High dimensional Information Processing

Kamiar Rahnama Rad

Submitted in partial fulfillment of the

Requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

ABSTRACT

High dimensional Information Processing

Kamiar Rahnama Rad

Part I: Consider the n-dimensional vector $y = X\beta + \epsilon$ where $\beta \in \mathbb{R}^p$ has only k nonzero entries and $\epsilon \in \mathbb{R}^n$ is a Gaussian noise. This can be viewed as a linear system with sparsity constraints corrupted by noise, where the objective is to estimate the sparsity pattern of β given the observation vector y and the measurement matrix X. First, we derive a non-asymptotic upper bound on the probability that a specific wrong sparsity pattern is identified by the maximum-likelihood estimator. We find that this probability depends (inversely) exponentially on the difference of $\|X\beta\|_2$ and the ℓ_2-norm of $X\beta$ projected onto the range of columns of X indexed by the wrong sparsity pattern. Second, when X is randomly drawn from a Gaussian ensemble, we calculate a non-asymptotic upper bound on the probability of the maximum-likelihood decoder not declaring (partially) the true sparsity pattern. Consequently, we obtain sufficient conditions on the sample size n that guarantee almost surely the recovery of the true sparsity pattern. We find that the required growth rate of sample size n matches the growth rate of previously established necessary conditions.

Part II: Estimating two-dimensional firing rate maps is a common problem, arising in a number of contexts: the estimation of place fields in hippocampus, the analysis of temporally nonstationary tuning curves in sensory and motor areas, the

estimation of firing rates following spike-triggered covariance analyses, etc. Here we introduce methods based on Gaussian process nonparametric Bayesian techniques for estimating these two-dimensional rate maps. These techniques offer a number of advantages: the estimates may be computed efficiently, come equipped with natural errorbars, adapt their smoothness automatically to the local density and informativeness of the observed data, and permit direct fitting of the model hyperparameters (e.g., the prior smoothness of the rate map) via maximum marginal likelihood. We illustrate the flexibility and performance of the new techniques on a variety of simulated and real data.

Part III: Many fundamental questions in theoretical neuroscience involve optimal decoding and the computation of Shannon information rates in populations of spiking neurons. In this paper, we apply methods from the asymptotic theory of statistical inference to obtain a clearer analytical understanding of these quantities. We find that for large neural populations carrying a finite total amount of information, the full spiking population response is asymptotically as informative as a single observation from a Gaussian process whose mean and covariance can be characterized explicitly in terms of network and single neuron properties. The Gaussian form of this asymptotic sufficient statistic allows us in certain cases to perform optimal Bayesian decoding by simple linear transformations, and to obtain closed-form expressions of the Shannon information carried by the network. One technical advantage of the theory is that it may be applied easily even to non-Poisson point process network models; for example, we find that under some conditions, neural populations with strong history-dependent (non-Poisson) effects carry exactly the same information as do simpler equivalent populations of non-interacting Poisson neurons with matched firing rates. We argue that our findings help to clarify some results from the recent literature on neural decoding and neuroprosthetic design.

Part IV: A model of distributed parameter estimation in networks is introduced,

where agents have access to partially informative measurements over time. Each agent faces a local identification problem, in the sense that it cannot consistently estimate the parameter *in isolation*. We prove that, despite local identification problems, if agents update their estimates recursively as a function of their neighbors' beliefs, they can consistently estimate the true parameter provided that the communication network is strongly connected; that is, there exists an information path between any two agents in the network. We also show that the estimates of all agents are asymptotically normally distributed. Finally, we compute the asymptotic variance of the agents' estimates in terms of their observation models and the network topology, and provide conditions under which the distributed estimators are as efficient as any centralized estimator.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my gratitude to Liam Paninski for his friendship, wisdom and mentorship. I would like to thank Martin Lindquist, David Madigan, Ken Miller and Chriss Wiggins for serving on my committee and providing me with critical comments. I want to thank the faculty, staff and students in the statistics department and the center for theoretical neuroscience for the nice environment they created. I want to deeply thank Dood Kalachiran for separating bureaucracy from research. I am extremely grateful to my parents, my sister Kimia, my brother Kian, and my partner Jeiran for their love.

