorithm. If we use the EL gradient to propose vectors, in principle we can still decrease the computation time required to run HMC. However, we find (based on simulations using an LNP model and flat prior using realistic values of $\frac{p}{N}(\frac{p}{N} >= 0.02)$) that these proposals are rejected at a higher rate than proposals that use the actual GLM gradient. This required us to increase the total number of MCMC iterations, compromising the computational advantage of using the EL gradient.

2.3 Discussion

We have demonstrated the computational advantage of using the EL to approximate the log-likelihood of a generalized linear model with canonical link. When making multiple calls to the GLM likelihood (or its gradient and Hessian), the EL can be computed approximately $O(Np/z)$ times faster, where N is the number of data samples collected, p is the dimensionality of the parameters and z is the cost of a one-dimensional integral. The expected log-likelihood approximation has appeared

previously in the literature for the Poisson model and for Gaussian processes (Park & Pillow 2011, Sollich & Williams 2005, Paninski 2004). In the Gaussian process regression literature, the estimator found by maximizing the EL has been used under the name of the equivalent kernel (Sollich & Williams 2005). To the best of our knowledge, this work is the first to show that this approximation can be easily extended to all generalized linear models. Previous work with ELs in the neuroscience literature has focused on using them with LNP models to derive and generalize the classical spike-triggered average/covariance methods for estimating neuronal filters (Park & Pillow 2011, Paninski 2004) and has highlighted the consistency properties of these estimators. We have focused on their computational advantage. We demonstrated how this approximation can speed up parameter estimation, under inhomogeneous Poisson and Gaussian response models by orders of magnitude compared with standard MAP estimates. In this sense, this work is closer to methods designed to alleviate the computational cost of maximum likelihood such as the Fisher scoring algorithm or stochastic gradient descent (Bottou 1998, Jennrich & Sampson 1976). The Fisher scoring algorithm is identical to the Newton-Raphson method with the inverse Hessian matrix replaced with the inverse of the Fisher Information matrix (Jennrich & Sampson 1976). Indeed, when responses are normally distributed the MELE considered here can be seen as the same as estimates obtained with the scoring method, if the Fisher Information matrix is calculated by taking the expectation over both responses r and random covariates x.

For linear Gaussian response models we also derived analytic expressions for the mean-squared error of the MELE and MPELE. By comparing these expressions to those of MAP estimates, we showed that the MPELE and MELE do not necessarily come with a large cost in accuracy and are most similar to MAP estimates when signal-to-noise is small. We also showed that the MELE can outperform Maximum Likelihood estimates in terms of mean-squared error when the number of samples is comparable to the number of parameters. To the best of our knowledge, this is

the first time explicit expressions for this type of risk analysis have been derived. In (Shaffer 1991) the author derives a special case of equations 2.2.41 and 2.2.42 (these equations with $\theta = 0$), however her main focus is different than ours. She uses these equations as proof that no unique, unbiased (averaged over x and r) linear estimator exists which has minimum variance for all values of θ for models in which the variance in r is constant and $\mathbf{E}\Big[r|x\Big] = x\theta$.

Throughout this article, we used the general idea of replacing the log-likelihood with the EL for the sake of developing computationally efficient, statistical inference procedures. For example, by optimizing the EL subject to constraints on the parameters, we showed that regularization can be easily incorporated with marginalized estimators. In a Bayesian framework, this corresponds to estimation of the posterior mode by replacing the GLM log-likelihood with the EL. There are a variety of other contexts where this idea can be explored in the future. For example, in (Park, et al. 2008) the authors develop a Gibbs sampler for the Bayesian Lasso, the posterior distribution corresponding to a Gaussian likelihood and Laplace prior. The time limiting step in their procedure is $O(p^3)$ and arises from the need to sample from a multivariate Gaussian distribution with inverse covariance $(D + X^T X)$, where D is a diagonal matrix. As we have shown in the results (see section marginal likelihood calculations) our approximation corresponds to substituting $X^T X$ with NC, where $C = \mathbf{E}\Big[xx^T\Big]$. If C is diagonal (or banded) then this substitution reduces the computation time to $O(p)$, which is much faster. Of course, to sample from the desired posterior (instead of the approximate posterior) a slight modification to the algorithm would be required. One can show that Gibbs sampling is a special case of the Metropolis-Hasting sampling algorithm, where the proposal density is chosen such that one always accepts samples (Robert & Casella 2005). Since we have changed the proposal density, this would no longer hold true and one would need to accept samples using the appropriate Metropolis-Hastings acceptance probability, a correction that costs $O(p^2)$. The idea of replacing the log-likelihood in the posterior with an

alternative function has been used before (Hooker & Vidyashankar 2012), however, while we use this replacement to develop computationally efficient estimators previous work has used it for the purposes of developing estimators that are not sensitive to outliers and efficient, in the sense that they minimize the Cramer-Rao bound.

Future work should also focus on using the inference procedures developed here in an online setting, similar to (Lewi et al. 2009a). By reducing the computation time for determining parameters the EL could be used to improve performance in applications such as BMI by allowing for high-dimensional stimuli and the incorporation of large numbers of observed neurons (Stevenson & Kording 2011, Santhanam, et al. 2006b). This will allow these online applications to make full use of available data for improving decoding performance or predicting connectivity between neurons.

2.4 Methods

2.4.1 Computing the mean-squared error for the MPELE and MAP

We begin by considering the general case where the MAP and MPELE are derived from the log-likelihood and EL with a quadratic regularizer of the form $-cp\frac{\theta^T\theta}{2}$ with scalar penalty c added. The non-regularized case can be recovered by setting the penalty to zero. Note that we allow the regularizer to scale with the dimensionality of the problem, p. The resulting MAP and MPELE are then found by

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta}\ \theta^T X^T r - \frac{1}{2}\theta^T(X^T X + cpI)\theta && (2.4.60)\\
&= (X^T X + cpI)^{-1} X^T r && (2.4.61)\\
\hat{\theta}_{\text{MPELE}} &= \arg\max_{\theta}\ \theta^T X^T r - \frac{1}{2}(N + cp)\theta^T\theta && (2.4.62)\\
&= \frac{X^T r}{N + cp}, && (2.4.63)
\end{aligned}$$

where we consider $X_{ij} \sim \mathcal{N}(0,1)\ \forall i,j$. For convenience of notation we define the quantity $\tilde{S} = X^T X + cpI$ and therefore write the MAP as $\hat{\theta}_{MAP} = \tilde{S}^{-1} X^T r$.

As usual the MSE can be written as the sum of a squared bias term and a variance term

$$\mathbf{E}\Big[\|\hat{\theta} - \theta\|_2^2\Big] = \|\mathbf{E}\Big[(\hat{\theta} - \theta)\Big]\|_2^2 + \mathbf{E}\Big[\|\hat{\theta} - \mathbf{E}\Big[\hat{\theta}\Big]\|_2^2\Big]. \qquad (2.4.64)$$

The bias of the MAP equals

$$\begin{aligned}
\mathbf{E}\Big[(\hat{\theta}_{MAP} - \theta)\Big] &= \mathbf{E}\Big[\hat{\theta}_{MAP}\Big] - \theta && (2.4.65)\\
&= \mathbf{E}\Big[\mathbf{E}\Big[\hat{\theta}_{MAP}|X\Big]\Big] - \theta && (2.4.66)\\
&= \mathbf{E}\Big[\tilde{S}^{-1}X^T\mathbf{E}\Big[r|X\Big]\Big] - \theta && (2.4.67)\\
&= \mathbf{E}\Big[\tilde{S}^{-1}X^TX\theta\Big] - \theta, && (2.4.68)
\end{aligned}$$

The second line follows from the law of total expectation (Johnson & Wichern 2007) and the fourth follows from the fact $\mathbf{E}[r|X] = X\theta$.

From the law of total covariance, the variance can be written as

$$\begin{aligned}
\mathbf{E}\Big[\|\hat{\theta}_{MAP} - \mathbf{E}\Big[\hat{\theta}_{MAP}\Big]\|_2^2\Big] &= \mathbf{tr}\Big(\mathrm{Cov}(\hat{\theta}_{MAP})\Big) && (2.4.69)\\
&= \mathbf{tr}\Big(\mathbf{E}\Big[\mathrm{Cov}(\hat{\theta}_{MAP}|X)\Big] + \mathrm{Cov}(\mathbf{E}[\hat{\theta}_{MAP}|X])\Big).
\end{aligned}$$

The term $\mathrm{Cov}(\hat{\theta}_{MAP}|X)$ equals

$$\begin{aligned}
\mathrm{Cov}(\hat{\theta}_{MAP}|X) &= \tilde{S}^{-1}X^T\mathrm{Cov}(r|X)X\tilde{S}^{-1} && (2.4.70)\\
&= \tilde{S}^{-1}X^TX\tilde{S}^{-1}. && (2.4.71)
\end{aligned}$$

In the second line we use the fact that $\mathrm{Cov}(r|X) = \mathbf{I}$. The term $\mathbf{E}[\hat{\theta}_{MAP}|X]$ was used to derive equation 2.4.68 and equals $\tilde{S}^{-1}X^TX\theta$.

Substituting the relevant quantities into equation 2.4.64, we find that the mean squared error of the MAP is

$$\begin{aligned}
\mathbf{E}\Big[\|\hat{\theta}_{MAP} - \theta\|_2^2\Big] &= \|(\mathbf{E}\Big[\tilde{S}^{-1}X^TX\Big] - I)\theta\|_2^2 + \\
&\quad \mathbf{tr}\Big(\mathbf{E}\Big[\tilde{S}^{-1}X^TX\tilde{S}^{-1}\Big] + \mathrm{Cov}(\tilde{S}^{-1}X^TX\theta)\Big). && (2.4.72)
\end{aligned}$$

The MSE of the MPELE can be computed in a similar fashion. The bias of the

MPELE equals

$$\begin{aligned}
\mathbf{E}\Big[(\hat{\theta}_{\text{MPELE}} - \theta)\Big] &= \mathbf{E}\Big[\hat{\theta}_{\text{MPELE}}\Big] - \theta && (2.4.73)\\
&= \mathbf{E}\Big[\mathbf{E}\Big[\hat{\theta}_{\text{MPELE}}|X\Big]\Big] - \theta && (2.4.74)\\
&= \mathbf{E}\Big[(N+cp)^{-1}X^T\mathbf{E}\Big[r|X\Big]\Big] - \theta && (2.4.75)\\
&= (N+cp)^{-1}\mathbf{E}\Big[X^TX\theta\Big] - \theta && (2.4.76)\\
&= (N+cp)^{-1}N\theta - \theta. && (2.4.77)
\end{aligned}$$

To derive the fourth line we have again used the fact $\mathbf{E}[r|X] = X\theta$ to show that $\mathbf{E}[\hat{\theta}_{\text{MPELE}}|X] = (N+cp)^{-1}X^TX\theta$. The fifth line follows by the definition $\mathbf{E}\Big[X^TX\Big] = N\mathbf{I}$. To compute the variance we again use the law of total covariance which requires the computation of a term $\mathbf{E}\Big[\text{Cov}(\hat{\theta}_{\text{MPELE}}|X)\Big]$,

$$\begin{aligned}
\mathbf{E}\Big[\text{Cov}(\hat{\theta}_{\text{MPELE}}|X)\Big] &= \mathbf{E}\Big[(N+cp)^{-1}X^T\text{Cov}(r|X)X(N+cp)^{-1}\Big] && (2.4.78)\\
&= \mathbf{E}\Big[(N+cp)^{-1}X^TX(N+cp)^{-1}\Big] && (2.4.79)\\
&= (N+cp)^{-2}N. && (2.4.80)
\end{aligned}$$

We use the fact that $\text{Cov}(r|X) = \mathbf{I}$ to derive the second line and the definition $\mathbf{E}\Big[X^TX\Big] = N\mathbf{I}$ to derive the third.

Using the bias-variance decomposition of the MSE, equation 2.4.64, we find that the mean squared error of the MPELE estimator is

$$\mathbf{E}\Big[\|\hat{\theta}_{\text{MPELE}} - \theta\|_2^2\Big] = \|(\frac{N}{N+cp} - 1)\theta\|_2^2 + \frac{1}{(N+cp)^2}\Big(Np + \mathbf{tr}\Big(\text{Cov}(X^TX\theta)\Big)\Big).$$

The term $\mathbf{tr}\Big(\text{Cov}(X^TX\theta)\Big)$ can be simplified by taking advantage of the fact that rows of X are i.i.d normally distributed with mean zero, so their fourth central moment can be written as the sum of outer products of the second central moments (Johnson

& Wichern 2007).

$$\mathbf{E}\Big[(X^TX)_{ij}(X^TX)_{kl}\Big] = N\mathbf{E}\Big[X_{1i}X_{1j}X_{1k}X_{1l}\Big] + N(N-1)\delta_{ij}\delta_{kl} \tag{2.4.81}$$

$$= N(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) + N(N-1)\delta_{ij}\delta_{kl}. \tag{2.4.82}$$

We then have

$$\mathbf{E}\Big[\|\hat{\theta}_{\text{MPELE}} - \theta\|_2^2\Big] = \|(\frac{N}{N+cp} - 1)\theta\|_2^2 + \frac{1}{(N+cp)^2}\Big(Np(1+\|\theta\|_2^2) + N\|\theta\|_2^2\Big). \tag{2.4.83}$$

Without regularization ($c = 0$) the MSE expressions simplify significantly

$$\mathbf{E}\Big[\|\hat{\theta}_{MAP} - \theta\|_2^2\Big] = \mathbf{tr}\Big(\mathbf{E}\Big[(X^TX)^{-1}\Big]\Big) \tag{2.4.84}$$

$$\mathbf{E}\Big[\|\hat{\theta} - \theta\|_2^2\Big] = \frac{\theta^T\theta + p(\theta^T\theta + 1)}{N}. \tag{2.4.85}$$

Noting that $(X^TX)^{-1}$ is distributed according to an inverse Wishart distribution with mean $\frac{1}{N-p-1}\mathbf{I}$ (Johnson & Wichern 2007), we recover equations 2.2.38 and 2.2.40.

For $c \neq 0$ we calculate the MSE for both estimators in the limit $N, p \to \infty$ but $\frac{p}{N}$ remains a finite scalar ρ. In this limit equation 2.4.83 reduces to

$$\mathbf{E}\Big[\|\hat{\theta}_{\text{MPELE}} - \theta\|_2^2\Big] \to \rho\frac{(1+\theta^T\theta)}{(1+c\rho)^2} + \theta^T\theta\Big(1 - \frac{1+2c\rho}{(1+c\rho)^2}\Big). \tag{2.4.86}$$

To calculate the limiting MSE value for the MAP we work in the eigenbasis of X^TX. This allows us to take advantage of the Marchenko-Pastur law (Marchenko 1967) which states that in the limit $N, p \to \infty$ but $\frac{p}{N}$ remains finite, the eigenvalues of $\frac{X^TX}{N}$ converge to a continuous random variable with known distribution. We denote the matrix of eigenvectors of X^TX by O

$$X^TX = OLO^T, \tag{2.4.87}$$

with the diagonal matrix L containing the eigenvalues of X^TX along the diagonal.

Evaluating the first and last term in the MAP MSE (equation 2.4.72) leads to the result

$$\begin{aligned}\|(\mathbf{E}\Big[\tilde{S}^{-1}X^TX\Big]-I)\theta\|_2^2+\mathbf{tr}\Big(\mathrm{Cov}(\tilde{S}^{-1}X^TX\theta)\Big) &= \|\theta\|_2^2-2\theta^T\mathbf{E}\Big[\tilde{S}^{-1}X^TX\Big]\theta \\ &\quad +\mathbf{E}\Big[\|\tilde{S}^{-1}X^TX\theta\|_2^2\Big]. \quad (2.4.88)\end{aligned}$$

Evaluating the last term in the above equation yields

$$\tilde{S}^{-1}X^TX\theta = O(L+cpI)^{-1}LO^T\theta, \quad (2.4.89)$$

$$\mathbf{E}\Big[\|\tilde{S}^{-1}X^TX\theta\|_2^2\Big] = \theta^T\mathbf{E}\Big[O\Big((L+cpI)^{-1}L\Big)^2O^T\Big]\theta \quad (2.4.90)$$

$$= \mathbf{tr}\Big(\mathbf{E}\Big[\Big((L+cpI)^{-1}L\Big)^2\Big]\mathbf{E}[O^T\theta\theta^TO|L]\Big) \quad (2.4.91)$$

In the last line we have used the law of total expectation and taken the expectation first conditioned on the eigenvalues and then with respect to the eigenvalues. Since the distribution of x is symmetric with identity covariance, the vector $O^T\theta$ is uniformly distributed on the sphere of radius $\|\theta\|_2^2$ given L and $\mathbf{E}[O^T\theta\theta^TO|L] = \|\theta\|_2^2/p\mathbf{I}$. So that

$$\mathbf{E}\Big[\|\tilde{S}^{-1}X^TX\theta\|_2^2\Big] = \|\theta\|_2^2\mathbf{tr}\Big(\mathbf{E}\Big[\Big((L+cpI)^{-1}L\Big)^2\Big]\Big).$$

Substituting the above result in equation 2.4.88 we find

$$\|(\mathbf{E}\Big[\tilde{S}^{-1}X^TX\Big]-I)\theta\|_2^2+\mathbf{tr}\Big(\mathrm{Cov}(\tilde{S}^{-1}X^TX\theta)\Big) = \|\theta\|_2^2/p\mathbf{E}\Big[\mathbf{tr}\Big((L+cpI)^{-1}L-1)^2\Big)\Big].$$

Using the result given above and noting that $\mathbf{tr}\Big(\mathbf{E}\Big[\tilde{S}^{-1}X^TX\tilde{S}^{-1}\Big]\Big) = \mathbf{tr}\Big(\mathbf{E}\Big[(L+$

$cp\mathbf{I})^{-2}L\Big]\Big)$, the MAP can then be written as

$$\mathbf{E}\Big[\|\hat{\theta}_{MAP}-\theta\|_2^2\Big] = \mathbf{tr}\Big(\mathbf{E}\Big[(L+cp\mathbf{I})^{-2}L\Big]\Big)+\|\theta\|_2^2\mathbf{E}\Big[\mathbf{tr}\Big((L+cpI)^{-1}L-1)^2\Big)\Big]. \tag{2.4.92}$$

Taking the limit $N,p\to\infty$ with $\frac{p}{N}$ finite,

$$\mathbf{E}\Big[\|\hat{\theta}_{MAP}-\theta\|_2^2\Big] \to \rho\mathbf{E}\Big[\frac{l}{(l+c\rho)^2}\Big]+\|\theta\|_2^2\mathbf{E}\Big[\Big(\frac{l}{(l+c\rho)}-1\Big)^2\Big], \tag{2.4.93}$$

where l is a continuous random variable with probability density function $\frac{d\mu}{dl}$ found by the Marchecko-Pastur law

$$\frac{d\mu}{dl} = \frac{1}{2\pi l\rho}\sqrt{(b-l)(l-a)}\mathcal{I}_{[a,b]}(l) \tag{2.4.94}$$

$$a\Big(\rho\Big) = (1-\sqrt{\rho})^2 \tag{2.4.95}$$

$$b\Big(\rho\Big) = (1+\sqrt{\rho})^2. \tag{2.4.96}$$

Using equation 2.4.94 we can numerically evaluate the MAP MSE. The results are plotted in figure 2.2. To show the accuracy of approximating real data, with finite N and p by the limit $N,p\to\infty$ with$\frac{p}{N}$, figure 2.7 compares the output of equations 2.2.38 and 2.2.40 as a function of N with their limiting values (equations 2.2.41 and 2.2.42). The quality of the approximation depends on the estimator used, however for values of N and p often used in real data sets ($\frac{p}{N}\sim 0.01, N\sim 10000$) this is an excellent approximation.

2.4.2 Computing the MPELE for an inhomogeneous Poisson process

We consider priors of the form $\log(f(\theta)) = -\frac{\theta^T R\theta}{2}$ and stimuli such that $\mathbf{E}\Big[x\Big] = 0, \mathbf{E}\Big[x(x)^T\Big] = C$. The MPELE is given by solving equation 2.2.30, which in this case

equals

$$\hat{\theta}_{\text{MPELE}} = \arg\max_{\theta,\theta_0} \; \theta_0 \sum_{n=1}^{N} r_n + (X\theta)^T r - N\mathbf{E}\Big[\exp(\theta_0)\exp((x)^T\theta)\Big] - \frac{(\theta)^T R\theta}{2}.$$

Since the covariance of x is C and we approximate $x\theta$ as Normally distributed, $x\theta \sim \mathcal{N}(0, \theta^T C\theta)$, we can analytically calculate the expectation yielding

$$\hat{\theta}_{\text{MPELE}} = \arg\max_{\theta,\theta_0} \; \theta_0 \sum_{n=1}^{N} r_n + (X\theta)^T r - N\exp(\theta_0)\exp\left(\frac{\theta^T C\theta}{2}\right) - \frac{\theta^T R\theta}{2} \tag{2.4.97}$$

Optimizing with respect to θ_0 we find

$$\exp(\theta_0^*) = \frac{\sum_{n=1}^{N} r_n}{N}\exp(-\frac{\theta^T C\theta}{2}). \tag{2.4.98}$$

Inserting θ_0^* into equation 2.4.97 leaves the following quadratic optimization problem

$$\hat{\theta}_{\text{MPELE}} = \arg\max_{\theta} \; -\frac{\theta^T\Big(C\sum_{n=1}^{N} r_n + R\Big)\theta}{2} + \theta^T X^T r. \tag{2.4.99}$$

The solution of which is equation 2.2.31.

2.4.3 Calculating $\hat{R}$ for the LNP model

For the LNP model with $\mathbf{E}\Big[G(x^T\theta)\Big]$ approximated as in equation 2.2.22 and $f(\theta|R) = \mathcal{N}(0, R^{-1})$, the Laplace approximation (2.2.52) to the marginal likelihood yields

$$\begin{aligned}
\log F(R) &\approx \sum_{n=1}^{N} \hat{\theta}_0^* r_n + (X\hat{\theta}_{\text{MPELE}})^T r - N\Delta t \exp(\hat{\theta}_0^*) \exp\left(\frac{(\hat{\theta}_{\text{MPELE}})^T C \hat{\theta}_{\text{MPELE}}}{2}\right) \\
&\quad + \log\Big(f(\hat{\theta}_{\text{MPELE}}|R)\Big) - \frac{1}{2}\log(\det(-H(\hat{\theta}_{\text{MPELE}}))) \qquad (2.4.100) \\
&= -\frac{(\hat{\theta}_{\text{MPELE}})^T C \hat{\theta}_{\text{MPELE}}}{2} \sum_{n=1}^{N} r_n + (X\hat{\theta}_{\text{MPELE}})^T r - \frac{(\hat{\theta}_{\text{MPELE}})^T R \hat{\theta}_{\text{MPELE}}}{2} \\
&\quad + \frac{1}{2}\log(\det(R)) - \frac{1}{2}\log(\det(-H(\hat{\theta}_{\text{MPELE}}))) \qquad (2.4.101) \\
&= -\frac{(\hat{\theta}_{\text{MPELE}})^T (CN_s + R) \hat{\theta}_{\text{MPELE}}}{2} + (X\hat{\theta}_{\text{MPELE}})^T r + \frac{1}{2}\log(\det(R)) \\
&\quad - \frac{1}{2}\log(\det(-H(\hat{\theta}_{\text{MPELE}}))), \qquad (2.4.102)
\end{aligned}$$

where the second line follows from substituting in equation 2.4.98 and we have denoted $\sum_{n=1}^{N} r_n$ as N_s in the third line. Note that from the definition of the L2 regularized MPELE (equation 2.2.31) we can write

$$X^T r = (CN_s + R)\hat{\theta}_{\text{MPELE}}, \qquad (2.4.103)$$

and simplify the first two terms

$$\begin{aligned}
-\frac{(\hat{\theta}_{\text{MPELE}})^T (CN_s + R) \hat{\theta}_{\text{MPELE}}}{2} + (X\hat{\theta}_{\text{MPELE}})^T r &= \frac{(\hat{\theta}_{\text{MPELE}})^T (CN_s + R) \hat{\theta}_{\text{MPELE}}}{2} \\
&= \frac{(X^T r)^T (CN_s + R)^{-1} X^T r}{2}.
\end{aligned} \qquad (2.4.104)$$

Noting that $C = \mathbf{I}, R = \beta\mathbf{I}$ equation 2.4.104 simplifies further to $\frac{1}{2}\Big(N_s + \beta\Big)^{-1} q$ with $q \equiv \|X^T r\|_2^2$. Using the fact that the profile Hessian is $-H(\hat{\theta}_{\text{MPELE}}) = CN_s + R$ and

the assumptions $C = \mathbf{I}, R = \beta\mathbf{I}$ the last two terms in equation 2.4.102 reduce to

$$\frac{1}{2}\log(\det(R)) - \frac{1}{2}\log(\det(-H(\hat{\theta}_{\text{MPELE}}))) \;=\; \frac{p}{2}\log(\beta) - \frac{p}{2}\log(N_s + \beta). \tag{2.4.105}$$

Combining these results we find

$$\log F(R) \;\approx\; \frac{1}{2}\Big(N_s + \beta\Big)^{-1} q + \frac{p}{2}\log(\beta) - \frac{p}{2}\log(N_s + \beta). \tag{2.4.106}$$

Taking the derivative of equation 2.4.106 with respect to β we find that the critical points, $\hat{\beta}_c$ obey

$$0 \;=\; -\frac{q}{(N_s + \hat{\beta}_c)^2} + p\frac{N_s}{(N_s + \hat{\beta}_c)\hat{\beta}_c}, \tag{2.4.107}$$

$$\hat{\beta}_c \;=\; \Big(\frac{p}{\frac{q}{N_s^2} - \frac{p}{N_s}}, \infty\Big). \tag{2.4.108}$$

If $p \geq \frac{q}{N_s}$, the only critical point is ∞ since β is constrained to be positive. When $p < \frac{q}{N_s}$, the critical point $\hat{\beta}_c = \frac{p}{\frac{q}{N_s^2} - \frac{p}{N_s}}$ is the maximum since $\log F$ (equation 2.4.106) evaluated at this point is greater than $\log F$ evaluated at ∞

$$\log F\Big(\frac{p}{\frac{q}{N_s^2} - \frac{p}{N_s}}\Big) = \frac{1}{2}\Big(\frac{q}{N_s} - p + p\log\Big(\frac{p}{q}N_s\Big)\Big) \geq 0 = \lim_{\beta\to\infty}\log F. \tag{2.4.109}$$

Therefore $\hat{R}$ satisfies equation 2.2.57 in the text.