Kamiar Rahnama Rad
New York, August 2011.

To my parents Jafar and Azam

Chapter 1

Introduction

"Nothing seems to me less likely than that a scientist or mathematician who reads me should be seriously influenced in the way he works.'

Ludwig Josef Johann Wittgenstein

The different chapters of this thesis regard problems in various field as essentially the same. Many of these problems share similar challenges: There is a tremendous amount of information in the presence of excessive irrelevant signals. How can a neuron in the brain or an agent in a social network efficiently process the constant flow of information when each bit of information in isolation is relatively void of information? There are two fundamental questions: What are limits of any method to decipher the relevant information? What are the practical methods whose performance reaches those limits?

In the second chapter, we discuss the fundamental limits imposed on any sparsity patter recovery problem. Finding solutions to underdetermined systems of equations arises in a wide array problems in science and technology; examples include array signal processing (Zibulevsky and Pearlmutter, 2001), neural (Vinje and Gal-

lant, 2000) and genomic data analysis (di Bernardo et al., 2005), to name a few. In many of these applications, it is natural to seek for *sparse* solutions of such systems, i.e., solutions with few nonzero elements. A common setting is when we believe or we know *a priori* that only a *small subset* of the candidate sources, neurons, or genes influence the observations, but their location is unknown. How well can any method recover the location of the influential factors, assuming that in fact only a few factors are responsible for the output? The answer is complicated. Thus, we look at it in it's simplest form: observations are available from a linear model with additive noise where the vector of interest is sparse. We characterize the phase-diagram of the error probability in terms of a minimal number of parameters.

In the third chapter, we discuss a more practical point of view on high dimensional information processing. Imagine by looking at the activity of a single neuron we want to estimate the location of a rat. At any moment we observe at most a single bit of information whereas the location is a point in a two dimensional space. How can a sequence of single bits help us follow the temporally varying location of the rat? The simplest idea is to accumulate information, that is, don't look at the single bits in isolation because they are almost void of information; rather, look at the temporal pattern. Further, rats can not change locations instantly; if the rat is now "here", it is also very likely that the rat will be around "here" during the next time step. Therefore, observing a single bit at this moment is also partially informative about the rat's location a few moment before and after. How to practically implement it, so that it can perform optimally given the current computational constraints is the subject of this chapter.

In the fourth chapter, we discuss again the fundamental limits imposed on any biological system that aims to decipher the sensory input from the activity of a coupled neural network. The fundamental limits are achieved by an ideal Bayesian observer. Additionally, the question of how to practically achieve that limit is also

discussed; we show that the ideal observer can also be realistic and biologically plausible.

Finally, in the last chapter information aggregation is regarded as a fundamental problem in multi-agent systems. In many scenarios, observations are distributed throughout the network in such a way that no agent has access to enough data to learn a relevant parameter in isolation, and therefore, agents face the task of recovering the truth by engaging in communication with one another. How should they communicate so that the wisdom of the crowd is transmitted across every individual; that is, can agents share information such that eventually they reach a consensus about the state of the world which is as close as possible to the truth? The final chapter answers this questions and provides details on a communication scheme that guarantees that the fundamental ideal global observer's performance is achieved by every individual.

Chapter 2

Nearly Sharp Sufficient Conditions on Sparsity Pattern Recovery

This chapter is based on the paper "Nearly Sufficient Conditions on Exact Sparsity Pattern Recovery " (Rahnama Rad, 2011).

2.1 Introduction

Finding solutions to underdetermined systems of equations arises in a wide array problems in science and technology; examples include array signal processing (Zibulevsky and Pearlmutter, 2001), neural (Vinje and Gallant, 2000) and genomic data analysis (di Bernardo et al., 2005), to name a few. In many of these applications, it is natural to seek for *sparse* solutions of such systems, i.e., solutions with few nonzero elements. A common setting is when we believe or we know *a priori* that only a *small subset* of the candidate sources, neurons, or genes influence

the observations, but their location is unknown.