We can derive similar results for a more general case if C and R are diagonalized by the same basis. If we denote this basis by M, we then have the property that the profile Hessian $CN_s + R = M(D^c N_s + D^r)M^T$ where D^c and D^r are diagonal matrices containing the eigenvalues of C and R. In this case the last two terms of equation 2.4.102 reduce to

$$\frac{1}{2}\log(\det(R)) - \frac{1}{2}\log(\det(-H(\hat{\theta}_{\text{MPELE}}))) \;=\; \frac{1}{2}\sum_{i=1}^{p}\log\Big(\frac{D^r_{ii}}{D^c_{ii}N_s + D^r_{ii}}\Big).$$

Defining $X^T r$ rotated in the coordinate system specified by M as $\tilde{q} = M^T X^T r$, equation 2.4.104 simplifies to

$$\frac{(X^T r)^T (CN_s + R)^{-1} X^T r}{2} = \frac{1}{2} \sum_{i=1}^{p} \tilde{q}_i^2 (D_{ii}^c N_s + D_{ii}^r)^{-1}. \tag{2.4.110}$$

Combining terms we find

$$\log F(R) \approx \frac{1}{2}\Big(\sum_{i=1}^{p} \tilde{q}_i^2 (D_{ii}^c N_s + D_{ii}^r)^{-1} + \log\Big(\frac{D_{ii}^r}{D_{ii}^c N_s + D_{ii}^r}\Big)\Big). \tag{2.4.111}$$

Taking the gradient of the above equation with respect to the eigenvalues of R, D_{jj}^{r*}, we find that the critical points obey

$$0 = -\frac{\tilde{q}_j^2}{(D_{jj}^c N_s + D_{jj}^{r*})^2} + \frac{D_{jj}^c N_s}{D_{jj}^{r*}(D_{jj}^c N_s + D_{jj}^{r*})} \tag{2.4.112}$$

$$D_{jj}^{r*} = \Big(\frac{(D_{jj}^c N^s)^2}{\tilde{q}_j^2 - D_{jj}^c N_s}, \infty\Big). \tag{2.4.113}$$

If $D_{jj}^c N_s \geq \tilde{q}_j^2$, the only critical point is ∞ since D_{jj}^{r*} is constrained to be positive (R is constrained to be positive definite).

2.4.4 Calculation of Bayesian credible regions

For each element of θ we define the 95% credible region, (A_i, B_i) for $i = 1, ..., p$ by

$$\int_{-\infty}^{A_i} p(\theta|X, r) d\theta_i = 0.05/2 \tag{2.4.114}$$

$$\int_{B_i}^{\infty} p(\theta|X, r) d\theta_i = 0.05/2. \tag{2.4.115}$$

We estimated the credible regions using sample quantiles. After running the HMC algorithm for T iterations, the samples $\theta^{(1)}, ..., \theta^{(T)}$ were sorted element-wise such that $\theta_i^{(1)} \leq \theta_i^{(2)} ... \leq \theta_i^{(T)}$ for $i = 1, ..., p$. We then take $\theta^{([T0.05/2]+1)}$ and $\theta^{([T(1-0.05/2)]+1)}$ as estimates for A and B ([y] denotes the largest integer less than

or equal to y).

2.4.5 Simulated and real neuronal data applications

Stimuli are refreshed at a rate of 120 Hz and responses are binned at this rate (figure 2.3) or at 10 times (figure 2.4) this rate. Stimulus receptive fields are fit with 81(figure 2.3) or 25(figure 2.4) spatial components and ten temporal basis functions, giving a total of 81x10 = 810 or 25x10=250 stimulus filter parameters. Five basis functions are delta functions with peaks centered at the first 5 temporal lags while the remaining 5 are raised cosine 'bump' functions ((Pillow et al. 2008)). Before running PCG in figure 2.3, we optimize the location and scale of the MELE stimulus filter using the corresponding GLM log-likelihood. The self-history filter shown in figure 2.4 is parameterized by 4 cosine 'bump' functions and a refractory function that is negative for the first stimulus time bin and zero otherwise. The coupling coefficient temporal components are modeled with a decaying exponential of the form, $\exp(-b\tau)$, with b set to a value, 0.5, which captures the time-scale of cross-correlations seen in the data. We use the activity of 100 neighboring cells yielding a total of 100 coupling coefficient parameters, 5 self-history parameters, 250 stimulus parameters, and 1 offset parameter (356 parameters in total). The regularization coefficients used in figure 2.3B and 2.4 are found via cross-validation on a novel two minute (14,418 samples) data set. Model performance is evaluated using 2 minutes of data not used for determining model parameters or regularization coefficients. To calculate the information rate we take the difference of the log-likelihood under the model and log-likelihood under a homogeneous Poisson process divided by the total time ($\frac{14{,}418}{120}$ seconds) (Paninski, et al. 2004b).

The true filter shown in figure 2.5 is a difference between two zero-mean Gaussian functions with standard deviations of 1 and 3 and discretized using 250 bins. The true filter norm is set to 10. Stimuli are simulated from a unit variance, zero-mean normal distribution and responses are simulated using an LNP model.

Acknowledgements

We thank Eftychios A. Pnevmatikakis and Ari Pakman for helpful comments and discussions. We thank Columbia University Information Technology and the Office of the Executive Vice President for Research for providing the computing cluster used in this study. LP is funded by a McKnight scholar award, an NSF CAREER award, and grants from the NEI and DARPA.

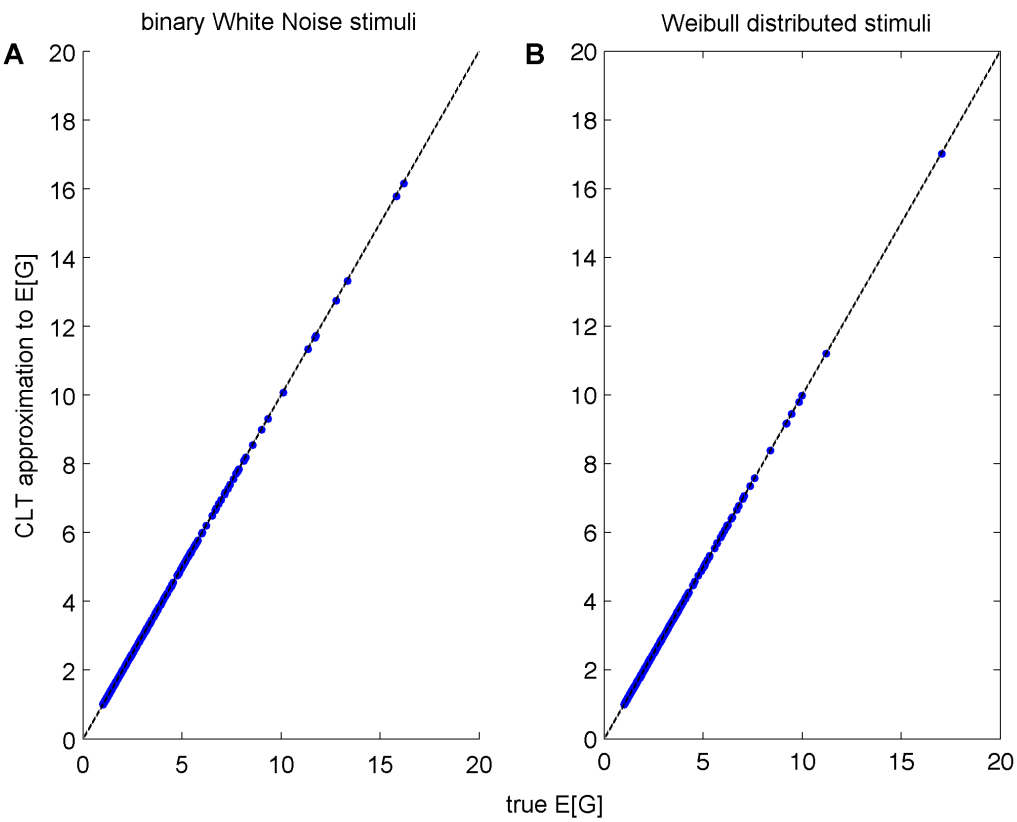

Figure 2.1: Numerical examples showing the accuracy of approximating the projection $x^T\theta$ as a normaly distributed random variable. We plot the output of $\mathbf{E}\Big[G\Big]$ (equation 2.2.18 in the text) with (vertical axis) and without (horizontal axis) approximating $p(q)$ as normal. G is taken to be exponential corresponding to Poisson responses. x has a binary white noise distribution (A) or Weilbull distribution (B). Without the normal approximation, $\mathbf{E}\Big[G\Big]$ was found by Monte-Carlo sampling using 12000 simulated samples of x for each projection. x is a 600 dimensional vector with mean 0.36 in A or drawn from a Weilbull distribution (B) with scale and shape parameters 0.15, 0.5. θ was restricted to have a Gaussian profile, in order to approximate filter shapes often found in applications, with mean 0, and with random variance and random amplitude.

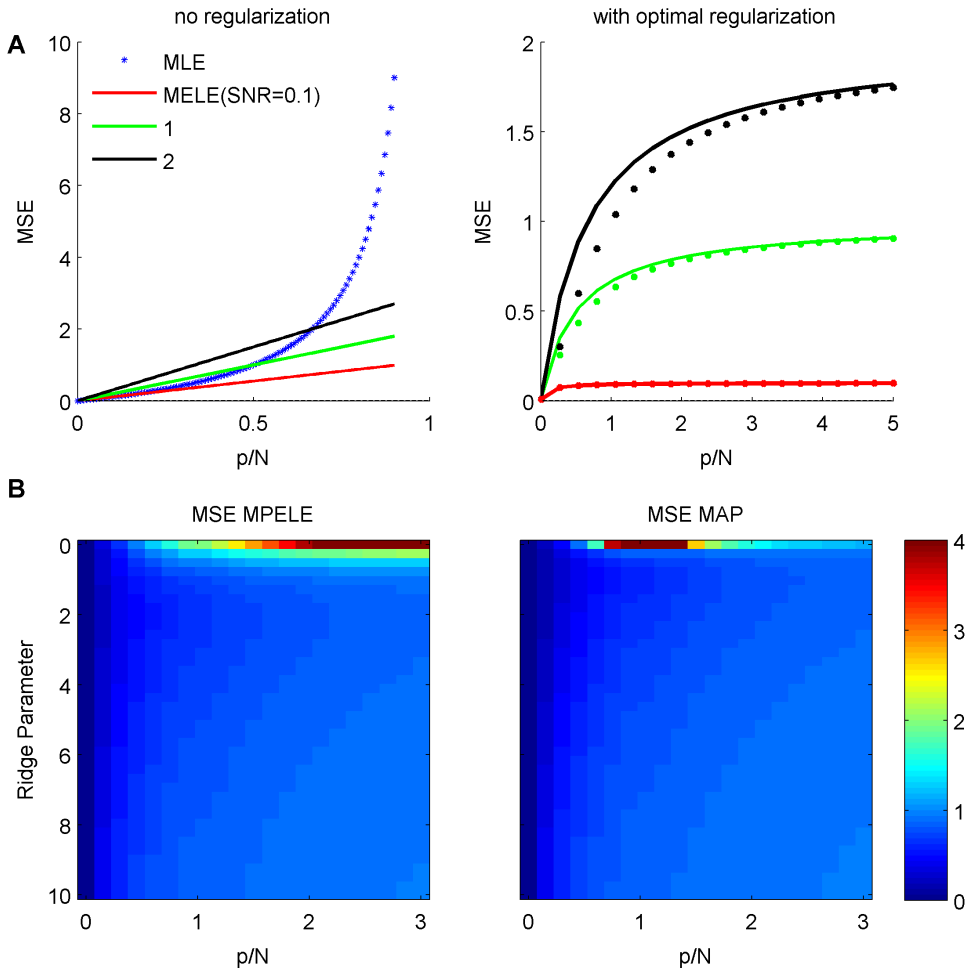

Figure 2.2: The MELE comes with a minimal cost in accuracy (and sometimes improved accuracy) compared to maximum-likelihood estimators (MLE). A.) (left) The mean squared error (MSE) for the MELE (solid lines) and the MLE (dotted line) is shown as a function of the ratio of the number of parameters to number of samples. We plot results for the asymptotic case where the number of samples and dimensions goes to infinity but the ratio remains finite. Different colors denote different values for the true filter norm. The MLE mean squared error is larger than the MELE MSE when the number of samples is close to the number of dimensions. B.) MSE for both estimators when L2 regularization is added. MSE is similar for both estimators for a large range of ridge parameters and values of $\frac{p}{N}$. A.) (right) For each value of $\frac{p}{N}$, separate ridge parameters are chosen for the MPELE and MAP estimators to minimize their respective mean squared errors. Even at the optimal choice of regularization, the difference in performance between the two estimators remains small for a wide range of values of SNR and $\frac{p}{N}$. In figure 2.3, we show that when the MELE and MPELE is used on real data small differences in performance can be eliminated using fast gradient based optimization methods.

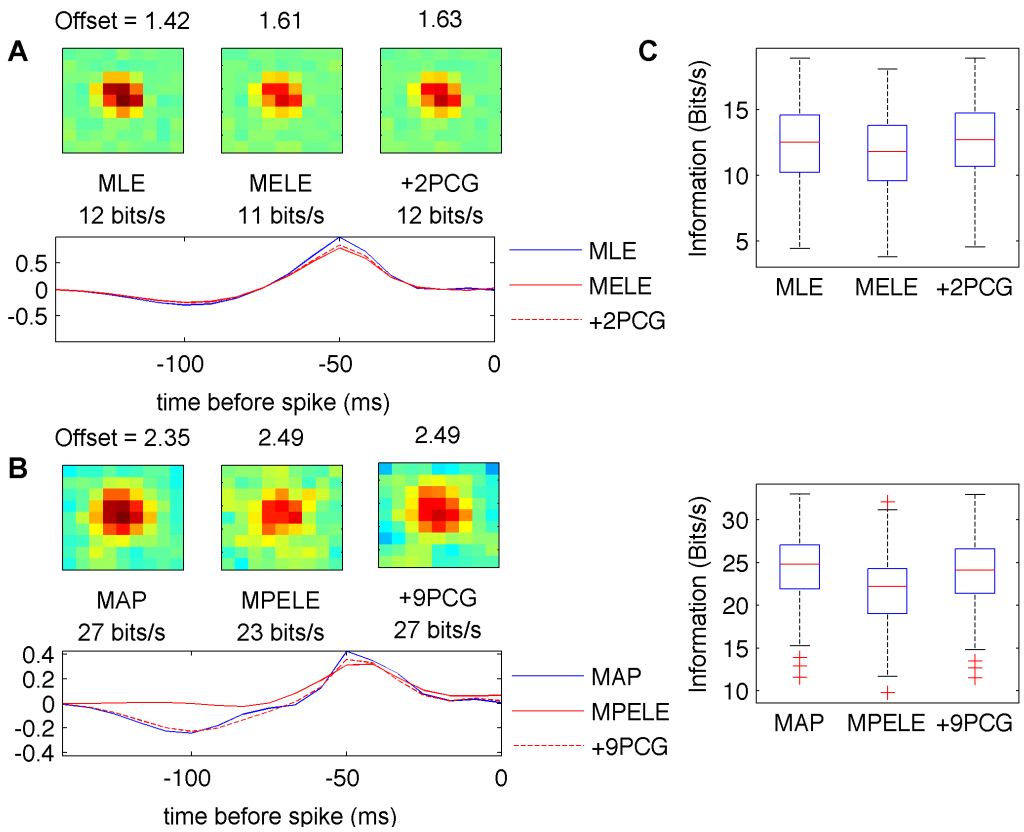

Figure 2.3: The MELE is a good initializer for finding maximum a posteriori (MAP) estimators using fast gradient based optimization methods. A.) The spatial-temporal receptive field of a randomly chosen Retinal Ganglion cell (RGC) responding to binary white-noise stimuli and fit with a linear-nonlinear Poisson (LNP) model with exponential non-linearity is shown (labeled MLE). The receptive field of the same cell fit using theMELE under an LNP model is also plotted (see label marg). The goodness-of-fit of the MLE (measured in terms of information rate, see methods) is slightly higher than that of the MELE (12 versus 11 bits/s). However this difference disappears after a couple pre-conditioned conjugate gradient (PCG) iterations using the true likelihood initialized at the MELE (see label +2PCG). Note that we are only showing spatial and temporal slices of the 9x9x10 dimensional receptive field for illustrative purposes. B.) Similar results hold when the same cell responds to 1/f correlated Gaussian noise stimuli (for correlated Gaussian responses, the MLE and MELE are both fit using an L2 regularizer). However, 9 PCG iterations instead of 2 were used to find an estimator with a goodness-of-fit equal to the MLE. C.) These conclusions hold for many cells. (top) Boxplot of information rates across a population of 91 cells, responding to binary white-noise stimuli, each fit independently using the MLE, MELE or a couple PCG iterations using the true likelihood and initialized at the MELE. (bottom) Information rates for the same population, responding to 1/f Gaussian noise, fit independently using an L2 regularized MLE, MPELE or 9 PCG iterations using the true regularized likelihood initialized at the MPELE.

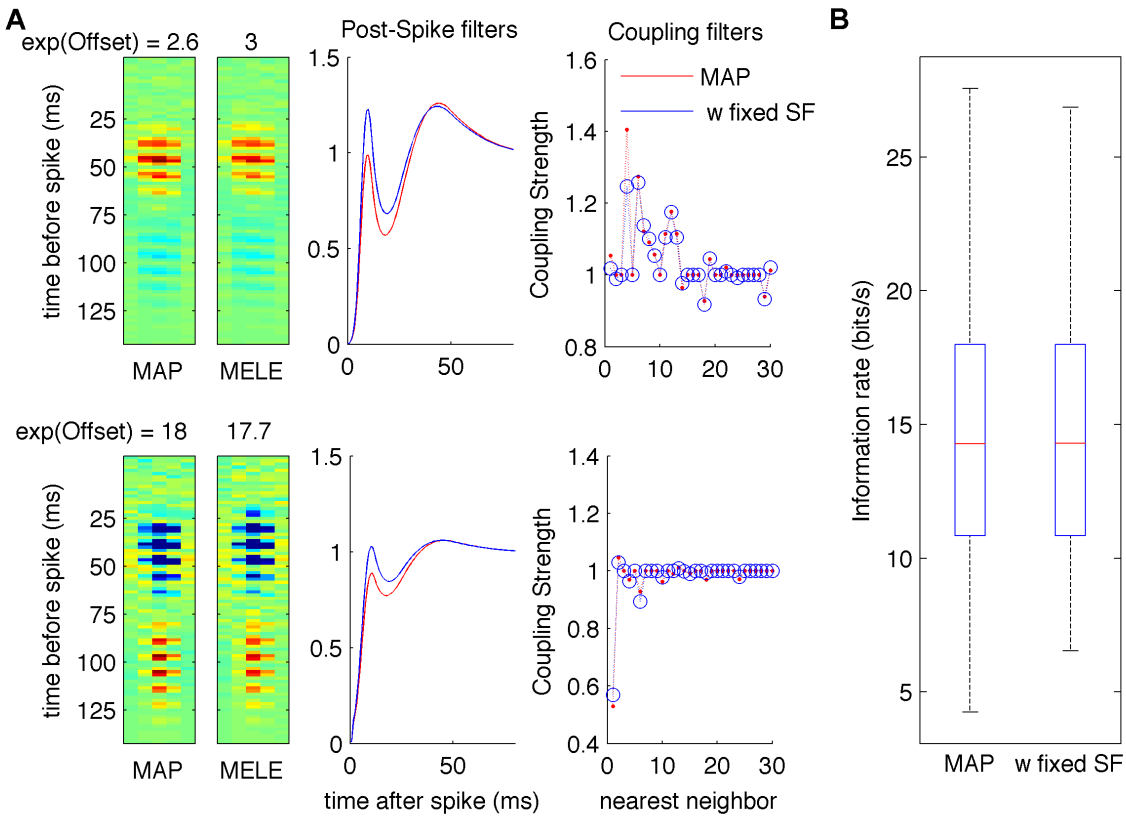

Figure 2.4: The MELE of an LNP model combined with a coordinate descent algorithm quickly determines post-spike L1 penalized coupling parameters. A.) Example stimulus, self-history, and coupling filters for two different RGC cells (top and bottom rows). MAP parameters are found using coordinate descent to optimize the true GLM log-likelihood with L1 penalized coupling filters (labeled MAP). Fast estimates of the self-history and coupling filters are found by running the same coordinate descent algorithm with the stimulus filter (SF) fixed, up to a gain, to the MELE of an LNP model(labeled 'w fixed SF'; see text for full procedure used in determining history parameters). By fixing the stimulus current using the MELE, the correct non-zero coupling parameters are found 16 times faster than the MLE. B.)Boxplots comparing the cross-validated information rate of 100 different RGC cells show that the small differences in parameters between the two estimators do not affect goodness-of-fit.

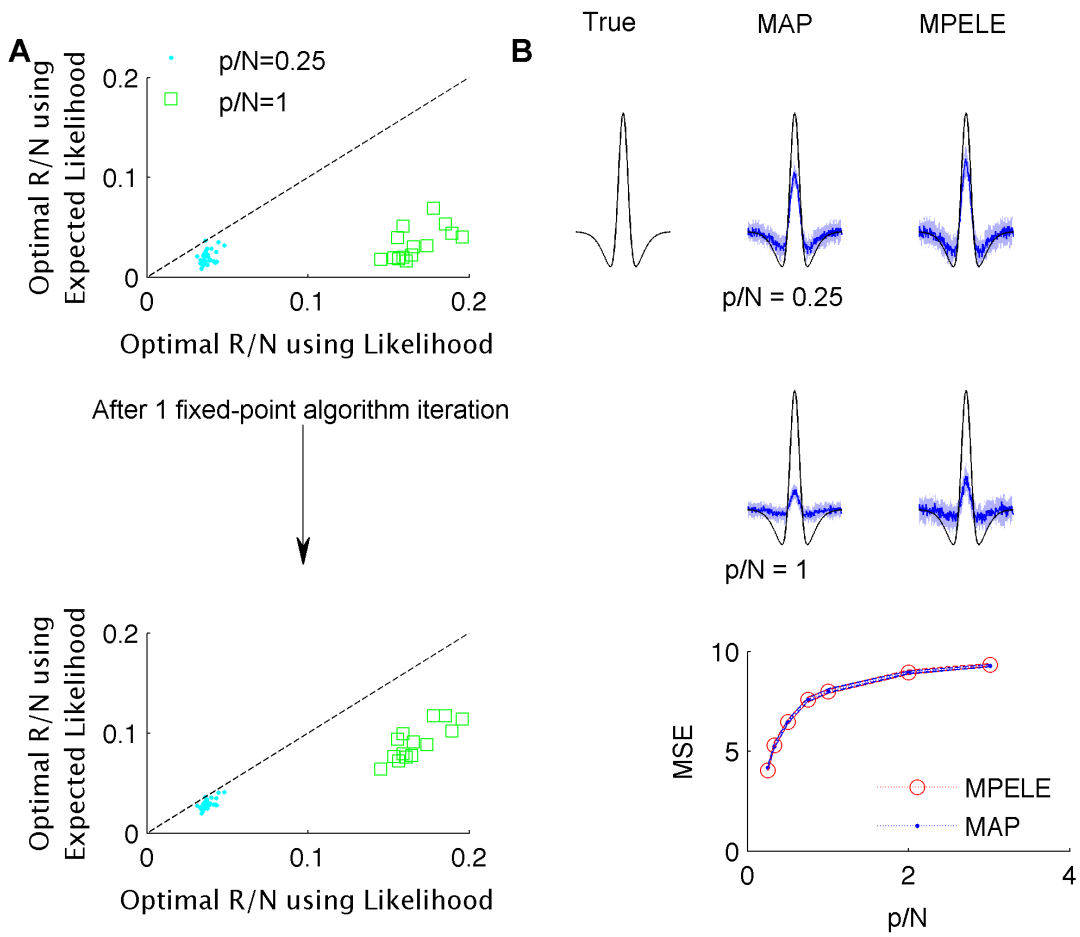

Figure 2.5: The EL can be used to quickly find Type II Maximum Likelihood filters. Thirty simulated neural responses were drawn from a linear-nonlinear Poisson (LNP) model with stimuli drawn from an independent white-noise Gaussian. The true filter (shown in black in B) has 250 parameters. A.) Optimal hyper-parameters which maximize the evidence function using the EL (top left column, vertical axis) are similar to those which maximize the GLM evidence function (top left column, horizontal axis). After a single iteration of the fixed point algorithm used to maximize the GLM evidence (see text, section Marginal likelihood calculations), the two sets of hyper-parameters match to sufficient accuracy (bottom left column). B.) Median filter estimates (blue), and absolute median deviation (light blue), using either set of hyper-parameters match for a wide range of values of $\frac{p}{N}$. Below the filter estimates we plot the mean squared error (MSE) of both filters which also matches for a wide range of values of $\frac{p}{N}$.

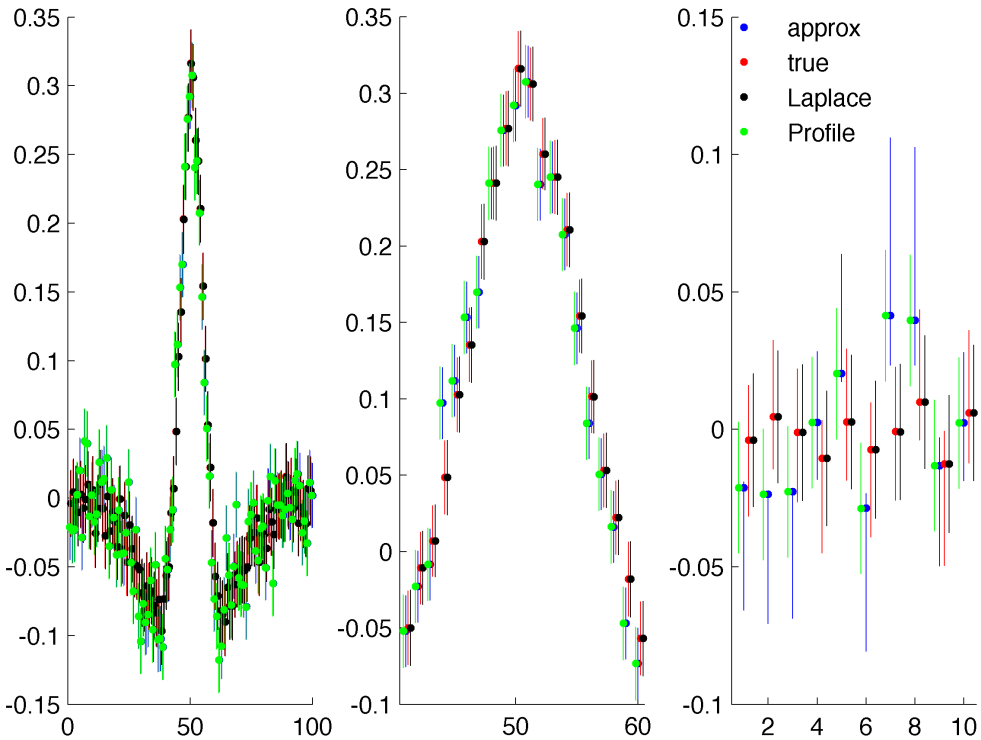

Figure 2.6: Statistics from the posterior parameter distribution approximated use the EL can be quickly computed and are in close agreement with those from the true posterior. 4000 responses were simulated from an 100 dimensional linear-nonlinear Poisson model using independent white-noise Gaussian stimuli ($\frac{p}{N} = 0.025$). 10^6 Markov Chain Monte Carlo samples, computed using Hybrid Monte Carlo, were then drawn from the posterior assuming a flat prior and using either a Poisson likelihood or the EL approximation to the Poisson. Both columns show the median vector along with 95% credible region (defined in the text) of the approximate (blue) and true (red) posterior distribution. In the middle and right column we have zoomed in around different elements for visual clarity. Statistics from both distributions are in close agreement for many elements, however some directions show discrepancies. It took 34 seconds to sample using the EL and 83 seconds using the true likelihood. Replacing the likelihood with the profile likelihood of the EL results yielded similar results to the approximate distribution (green). The Laplace approximation also provided a good approximation to the posterior (black).

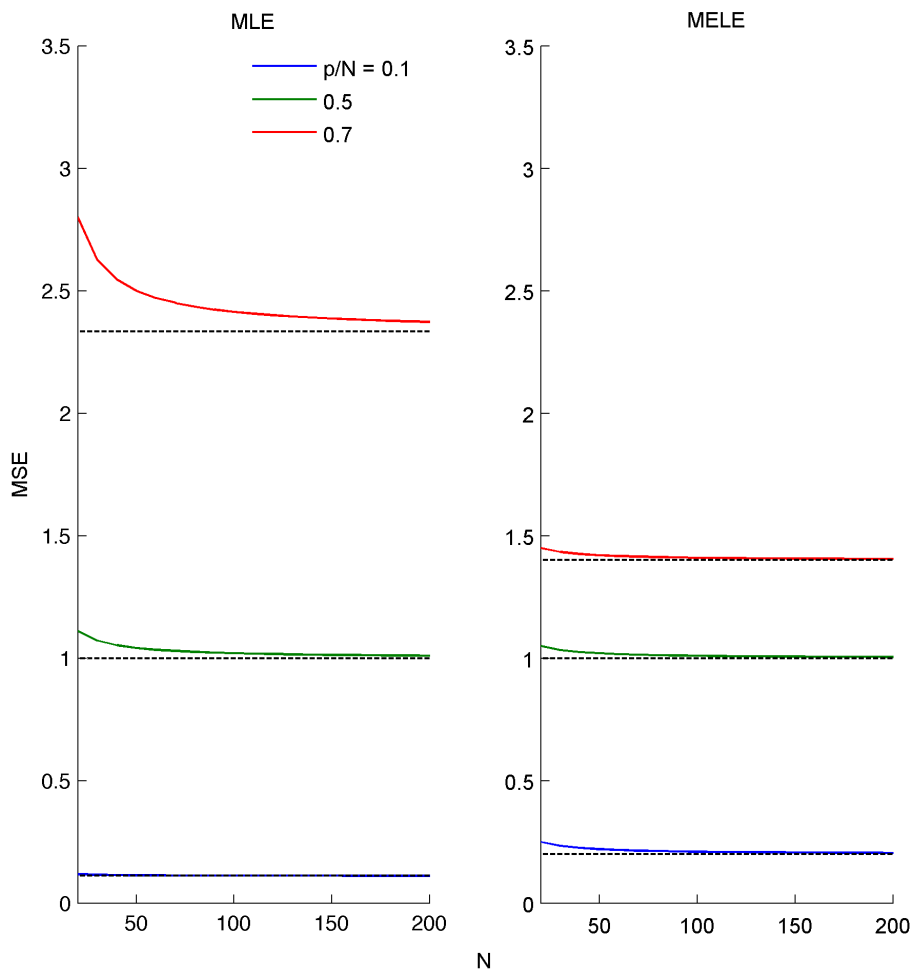

Figure 2.7: Finite sample approximations to the limit $p, N \to \infty$, $\frac{p}{N} \to \rho$. We plot the true mean-squared error (MSE) (colored lines) as a function of finite sample size for different values of $\frac{p}{N}$ for both the MLE (left column) and MELE (right column). Black dashed lines show MSE when $p, N \to \infty$, $\frac{p}{N} \to \rho$. The quality of the approximation does not seem to depend on $\frac{p}{N}$ for the MELE and is within 1% accuracy after about 100 samples. For the MLE, this approximation depends on $\frac{p}{N}$. However, while the quality of the approximation depends on the estimator used, for values often used in real data sets ($\frac{p}{N} \sim 0.01$, $N \sim 10000$) this is an excellent approximation.

Chapter 3

Incorporating Naturalistic Correlation Structure Improves Spectrogram Reconstruction From Neuronal Activity in the Songbird Auditory Midbrain

The work described in this chapter was a collaboration with many people, whose contributions I gratefully acknowledge. The computational and theoretical work was done by myself and Liam Paninski with helpful advice and feedback by Yashar Ahmadian and Joseph W. Schumacher. All the experiments described herein were conducted in collaboration with Sarah Woolley and her lab members David Schneider and Joseph W. Schumacher. The published manuscript was written by myself, Liam Paninski and Sarah Woolley .

This work was published as:

A.D. Ramirez, Y. Ahmadian, J.W. Schumacher, D. Schneider, S. Woolley and L.

Paninski (2011). 'Incorporating Naturalistic Correlation Structure Improves Spectrogram Reconstruction From Neuronal Activity in the Songbird Auditory Midbrain'. *J. Neuroscience* **31**(10):38283842.

Abstract

Birdsong is comprised of rich spectral and temporal organization which might be used for vocal perception. To quantify how this structure could be used, we have reconstructed birdsong spectrograms by combining the spike trains of zebra finch auditory midbrain neurons with information about the correlations present in song. We calculated maximum a posteriori (MAP) estimates of song spectrograms using a generalized linear model of neuronal responses and a series of prior distributions, each carrying different amounts of statistical information about zebra finch song. We found that spike trains from a population of MLd neurons combined with an uncorrelated Gaussian prior can estimate the amplitude envelope of song spectrograms. The same set of responses can be combined with Gaussian priors that have correlations matched to those found across multiple zebra finch songs to yield song spectrograms similar to those presented to the animal. The fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations than temporal correlations. However the best reconstructions combine MLd responses with both spectral and temporal correlations.

3.1 Introduction

Understanding the neural mechanisms that subserve vocal perception and recognition remains a fundamental goal in auditory neuroscience (Eggermont 2001, Theunissen & Shaevitz 2006). The songbird has emerged as a particularly useful animal model for pursuing this goal because its complex vocalizations are used for communication

(Catchpole & Slater 1995, Simmons, et al. 2002). Behavioral experiments have shown that songbirds can discriminate between similar, behaviorally relevant sounds (Lohr & Dooling 1998, Shinn-Cunningham, et al. 2007), use song for establishing territorial boundaries (Peek 1972, Godard 1991) and in mate preference (O'Loghlen & Beecher 1997, Hauber, et al. 2010). While the ethological importance of songbird vocalizations is well known, the neural basis underlying vocal recognition remains unknown.

The idea that song is processed by neurons which selectively respond to features of the song's time-varying amplitude spectrum (spectrogram) has been quantified by modeling neuronal responses using spectrotemporal receptive fields (STRFs) (Eggermont, et al. 1983, Decharms, et al. 1998, Theunissen, et al. 2000, Sen, et al. 2001, Woolley, et al. 2006, Calabrese et al. 2011). These models can successfully predict neuronal responses to novel stimuli with a high degree of accuracy. In particular, neurons in the auditory midbrain, the mesencephalicus lateral dorsalis (MLd), region have STRFs that can be categorized into independent functional groups which may function in detecting perceptual features in song such as pitch, rhythm and timbre (Woolley, et al. 2009). Midbrain responses from single and multiple neurons have also been used, without the STRF model, to discriminate among conspecific songs (Schneider & Woolley 2010).

These results provide compelling evidence that zebra finch auditory midbrain neurons are tuned to specific spectrotemporal features that could be important for song recognition. Here, we tested whether responses encode enough information about song so that an 'ideal observer' of MLd spike trains could reconstruct song spectrograms. This method of assessing the information about stimuli preserved in neural responses by reconstructing the stimulus is well studied (Hesselmans & Johannesma 1989, Bialek, et al. 1991, Rieke, et al. 1995, Rieke et al. 1997, Mesgarani, et al. 2009, Pillow, et al. 2011, Koyama, et al. 2010) and some of the earliest applications have been in the auditory system. Hesselmans and Johannesma (1989) created coarse reconstructions of a grassfrog mating call, represented using a trans-

formation known as the Wigner coherent spectrotemporal intensity density, using neural responses from the frog auditory midbrain. Rieke et al (1995) used stimulus reconstruction to show that auditory nerve fibers in the frog encode stimuli with naturalistic amplitude spectra more efficiently than broadband noise and Mesgarani et al (2009) have recently used stimulus reconstruction to study the effects of behavioral state on responses properties of ferret auditory cortex.

Like most natural sounds, zebra finch songs have highly structured correlations across frequency and time, statistical redundancies that the nervous system might use for perceiving sound (Singh & Theunissen 2003). To test how these statistical redundancies could be used in song recognition we asked whether reconstructions based on MLd responses, and a novel Generalized Linear Model of these responses (Calabrese et al., 2010), improve when responses are combined with prior knowledge of correlations present across zebra finch songs. We tested whether the fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations rather than temporal correlations and we examined how the filtering properties of MLd neurons affect reconstruction. Finally we compare spectrogram reconstructions under a Generalized Linear Model of responses to reconstructions based on the more common method of linear regression.

3.2 Methods

All procedures were in accordance with the NIH and Columbia University Animal Care and Use Policy. Thirty-six Adult male zebra finches (Taeniopygia guttata) were used in this study.

3.2.1 Electrophysiology

The surgical and electrophysiological procedures used have been described elsewhere (Schneider & Woolley 2010). Briefly, zebra finches were anesthetized two days prior

to recording with a single injection of 0.04 ml Equithesin. After administration of lidocaine, the bird was placed in a stereotaxic holder with its beak pointed 45 degrees downward. Small openings were made in the outer layer of the skull, directly over the electrode entrance locations. To guide electrode placement during recordings, ink dots were applied to the skull at stereotaxic coordinates (2.7 mm lateral and 2.0 mm anterior from the bifurcation of the sagittal sinus). A small metal post was then affixed to the skull using dental acrylic. After surgery the bird recovered for two days.

Prior to electrophysiological recording, the bird was anesthetized with three injections of 0.03 ml of 20 percent urethane, separated by 20 minutes. All experiments were performed in a sound-attenuating booth (IAC) where the bird was placed in a custom holder 23 cm away from a single speaker. Recordings were made from single auditory neurons in the MLd using either glass pipettes filled with 1M NaCl (Sutter Instruments) or tungsten microelectrodes (FHC, Inc.) with a resistance between three and 10 MΩ (measured at 1kHz). The duration of the recording sessions ranged from 4 to 15 hours. Awake recording sessions were no longer than 6 hours. For a single animal, awake recordings were performed over a period of approximately two weeks. Electrode signals were amplified (1000x) and filtered (300-5000 Hz; A-M Systems). A threshold discriminator was used to detect potential spike times. Spike waveforms were upsampled 4x offline using a cubic spline function, and action potentials were separated from non-spike events by cluster sorting the first three principal components of the action potential waveforms (custom software, Matlab). The number of neurons used in the analysis varied from 1 to 189.

3.2.2 Auditory Stimuli

Stimuli consisted of a set of 20 different adult male zebra finch songs sampled at 48,828 Hz, and frequency filtered between 250 and 8000 Hz. Each song was presented, in a pseudorandom order, 10 times at an average intensity of 72 dB SPL. Song duration ranged from 1.62 to 2.46 seconds, and a silent period of 1.2 to 1.6 seconds separated

the playback of subsequent songs. All songs were unfamiliar to the bird from which recordings were made.

3.2.3 Bayesian Decoding

In the Bayesian framework, the spectrogram decoding problem is equivalent to determining the posterior probability distribution, $p(\mathbf{x}|\mathbf{r},\theta)$, for observing a spectrogram, $\mathbf{x}$, given the measured neural responses, $\mathbf{r}$, and parameters θ. In principle, the posterior contains all available information about $\mathbf{x}$. We use different statistics from this distribution, for example the mode or mean, to reconstruct the particular stimulus presented to the animal.

The encoding model specifies the likelihood, $p(\mathbf{r}|\mathbf{x},\theta)$, which assigns probabilities to spike trains given the stimulus and parameters. The posterior distribution is related to the encoding model by Bayes rule,

$$p(\mathbf{x}|\mathbf{r},\theta) = \frac{p(\mathbf{r}|\mathbf{x},\theta)p(\mathbf{x})}{p(\mathbf{r}|\theta)}, \tag{3.2.1}$$

where $p(\mathbf{x})$ is the prior distribution over song spectrograms. Here, we reconstruct song spectrograms using single and multiple neurons and different prior distributions (see below) that systematically add information about the birdsong spectrotemporal statistics.

3.2.4 Encoding Model

For a population of N midbrain neurons, we model the number of spikes fired by neuron i at time t by a random variable $\mathbf{r}_{it}$, where i can range from 1 to N and t from 1 to T. We must assume that neurons are conditionally independent given the stimulus since we recorded cells one by one. Under this assumption, the likelihood in

eqn. 3.2.1 is given by

$$p(\mathbf{r}|\mathbf{x},\theta) \quad = \quad \prod_{t=1}^{T}\prod_{i=1}^{N} p(\mathbf{r}_{it}|s,\theta,\mathbf{r}_{i1},...,\mathbf{r}_{i,t-1}). \tag{3.2.2}$$

We discretize time into bins of width dt, and model the conditional distribution for $\mathbf{r}_{it}$ given the spectrogram, spike-history up to time t and parameters, θ, as Poisson

$$p(\mathbf{r}_{it}|\mathbf{x},\theta,\mathbf{r}_{i1},...,\mathbf{r}_{i,t-1}) \quad = \quad \exp\left(-\lambda_{it}dt\right)\frac{(\lambda_{it}dt)^{\mathbf{r}_{it}}}{\mathbf{r}_{it}!}, \tag{3.2.3}$$

where λ_{it} is the instantaneous firing rate of the ith neuron at time t. λ_{it} is given as the output of a generalized linear model (GLM). The GLM, and its application to neural data, has been described in detail elsewhere (Brillinger 1988, McCullagh & Nelder 1989, Paninski 2004, Truccolo et al. 2005, Calabrese et al. 2011) and we only give a brief overview. The GLM for λ_{it} applies a non-linearity (we use an exponential) to a linear mapping of input stimuli. As discussed in a recent paper (Calabrese et al., 2010) the model's ability to predict spikes slightly improves with this nonlinearity. In addition, the exponent prevents the model firing rate from taking on negative values and allows us to tractably fit the model to experimental data. The linear mapping is characterized by b_i, a stimulus-independent parameter which models baseline firing, k_i which will be referred to as the spectrotemporal receptive field (STRF) as it performs a linear mapping of stimulus to response, and a "spike-history" filter, $h_i(\tau)$, which allows us to model neuronal effects such as firing-rate saturation, refractory periods, and/or bursting behavior. Even though the GLM conditional distribution, $p(\mathbf{r}_{it}|\mathbf{x},\theta,\mathbf{r}_{i1},...,\mathbf{r}_{i,t-1})$, is Poisson, the joint spike train, $\mathbf{r}_{i1},...,\mathbf{r}_{i,T}$, does not follow a Poisson process because of the feedback from the spike-history filter. This procedure for mapping stimuli onto neural responses is schematized in Figure 3.1, which shows STRFs derived from data and shows simulated spike responses produced by the GLM.

Denoting the spectrogram by $\mathbf{x}(f,t)$ (f indicates the spectral bin number and

t denotes the temporal bin number), the firing rate r_{it} is modeled as

$$\lambda_{it} = \exp\left(b_i + \sum_{f'=0}^{F-1}\sum_{\tau'=0}^{M-1} k_i(f',\tau')\mathbf{x}(f',t-\tau') + \sum_{j=1}^{J} h_i(j)\mathbf{r}_{i,t-j}\right), \quad (3.2.4)$$

where F is the total number of frequency bins in the spectrogram, M is the maximal lag-time of the STRF, and J is the maximal lag-time of the spike-history filter. Unless explicitly stated otherwise, spectrograms were temporally binned at 3 ms with 35 linearly spaced frequency bins ($F = 35$) from 400 to 6000 Hz. The power density of all spectrograms is log transformed so that units of power are expressed in decibels. We set $M = 7$ (21 ms) and $J = 10$ (30 ms). Model parameters, $\theta_i = \{b_i, k_i, h_i\}$ for $i = 1...N$ are fit from MLd responses to conspecific song using L1-penalized Maximum Likelihood (Lee, et al. 2006). See Calabrese et al., (2010) for full details about the penalized fitting procedures.

3.2.5 Birdsong Priors

Eqn. 3.2.1 shows that song reconstruction depends on $p(\mathbf{x})$, the prior distribution of power spectral densities present in spectrograms. We test how song reconstruction depends on prior distributions that have the same spectrotemporal covariations present in song. We used several Gaussian priors because these distributions only depend on the covariance and mean. Other distributions might lead to better reconstructions by providing information about higher-order statistics in song, but are much more complicated to fit and optimize over. All Gaussians had the same frequency dependent mean but each had its own covariance matrix. All prior parameters were computed using the same songs as those used to collect the data (see Auditory Stimuli above). These songs appear to be sufficient to estimate the prior parameters under the Gaussian models presented below. Estimating the prior parameters for more complicated models may require more data which can be obtained by using more bird songs than the ones used to collect the neural data. Each song is reconstructed with a prior

whose parameters are estimated from all songs in the data set, except the one being reconstructed. The prior mean, $\hat{\mu}_f$, was found by assuming temporal stationarity of song and computing the empirical average power density across all temporal bins in the song data set.

Non-Correlated Gaussian Prior

To measure how well a population of midbrain neurons alone could reconstruct the spectrogram, we used a minimally informative prior. The least informative prior we used is an uncorrelated Gaussian

$$p(\mathbf{x}) = \prod_{f=1}^{F}\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\hat{\sigma}_f^2}}\exp\left(-\frac{(\mathbf{x}(f,t)-\hat{\mu}_f)^2}{2\hat{\sigma}_f^2}\right), \tag{3.2.5}$$

where $\hat{\mu}_f$ is the empirical average power density discussed above, and $\hat{\sigma}_f^2$ is the empirical variance of songs in our data set at each frequency bin f. This prior does not provide information about spectral and/or temporal correlations in song. The prior variance is estimated by the empirical variance of songs in our data set. Figure 3.2 shows a histogram of spectrogram power density values across all spectrogram bins in the song data set (blue dots) and a univariate Gaussian with mean and variance equal to that found in the data.