More concretely, the problem we consider is that of estimating the support of $\beta \in \mathbb{R}^p$ given the *a priori* knowledge that only k of its entries are nonzero based on the observational model

$$y = X\beta + \epsilon \tag{2.1}$$

where $X \in \mathbb{R}^{n\times p}$ is a collection of input measurement vectors, $y \in \mathbb{R}^n$ is the output measurement and $\epsilon \in \mathbb{R}^n$ is the additive measurement noise, assumed to be zero mean and with known covariance equal to $I_{n\times n}$ [1]. Each row of X and the corresponding entry of y are viewed as an input and output measurement, respectively.

The output of the optimal (sparsity) decoder is defined as the support set of the sparse solution $\hat{\beta}$ with support size k that minimizes the residual sum of squares where

$$\hat{\beta} = \underset{|\text{support}(\theta)|=k}{\arg\min} \|y - X\theta\|_2^2 \tag{2.2}$$

is the optimal estimate of β given the *a priori* information of sparseness. The support set of $\hat{\beta}$ is optimal in the sense of minimizing the probability of identifying a wrong sparsity pattern.

First, we are concerned with the likelihood of the sparsity pattern of $\hat{\beta}$ as a function of X and β. We obtain an upper bound on the probability that $\hat{\beta}$ has any specific sparsity pattern and find that this bound depends (inversely) exponentially on the difference of $\|X\beta\|_2$ and the ℓ_2-norm of $X\beta$ projected onto the range of columns of X indexed by the wrong sparsity pattern.

Second, when the entries of X are independent and identically distributed (i.i.d.) random variables we are concerned with establishing sufficient conditions that guarantee the reliability of sparsity pattern recovery. Ideally, we would like to charac-

[1] This entails no loss of generality, by standard rescaling of β.

terize such conditions based on a minimal number of parameters including the sparsity level k, the signal dimension p, the number of measurements n and the signal-to-noise ratio(SNR) which is equal to

$$\mathrm{SNR} = \frac{\mathrm{E}\left[\|X\beta\|_2^2\right]}{\mathrm{E}\left[\|\epsilon\|_2^2\right]}. \tag{2.3}$$

Assume that the absolute value of the non zero entries of β are lower bounded by $\beta_{\min}$ [2]. Further, suppose that the variance of the entries of X is equal to one [1]. Hence,

$$\mathrm{SNR} \geq k\beta_{\min}^2$$

and therefore it is natural to ask, how does the ability to reliably estimate the sparsity pattern depend on $(n, p, k, \beta_{\min}^2)$.

We find that a non-asymptotic upper bound on the probability of the maximum-likelihood decoder not declaring the true sparsity pattern can be found when the entries of the measurement matrix are independent and identically distributed (i.i.d.) normal random variables. This allows us to obtain sufficient conditions on the number of measurements n as a function of $(p, k, \beta_{\min}^2)$ for reliable sparsity recovery. We show that our results strengthen earlier sufficient conditions (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Karbasi et al., 2009), and we show that the sufficient conditions on n match the growth rate of the necessary conditions in both the linear, i.e., $k = \Theta(p)$, and the sub-linear, i.e., $k = o(p)$, regimes, as long as $\beta_{\min}^2$ is $\Omega(\frac{1}{k})$ and $O(1)$.

2.1.1 Previous Work

A large body of recent work, including (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Karbasi et al., 2009; Wang et al., 2008; Reeves and Gast-

[2] To the best of our knowledge, Wainwright (Wainwright, 2007) was the first to formulate the information theoretic limitations of sparsity pattern recovery using $\beta_{\min}$ as one of the key parameters.

par, 2008; Wainwright, 2009), analyzed reliable sparsity pattern recovery exploiting optimal and sub-optimal decoders for large random Gaussian measurement matrices. The average error probability, necessary and sufficient conditions for sparsity pattern recovery for Gaussian measurement matrices were analyzed in (Wainwright, 2007) in terms of $(n, p, k, \beta^2_{\min})$. As a generalization of the previous work, using the Fano inequality, necessary conditions for general random and sparse measurement matrices were presented in (Wang et al., 2008). The sufficient conditions in (Fletcher et al., 2008) were obtained based on a simple maximum correlation algorithm and a closely related thresholding estimator discussed in (Rauhut et al., 2008). In addition to the well known formulation of the necessary and sufficient conditions based on $(n, p, k, \beta^2_{\min})$, Fletcher et al. (Fletcher et al., 2008) included the maximum-to-average ratio[3] of β in their analysis. Necessary and sufficient conditions for fractional sparsity pattern recovery were analyzed in (Akcakaya and Tarokh, 2008; Reeves and Gastpar, 2008).

We will discuss the relationship to this work below in more depth, after describing our analysis and results in more detail.

2.1.2 Notation.

The following conventions will remain in effect throughout this paper. Calligraphic letters are used to indicate sparsity patterns defined as a set of integers between 1 and p, with cardinality k. We say $\beta \in \mathbb{R}^p$ has sparsity pattern $\mathcal{T}$ if the entries with indices $i \in \mathcal{T}$ are nonzero. $\mathcal{T} - \mathcal{F}$ stands for the set of entries that are in $\mathcal{T}$ but not in $\mathcal{F}$ and $|\mathcal{T}|$ for the cardinality of $\mathcal{T}$. We denote by $X_{\mathcal{T}} \in \mathbb{R}^{n \times |\mathcal{T}|}$, the matrix obtained from X by extracting $|\mathcal{T}|$ columns with indices obeying $i \in \mathcal{T}$. Let $\mathcal{S}(\beta)$ stand for the sparsity pattern or support set of β. The matrix norm $\|.\|_{a,b}$ of a

[3]The maximum-to-average ratio of β was defined as $k\beta^2_{\min}/\|\beta\|^2_2$.

matrix A defined as

$$\|A\|_{a,b} := \max_{x \neq 0} \frac{\|Ax\|_a}{\|x\|_b}.$$

Note that if A is a positive semi-definite matrix then $\|A\|_{2,2}$ is equal to the top eigenvalue of A. Except for the matrix norm $\|.\|_{2,2}$ all vector norms are ℓ_2, $\|\cdot\| = \|\cdot\|_2$. Finally, let the orthonormal operator projecting into the subspace spanned by the columns of $X_{\mathcal{F}}$ be defined as $\Pi_{\mathcal{F}} = X_{\mathcal{F}}(X_{\mathcal{F}}^T X_{\mathcal{F}})^{-1} X_{\mathcal{F}}^T$.

2.2 Results

For the observational model in equation (2.1), assume that the true sparsity model is $\mathcal{T}$; as a result,

$$y = X_{\mathcal{T}} \beta_{\mathcal{T}} + \epsilon. \tag{2.4}$$

We first state a result on the probability of the event $\mathcal{S}(\hat{\beta}) = \mathcal{F}$, i.e. $\Pr[\mathcal{S}(\hat{\beta}) = \mathcal{F}|X, \beta, \mathcal{T}]$, for any $\mathcal{F} \neq \mathcal{T}$ and any measurement matrix X.

Theorem 1. *For the observational model of equation (2.4) and estimate $\hat{\beta}$ in equation (2.2), the following bound holds:*

$$\Pr\left[\mathcal{S}(\hat{\beta}) = \mathcal{F}|X, \beta, \mathcal{T}\right] \leq \exp\left\{-\frac{C}{2}\left\|(I - \Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\right\|^2 + \frac{|\mathcal{T} - \mathcal{F}|}{2}\right\},$$

where $C = 3 - 2\sqrt{2}$.

The proof of Theorem 1, given in Section 2.3, employs the Chernoff technique and the properties of the eigenvalues of the difference of projection matrices, to bound the probability of declaring a wrong sparsity pattern $\mathcal{F}$ instead of the true one $\mathcal{T}$ as function of the measurement matrix X and the true parameter β. The error rate decreases exponentially in the norm of the projection of $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ on the orthogonal subspace spanned by the columns of $X_{\mathcal{F}}$. This is in agreement with

the intuition that the closer different subspaces corresponding to different sets of columns of X are the harder it is to differentiate them, and hence the higher the error probability will be.