Spectrally-Correlated Gaussian Prior

Next we measured how well spectrograms can be reconstructed when midbrain neuronal responses are combined with prior knowledge of spectral correlations across multiple conspecific songs. To do this we used a Gaussian prior whose covariance matrix only depended on frequency. Writing the covariance in spectrogram power between one time and frequency bin, $\{t, f\}$, and another, $\{t', f'\}$, as $C(\{t, f\}, \{t', f'\})$, this

prior covariance is written as:

$$C(\{t,f\},\{t',f'\}) \;=\; \Phi(f,f')\delta(t-t'), \tag{3.2.6}$$

where $\delta()$ is the Dirac delta function. The prior distribution is given by

$$p(\mathbf{x}) \;=\; \prod_{t=1}^{T} \frac{1}{(2\pi)^{\frac{F}{2}}|\Phi|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x}(.,t)-\hat{\mu})^T\Phi^{-1}(\mathbf{x}(.,t)-\hat{\mu})}{2}\right), \tag{3.2.7}$$

where we use $\mathbf{x}$(.,t) to denote the column vector of power density across frequencies at time t. The Φ matrix is empirically fit from example songs:

$$\Phi(f,f') \;=\; \frac{1}{N_t-1}\sum_{n=1}^{N_t}\Big(\mathbf{x}(f,n)-\hat{\mu}_f\Big)\Big(\mathbf{x}(f',n)-\hat{\mu}_{f'}\Big), \tag{3.2.8}$$

where N_t is the total number of time-bins in the data set. N_t can be different from T, because T refers to the number of time-bins in the spectrogram being reconstructed, whereas N_t is the number of time bins in the entire data set used for training. N_t =13,435 for the data set used here. Figure 3.3A (upper panel) plots the Φ matrix. The spectral correlations averaged across all songs are larger at higher frequencies.

Temporally-Correlated Gaussian Prior

In order to measure how well songs can be reconstructed when midbrain responses are combined with prior knowledge of temporal correlations across conspecific songs, we reconstructed spectrograms with a prior containing temporal correlations but no spectral correlations

$$C(\{t,f\},\{t',f'\}) \;=\; C_T(t,t')\delta(f-f'). \tag{3.2.9}$$

The prior distribution is given by

$$p(\mathbf{x}) = \prod_{f=1}^{F} \frac{1}{(2\pi)^{\frac{T}{2}}|C_T|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{x}(f,.)-\hat{\mu_f})^T C_T^{-1}(\mathbf{x}(f,.)-\hat{\mu_f})}{2}\right), \quad (3.2.10)$$

where $\mathbf{x}(f,.)$ denotes the column vector of power density across time at frequency bin f.

We estimated the covariance matrix C_T by modeling the temporal changes in power density at a given frequency bin f as a stationary, order p, Autoregressive (AR) process:

$$\mathbf{x}'(f,t) \equiv \mathbf{x}(f,t) - \hat{\mu}_f, \quad (3.2.11)$$

$$\mathbf{x}'(f,t) = \sum_{i=1}^{p} a_i \mathbf{x}'(f,t-i) + \hat{\sigma}'\epsilon_t, \quad (3.2.12)$$

the constant terms, $a_i, \hat{\sigma}'$, are model coefficients and ϵ_t is a white noise, Gaussian random variable with unit variance. We used the covariance of this AR process instead of the empirical temporal covariance matrix to construct C_T. This is beneficial because it allowed us to approximate song correlations with far fewer parameters. Without an AR model the number of nonzero values in the matrix C_T^{-1} would grow quadratically with T, the temporal size of the spectrogram. This is troubling because each matrix element must be estimated from data, and therefore the amount of data required for accurately estimating C_T^{-1} grows with T. The inverse covariance matrix, C_T^{-1}, under an AR model is given by the square of a sparse Toeplitz matrix, A (Percival

& Walden 1993)

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a_1 & \ddots & 0 & 0 & 0 & 0 \\ \vdots & a_1 & \ddots & 0 & 0 & 0 \\ a_p & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & a_1 & 1 & 0 \\ 0 & 0 & a_p & \cdots & a_1 & 1 \end{pmatrix}, \tag{3.2.13}$$

$$C_T^{-1} = \frac{A^T A}{\hat{\sigma}'^2}. \tag{3.2.14}$$

As seen in eqn. 3.2.14, when we estimate the correlations using an AR model the number of nonzero values in the matrix C_T^{-1} only depends on the parameters $a_i, \hat{\sigma}'$ and, importantly, is independent of T. Thus the amount of data required to accurately estimate C_T^{-1} using an AR model is independent of T. To fit the AR coefficients we used the Burg method to minimize the sum of squares error between the original and AR model power density (Percival & Walden 1993). We combined the temporal changes across all songs and spectral bins to fit the AR coefficients. Figure 3.3A (lower panels) compares the correlations of a twenty six order (p=26) AR model with empirical temporal correlations, averaged across songs and spectral bins. There is a trade-off between increasing the order of the AR model for obtaining good fits to the birdsong correlation function and the memory required to store the inverse covariance matrix/computational time to reconstruct spectrograms. We set p=26 because lower order models did not do a sufficient job of capturing the dip and rise present in the correlation function visible between 0 and 100 ms (see Figure 3.3A). Note that we do not show a covariance matrix because an AR process assumes that the covariance between time points t and t' only depends on the absolute difference or lag time between these points: $C_T(t, t') = C_T(|t - t'|)$; i.e. all necessary information is contained in the correlation function. It is clear from Figure 3.3A that the temporal correlation

function of the AR model closely matches the empirical correlation function found directly from the data.

Gaussian Prior with Spectrotemporal Correlations

Finally, we measured how well songs can be reconstructed when midbrain responses are combined with the spectral and temporal correlations across conspecific songs. To do this we reconstructed songs using a Gaussian prior with covariance equal to the separable product of the previously described AR covariance matrix (we will also study temporal covariance matrices that are not time-translation invariant) and the Φ matrix:

$$C(\{t,f\},\{t',f'\}) \;=\; \alpha\Phi(f,f')C_T(|t-t'|). \tag{3.2.15}$$

The factor α is set so that the marginal variance of $C(\{t',f'\},\{t',f'\})$ is matched to the average variance of the song spectrograms, $\sum_f \hat{\sigma}_f^2$.

The prior distribution is given by

$$p(\mathbf{x}) \;=\; \frac{1}{(2\pi)^{\frac{FT}{2}}|C|^{\frac{1}{2}}}\exp\left(-\frac{(\vec{\mathbf{x}}-\hat{\mu})^T C^{-1}(\vec{\mathbf{x}}-\hat{\mu})}{2}\right). \tag{3.2.16}$$

Equation 3.2.15 shows that C has a particular block structure in which each element of the Φ matrix is multiplied by the C_T matrix. This structure is known as a Kronecker product and leads to computational advantages when manipulating the C matrix. For example, the inverse matrix C^{-1} also has a Kronecker product form. This is particularly advantageous because we can use this fact to compute the required matrix multiplication $C^{-1}(\vec{\mathbf{x}}-\hat{\mu})$ in a time that scales linearly with the dimension T ($O(F^3T)$ instead of the usual $O((FT)^3)$ time). In order to do so, we must construct the spectrogram vector $\vec{s}$ so that same-time frequency bands are contiguous, $\vec{\mathbf{x}} = \Big(\mathbf{x}(.,1),\mathbf{x}(.,2),...,\mathbf{x}(.,T)\Big)^T$.

The matrix C does not exactly match the correlations in birdsong because it assumes that spectral and temporal correlations can be separated. Using a separable covariance matrix and our AR model is beneficial because we do not need to estimate and store the full (FT) x (FT) covariance matrix, a task which becomes infeasible as we increase the number of time bins in our reconstruction. Importantly, we wanted to find reconstruction algorithms that could be performed in a computationally efficient manner. As discussed below, the separability approximation allows us to reconstruct spectrograms in a manner that is much more efficient than using a non-separable matrix. To examine the validity of the separability assumption we computed an empirical covariance matrix, $\hat{C}(f, f', |t - t'|)$ without assuming separability:

$$\hat{C}(f, f', \tau) = \frac{1}{N_t} \sum_{i=1}^{N_t - |\tau| + 1} \Big(\mathbf{x}(f, i) - \hat{\mu}_f\Big)\Big(\mathbf{x}(f', i + \tau) - \hat{\mu}_{f'}\Big), \qquad (3.2.17)$$

where N_t is again the total number of time-bins in the data set. In Figure 3.3B (middle panel under the title 'True Covariance') we plot the matrix $\hat{C}$ and compare it with the separable matrix used in this study (Figure 3.3B bottom panel under the title 'Approximate Covariance'). Each lag, τ, can be thought of as an index for an FxF frequency matrix. For example, the upper panels in Figure 3.3B plot these FxF matrices when the lag equals zero. The matrix $\hat{C}$ and its separable approximation plot these FxF matrices, one for each lag, next to each other. The two matrices fall to zero power at the same rate and are closely matched near zero lags. The separable approximation has less power in the off-diagonal frequency bands at intermediate lags but overall the separable approximation is fairly accurate.

To visualize the information about song provided by this prior, Figure 3.3C (bottom panel) shows a sample spectrogram drawn from this Gaussian. The differences between this sample and a typical song spectrogram (upper panel) are due to the separable approximation to the song covariance matrix and the Gaussian prior model for the distribution of song spectrograms. Comparing the two-dimensional

power spectra (also called the modulation spectra) of song spectrograms and of this prior is another method for assessing the effects of assuming a separable matrix. Figure 3.3D shows that the prior distribution lacks the peak across spectral modulations at temporal modulations close to zero, but otherwise has a similar spectrum.

Hierarchical Model Prior

One clear failure of the previous prior models is that real songs have silent and vocal periods. We can capture this crudely with a two-state model prior. This prior consists of a mixture of two correlated Gaussian priors, and a time-dependent, latent, binary random variable, q_t, that infers when episodes of silence and vocalization occur. We refer to this model as the hierarchical prior. One of the Gaussian distributions has mean power and spectral covariance determined by only fitting to the silent periods in song, while the other has mean power and spectral covariance fit to the vocalization periods. The two covariance matrices are shown in Figure 3.4B (upper panel).

Vocalization periods and silent episodes are extracted from the spectrogram data set by using a slightly ad hoc method that works well for our purposes here. A hard threshold is placed on the total power density summed across spectral bins, a variable we call y_t, and on the power density variance, $\hat{\sigma}_t^2$, across spectral bins:

$$y_t = \sum_{f=1}^{F} \mathbf{x}(f,t), \tag{3.2.18}$$

$$\hat{\sigma}_t^2 = \frac{1}{F-1} \sum_{f=1}^{F} \left(\mathbf{x}(f,t) - \frac{y_t}{F} \right)^2, \tag{3.2.19}$$

$$q_t = \begin{cases} 1(\text{vocalization period}) & y_t \geq q_1^*,\ \hat{\sigma}_t^2 \geq q_2^* \\ 0(\text{silent period} & \text{otherwise} \end{cases} \tag{3.2.20}$$

Figure 3.4A shows an example spectrogram and associated state transitions found using the above thresholding procedure, with q_1^* set to one standard deviation below

the mean power density in the song and q_2^* set to an empirically determined value of 90 dB^2.

We model q_t as a Markov process and fit the transition matrix using maximum likelihood with training data found by taking state transitions from real song. The data set used here consisted of 13,435 state samples. This procedure leads to the transition rates displayed in Figure 3.4B (lower panel). Temporal correlations come from the AR model covariance matrix (described above) and from the temporal correlations induced by the Markovian model for q. Modeling q_t as a Markov process captures features of state transitions found in song and allows us to decode spectrograms using well-known algorithms (see below). However, by using a Markov model we assume that state durations are exponentially distributed, which only approximates the distribution of durations found in birdsong.

A sample from this prior is shown in Figure 3.4C. The differences between vocal and silent periods are more clearly pronounced in this sample than that of the correlated Gaussian prior (Figure 3.3A). Because of the large differences in spectral correlations between vocal and silent periods, samples from this model also show spectral correlations closer to those found in song.

3.2.6 Song Reconstructions

Most of our reconstructions will be given by the spectrogram matrix that maximizes the log of the posterior distribution (the *maximum a posteriori* or MAP estimate). Substituting eqns. 3.2.2 and 3.2.3 into eqn. 3.2.1, the objective function that we maximize is then

$$\begin{aligned} L(\mathbf{x},\theta) = \log p(\mathbf{x}|\mathbf{r},\theta) &= \sum_{i=1}^{N}\sum_{t=1}^{T} \log p(\mathbf{r}_{it}|\mathbf{x},\theta,\mathbf{r}_{i1},...,\mathbf{r}_{i,t-1}) + \log p(\mathbf{x}) + \text{const} \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{r}_{it}\log\lambda_{it} - \lambda_{it}dt + \log p(\mathbf{x}) + \text{const}, \end{aligned} \tag{3.2.21}$$

where N is the total number of neurons used in decoding, θ refers to the encoding model parameters, λ_{it} is the firing rate for the ith neuron at time t (computed via eqn. 3.2.4), and $p(\mathbf{x})$ denotes the prior distribution. We write the term $\log p(\mathbf{r}|\theta)$ as 'const' because it is constant with respect to the stimulus. In general, MAP estimates are found by searching over probabilities for all combinations of power density in a spectrogram and determining the most probable configuration. This task can be extremely computationally difficult as the number of spectral and temporal bins in the estimate grows. However, this problem is computationally tractable using standard Newton-Raphson (NR) optimization methods with the likelihood and prior distributions discussed above (Paninski et al. 2009, Pillow et al. 2011). In general, NR optimization computes the optimal configuration in a time that is on the order of d^3 (written as $O(d^3)$), where d is the dimensionality of the quantity being optimized (in our case $d = FT$). This is because the rate-limiting step in NR optimization is the time required to solve a linear equation involving the matrix of second derivatives of the objective function, L, in eqn.3.2.21, which requires $O(d^3)$ time in general. The likelihood and AR model used here yield sparse, banded Hessian matrices which reduces the time for optimization to $O(F^3T)$ (Paninski et al. 2009, Pillow et al. 2011). This speedup is critical since the dimensionality of the decoded spectrograms is around $d \sim 7000$.

Song reconstructions under the hierarchical prior are created using the posterior mean, $E[\mathbf{x}|\mathbf{r}]$. The posterior mean is an optimal statistic to use for reconstruction as it is the unique estimate which minimizes the averaged squared error between the reconstruction and presented spectrogram. Using a Gaussian prior we decoded spectrograms with the MAP estimate because it is computationally efficient and because $E[\mathbf{x}|\mathbf{r}] \approx MAP$ in this case (Ahmadian, et al. 2011, Pillow et al. 2011). It is easier to compute $E[\mathbf{x}|\mathbf{r}]$ using Markov Chain Monte-Carlo (MCMC) sampling when we decode using the hierarchical prior. The idea behind MCMC is that if we can generate samples from the posterior distribution we can use these samples to estimate the

mean (Robert & Casella 2005). It is difficult to sample directly from the posterior distribution using the hierarchical prior described above. However, it is possible to generate samples from the joint distribution, $p(\mathbf{x}, q|\mathbf{r}, \theta)$ which can then be used to estimate $E[\mathbf{x}|\mathbf{r}]$ (q again refers to the vocalization state). By definition $E[\mathbf{x}|\mathbf{r}]$ is given by the following multi-dimensional integral:

$$E[\mathbf{x}|\mathbf{r}] = \int p(\mathbf{x}|\mathbf{r}, \theta)\mathbf{x}d\mathbf{x} \tag{3.2.22}$$

$$= \sum_{q_1=0}^{1} ... \sum_{q_T=0}^{1} \int p(\mathbf{x}, q_1, ..., q_T|\mathbf{r}, \theta)\mathbf{x}d\mathbf{x}. \tag{3.2.23}$$

The relationship in eqn.3.2.23 shows how $E[\mathbf{x}|\mathbf{r}]$ is related to the joint distribution. We do not compute the sum in eqn.3.2.23 directly but instead use samples from the joint distribution $p(\mathbf{x}, q|\mathbf{r}, \theta)$ to evaluate $E[\mathbf{x}|\mathbf{r}]$. The details are given in the appendix.

3.2.7 Simulating STRFs

We also examined how our results depended on the spectral filtering properties of the STRF. We compute MAP estimates using simulated STRFs that have no spectral blur, no history dependence and peak frequency locations sampled from a distribution fit to real STRFs. This distribution was empirically constructed by creating a histogram of spectral bins at which STRFs obtained their maximal value. Denoting the simulated STRF of the ith neuron at frequency bin f and temporal lag τ by $k'_{if\tau}$, our simulated STRFs take the form:

$$k'_{if\tau} = a'_i \delta_{f,f'_i} \delta_{\tau 0}, \tag{3.2.24}$$

where δ_{ij} is the Kronecker-delta function. We choose the values of a'_i and the new encoding model bias parameters, b'_i, to obtain model responses whose first and second moment are roughly matched to those of the true responses. For each neuron, we use the true filters and offset, k and b, to compute the linear mapping:

$$\rho_{it} = b_i + \sum_{f'=0}^{F-1} \sum_{\tau'=0}^{M-1} k_i(f', \tau')\mathbf{x}(f', t - \tau'). \tag{3.2.25}$$

Then we compute the median, $\text{med}(\rho_i)$, and absolute median deviation, $\text{med}(|\rho_i - \text{med}(\rho_i)|)$ across time of ρ_i. Given $\text{med}(\rho_i)$ and $\text{med}(|\rho_i - \text{med}(\rho_i)|)$, we algebraically determine values of a'_i and b'_i which yield equivalent linear medians and absolute median deviations when convolved with the input spectrogram $\mathbf{x}$. In other words, we solve the following linear equations for a'_i and b'_i:

$$\text{med}(\rho_i) = b'_i + a'_i \text{med}(\mathbf{x}_{f'_i}) \tag{3.2.26}$$

$$\text{med}(|\rho_i - \text{med}(\rho_i)|) = |a'_i| \text{med}(|\mathbf{x}_{f'_i} - \text{med}(\mathbf{x}_{f'_i})|). \tag{3.2.27}$$

Spike trains generated using simulated STRFs with parameters fit as described have first and second moments roughly matched to spikes generated from real STRFs (compare the raster plot in the middle panel of Figure 3.10B with the raster in the bottom panel of Figure 3.10B).

3.2.8 Optimal Linear Estimator

We compare our estimates with the optimal linear estimator (OLE) (Bialek et al. 1991, Warland, et al. 1997, Mesgarani et al. 2009). In brief, this model estimates the spectrogram by a linear filter g_{ift} which linearly maps a population of spike responses n_{it} to a spectrogram estimate $\hat{\mathbf{x}}(f, t)$:

$$\mathbf{r}'_{it} = \mathbf{r}_{it} - \frac{1}{T} \sum_{j}^{T} \mathbf{r}_{ij} \tag{3.2.28}$$

$$\hat{\mathbf{x}}(f, t) = \sum_{i=1}^{N} \sum_{j=0}^{\tau-1} g_{ifj} \mathbf{r}'_{it-j} + \frac{1}{T} \sum_{j}^{T} \hat{\mathbf{x}}(f, j), \tag{3.2.29}$$

where N is again the total number of neurons used in decoding and τ is the maximal lag used in the decoding filter. $\mathbf{r}'$ is the mean subtracted spike responses and is used to ensure that the OLE and spectrogram have the same means. The function g is found by minimizing the average, mean-squared error between $\hat{\mathbf{x}}$ and the spectrogram $\mathbf{x}$ at each frequency bin. The solution to this problem (Warland et al. 1997, Mesgarani et al. 2009) is given by

$$g_f = C_{\mathbf{rr}}^{-1} C_{\mathbf{rx}(f)}, \tag{3.2.30}$$

where $C_{\mathbf{rr}}$ denotes the auto-covariance of neural responses and $C_{\mathbf{rx}(f)}$ denotes the cross-covariance of the response with the temporal changes in bin f of the spectrogram. The amount of data required to accurately estimate the matrices $C_{\mathbf{rr}}$ and $C_{\mathbf{rx}(f)}$ increases as the filter length and the number of neurons used in the estimation increases. We did not implement any regularization on the matrices $C_{\mathbf{rr}}$ and $C_{\mathbf{rx}(f)}$ to deal with this problem (see Pillow et al (2010) for further discussion). As is customarily done (Theunissen, et al. 2001), we assume stimuli and responses are stationary so that temporal correlations between two points in time, say t and t', only depend on the distance or lag between these points, $t - t'$. We compute the covariances up to a maximal lag of 18 ms using spectrograms with time binned into 3 ms intervals with 35 linearly spaced frequency bins from 250 to 8000 Hz. These values were chosen in an attempt to maximize OLE performance.

3.2.9 Measuring Reconstruction Accuracy

The quality of reconstructions is measured using the Signal-to-Noise ratio (SNR) which is defined as the variance in the original spectrogram divided by the mean-squared error between the original and estimated spectrograms. Each song is reconstructed 4 times using the responses to different presentations of the same song. Since there are 20 songs in the data set, we obtain 80 different samples of mean-squared

error between the estimated spectrograms and original. The mean-squared error is estimated by averaging these estimates together. The estimator's stability is measured using the standard error, which is the sample standard deviation of these estimates divided by the square-root of our sample size (80). Songs were reconstructed using different numbers of neurons. The neurons used for reconstruction were chosen by randomly sampling without replacement from the complete data set of neural responses.

We also examined reconstruction quality in the Fourier domain. For each prior used, we computed the coherence between the estimated spectrogram and the original. The coherence between the original spectrogram, $\mathbf{x}$, and the reconstructed spectrogram $\hat{\mathbf{x}}$ is defined as

$$C(\nu_1, \nu_2) = \frac{|R_{\mathbf{x},\hat{\mathbf{x}}}(\nu_1, \nu_2)|}{[R_{\mathbf{x},\mathbf{x}}(\nu_1, \nu_2) R_{\hat{\mathbf{x}},\hat{\mathbf{x}}}(\nu_1, \nu_2)]^{\frac{1}{2}}}, \tag{3.2.31}$$

$$R_{XY}(\nu_1, \nu_2) = \sum_{u=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \exp(-i2\pi u\nu_1) \exp(-i2\pi m\nu_2) \mathrm{cov}\{X(u,m)Y(0,0)\} \tag{3.2.32}$$

The cross-spectral density function, R, for each reconstruction-spectrogram pair was estimated using Welch's modified periodogram method with overlapping segments (Percival & Walden 1993). For each pair, the spectrograms were divided into segments that overlap by 25% and whose length, L, was $\frac{1}{8}$ the number of time bins in the spectrogram. Within each segment we computed a modified periodogram of the form:

$$\hat{X}(\nu_1, \nu_2) = \sum_{f=1}^{U} \sum_{t=1}^{U} \exp(-i2\pi f\nu_1) \exp(-i2\pi t\nu_2) h(f,t) X(f,t) \tag{3.2.33}$$

$$\hat{R}_{XY}(\nu_1, \nu_2) = \frac{\hat{X}(\nu_1, \nu_2)\hat{Y}^*(\nu_1, \nu_2)}{\sum_{u=1}^{N} \sum_{m=1}^{N} h^2(u,m)}, \tag{3.2.34}$$

where Y^* denotes the complex conjugate of Y, h is known as a data taper, and U

denotes the window size. Data tapers and the use of overlapping windows were used because they can reduce the bias and variance associated with estimating R by a naive periodogram using data of finite length (Percival & Walden 1993). Data was zero-padded so that U equaled 256 even though the window length was variable. h was chosen to be the product of two Hanning windows:

$$h(u,m) \quad = \quad 0.25(1-\cos(2\pi u/L))(1-\cos(2\pi m/L)) \;\; \text{for}\, u=1,...,L\,; m=1,...,L. \tag{3.2.35}$$

R was estimated by averaging the estimates across segments. C was computed as in eqn. 3.2.31 substituting in the estimated R for the cross-spectral density function. We plot coherence values in decibels, given by the base-10 logarithm of the coherence multiplied by a factor of 10.

3.3 Results

3.3.1 Two-alternative forced choice discrimination

We begin by asking if an ideal observer of MLd neural responses could discriminate between conspecific songs using the GLM encoding model. Zebra finches can accurately discriminate between songs based on a single presentation, so it is interesting to determine if MLd responses can also discriminate between songs. We performed multiple two-alternative forced choice tests using an optimal decision rule and counted the fraction of correct trials from the different runs. Each test consisted of two segments of different conspecific songs and the spike trains, from multiple neurons, produced in response to one of the segments (Figure 3.5A). Using the log-likelihood eqns(3.2.2 - 3.2.4), we evaluated the probability of observing the spike trains given both segments and chose the segment with the higher likelihood. This procedure is optimal for simple hypothesis tests (Lehmann & Romano 2005). The likelihood depends on

the STRFs of the neurons used for discrimination and thus this test directly measures if the information about song provided by these STRFs can be used for song discrimination.

Figure 3.5B shows the fraction of correct trials for four different response/segment durations, when single spike trains produced by 1, 104, and 189 neurons are used for decision-making. As expected, the probability of responding correctly increases with longer response durations and as more neurons are used for discrimination. Given the very short response duration of 3 ms single neurons discriminated at chance level. As the response duration increased to 102 ms the average fraction correct increased to around 70%. Combining the responses of 189 neurons led to a discriminability accuracy as great as that seen in behavioral experiments, 90-100% (Shinn-Cunningham et al. 2007), after response durations around 27 ms and perfect discrimination after durations of 100 ms. These results show that MLd responses can be used for single presentation conspecific song discrimination.

3.3.2 Decoding Song Spectrograms Using Single Neurons

The results discussed above are in agreement with previous studies showing that responses from auditory neurons in the forebrain area field L (Wang, et al. 2007) and MLd (Schneider & Woolley 2010) can be used for song discrimination. As in previous studies, MLd responses and our encoding model could be used to determine the most likely song identity, given a pre-defined set of possible songs. Instead of directly comparing our results with previous methods we focused on a different problem; we asked if these responses contain enough information to reconstruct song spectrograms. Spectrogram reconstruction is a more computationally demanding task than song discrimination and is a better test of the information about song encoded by neuronal responses. As explained in the methods we use the *maximum a posteriori* (MAP) value to estimate the spectrogram. We first compute spectrogram reconstructions using single-trial responses from single MLd neurons to understand how MAP estimates

depend on the STRF and prior information.

The upper panel of Figure 3.6A shows 250 miliseconds of a spectrogram which elicited the two spikes shown below the spectrogram. The spikes are plotted at the frequency at which that neuron's STRF reaches a maximum (the BF or best frequency). Below the evoked spike-train is the MAP estimate (see methods) computed without prior information of spectrotemporal correlations in song power (Figure 3.6B, upper panel) and with prior information (Figure 3.6B, lower panel).

When the stimulus is only weakly correlated with the neuronal response, i.e. when the convolved spectrogram, computed by $\sum_{f'=0}^{F-1}\sum_{\tau'=0}^{M-1} k(f',\tau')\mathbf{x}(f',t-\tau')$, is much less than one, it is possible to calculate the MAP solution analytically (Pillow et al. 2011). As discussed in Pillow et al.(2010), the analytic MAP solution when a spike occurs is approximately equal to the neuron's STRF multiplied by the prior covariance matrix. Under our uncorrelated prior (see methods section 3.2.5) this is equivalent to estimating the spectrogram by the STRF scaled by the song variance. In the absence of spiking the analytic MAP solution is equal to the prior mean. We plot the frequency averaged prior mean, $\sum_f \hat{\mu}_f$, in green and adjacent to the MAP estimate we plot the STRF for this particular neuron. Comparing the STRF with the MAP estimate we see that the analytic solution for the weakly correlated case is valid for this example neuron. This solution is intuitive because it reflects the fact that the only information a MLd neuron has about the spectrogram is from its song filtering properties. It also illustrates a fundamental difficulty in the problem of song estimation; an MLd neuron only responds to a small spectral and temporal area of song power. Because of this, spikes from a single neuron, without prior correlation information, can only contribute information on the small spectrotemporal scales encoded by that neuron.

Information independent of MLd spike responses can aid in song reconstruction by using spectral and temporal correlations to interpolate bands filtered out by the STRF. The MAP solution using our correlated Gaussian prior displays this intuitive

behavior. Next to the MAP solution using a correlated prior we plot the neuron's STRF temporally convolved and spectrally multiplied with the prior covariance matrix

$$k'(f,\tau) = \sum_{f',\tau'} C_T(|\tau-\tau'|)\Phi(f,f')k(f',\tau'). \tag{3.3.36}$$

Comparing the MAP solution with k' we see that when a spike occurs, to a good approximation, the MAP estimates the spectrogram with a covariance multiplied STRF. The MAP estimate shows values of power after a spike occurs because it uses prior knowledge of the temporal correlations in song to infer spectrogram power at these times.

3.3.3 Population Decoding of Song Spectrograms

We expect reconstruction quality to improve as more neurons are used in decoding. From the intuition developed performing single neuron reconstructions, we guessed that each neuron, without prior information of song correlations, would estimate a STRF each time it spikes. With the diverse array of STRF patterns known to exist in MLd (Woolley et al. 2009), a population of MLd neurons might faithfully estimate a spectrogram, without prior information, by having each neuron estimate a small spectrotemporal area determined by its STRF.

In Figure 3.7A (upper panel) we plot 1.25 seconds of an example song that we attempt to reconstruct given the spike trains from 189 neurons. In Figure 3.7A (lower panel) we have plotted these responses, with each neuron's spikes plotted at the best frequency (BF) at which that neuron's receptive field reaches a maximal value. Neurons with the same BF are plotted on the same row. Figure 3.7B shows the MAP estimate using the uncorrelated prior. As in the single neuron case, during segments of song that result in few spikes from the population the song is roughly estimated by the prior mean (green segments). This MAP estimate does a good job of

distinguishing areas of high power from silent periods. Examination of the population spike responses show that this property is due to the population's ability to respond in a manner temporally-locked to vocalization episodes.

3.3.4 Effect of Prior Information on Song Estimation

Reconstructions without prior information show discontinuous gaps in power during vocal periods. They also show sparse spectrotemporal correlations at frequencies above 4 kHz. As in the single-neuron case, these features probably reflect the fact that each neuron only filters small spectrotemporal segments of song. In addition, most STRFs have peak frequencies below four kHz (this is evident in the plot of spike responses ordered by BF). Intuitively we expect MAP estimates constructed with prior information of song correlations to enhance reconstructions by filling in these 'gaps' in power density.

Figure 3.7 shows how the MAP estimate given the responses from 189 neurons changes as prior information is added. We plot MAP estimates using a prior covariance matrix that contains temporal information but no spectral information (Figure 3.7C), spectral information but no temporal information (Figure 3.7D), and both spectral and temporal correlations (Figure 3.7E; see section 3.2.5 in the methods for details and Figure 3.10 for a plot of preferred frequency tuning across the neural population). Comparing these estimates with the estimate using an uncorrelated prior (Figure 3.7B) shows that information about the second-order statistics in zebra finch song enhances reconstructions by interpolating correlations in spectrogram areas not covered by MLd STRFs.

A clear improvement in estimates using the spectrally correlated priors occurs at times where spiking activity is sparse. At these times the MAP estimate using the uncorrelated prior equals the prior mean. When given knowledge of song correlations, the MAP uses the sparse activity in MLd to correctly infer the deflections from the mean that occur in the original song spectrogram. The correlations also help the

MAP infer that, during episodes of high spiking activity, spectral bands above four kHz should have similar levels of power as below four kHz. In the supplementary material we provide the reconstruction in the time-domain created by combining the spectrogram in Figure 3.7E with random phase. For comparison purposes we also provide the original song constructed with randomized phase.

In Figure 3.8 we quantify how reconstruction quality improves as a function of the number of neurons used in decoding. We use the signal-to-noise ratio (SNR) (see methods for a definition of the SNR) as a quantitative method for evaluating reconstruction accuracy. As described in the methods, the neurons chosen for reconstruction were randomly chosen from the population. Figure 3.8A plots example MAP estimates using the Gaussian prior with spectrotemporal correlations as a function of the number of neurons for a single example song. The associated value of SNR is given above the MAP estimate. The solid lines in Figure 3.8B show the SNR averaged across all songs and dashed lines show the standard error about these lines. A SNR value of one corresponds to estimating the spectrogram by a single number, the mean. Improvements in SNR reflect improved estimates in the correlations of song power about the mean. The colors denote which prior was used in computing the MAP estimate. As expected, the SNR from MAP estimates using prior spectrotemporal information (black line) grows the fastest, followed by the SNR from MAP estimates which only use spectral prior information (green line). The faster growth in SNR only using spectral prior information versus temporal information is probably due to the facts that MLd population responses already capture a good deal of temporal information, spectral information helps infer deflections from the mean at times of sparse activity, and most MLd neurons have STRFs with peak frequencies below 4 kHz.

Figure 3.8C plots the coherence between the reconstructions and original spectrograms. The coherence is a normalized measure of the cross-correlation between the original two-dimensional signal and estimate in the frequency domain. In all the

plots the vertical axis is spectral modulations (in units of cycles/kHz). These frequencies are often referred to as the ripple density. The horizontal axis is temporal modulations (in units of Hz). We note that the coherence plot is not the same as the modulation power spectrum shown in Figure 3.3D. In all plots, the range of the coherence is limited from -10 decibels (dark blue), a coherence of 0.1, to 0 decibels, i.e. perfect coherence (shown in red). With the exception of the non-correlated prior, we see a high coherence for temporal modulations between -50 and 50 Hz and ripple densities between 0 and 0.6 cycles/kHz. When we analyzed the coherence within these frequencies we found that the average coherence is highest for the spectrotemporal prior, second highest for the spectral prior and smallest for the prior without covariance information. From this plot we conclude that prior knowledge of the stimulus correlations primarily aids in reconstructing lower temporal modulations and ripple densities.

It is interesting to compare the decoding performance just described with the optimal linear estimator (OLE), a simpler and more commonly used decoder (Mesgarani et al. 2009). As discussed in the methods, the OLE finds the estimate that minimizes the average-squared euclidean distance between the spectrogram being estimated and a linear combination of the responses. Figure 3.8 (magenta line) shows the growth in the SNR of the OLE using the same real responses as those used for the non-linear, Bayesian model. The OLE depends on spectrotemporal correlations in the stimulus so we compare its performance with the prior that contains both spectral and temporal correlations (black line). Comparing these two shows that when the number of neurons is low the two estimates perform similarly. As more neurons are added to the estimate, the MAP estimator outperforms the OLE. Recent work (Pillow et al. 2011) has shown that this behavior is expected if the encoding model is a good model for spike responses to stimuli and if the prior model does a good job of capturing stimulus correlations. Pillow et al.(2010) showed that when the number of neurons used for estimation is low, the MAP estimate and OLE are equivalent. As the number of neu-

rons grows, the MAP estimate can outperform the OLE because the MAP estimator is not restricted to be a linear function of spike responses.

Hierarchical Prior Model

We observed visible differences in power density covariance and mean during silent and vocal periods (see the covariance matrices plotted in Figure 3.4 and the differences in color between silent and vocal periods in the plotted spectrograms). These differences are averaged together when constructing the covariance matrix and mean used in the single Gaussian prior. Averaging together the correlation information from these two different periods smooths the spectral correlation information (compare the covariance matrix in Figure 3.3A with that of Figure 3.4B (left panel)). We reconstructed songs using a hierarchical prior (see methods sec 3.2.5) to test whether this smoothing hinders the reconstruction performance. This prior includes a state variable which determines the mean and spectral covariance. We first study the case where all possible state trajectories are used for decoding, with trajectory probabilities determined by neural responses and the transition probabilities in our model (see methods and Figure 3.4). Each trajectory yields a different reconstructed spectrogram and the final estimate is determined by averaging across these reconstructions. This is equivalent to estimating the song using the posterior mean. This estimate should be better than an estimate using a single Gaussian if the neural responses provide sufficient information to properly infer state transitions.

In Figure 3.9A (left column) we plot an example song spectrogram with evoked single-neuron, single-trial responses immediately below. We have again plotted the responses at the neuron's best frequency, which in this case is 1.8 kHz. Below this we have plotted the MAP estimate using these responses and a single correlated Gaussian prior (Figure 3.9B upper panel) and the posterior mean using the hierarchical prior (Figure 3.9B lower panel). The estimates show surprisingly similar behavior. Under the hierarchical prior we see power densities slightly closer to those in song, around

the neuron's BF, compared to the estimate using a single Gaussian prior. Otherwise no large differences between the two estimators are seen.

It is possible that estimates based on the hierarchical model are not much better than using a single Gaussian because single neuron responses do not provide enough information to infer the state transitions. Figure 3.9C shows the average state transition given the neural response. We see that this is indeed the case and on average the inferred state transitions do not match those in the song being estimated. Given the above result we asked if the hierarchical model would outperform the single Gaussian prior when more neurons are used for decoding. In the right column of Figure 3.9A we plot the responses of 49 additional neurons (for a total of 50 neurons) with BFs slightly greater than the single neuron used in the left column. These responses are again plotted below the spectrogram being estimated. Examining the average state changes given responses in Figure 3.9C we see a closer resemblance between the inferred state transitions to those present in the estimated song. In the right column of Figure 3.9B we plot the posterior mean under the hierarchical prior and the MAP estimate using the same subset of neural responses combined with a single Gaussian (non-hierarchical) prior. The two estimators do not show any prominent differences. Adding more neurons to the estimation should only cause the two estimators to look more similar since the reconstructions will have less dependence on prior information when more data is included. Therefore we did not compute estimates with more than 50 neurons using the Hierarchical prior. Finally, we eliminated the portion of reconstruction error due to problems associated with estimating the state transitions by computing the MAP estimate of the Hierarchical prior given the true underlying state in the song being estimated. We compared this estimate, which has perfect knowledge of the underlying song transitions, to the MAP estimate using a single Gaussian prior. Even in this case we do not see any large differences between the estimators (data not shown). These results demonstrate that spectrogram estimates do not necessarily improve as more complicated prior information of song is included

in the posterior distribution. While samples from the hierarchical prior contain more statistical information of song and arguably show more resemblance to song than samples from a single, correlated Gaussian prior (compare Figure 3.4C with Figure 3.3C), this advantage does not translate into better spectrogram reconstructions.

3.3.5 Reconstruction Dependence On STRF Frequency Tuning

The information for reconstruction provided by an individual MLd neuron depends on its STRF. Neurons that have STRFs which overlap in their spectrotemporal filtering properties will provide redundant information. While this redundancy is useful for reducing the noise associated with the spike generation process (Schneider & Woolley 2010), good spectrogram reconstructions also require enough neurons that provide independent information. We asked if our results would improve if we used neurons that had either no overlap in their filtering properties or complete overlap. We computed MAP estimates using simulated STRFs, which we will refer to as point STRFs, that have no spectral blur and no history dependence (see methods for how these receptive fields were constructed). Figure 3.10A plots a STRF (upper, left panel) calculated from real responses using the method of maximum likelihood ('full' STRF) and a point STRF (upper, right panel) with an equivalent peak frequency. In Figure 3.10A (lower, left panel), we show the extent of the blurring behavior in our neuronal population. For each neuron, we plot the spectral axis of its STRF at the latency which that STRF reaches its maximum value. The right panel shows the same information for the point STRFs (see methods for our determination of the number of neurons with a particular peak frequency).

Figure 3.10B (upper panel) shows an example spectrogram we attempt to reconstruct. For both STRFs we reconstructed songs using simulated responses. We did not use real responses because we wanted to reduce the differences in reconstruction performance caused by the poorer predictive performance of point STRFs on real

data. Using simulated responses allowed us to better control for this effect and focus on differences in reconstruction performance due to spectral blurring. For comparison purposes, we plot the real responses of 189 neurons, aligned according to their BF, immediately below this spectrogram. The middle panel shows simulated responses to this example song created using the generalized linear model with point STRFs. The lower panel shows simulated responses using full STRFs. Using a correlated Gaussian prior, we reconstructed the spectrogram using the point STRFs and the simulated responses generated from them (middle panel) and using the full STRFs and their associated simulated responses (lower panel).

Stimulus reconstructions using point STRFs show slightly finer spectral detail compared to reconstructions using full STRFs. However, overall we do not find that spectral blurring of the full STRFs leads to much degradation in stimulus reconstructions. The growth in SNR for point STRFs and full STRFs as a function of the number of neurons is shown in Figure 3.10C. On average point STRFs have slightly higher signal-to-noise ratios as the number of neurons increases; however the difference between the two curves is not too great. It is important to point out that these results depend on the fact that reconstructions were performed using a correlated prior trained on natural stimuli. The spectrotemporal width of the covariance is broad compared to that of the full STRFs. When we reconstructed songs using a prior with no correlations, we found that full STRFs decode slightly better than the point STRFs (data not shown). Also, for the reasons stated above, we used simulated responses which also influences the results. Reconstructions using point STRFs are slightly worse than reconstructions with full STRFs when real data is used. We attribute this difference to the better predictive performance of the full STRFs on real data.

3.4 Discussion

We asked if the responses of zebra finch auditory midbrain neurons to song encode enough information about the stimulus so that an 'ideal observer' of MLd spike trains could recognize and reconstruct the song spectrogram. We found that 189 sequentially recorded MLd responses can be combined using a generalized linear model (GLM) to discriminate between pairs of songs that are 30 ms in duration with an accuracy equivalent to that found in behavioral experiments. These results are in agreement with prior studies showing that responses from auditory neurons in the forebrain area field L (Wang et al. 2007) and MLd (Schneider & Woolley 2010) can be used for song discrimination. Importantly, this previous work did not use the GLM to evaluate the discriminability and thus provides an independent benchmark to compare with our GLM-dependent results.

We tested the hypothesis that the statistics of zebra finch song can be used to perform vocal recognition by decoding MLd responses to conspecific song using *a priori* knowledge of the joint spectrotemporal correlations present across zebra finch songs. We explicitly used prior information lacking higher-order information of song to test if MLd responses only require knowledge of correlations to be used for spectrogram reconstruction. When we evaluated the reconstructed spectrograms in the Fourier domain, we found that these responses do a fair job of reproducing temporal and spectral frequencies, i.e. temporal modulations and ripple densities, between -50 and 50 Hz and below 0.6 cycles per kHz. When combined with the joint spectrotemporal correlations of zebra finch song we found an improvement in the coherence in these regions. These results did not change greatly when we used STRFs with non-overlapping best frequencies, suggesting that the spectral blur or 'bandwidth' limitations of the STRF did not strongly affect reconstruction performance using these responses combined with spectrotemporal correlations in zebra finch song.

None of the reconstructions using MLd neurons and the correlations present in song reproduced all the details of a particular spectrogram. These results are

qualitatively similar to previous findings showing that the auditory system of zebra finch, as well as other songbirds, can recognize songs even when some of the fine details of the song signal have been degraded by various types of background noise (Bee & Klump 2004, Appeltants, et al. 2005, Narayan, et al. 2006, Knudsen & Gentner 2010). This may be similar to the finding that humans can recognize speech even after the spectral and temporal content has been degraded (Drullman, et al. 1994, Shannon, et al. 1995).

It is interesting to speculate if the song features that were reproduced in this study are relevant to the bird for song recognition. For example, we found that reconstructions were most accurate at low ripple densities and temporal modulations. Song recognition based on these features would be consistent with existing evidence that zebra finch are better able to discriminate auditory gratings with lower ripple density/temporal modulations (Osmanski, et al. 2009). Because of the complexity of song it is difficult to quantify behaviorally relevant song features birds use for recognition and communication (Osmanski et al. 2009, Knudsen & Gentner 2010). The spectrogram reconstructions reported here may serve as a useful probe for future discrimination studies. For example, one could compare discrimination thresholds between songs whose amplitude spectrums have been degraded according to the regions where reconstructions have low coherence with songs whose amplitude spectrums are randomly degraded. If the MAP reconstructions are relevant to the bird, we would expect performance to be worse on songs with randomly degraded amplitude spectrums. This idea is similar to the previously mentioned Osmanski et al. (2009) study testing discrimination of auditory gratings in birds; however the ripple density/temporal modulations used for probes would be more complex than simple gratings. Working with ferret auditory cortical neurons, Mesgarani et al.(2009) have also recently examined the effects of stimulus correlation on spectrogram decoding. Similar to our findings, this group finds improvements in reconstruction quality when they use prior information of sound correlations. This suggests that the use of natural sound cor-

relations for vocal recognition might be a general strategy employed across species. However, there are important distinctions between the Bayesian approach used here for reconstruction and the optimal linear decoder used in this previous work. The optimal linear decoder incorporates stimulus correlations via the stimulus-response cross-covariance matrix and the response auto-covariance matrix. The Bayesian decoder incorporates stimulus statistics using a prior distribution that is independent of the likelihood distribution used to characterize neural responses (eqn. 3.2.1). Therefore this decoder allows one to estimate song correlations independent of the amount of neural data available. This is beneficial for obtaining good estimates of song correlations when it is easier to obtain song samples than physiological data. Another important distinction between the linear decoder and the Bayesian method is that the Bayesian decoder does not have to be a linear function of the spike responses. This seems to be the reason for the Bayesian method's slight improvement over the linear decoder. When we decode songs using a linear, Gaussian, Bayesian decoder with the same correlated Gaussian prior as the one in this study we find worse reconstruction performance than the GLM. This suggests that the nonlinearity is an important factor in the GLM's improved performance.

Another advantage of separating prior information from neural responses is that we could systematically change the prior to study which statistical properties of song are most important for stimulus reconstruction without refitting the filters applied to the observed spike trains. We found that reconstructions based on MLd responses with *a priori* information of spectral correlations yielded better estimates of song than did reconstructions using temporal correlations present in song. While we cannot conclude from this study whether or not the bird actually uses a prior, we speculate that these results suggest what information, in addition to MLd responses, maybe used when the bird recognizes song. These results suggest that there is a greater benefit to the bird, in terms of vocal recognition capabilities, if MLd responses are processed by neuronal circuits that have access to the joint spectrotemporal or spectral

song correlations rather than temporal correlations. This interpretation would be consistent with recent work showing that zebra finch appear to be more sensitive to frequency cues than temporal cues when categorizing songs belonging to one of two males (Nagel, et al. 2010). However, even though much work has been done relating information encoded within a prior distribution to neuronal spiking properties (Zemel, et al. 1998, Beck & Pouget 2007, Litvak & Ullman 2009), it is unclear how to predict response properties of cells based on the statistical information about a stimulus they may be encoding. To better understand this relationship future experiments could perform a similar decoding analysis using the responses from other brain areas to look for spiking activity in which it is more beneficial to store temporal correlations rather then spectral correlations. If such activity exists, these responses could be combined with MLd spike trains to perform reconstructions which presumably would only show marginal improvement when combined with prior knowledge of either temporal or spectral correlations.

There has been much recent interest in determining good priors for describing natural sounds and stimuli(Singh & Theunissen 2003, Karklin & Lewicki 2005, Cavaco & Lewicki 2007, McDermott, et al. 2009, Berkes, et al. 2009). With the two-state model we briefly explored the effects on reconstruction quality of prior distributions which contain more information than just the mean and covariance of birdsong, however none of the priors used in this study explicitly contain information about the subunits such as song notes, syllables or motifs typically used to characterize song (Catchpole & Slater 1995, Marler & Slabbekoorn 2004). Future work could examine if reconstruction quality changes using more realistic, non-gaussian prior distributions of birdsong which contain higher-order information. For example, neurons in the songbird forebrain nucleus HVc are known to be sensitive to syllable sequence (Margoliash & Fortune 1992, Lewicki & Arthur 1996, Nishikawa, et al. 2008) suggesting that there are neural circuits which could provide prior information of sound categories such as syllables and motifs. One could therefore reconstruct songs using

this prior information, for example by using a Hidden Markov Model (HMM) with the hidden states trained on sound categories (Kogan & Margoliash 1998). While we didn't find much of an improvement in reconstruction quality using the two-state prior compared to a Gaussian prior, more realistic priors may yield better reconstructions. If so, one could determine additional statistical information about song stimuli, other than stimulus correlations, also useful for song recognition.