The theorem below gives a non-asymptotic bound on the probability of the event that the declared sparsity pattern $\mathcal{S}(\hat{\beta})$ differs from the true sparsity pattern $\mathcal{T}$ in no more than d indices, when the entries of the measurement matrix X are drawn i.i.d. from a standard normal distribution. It is clear that by letting $d = 1$ we obtain an upper bound on the error probability of exact sparsity pattern recovery.

Theorem 2. *Suppose that for the observational model of equation (2.4) and the estimate $\hat{\beta}$ in equation (2.2) the entries of X are i.i.d. $\mathcal{N}(0,1)$ and $p > 2k$. If*

$$n - k > \max\Big\{B\ f_0(d,p,k,\beta_{\min}), B\ f_0(k,p,k,\beta_{\min}), f_1(k,\beta_{\min}), f_2(p,k,\beta_{\min})\Big\}$$

where

$$\begin{aligned} f_0(d,p,k,\beta_{\min}) &:= \frac{d\log(\frac{k(p-k)}{d^2}) + d}{\log(1 + Cd\beta_{\min}^2)} \\ f_1(k,\beta_{\min}) &:= 4k\left(1 + \frac{1}{Ck\beta_{\min}^2}\right)^2 \\ f_2(p,k,\beta_{\min}) &:= \left(1 + \frac{1}{C\beta_{\min}^2}\right)\left[1 + 2\log(k(p-k))\right] \end{aligned}$$

then

$$\Pr\Big[\big|\mathcal{S}(\hat{\beta}) - \mathcal{T}\big| \geq d\Big] < k\max\left\{\left[\frac{ek(p-k)}{d^2}\right]^{-B^\star d}, \left[\frac{e(p-k)}{k}\right]^{-B^\star k}\right\}$$

where $B^\star = \frac{B-5}{2}$ and $C = 3 - 2\sqrt{2}$.

The key elements in the proof include Theorem 1, application of union bounds (a fairly standard technique which has been used before for this problem (Wainwright, 2007; Akcakaya and Tarokh, 2008; Karbasi et al., 2009)), asymptotic behavior of binomial coefficients and properties of convex functions.

Note that in the linear regime, i.e. $k = \Theta(p)$, with $n = \Theta(p)$ and $k\beta_{\min}^2 = \Theta(1)$ the probability of misidentifying more than any fraction (less than one) goes to zero exponentially fast as $p \to \infty$. In words, if the SNR is fixed while the dimension of the signal increases unboundedly, it is still possible to recover reliably some fraction of the support. This is in agreement with previous results on partial sparsity pattern recovery (Akcakaya and Tarokh, 2008; Reeves and Gastpar, 2008).

If we let $n(p)$, $k(p)$ and $\beta_{\min}(p)$ scale as a function of p, then the upper bound of $\Pr[\mathcal{S}(\hat{\beta}) \neq \mathcal{T}]$ scales like $k(p-k)^{-B^\star}$. For $B^\star > 2$ or equivalently $B > 9$, the probability of error as $p \to \infty$ is bounded above by p^{-D} for some $D > 1$. Therefore

$$\sum_{p=1}^{\infty} \Pr[\mathcal{S}(\hat{\beta}_{p\times 1}) \neq \mathcal{T}_p] \tag{2.5}$$

is finite and as a consequence of the Borel-Cantelli Lemma, for large enough p, the decoder declares the true sparsity pattern almost surely. In other words, the estimate $\hat{\beta}$ based on (2.2) achieves the same loss as an oracle which is supplied with perfect information about which coefficients of β are nonzero. The following corollary summarizes the aforementioned statements.