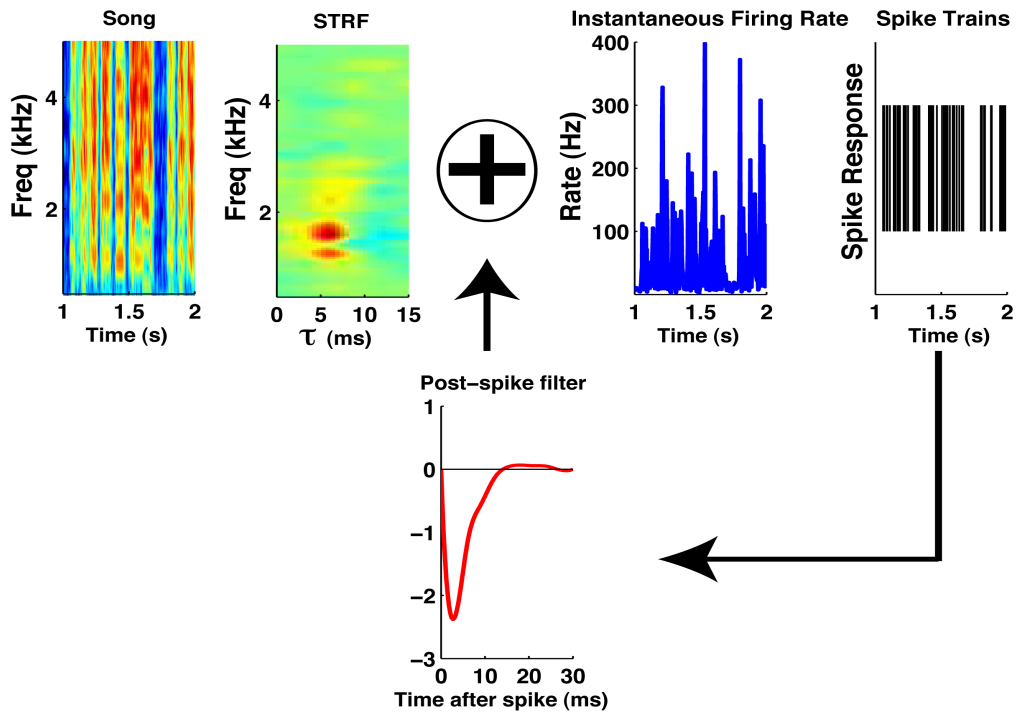

Figure 3.1: Encoding model and parameters. In the encoding model, each neuron is modeled with a spectrogram filter (STRF) and post-spike filter that captures stimulus-independent spiking properties. The stimulus is temporally convolved and frequency multiplied with the STRF and then exponentiated to obtain the instantaneous firing rate used for generating spikes. The spikes are convolved with the post-spike filter and used in the model as a feedback signal that affects future spike generation.

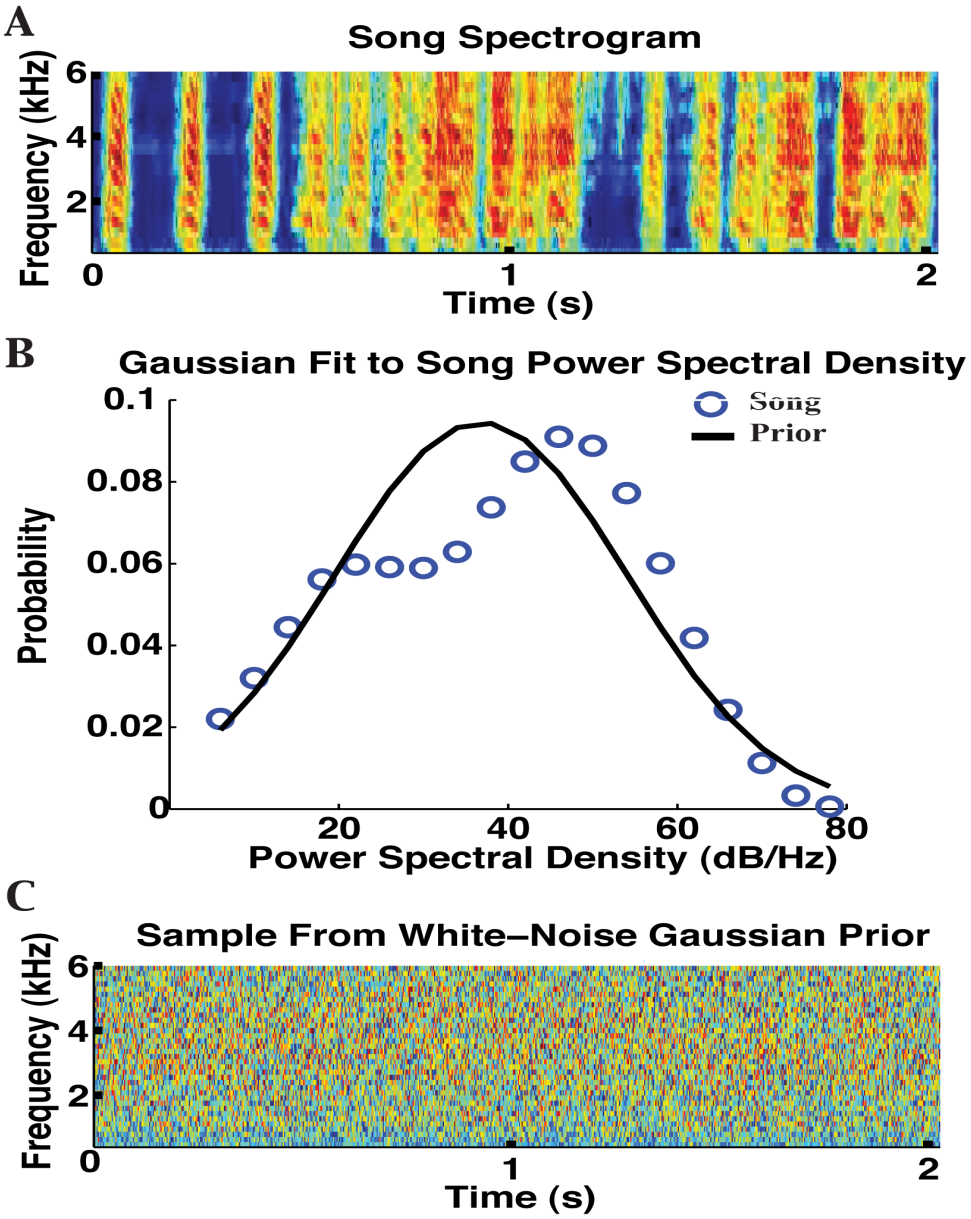

Figure 3.2: Least informative prior: uncorrelated Gaussian distribution. **A**, An example spectrogram with power spectral density indicated by color. **B**, Normalized histogram of power spectral density values across all songs and spectrogram bins (blue dots). The mean and variance of these power values is used to construct a Gaussian prior (black line) that confines estimated values of power spectral density to regions found in actual song spectrograms. **C**, To visualize the information provided by the prior, a sample spectrogram drawn from this prior is plotted. This prior does not provide information on spectrotemporal correlations in spectrograms, as demonstrated by this sample.

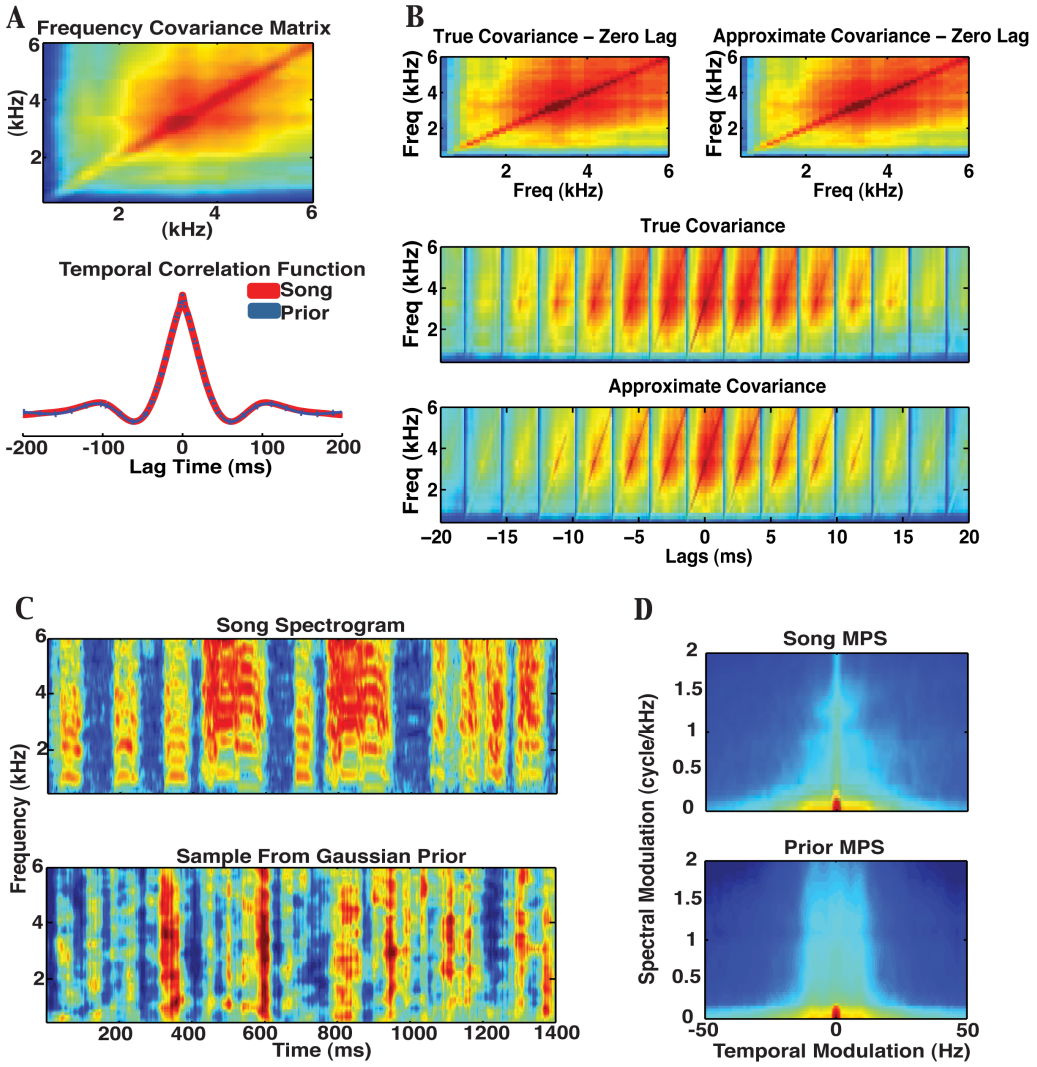

Figure 3.3: Spectrotemporaly Correlated Gaussian prior. **A**, The spectrotemporal covariance matrix is modeled as separable in frequency and time. The frequency component is the spectral covariance matrix (upper panel). The temporal component is fully described by the temporal autocorrelation function in song spectrogram power density (bottom panel, red line). The prior uses an approximation to this function using an Autoregressive model (blue line). **B**, The full spectrotemporal covariance matrix is a concatenation of several spectral covariance matrices, like those shown in the upper panel, each corresponding to the covariance at a different temporal lag. The bottom panel labeled 'Approximate Covariance' plots the separable covariance matrix and the middle panel labeled 'True Covariance' plots the non-separable covariance matrix. **C**, (Top) An example spectrogram used in determining song statistics for constructing the Gaussian prior. (Bottom) Sample spectrogram drawn from the Correlated Gaussian prior. **D**, Two-dimensional power spectra, also called the modulation power spectra (MPS), for song spectrograms (top) and for the prior (bottom); the prior does a good job of capturing information about spectrotemporal modulations except at joint regions of high spectral modulations and temporal modulations near zero.

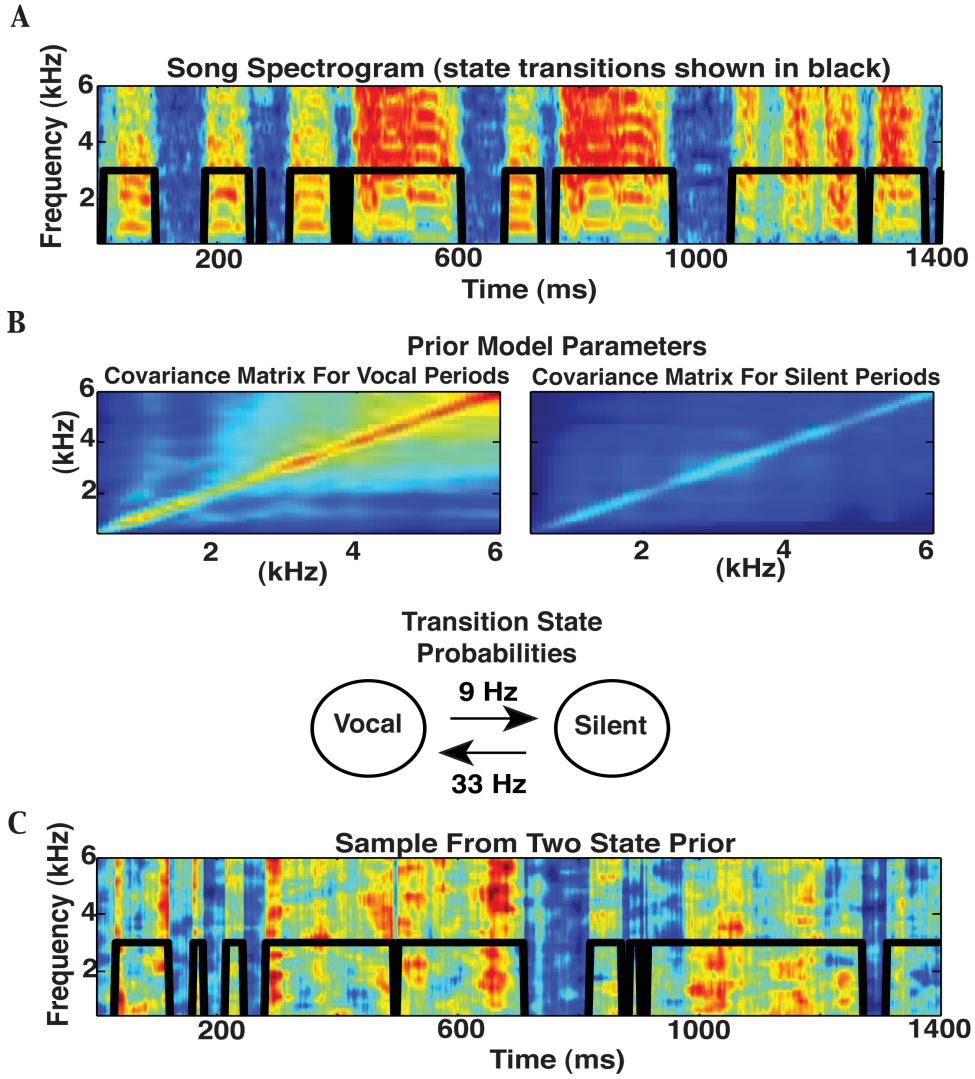

Figure 3.4: Most informative prior: hierarchical model with a two-state hidden variable that infers whether the spectrogram is in a vocalization or silent period. These periods have different statistical properties not captured by a single Gaussian prior. The state variable determines which spectral covariance matrix and mean the prior uses to inform reconstructions. **A**, Example spectrogram overlaid with vocalization and silent states (black line). **B**, (Top left) Spectral covariance matrix used during vocal periods. (Top right) Spectral covariance matrix used for silent periods. (Bottom) Prior information of transition rates between silent and vocal periods determined from song spectrograms. **C**, Sample spectrogram drawn from this prior; the sharp transitions in song statistics during vocal and silent periods better matches song spectrograms.

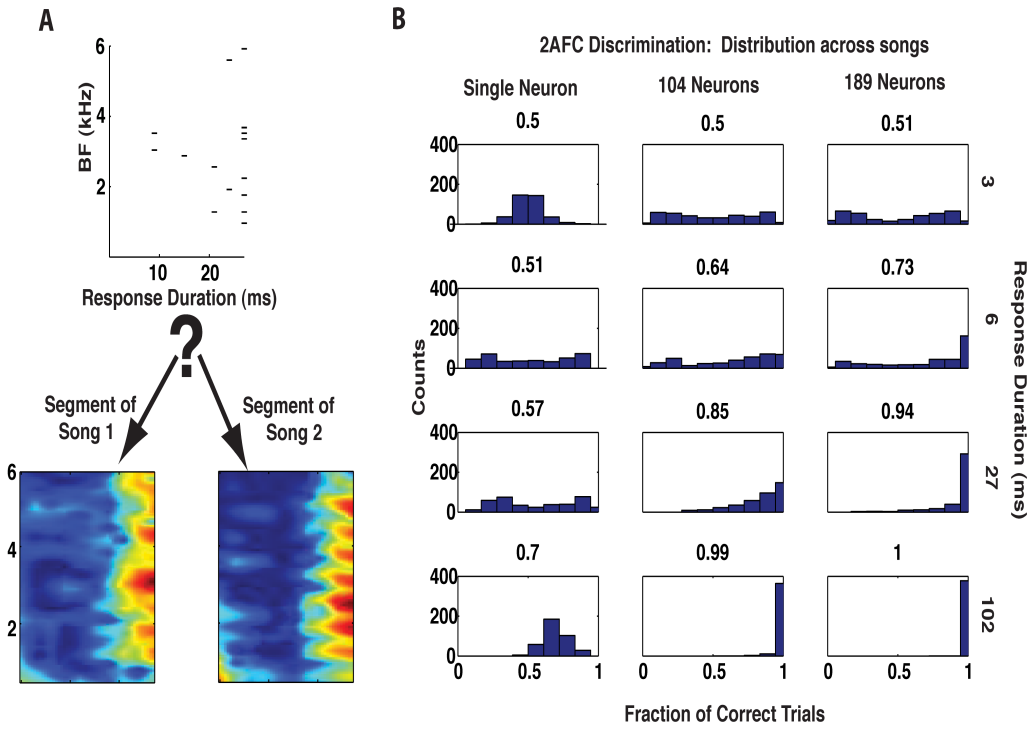

Figure 3.5: Conspecific song discrimination based on likelihood of spike-trains from multiple neurons. **A**, Spike trains from multiple neurons in response to presentation of song segment 1. Under a Two-alternative forced choice (2AFC) test, song discrimination is performed by choosing the song which leads to a greater likelihood of observing the given spikes. spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches maximal value. Neurons with the same BF are plotted on the same row. **B**, 2AFC results as a function of response duration and the number of neurons used for discrimination. 2AFC was performed multiple times for each possible pairing of the twenty songs in the data set. Each panel shows the frequency of correct trials across all possible song pairings. Above each panel is reported the average of the histogram. On average, neurons performed at chance level when stimulus segments were only 3 ms in duration. Near perfect song discrimination can be achieved using 189 responses and response durations at least around 30 ms, or 104 neurons and durations of about 100 ms.

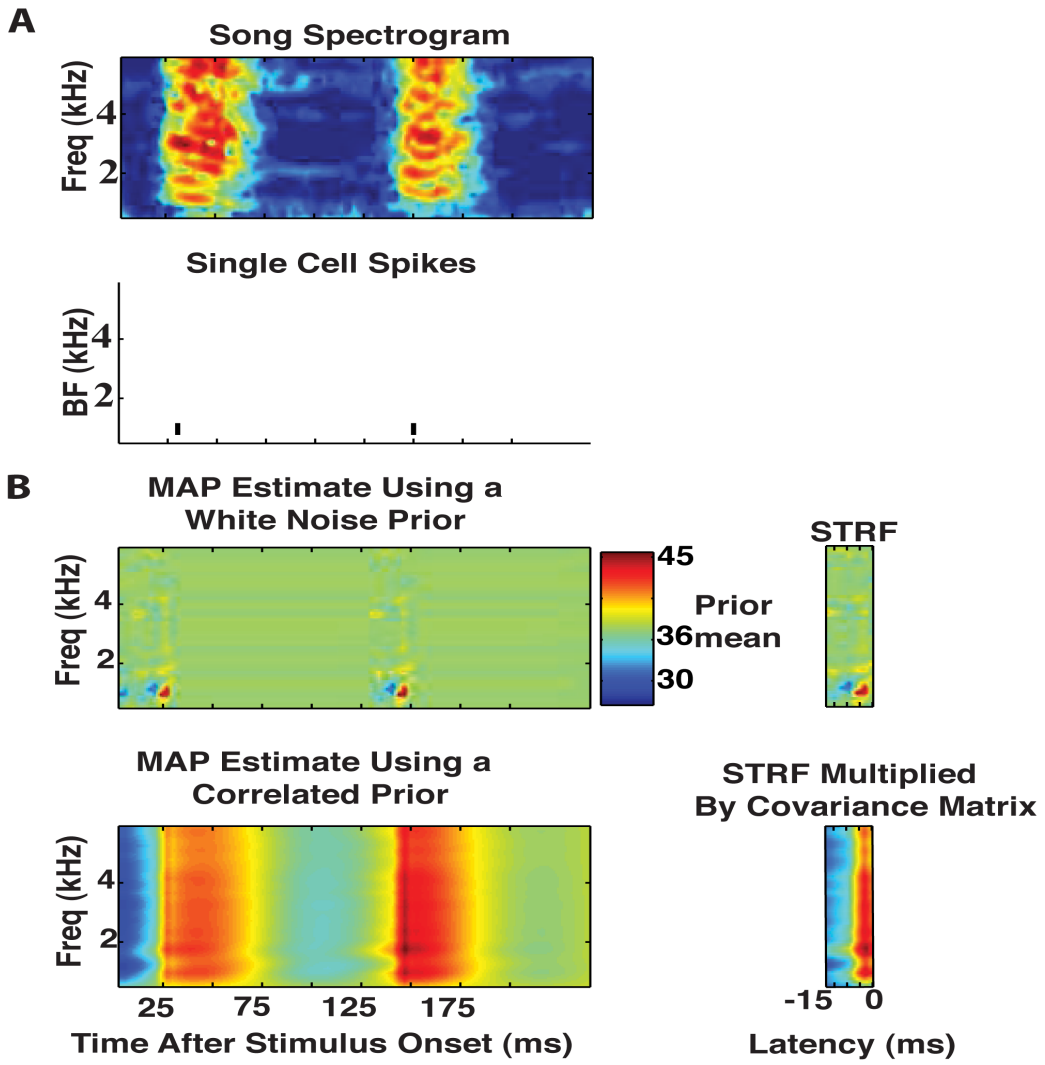

Figure 3.6: Single cell decoding of song spectrogram. **A**, (Top) Spectrogram of birdsong that elicited the two spikes shown immediately below. spikes are plotted at the frequency at which this neuron's receptive field reaches maximal value. **B**, (Top, left) The most probable spectrogram from the posterior distribution (MAP Estimate) given the two spikes shown in A and using an uncorrelated prior. When a single spike occurs the MAP is determined by the neuron's spectrotemporal receptive field (STRF, shown to the right). In the absence of spikes, the MAP is determined by the prior mean. (Bottom, left) MAP estimate using the correlated Gaussian prior; when a spike occurs the MAP is determined by the neuron's STRF multiplied by the prior covariance matrix (shown to the right). Immediately after a spike the MAP infers spectrogram values using prior knowledge of stimulus correlations.