Corollary 3. *For the observational model of equation (2.4) and the estimate $\hat{\beta}$ in equation (2.2), let n, k and $\beta_{\min}^2$ scale as a function of p. Then there exists a constant $C^\star$ such that if $\beta_{\min}^2$ is $\Omega(\frac{1}{k})$ and $O(1)$, and*

$$n > C^\star \max\left\{\frac{\log(p-k)}{\log\left(1+\beta_{\min}^2\right)}, \frac{k\log(\frac{p}{k})}{\log(1+k\beta_{\min}^2)}, k\right\}$$

then a.s. for large enough p, $\hat{\beta}$ achieves the same performance loss as an oracle which is supplied with perfect information about which coefficients of β are nonzero and $\mathcal{S}(\hat{\beta}) = \mathcal{T}$.

Remarks:

- $\beta_{\min}^2 = O(1)$ is required to ensure that for a sufficiently large $C^\star$ we have $C^\star f_0(1, p, k, \beta_{\min}) > f_2(p, k, \beta_{\min})$ where f_0 and f_2 are defined in Theorem 1.
- $\beta_{\min}^2 = \Omega(\frac{1}{k})$ is required to ensure that for a sufficiently large $C^\star$ we have $C^\star k > f_1(k, \beta_{\min})$ where f_1 is defined in Theorem 1.

The sufficient conditions in Corollary 3 can be compared against similar conditions for exact sparsity pattern recovery in (Wainwright, 2007; Fletcher et al., 2008; Akcakaya and Tarokh, 2008; Karbasi et al., 2009); for example, in the sub-linear regime $k = o(p)$, when $\beta_{\min}^2 = \Theta(1)$, (Wainwright, 2007; Karbasi et al., 2009) proved that $n = \Theta(k \log(\frac{p}{k}))$ is sufficient, and (Akcakaya and Tarokh, 2008; Fletcher et al., 2008) proved that $n = \Theta(k \log(p - k))$ is sufficient. In that vein, according to Corollary 3

$$n = \max\left\{\Theta\left(\frac{k \log(\frac{p}{k})}{\log k}\right), \Theta(k)\right\}$$

suffices to ensure exact sparsity pattern recovery; therefore, it strengthens these earlier results.

What remains is to see whether the sufficient conditions in Corollary 3 match the necessary conditions proved in (Wang et al., 2008) :

Theorem 4. *(Wang et al., 2008): Suppose that the entries of the measurement matrix $X \in \mathbb{R}^{n \times p}$ are drawn i.i.d. from any distribution with zero-mean and variance one. Then a necessary condition for asymptotically reliable recovery is that:*

$$n > \max\{f_1(k, p, \beta_{\min}^2), f_2(k, p, \beta_{\min}^2), k - 1\},$$

where

$$\begin{aligned} f_1(k, p, \beta_{\min}^2) &= \frac{\log\binom{p}{k} - 1}{\frac{1}{2}\log(1 + k\beta_{\min}^2(1 - \frac{k}{p}))} \\ f_2(k, p, \beta_{\min}^2) &= \frac{\log(p - k + 1) - 1}{\frac{1}{2}\log(1 + \beta_{\min}^2(1 - \frac{1}{p-k+1}))}. \end{aligned}$$

Scaling	Sufficient condition Corollary 3	Necessary condition Theorem 4 (Wang et al., 2008)
$k = \Theta(p)$ $\beta^2_{\min} = \Theta(\frac{1}{k})$	$n = \Theta(p \log p)$	$n = \Theta(p \log p)$
$k = \Theta(p)$ $\beta^2_{\min} = \Theta(\frac{\log k}{k})$	$n = \Theta(p)$	$n = \Theta(p)$
$k = \Theta(p)$ $\beta^2_{\min} = \Theta(1)$	$n = \Theta(p)$	$n = \Theta(p)$
$k = o(p)$ $\beta^2_{\min} = \Theta(\frac{1}{k})$	$n = \Theta(k \log(p-k))$	$n = \Theta(k \log(p-k))$
$k = o(p)$ $\beta^2_{\min} = \Theta(\frac{\log k}{k})$	$n = \max\left\{\Theta(\frac{k \log(p-k)}{\log k}), \Theta\left(\frac{k \log(\frac{p}{k})}{\log \log k}\right)\right\}$	$n = \max\left\{\Theta(\frac{k \log(p-k)}{\log k}), \Theta\left(\frac{k \log(\frac{p}{k})}{\log \log k}\right)\right\}$
$k = o(p)$ $\beta^2_{\min} = \Theta(1)$	$n = \max\left\{\Theta\left(\frac{k \log(\frac{p}{k})}{\log k}\right), \Theta(k)\right\}$	$n = \max\left\{\Theta\left(\frac{k \log(\frac{p}{k})}{\log k}\right), \Theta(k)\right\}$