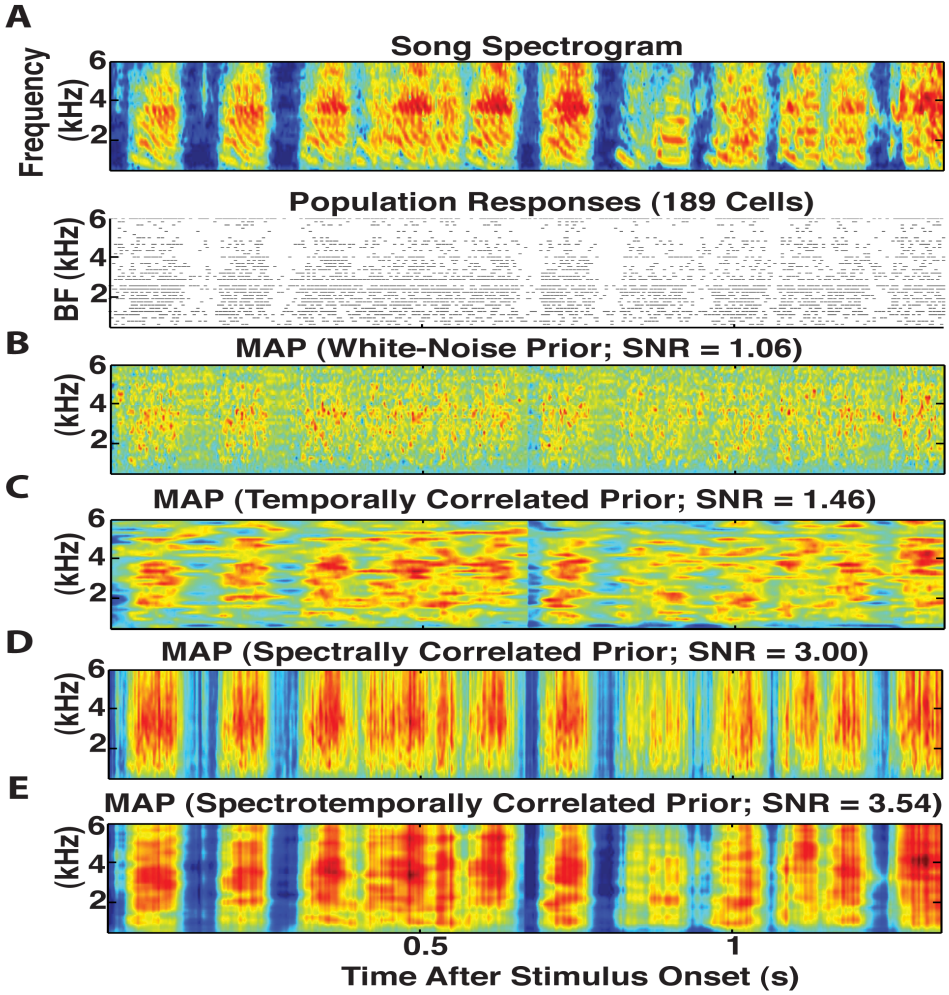

Figure 3.7: Population decoding of song spectrogram with varying degrees of prior information of song statistics. **A**, (Top) Spectrogram of birdsong played to 189 different neurons leading to the spike responses shown immediately below. spikes from a given neuron are plotted at the best frequency (BF) at which that neuron's receptive field reaches its maximal value. Neurons with the same BF are plotted on the same row. MAP estimate given the responses in A, using an uncorrelated prior **B**, a prior with temporal correlations and no spectral correlations **C**, a prior with spectral correlations and no temporal correlations **D**, and a prior with spectral and temporal correlations **E**. Combining the spike train with spectral information is more important for reconstructing the original spectrogram than combining the spike train with temporal information. However, combining spikes with joint spectrotemporal information leads to the best reconstructions.

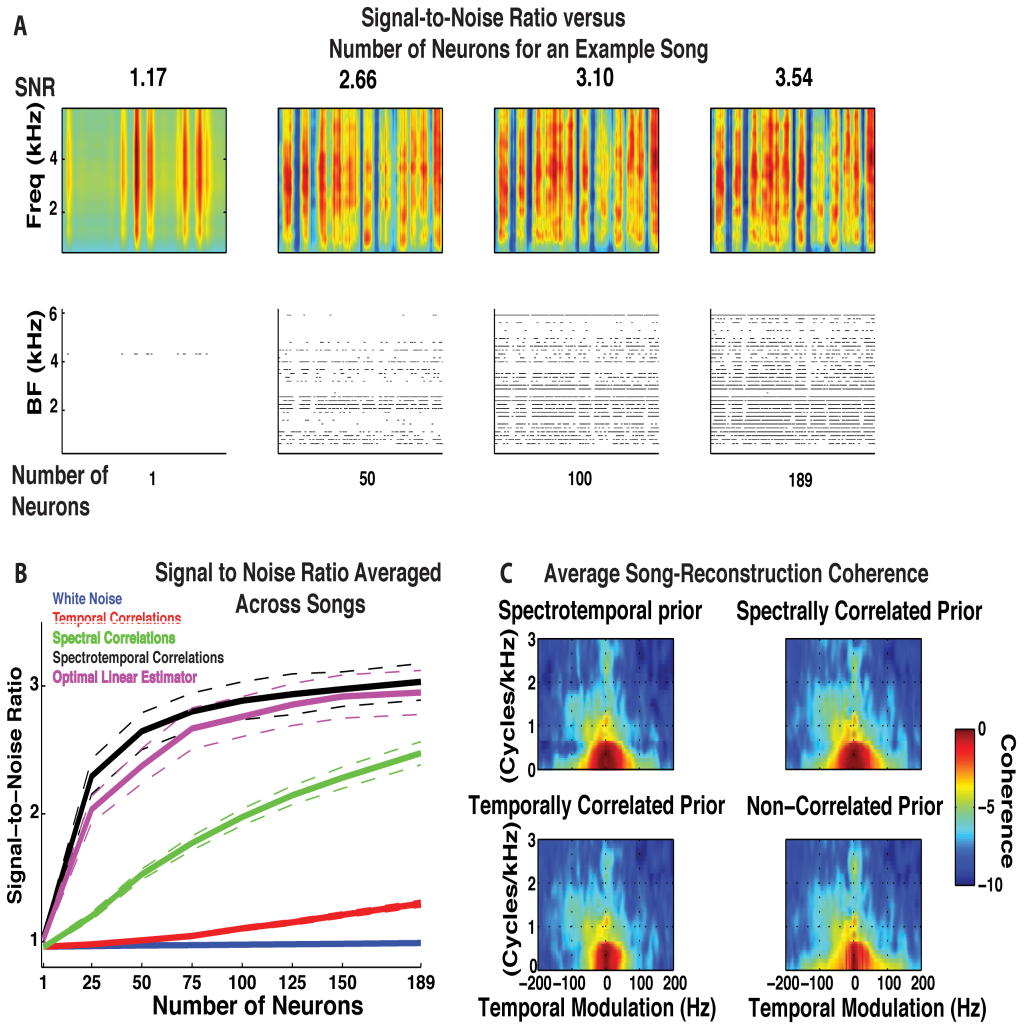

Figure 3.8: Decoding performance given different amounts of prior information and numbers of neurons. **A**, (upper row) Spectrogram reconstructions for an example song (Figure 3.7A) using a Gaussian prior with spectrotemporal correlations and using varying numbers of neuronal responses (plotted in the lower row). Above each reconstruction is the signal-to-noise (SNR) used to measure similarity between the reconstructed song and the song presented to the bird. **B**, Solid lines show the signal-to-noise (SNR) ratio averaged across all decoded songs while dashed lines show one standard error. The prior used for decoding is denoted by color. Spectral prior information leads to faster growth in the SNR than temporal information. For reference, the magenta line shows the growth in SNR for the commonly used optimal linear estimator (OLE). The OLE has access to both spectral and temporal correlations. **C**, Coherence between spectrograms and reconstructions under the four different priors. The horizontal axis reports temporal modulations and the vertical axis reports spectral modulations. All plots display the highest coherence at low spectral and temporal modulations. The primary effect of adding spectrotemporal prior information is to improve reconstructions at lower spectral and temporal modulations.

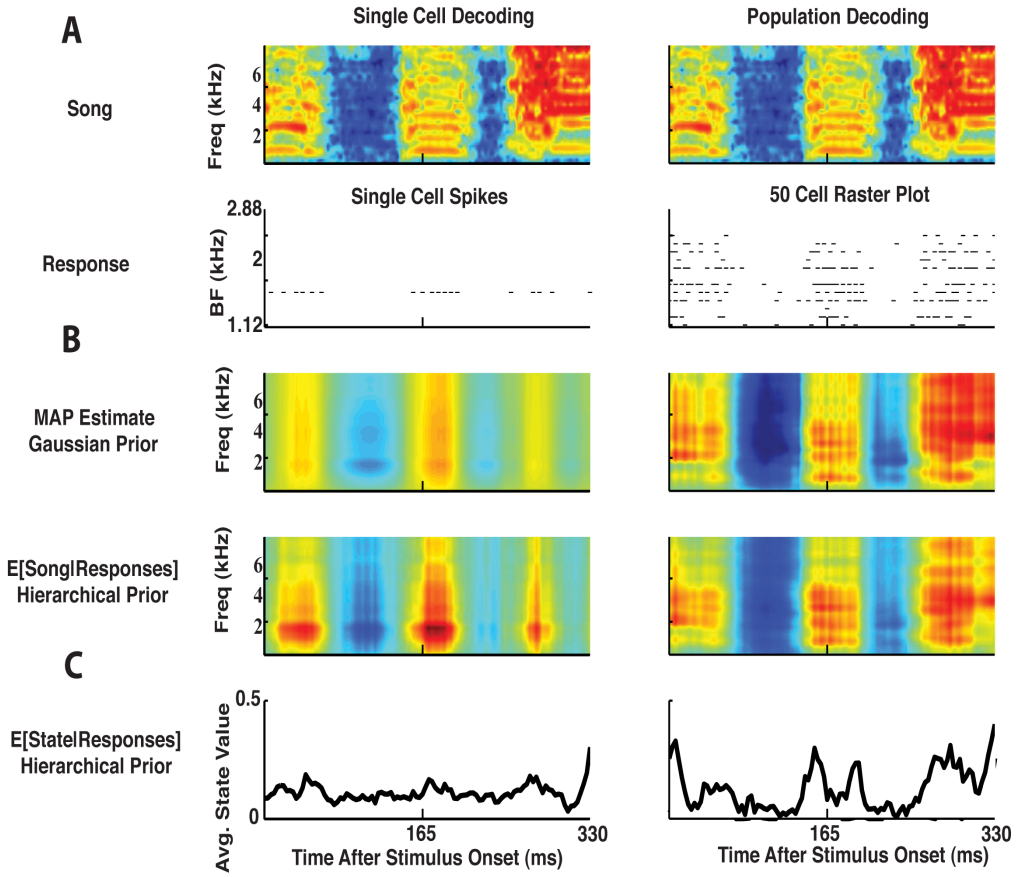

Figure 3.9: Single neuron and population decoding using a hierarchical prior. **A**, Song spectrogram along with a single cell's response to this song (left column) and the response of this cell plus forty nine other cells with nearby characteristic frequencies (right column). **B**, MAP estimates using a single, correlated Gaussian prior (top row) are compared with estimates using the posterior mean and the hierarchical prior (bottom row); in both the single neuron and population decoding case, the estimate using a hierarchical prior looks similar to the MAP with a Gaussian prior. **C**, The expected value for vocalization state given responses; single cell responses do not yield enough information to accurately infer the spectrogram's vocalization state, however as the number of neurons used for inference increases the vocalization state becomes more pronounced.

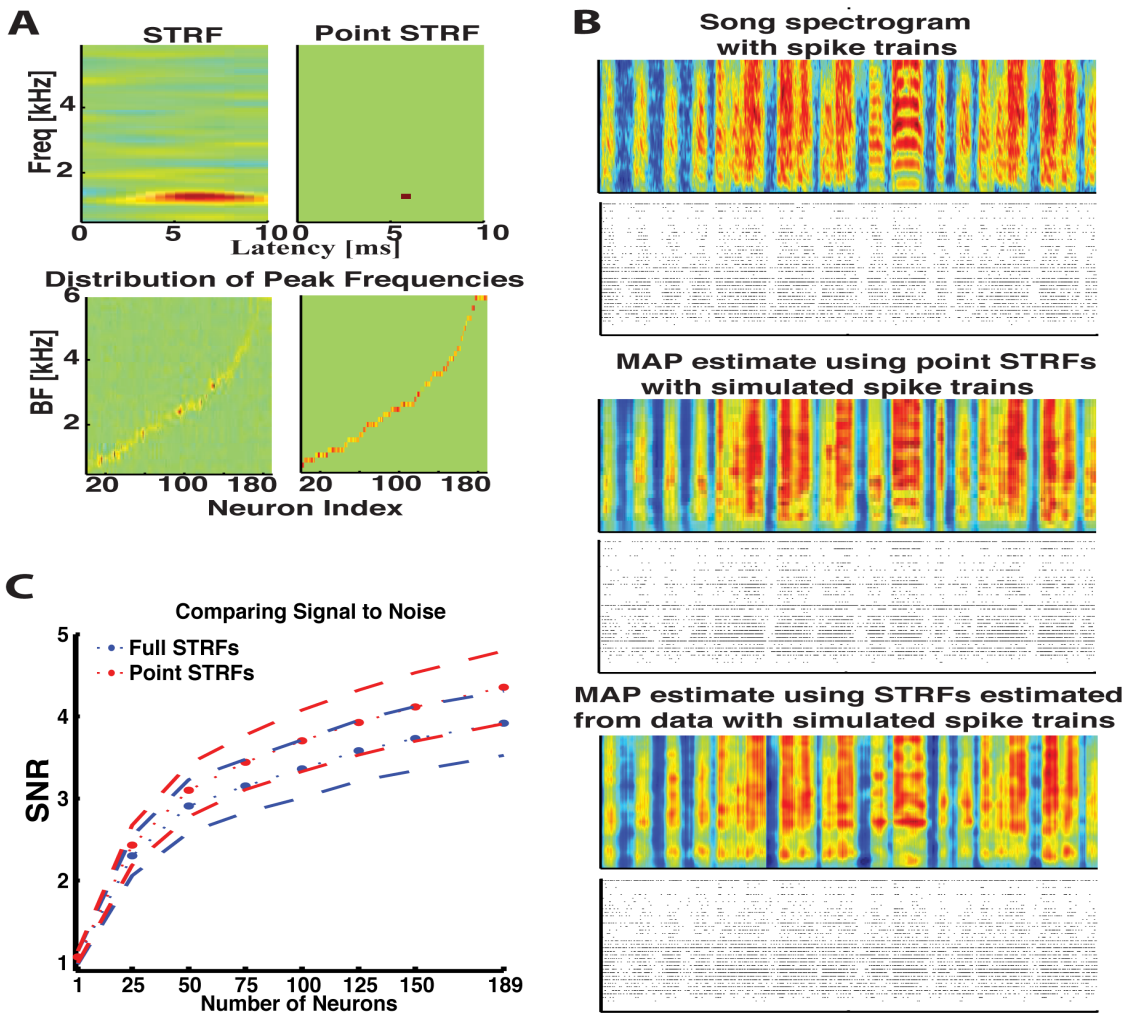

Figure 3.10: Spectral blur of STRFs causes a small loss of information for reconstructions. **A**, (upper, left panel) Example STRF and localized point STRF (upper, right panel) with equivalent peak frequency. (lower, left panel) Frequency vectors at latency where STRF obtains maximal value for the population of neurons used in this study. The equivalent plot for point STRFs (lower, right panel). Point STRF peak locations were randomly drawn from a distribution constructed using the peak locations of real STRFs. **B**, (first two rows) Song spectrogram and evoked responses of 189 real neurons. (middle rows) Reconstructed song spectrogram given simulated responses using a point STRF model. Simulated responses are shown immediately below the reconstruction. (bottom two rows) Reconstructed song spectrogram given simulated responses using full STRFs. Responses are shown immediately below the reconstruction. Reconstructions with full STRFs show slightly different spectral details but otherwise look *very* similar to reconstructions using point STRFs. **C**, SNR growth (plus and minus one standard error) as a function of the number of neurons used in decoding for point STRFs and full STRFs; on average the point STRFs have higher SNR than full STRFs.

3.5 Acknowledgements

We thank Ana Calabrese and Stephen David for helpful comments and discussions. We thank Columbia University Information Technology and the Office of the Executive Vice President for Research for providing the computing cluster used in this study. ADR is supported by an NSF Graduate Research Fellowship. LP is funded by a McKnight scholar award, and an NSF CAREER award. SMNW is funded by NSF, NIDCD, and the Searle Scholars Fund. LP and SMNW are supported by a Gatsby Initiative in Brain Circuitry pilot grant.

3.6 Appendix

3.6.1 Generating Samples from $p(\mathbf{x}, q|\mathbf{r}, \theta)$

To generate samples from the joint distribution, $p(\mathbf{x}, q|\mathbf{r}, \theta)$ described in section 3.2.5, we use a technique known as Gibbs sampling (Geman & Geman 1984, Robert & Casella 2005). Gibbs sampling works by iteratively sampling from the conditional distributions $p(\mathbf{x}|q, \mathbf{r}, \theta)$ and $p(q|\mathbf{x}, \mathbf{r}, \theta)$. First we initialize $\mathbf{x}$ to a matrix $\mathbf{x}_0$; we take $\mathbf{x}_0$ to be a matrix where each bin equals the vocalization period prior mean. Then we draw a sample, Q_0, from the conditional distribution $p(q|\mathbf{x} = \mathbf{x}_0, \mathbf{r}, \theta)$ (note that our hierarchical model does not depend on $\mathbf{r}$ or θ so that $p(q|\mathbf{x} = \mathbf{x}_0, \mathbf{r}, \theta) = p(q|\mathbf{x} = \mathbf{x}_0)$. From now on we will write this conditional distribution as $p(q|\mathbf{x})$). We then draw a sample, $\mathbf{x}_1$, from the conditional distribution $p(\mathbf{x}|Q_0, \mathbf{r}, \theta)$. Iterating this process by drawing alternating samples

$$\mathbf{x}_{i+1} \sim p(\mathbf{x}|Q_i, \mathbf{r}, \theta), \tag{3.6.37}$$

$$Q_{i+1} \sim p(q|\mathbf{x} = \mathbf{x}_{i+1}), \tag{3.6.38}$$

(where the notation $\mathbf{x} \sim p(\mathbf{x})$ indicates that $\mathbf{x}$ is drawn from the distribution p) results in samples, $(Q_i, \mathbf{x}_i)$, which converge to samples drawn from the joint distribution $p(\mathbf{x}, q|\mathbf{r}, \theta)$ (Geman & Geman 1984, Robert & Casella 2005).

The hierarchical prior models a spectrogram, $\mathbf{x}_i$, as a Gaussian vector with a mean, μ_q, and spectral covariance matrix, Φ'_q, that depend on the vector of vocalization states, q, as well as the previously defined (see methods) temporal covariance matrix, C_T. Such a Gaussian vector can be written as a linear transformation of an uncorrelated standard Gaussian vector. More precisely, the spectrogram can be written as

$$\mathbf{x}_i = \Phi_q'^{\frac{1}{2}} Z_i C_T^{\frac{1}{2}} + \mu_q, \tag{3.6.39}$$

where $\Phi^{1/2}$ and $C_T^{1/2}$ are the matrix square roots of Φ and C_T and Z_i is sampled from an uncorrelated standard Gaussian distribution

$$Z_i \sim \prod_{f=1}^{F} \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z_i(f,t)^2}{2}\right). \tag{3.6.40}$$

To determine $\mu_{q=0}$ and $\Phi'_{q=0}$ ($q = 0$ denotes a silent period) we construct a spectrogram, $\mathbf{x}'$, composed of all silent periods extracted from the data set of birdsongs (determined as described in the methods) and determine the empirical mean and covariance of $\mathbf{x}'$

$$\mu_0 = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{F} \sum_{f=1}^{F} \mathbf{x}'(f,n) \tag{3.6.41}$$

$$\Phi_0(f,f') = \frac{1}{N-1} \sum_{n=1}^{N} \left(\mathbf{x}'(f,n) - \frac{1}{N} \sum_{n'=1}^{N} \mathbf{x}'(f,n')\right)\left(\mathbf{x}'(f',n) - \frac{1}{N} \sum_{n'=1}^{N} \mathbf{x}'(f,n')\right), \tag{3.6.42}$$

where N is the total number of time-bins in the data set. The same is done to construct $\mu_{q=1}$ and $\Phi'_{q=1}$ except using a spectrogram composed of all vocal periods.

Although eqn. 3.6.39 depends on $C_T^{\frac{1}{2}}$ we will show that sampling only requires computation of the matrix $C_T^{-\frac{1}{2}}$. As discussed in section 3.2.5, we construct C_T^{-1} from a sparse matrix of Autoregressive coefficients (eqn. 3.2.14). As evidenced by eqn. 3.2.14, the bandwidth of the matrix C_T^{-1} does not grow with T allowing us to sample even for large values of T.

To draw a sample from $p(q|\mathbf{x} = \mathbf{x}_i)$ we first multiply $\mathbf{x}_i$ by $C_T^{-\frac{1}{2}}$, from eqn. 3.6.39

$$\mathbf{x}_i \rightarrow Y_i \quad = \quad \mathbf{x}_i C_T^{-\frac{1}{2}} = \Phi_q^{'\frac{1}{2}} Z_i + \mu_q C_T^{-\frac{1}{2}}. \tag{3.6.43}$$

This is done to create a collection of spectral vectors, $Y_i(.,t)$, that are conditionally independent given the latent variable q. At time t, the spectral vector $Y_i(.,t)$ is a Gaussian random variable drawn from a distribution whose mean and variance is determined by the value of q_t. Since q is a Markov process, Y_i forms a collection of spectral vectors that are emissions from a hidden-Markov model. We sample the $q(t)$ element of the vector q from the distribution $p(q(t)|q(t+1)...q(T), Y)$ using the forward filter-backward sample algorithm (de Gunst, et al. 2001). The algorithm uses the relation $p(q(t)|q(t+1)...q(T), Y) \propto p(q(t)|q(t-1))p(Y(.,1)...Y(.,t), q_t = n)$ to compute $p(q(t)|q(t+1)...q(T), Y)$. This relation is helpful because $p(q(t)|q(t-1))$ is the known transition matrix and the forward probabilities $\alpha_n(t) = p(Y(.,1)...Y(.,t), q_t = n)$ can be computed recursively using the conventional forward algorithm (Rabiner 1989). Given a method for computing the probabilities $p(q(t)|q(t+1)...q(T), Y)$, q(t) can then be sampled using inverse transform sampling.

Samples from $p(\mathbf{x}|q = Q_i, \mathbf{r}, \theta)$ are generated using a modified form of the Metropolis-Hastings (MH) algorithm known as Hybrid Monte-Carlo (HMC) (Roberts & Rosenthal 2001, Duane et al. 1987). A step-by-step description for implementing the HMC and its convergence properties can be found elsewhere (Ahmadian et al. 2011, Roberts & Rosenthal 2001). We simply note that a key step in effi-

ciently sampling the distribution $p(\mathbf{x}|q, \mathbf{r}, \theta)$ is using the inverse Hessian matrix of the log posterior, $\log p(\mathbf{x}|q, \mathbf{r}, \theta)$, evaluated at the MAP to construct the proposal distribution (again see Ahmadian (2010)). As previously noted (see methods) taking the inverse of the Hessian would be computationally expensive (scaling like $O(d^3)$ where d=FT) if the Hessian was not banded. In our case, the Hessian of the GLM log-likelihood, J, and Hessians of our Gaussian distributions, $C^{-1}(q) = \Phi_q'^{-1} C_T^{-1}$, are banded, making the Hessian of the log-posterior, given by $J + C^{-1}(q)$, also banded. In addition to the matrix $C^{-1}(q)$, the HMC algorithm specifies the proposal distribution with two parameters, the number of 'leapfrog iterations', L, and proposal distribution jump size, σ, which must be set by the user. We set L=1 and $\sigma = 0.9$ because we find that these values lead to relatively quick convergence rates for spectrogram estimates with 111 time bins (0.3 s). At each step of the Gibbs sampler we run the HMC for 100 iterations and keep the last sample.