Table 2.1: Necessary and sufficient conditions on the number of measurements n required for reliable support recovery in the linear and the sublinear regime. The sufficient conditions presented in the first four rows are a consequence of past work (Wainwright, 2007), also recovered by Corollary 3. The new stronger result in this paper provides the sufficient conditions in row 5 and 6, which did not appear in previous studies (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Karbasi et al., 2009), and match the necessary conditions presented in (Wang et al., 2008).

The necessary condition in Theorem 4 asymptotically resembles the sufficient condition in Corollary 3; recall that $\log \binom{p}{k} < k \log(\frac{ep}{k})$. The sufficient conditions of Corollary 3 can be compared against the necessary conditions in (Wang et al., 2008) for exact sparsity pattern recovery, as shown in Table 2.1. The first paper to establish the sufficient conditions in row 1 and row 4 of Table 2.1 is (Wainwright, 2009). The sufficient conditions presented in the first four rows of Table 2.1 are a consequence of past work (Wainwright, 2007), also recovered by Corollary 3. The new stronger result in this paper provides the sufficient conditions in row 5 and 6, which did not appear in previous studies (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Karbasi et al., 2009), and match the previous necessary conditions presented in (Wang et al., 2008). (It is worth reminding that these results are restricted to $\beta^2_{\min} = O(1)$ and $\beta^2_{\min} = \Omega(\frac{1}{k})$.)

2.3 Proof of Theorem 1

We first state three basic lemmas.

Lemma 5. *If any $2k$ columns of the $n \times p$ matrix X are linearly independent then for any sparsity pattern $\mathcal{T}$ and $\mathcal{F}$ such that $|\mathcal{T}| = |\mathcal{F}| = k$ the difference of projection matrices $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $d = |\mathcal{T} - \mathcal{F}|$ pairs of nonzero positive and negative eigenvalues, bounded above by one and bounded below by negative one, respectively, and equal in magnitude.*

Lemma 6. *For $y \sim \mathcal{N}(\mu, I)$ and $\|2t\Psi\|_{2,2} < 1$ we have:*

$$\mathrm{E}[e^{ty^T\Psi y}] = \frac{e^{t\mu^T\Psi\mu + 2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu}}{\det(I-2t\Psi)^{\frac{1}{2}}}.$$

Lemma 7. *For $\Psi = \Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ and $d = |\mathcal{T} - \mathcal{F}|$ we have:*

$$\begin{aligned} \log\det(I-2t\Psi) &\geq d\log(1-4t^2), \\ \|(I-2t\Psi)^{-1/2}\|_{2,2}^2 &\leq (1-2t)^{-1}. \end{aligned}$$

We defer the proofs of the lemmas 5 and 7 to after the proof of Theorem 1. Lemma 6 follows standard Gaussian integrals (Severini, 2005).

2.3.1 Proof of Theorem 1

For a given sparsity pattern $\mathcal{F}$, the minimum residual sum of squares is achieved by

$$\min_{\theta_{\mathcal{F}} \in \mathbb{R}^k} \|y - X_{\mathcal{F}}\theta_{\mathcal{F}}\|^2 \quad = \quad \|y - \Pi_{\mathcal{F}} y\|^2$$

where $\Pi_{\mathcal{F}}$ denotes the orthogonal projection operator into the column space of $X_{\mathcal{F}}$; that is, among all sparsity patterns with size k, the optimum decoder declares

$$\hat{\mathcal{T}}(y, X) = \arg\min_{|\mathcal{F}|=k} \|y - \Pi_{\mathcal{F}} y\|^2$$

as the optimum estimate of the true sparsity pattern in terms of minimum error probability. Recall the definition of $\hat{\beta}$ in equation (2.2) and note that $\mathcal{S}(\hat{\beta}) = \hat{\mathcal{T}}(y, X)$. If the decoder incorrectly declares $\mathcal{F}$ instead of the true sparsity pattern (namely $\mathcal{T}$), then

$$\|y - \Pi_{\mathcal{F}} y\|^2 < \|y - \Pi_{\mathcal{T}} y\|^2$$

or equivalently

$$Z_{\mathcal{F}} := y^T(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})y > 0.$$