We reconstruct spectrograms using $E[\mathbf{x}|\mathbf{r}]$, which we approximate by averaging the conditional means $E[\mathbf{x}|q = Q_i, \mathbf{r}, \theta]$ because, by Rao-Blackwellisation (Doucet, et al. 2000), this leads to a better estimate of $E[\mathbf{x}|\mathbf{r}]$ then averaging the samples $\mathbf{x}_i$. For Gaussian priors, the MAP is often a good approximation for the posterior mean (Ahmadian et al. 2011)

$$\hat{\mathbf{x}}_i = \arg\max_{\mathbf{x}} p(\mathbf{x}|q = Q_i, \mathbf{r}, \theta) \approx E[\mathbf{x}|q = Q_i, \mathbf{r}, \theta], \tag{3.6.44}$$

therefore each time we sample from $p(\mathbf{x}|q = Q_i, \mathbf{r}, \theta)$ we also calculate and store the most probable spectrogram, $\hat{\mathbf{x}}_i$, under this distribution. Using the approximation in eqn. 3.6.44, we compute $E[\mathbf{x}|\mathbf{r}]$ by averaging different values of $\hat{\mathbf{x}}_i$

$$E[\mathbf{x}|\mathbf{r}] \approx \frac{1}{N_{samp}} \sum_{j=1}^{N_{samp}} \hat{\mathbf{x}}_i. \tag{3.6.45}$$

We average together 100 independent Gibbs sampler chains. Each chain was

created by iterating the Gibbs sampler 100 times, burning the first 50 iterations and keeping the last 50 samples. Each chain was created on a separate machine using Columbia University's Hotfoot cluster.

Chapter 4

Concluding remarks

I have studied neural coding problems in the context of both the auditory and visual systems. In both cases, I focused on developing and applying accurate and computationally efficient methods for computing statistics of $p(\mathbf{r}|\mathbf{x})$ or $p(\mathbf{x}|\mathbf{r})$ when $\mathbf{r}$ and $\mathbf{x}$ were high-dimensional (hundreds to thousands of elements). In both cases the conditional response distribution was modeled with a GLM parameterized by a high-dimensional, necessitated by the high-dimensionality of the stimulus, vector θ. This was key for developing and applying the two principal methods described in this thesis.

The first method combines the structure of the exponential family with the law of large numbers, and often the central limit theorem, to average away the computationally expensive portion of a generalized linear model (see equation 2.2.16). I showed that taking this average reduces the computational cost of evaluating the GLM log-likelihood by a factor proportional to the number of parameters in the model times the number of observations. The power of this lies in the numerous applications of the likelihood in statistical inference. Importantly, I showed that for many applications this increase in efficiency did not come with a large sacrifice in accuracy.

The second method relies on recent connections made between high-dimensional decoding problems and well-known results from applied mathematics and engineering

for quickly solving linear systems of equations involving matrices with special structure (Paninski et al. 2009). The fundamental idea is that inferring the mode of $p(\mathbf{x}|\mathbf{r})$ in many situations, e.g. when the Newton-Raphson method is used, is equivalent to solving a series of systems of linear equations. Under a GLM model of $p(\mathbf{r}|\mathbf{x})$ and with a careful choice of the prior distribution $p(\mathbf{x})$, we showed that the matrices, $H_{\mathbf{x}}$,

$$H_{\mathbf{x}} = \nabla_{\mathbf{x}}\nabla_{\mathbf{x}}p(\mathbf{r}|\mathbf{x},\theta) + \nabla_{\mathbf{x}}\nabla_{\mathbf{x}}p(\mathbf{x}), \tag{4.0.1}$$

involved in these systems of linear equations had the requisite special structure. The speed of the method made it easier to investigate the posterior's $p(\mathbf{x}|\mathbf{r})$ dependence on the prior. In turn, this allowed my collaborators and I to discover that the fidelity of spectrogram reconstructions from MLd responses relies more heavily on prior knowledge of spectral correlations than temporal correlations.

Throughout this thesis we have emphasized computationally efficient methods. In this author's opinion, models whose hypotheses can be tested in a computationally efficient manner will become increasingly important for systems neuroscience. Similar sentiments have started to appear in the literature (Stevenson & Kording 2011, Alivisatos et al. 2012). This will be important for understanding how the nervous system processes high-dimensional 'natural' stimuli under 'natural' conditions while observing, ever increasing numbers of, simultaneously observed neurons. While computing power continues to increase under Moore's law, efficient methods will be critical for asking context-dependent questions (Maimon, et al. 2010) which require 'closed-loop' experiments (Lewi, et al. 2009b, Harvey, et al. 2012) to perturb network dynamics. I hope that this research suggests constructive ways to develop these methods from statistical principles.

Bibliography

L. Abbott & P. Dayan (2001). *Theoretical Neuroscience*. MIT Press.

M. Abeles (1988). *Information Processing by the Brain*, chap. Neural codes for higher brain functions. Hans Huber.

E. Adrian (1964). *The Basis of Sensation: The Action of the Sense Organs*. Hafner Publishing Company.

E. Adrian & Y. Zotterman (1926). 'The Journal of physiology; The impulses produced by sensory nerve-endings: Part II. The response of a Single End-Organ.' **61**(2):151–171.

Y. Ahmadian, et al. (2011). 'Efficient Markov Chain Monte Carlo Methods For Decoding Neural Spike Trains'. *Neural Computation* **1**(23):46–96.

A. P. Alivisatos, et al. (2012). 'The brain activity map project and the challenge of functional connectomics'. *Neuron* **74**(6):970–974.

D. Appeltants, et al. (2005). 'The effect of auditory distractors on song discrimination in male canaries (Serinus canaria)'. *Behav Processes* **69**(3):331–41.

A. K. Bansal, et al. (2012). 'Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: spikes, multiunit activity, and local field potentials'. *J Neurophysiol* **107**(5):1337–1355.

A. R. Barron, et al. (2008). 'Approximation and Learning By Greedy Algorithms'. *The Annals of Statistics* **36**(1):64–94.

J. M. Beck & A. Pouget (2007). 'Exact inferences in a neural implementation of a hidden Markov model'. *Neural Computation* **19**(5):1344–1361.

M. A. Bee & G. M. Klump (2004). 'Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain'. *J Neurophysiol* **92**(2):1088–104.

P. Berkes, et al. (2009). 'A structured model of video reproduces primary visual cortical organisation.'. *PLoS Comput Biol* **5**(9):e1000495.

W. Bialek, et al. (1991). 'Reading a neural code'. *Science* **252**:1854–1857.

C. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer.

B. L. Bloodgood & B. L. Sabatini (2007). 'Ca(2+) signaling in dendritic spines'. *Curr Opin Neurobiol* **17**(3):345–51.

L. Bottou (1998). 'Online Algorithms and Stochastic Approximations'. In D. Saad (ed.), *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK.

D. Brillinger (1988). 'Maximum likelihood analysis of spike trains of interacting nerve cells'. *Biological Cyberkinetics* **59**:189–200.

D. Brillinger (1992). 'Nerve cell spike train data analysis: a progression of technique'. *Journal of the American Statistical Association* **87**:260–271.

E. Brown, et al. (2004). 'Multiple neural spike train data analysis: state-of-the-art and future challenges'. *Nature Neuroscience* **7**(5):456–461.

A. Calabrese, et al. (2011). 'A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds'. *PLoS One* **6**(1):e16104.

C. Catchpole & P. Slater (1995). *Bird song: Biological themes and variations*. Cambridge.

S. Cavaco & M. S. Lewicki (2007). 'Statistical modeling of intrinsic structures in impacts sounds.'. *J Acoust Soc Am* **121**(6):3558–3568.

E. Chornoboy, et al. (1988). 'Maximum likelihood identification of neural point process systems'. *Biological Cybernetics* **59**:265–275.

R. Cossart, et al. (2003). 'Attractor dynamics of network UP states in the neocortex'. *Nature* **423**:283–288.

M. de Gunst, et al. (2001). 'Statistical Analysis of Ion Channel Data Using Hidden Markov Models With Correlated State-Dependent Noise and Filtering'. *JASA* **96**:805–815.

R. C. Decharms, et al. (1998). 'Optimizing sound features for cortical neurons.'. *Science* **280**(5368):1439–1443.

P. Diaconis & D. Freedman (1984). 'Asymptotics of Graphical Projection Pursuit'. *The Annals of Statistics* **12**(3):793–815.

J. P. Donoghue (2002a). 'Connecting cortex to machines: recent advances in brain interfaces'. *Nature Neuroscience* **5**:1085–1088.

J. P. Donoghue (2002b). 'Connecting cortex to machines: recent advances in brain interfaces'. *Nat Neurosci* **5 Suppl**:1085–8.

A. Doucet, et al. (2000). 'On sequential Monte Carlo sampling methods for Bayesian filtering'. *Statistics and Computing* **10**:197–208.

R. Drullman, et al. (1994). 'Effect of temporal envelope smearing on speech reception'. *J Acoust Soc Am* **95**(2):1053–64.

S. Duane, et al. (1987). 'Hybrid Monte Carlo'. *Physics Letters B* **195**(2):216 – 222.

B. Efron, et al. (2004). 'Least Angle Regression'. *The Annals of Statistics* **32**(2):407–499.

J. Eggermont (2001). 'Between sound and perception: reviewing the search for a neural code'. *Hearing Research* **157**:1–42.

J. J. Eggermont, et al. (1983). 'Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field'. *Hearing Research* **10**(2):167–190.

K. T. Fang, et al. (1990). *Symmetric multivariate and related distributions*. CRC Monographs on Statistics and Applied Probability. Chapman and Hall.

G. D. Field, et al. (2010). 'Functional connectivity in the retina at the resolution of photoreceptors'. *Nature* **467**(7316):673–7.

J. H. Friedman, et al. (2010). 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. *Journal of Statistical Software* **33**(1):1–22.

A. Gelman, et al. (2003a). *Bayesian data analysis*. Chapman and Hall/CRC, 2nd ed. edn.

A. Gelman, et al. (2003b). *Bayesian data analysis*. Chapman and Hall/CRC, 2nd ed. edn.

S. Geman & D. Geman (1984). 'Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**:721–741.

T. Q. Gentner & D. Margoliash (2002). *Acoustic Communication*, chap. 7, The Neuroethology of Vocal Communication: Perception and Cognition. Springer-Verlag.

R. Godard (1991). 'Long-term memory for individual neighbors in a migratory songbird'. *Nature* **350**:228–229.

G. Golub & C. van Van Loan (1996). *Matrix Computations*. (Johns Hopkins Studies in Mathematical Sciences). The Johns Hopkins University Press, 3rd edn.

C. D. Harvey, et al. (2012). 'Choice-specific sequences in parietal cortex during a virtual-navigation decision task'. *Nature* **484**(7392):62–68.

M. Hauber, et al. (2010). 'Functional role and female perception of male song in zebra finches'. *Emu-Austral Ornithology* .

G. H. Hesselmans & P. I. Johannesma (1989). 'Spectro-temporal interpretation of activity patterns of auditory neurons'. *Mathematical Biosciences* **93**(1):31 – 51.

G. Hooker & A. Vidyashankar (2012). 'Bayesian Model Robustness via Disparities'. *under review* .

S. Iyengar (2001). 'The analysis of multiple neural spike trains'. In *Advances in Methodological and Applied Aspects of Probability and Statistics*, pp. 507–524. Gordon and Breach.

R. I. Jennrich & P. F. Sampson (1976). 'Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation'. *Technometrics* **18**(1):pp. 11–17.

R. A. Johnson & D. W. Wichern (2007). *Applied multivariate statistical analysis*. Pearson Prentice Hall.

Y. Karklin & M. S. Lewicki (2005). 'A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals.'. *Neural Comput* **17**(2):397–423.

R. Kass & A. Raftery (1995a). 'Bayes Factors'. *Journal of the American Statistical Association* **90**:773–795.

R. Kass & A. Raftery (1995b). 'Bayes Factors'. *Journal of the American Statistical Association* **90**:773–795.

D. P. Knudsen & T. Q. Gentner (2010). 'Mechanisms of song perception in oscine birds'. *Brain and Language* **1**(115):59–68.

J. A. Kogan & D. Margoliash (1998). 'Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study'. *Journal of the Acoustical Society of America* **103**:2185–2196.

S. Koyama, et al. (2010). 'Bayesian decoding of neural spike trains'. *Annals of the Institute of Statistical Mathematics* **62**(1):37–59.

S. I. Lee, et al. (2006). 'Efficient L1 Regularized Logistic Regression'. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*.

E. L. Lehmann & G. Casella (1998). *Theory of Point Estimation*. Springer, second edn.

E. L. Lehmann & J. P. Romano (2005). *Testing Statistical Hypotheses*. Springer.

J. Lewi, et al. (2009a). 'Sequential optimal design of neurophysiology experiments'. *Neural Computation* **21**:619–687.

J. Lewi, et al. (2009b). 'Sequential optimal design of neurophysiology experiments'. *Neural Computation* **21**:619–687.

M. S. Lewicki & B. J. Arthur (1996). 'Hierarchical organization of auditory temporal context sensitivity.'. *Journal of Neuroscience* **16**(21):6987–6998.

A. Litke, et al. (2004). 'What does the eye tell the brain? Development of a system for the large scale recording of retinal output activity'. *IEEE Trans Nucl Sci* pp. 1434–1440.

S. Litvak & S. Ullman (2009). 'Cortical Circuitry Implementing Graphical Models'. *Neural Computation* **21**(11):3010–3056.

B. Lohr & R. Dooling (1998). 'Detection of changes in timbre and harmonicity in complex sounds by zebra finches (Taeniopygia guttata) and budgerigars (Melopsittacus undulatus)'. *Journal of Comparative Psychology* **112**(1):36–47.

H. Lütcke, et al. (2010). 'Optical recording of neuronal activity with a genetically-encoded calcium indicator in anesthetized and freely moving mice'. *frontiers in Neural Circuits* **4**(9):1–12.

G. Maimon, et al. (2010). 'Active flight increases the gain of visual motion processing in Drosophila'. *Nat Neurosci* **13**(3):393–9.

W. Q. Malik, et al. (2011). 'Efficient decoding with steady-state Kalman filter in neural interface systems'. *IEEE Trans Neural Syst Rehabil Eng* **19**(1):25–34.

S. Mallat & Z. Zhang (1993). 'Matching pursuits with time-frequency dictionaries'. *IEEE Transactions on Signal Processing* **41**(12):3397–3415.

P. L. Marchenko, V.A (1967). 'Distribution of eigenvalues for some sets of random matrices'. *Mat.Sb* **72**.

D. Margoliash & E. S. Fortune (1992). 'Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc.'. *J Neurosci* **12**(11):4309–4326.

P. R. Marler & H. Slabbekoorn (2004). *Nature's Music: The Science of Birdsong*. Academic Press.

L. Martignon, et al. (2000). 'Neural coding: higher-order temporal patterns in the neuro-statistics of cell assemblies'. *Neural Computation* **12**:2621–2653.

P. McCullagh & J. A. Nelder (1989). *Generalized Linear Models*. Chapman and Hall/CRC, second edn.

J. H. McDermott, et al. (2009). 'Sound Texture Synthesis Via Filter Statistics'. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* .

N. Mesgarani, et al. (2009). 'Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex'. *Journal of Neurophysiology* **102**(6):3329–3339.

S. Murphy & A. van der Vaart (2000). 'On Profile Likelihood'. *Journal of the American Statistical Association* **95**(450):449–465.

K. I. Nagel, et al. (2010). 'Differential influence of frequency, timing, and intensity cues in a complex acoustic categorization task'. *J Neurophysiol* **104**(3):1426–37.

R. Narayan, et al. (2006). 'Distinct Time Scales in Cortical Discrimination of Natural Sounds in Songbirds'. *Journal of Neurophysiology* **96**(1):252–258.

Y. Nesterov (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 1 edn.

M. Nicolelis, et al. (2003). 'Chronic, multisite, multielectrode recordings in macaque monkeys'. *PNAS* **100**:11041–11046.

V. Nikolenko, et al. (2008). 'SLM Microscopy: Scanless Two-photon Imaging and Photostimulation using Spatial Light Modulators'. *Frontiers in Neural Circuits* **2**:5.

J. Nishikawa, et al. (2008). 'Population coding of song element sequence in the Bengalese finch HVC.'. *Eur J Neurosci* **27**(12):3273–3283.

D. Nykamp (2005). 'Revealing pairwise coupling in linear-nonlinear networks'. *SIAM Journal on Applied Mathematics* **65**:2005–2032.

K. Ohki, et al. (2005). 'Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex'. *Nature* **433**:597–603.

A. L. O'Loghlen & M. D. Beecher (1997). 'Sexual preferences for mate song types in female song sparrows'. *Animal Behaviour* **53**:835–841.

M. Osborne, et al. (2000). 'A new approach to variable selection in least squares problems'. *IMA Journal of Numerical Analysis* **20**(3):389–403.

M. S. Osmanski, et al. (2009). 'Discrimination of auditory gratings in birds'. *Hearing Research* **256**(1-2):11 – 20.

L. Paninski (2004). 'Maximum likelihood estimation of cascade point-process neural encoding models'. *Network: Computation in Neural Systems* **15**:243–262.

L. Paninski, et al. (2009). 'A new look at state-space models for neural data'. *Journal of Computational Neuroscience* **29**(1-2):107–126.

L. Paninski, et al. (2004a). 'Superlinear population encoding of dynamic hand trajectory in primary motor cortex'. *Journal of Neuroscience* **24**(39):8551–8561.

L. Paninski, et al. (2004b). 'Superlinear population encoding of dynamic hand trajectory in primary motor cortex'. *J. Neurosci.* **24**:8551–8561.

Park, et al. (2008). 'The Bayesian Lasso'. *Journal of the American Statistical Association* **103**(482):681–686.

I. M. M. Park & J. W. Pillow (2011). 'Bayesian Spike-Triggered Covariance Analysis'. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 1692–1700.

F. W. Peek (1972). 'An experimental study of the territorial function of vocal and visual display in the male red-winged blackbird (Agelaius phoeniceus)'. *Animal Behaviour* **20**:112–118.

D. B. Percival & A. T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press.

D. Perkel & D. Bullock (1969). 'Neural coding'. In F. Schmitt, T. Melenchuk, G. Quarton, & G. Adelman (eds.), *Neurosciences Research Symposium Summaries*, vol. 3, pp. 405–528. MIT Press.

J. Pillow, et al. (2011). 'Model-based decoding, information estimation, and change-point detection in multi-neuron spike trains.'. *Neural Computation* **1**(23):1–45.

J. W. Pillow, et al. (2008). 'Spatio-temporal correlations and visual signalling in a complete neuronal population'. *Nature* **454**(7207):995–999.

H. Plesser & W. Gerstner (2000). 'Noise in Integrate-and-Fire Neurons: From Stochastic Input to Escape Rates'. *Neural Computation* **12**:367–384.

L. Rabiner (1989). 'A tutorial on Hidden Markov Models and selected applications in speech recognition'. *Proceedings of the IEEE* **77**:257–286.

C. Rasmussen & C. Williams (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning series)*. The MIT Press.

F. Rieke, et al. (1995). 'Naturalistic Stimuli Increase the Rate and Efficiency of Information Transmission by Primary Auditory Afferents'. *Proceedings of the Royal Society B: Biological Sciences* **262**(1365):259–265.

F. Rieke, et al. (1997). *Spikes: Exploring the Neural Code*. The MIT Press.

F. Rigat, et al. (2006). 'Bayesian modelling and analysis of spatio-temporal neuronal networks'. *Bayesian Analysis* **1**:733–764.

C. Robert & G. Casella (2005). *Monte Carlo Statistical Methods*. Springer.

G. O. Roberts & J. S. Rosenthal (2001). 'Optimal Scaling for Various Metropolis-Hastings Algorithms'. *Statistical Science* **16**(4):351–367.

K. Sadeghi & L. Paninski (2012). 'Inferring retinal cone locations and functional connectivity from ganglion cell recordings'. *In preparation* .

M. Saleh, et al. (2012a). 'Encoding of Coordinated Reach and Grasp Trajectories in Primary Motor Cortex'. *Journal of Neuroscience* **32**(4):1220–1232.

M. Saleh, et al. (2012b). 'Encoding of Coordinated Reach and Grasp Trajectories in Primary Motor Cortex'. *Journal of Neuroscience* **32**(4):1220–1232.

G. Santhanam, et al. (2006a). 'A high-performance brain–computer interface'. *Nature* **442**(7099):195–198.

G. Santhanam, et al. (2006b). 'A high-performance brain–computer interface'. *Nature* **442**(7099):195–198.

D. Schneider & S. M. Woolley (2010). 'Discrimination of communication vocalizations by single neurons and groups of neurons in the auditory midbrain'. *Journal of Neurophysiology* **6**(103):3248–65.

E. Schneidman, et al. (2006). 'Weak pairwise correlations imply strongly correlated network states in a neural population'. *Nature* **440**:1007–1012.

M. Schnitzer & M. Meister (2003). 'Multineuronal firing patterns in the signal from eye to brain'. *Neuron* **37**:499–511.

K. Sen, et al. (2001). 'Feature analysis of natural sounds in the songbird auditory forebrain.'. *J Neurophysiol* **86**(3):1445–1458.

J. P. Shaffer (1991). 'The Gauss-Markov Theorem and Random Regressors'. *The American Statistician* **45**(4):269–273.

R. V. Shannon, et al. (1995). 'Speech recognition with primarily temporal cues'. *Science* **270**(5234):303–4.

B. G. Shinn-Cunningham, et al. (2007). *Behavioral and Neural Identification of Birdsong under Several Masking Conditions*. Springer Berlin Heidelberg.

E. Simoncelli, et al. (2004). 'Characterization of neural responses with stochastic stimuli'. In *The Cognitive Neurosciences*. MIT Press, 3rd edn.

N. Singh & F. Theunissen (2003). 'Modulation spectra of natural sounds and ethological theories of auditory processing'. *Journal of the Acoustical Society of America* **114**(6, Part 1):3394–3411.

P. Sollich & C. K. I. Williams (2005). 'Understanding gaussian process regression using the equivalent kernel'. In *Proceedings of the First international conference on Deterministic and Statistical Methods in Machine Learning*, pp. 211–228, Berlin, Heidelberg. Springer-Verlag.

I. H. Stevenson & K. P. Kording (2011). 'How advances in neural recording affect data analysis'. *Nature Neuroscience* **14**(2):139–142.

F. Theunissen, et al. (2001). 'Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli'. *Network: Computation in Neural Systems* **12**:289–316.

F. E. Theunissen, et al. (2000). 'Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds.'. *J Neurosci* **20**(6):2315–2331.

F. E. Theunissen & S. S. Shaevitz (2006). 'Auditory processing of vocal sounds in birds'. *Current Opinion in Neurobiology* **16**(4):400 – 407.

J. Tropp (2004). 'Greed is Good: Algorithmic Results for Sparse Approximation'. *IEEE Transactions on Information Theory* **50**(10):2231–2242.

W. Truccolo, et al. (2005). 'A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble and Extrinsic Covariate Effects'. *Journal of Neurophysiology* **93**:1074–1089.

K. Utikal (1997). 'A new method for detecting neural interconnectivity'. *Biological Cyberkinetics* **76**:459–470.

A. van der Vaart (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

M. Vidne, et al. (2011a). 'Modeling the impact of common noise inputs on the network activity of retinal ganglion cells'. *Journal of Computational Neuroscience* .

M. Vidne, et al. (2011b). 'Modeling the impact of common noise inputs on the network activity of retinal ganglion cells'. *J Comput Neurosci* .

L. Wang, et al. (2007). 'Cortical Discrimination of Complex Natural Stimuli: Can Single Neurons Match Behavior?'. *Journal of Neuroscience* **27**(3):582–589.

D. Warland, et al. (1997). 'Decoding visual information from a population of retinal ganglion cells'. *Journal of Neurophysiology* **78**:2336–2350.

S. Woolley, et al. (2005). 'Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds'. *Nature Neuroscience* **8**(10):1371–1379.

S. Woolley, et al. (2006). 'Stimulus-dependent auditory tuning results in synchronous population coding of vocalizations in the songbird midbrain'. *Journal of Neuroscience* **26**(9):2499–2512.

S. M. N. Woolley, et al. (2009). 'Functional Groups in the Avian Auditory System'. *Journal of Neuroscience* **29**(9):2780–2793.

W. Wu, et al. (2009). 'Neural decoding of hand motion using a linear state-space model with hidden states'. *IEEE Trans Neural Syst Rehabil Eng* **17**(4):370–378.

B. M. Yu, et al. (2007). 'Mixture of Trajectory Models for Neural Decoding of Goal-Directed Movements'. *J Neurophysiol* **97**(5):3763–3780.

R. Zemel, et al. (1998). 'Probabilistic interpretation of population codes'. *Neural Computation* **10**:403–430.

T. Zhang (2011). 'Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations'. *IEEE Transactions on Information Theory* **57**(7):4689–4708.