The probability that the optimal decoder declares wrongly the sparsity pattern $\mathcal{F}$ instead of the true sparsity pattern $\mathcal{T}$ is less than the probability that $Z_{\mathcal{F}} > 0$. With the aid of the Chernoff technique an upper bound on the probability that $Z_{\mathcal{F}} > 0$ is obtained:

$$\Pr[Z_{\mathcal{F}} > 0 | X, \mathcal{T}, \beta] \quad \leq \quad \inf_{|t|<1/2} \mathrm{E}[e^{Z_{\mathcal{F}} t} | X, \mathcal{T}, \beta].$$

Note that $Z_{\mathcal{F}}$ is a random variable that has a quadratic form in Gaussian random vectors. This allows us to use standard Gaussian integrals to calculate $\mathrm{E}[e^{Z_{\mathcal{F}} t} | X, \mathcal{T}, \beta]$. In order to bound the expectation, $|t|$ is required to be bounded which is a necessary condition in Lemma 6. From Lemma 6 we learned that

$$\log \mathrm{E}[e^{Z_{\mathcal{F}} t}] = 2t^2 \mu^T \Psi (I - 2t\Psi)^{-1} \Psi \mu + t\mu^T \Psi \mu - \frac{1}{2} \log \det(I - 2t\Psi) \tag{2.6}$$

where we made the following abbreviations:

$$\begin{aligned} \mu &= X_{\mathcal{T}} \beta_{\mathcal{T}} \\ \Psi &= \Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}. \end{aligned}$$

For Lemma 6 we need $\|2t\Psi\|_{2,2} < 1$ and we prove in Lemma 5 that the eigenvalues of Ψ are bounded in absolute value by one; consequently, equation (2.6) holds for $|2t| < 1$.

With the aid of the definition of the ℓ_2 norm of matrices and applying it to $\|(I - 2t\Psi)^{-1/2}\Psi\mu\|^2$ the first term in the r.h.s. of equation (2.6) can be bounded as follows:

$$2t^2\mu^T\Psi(I-2t\Psi)^{-1}\Psi\mu \leq 2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2\mu^T\Psi^2\mu. \tag{2.7}$$

Since μ lies in the subspace spanned by the columns of $X_{\mathcal{T}}$ we have

$$\begin{aligned}\Pi_{\mathcal{T}}\mu &= \mu \quad \text{and} \\ (\Pi_{\mathcal{T}} - \Pi_{\mathcal{F}})\mu &= (\Pi_{\mathcal{T}} - \Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}} = (I - \Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\end{aligned}$$

which yields the following:

$$\begin{aligned}\mu^T\Psi\mu &= -\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}}\beta_{\mathcal{T}}\|^2 \\ &= -\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2,\end{aligned}$$

and similarly,

$$\mu^T\Psi^2\mu = \|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2.$$

The aforementioned equations and the inequality (2.7) yields the following upper bound:

$$\begin{aligned}\log \mathrm{E}[e^{Z_{\mathcal{F}}t}] &\leq 2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2\mu^T\Psi^2\mu + t\mu^T\Psi\mu - \frac{1}{2}\log\det(I-2t\Psi) \\ &= \left\{2t^2\|(I-2t\Psi)^{-1/2}\|_{2,2}^2 - t\right\}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 - \frac{1}{2}\log\det(I-2t\Psi)\end{aligned} \tag{2.8}$$

Lemma 7 introduces an upper bound for $\|(I-2t\Psi)^{-1/2}\|_{2,2}^2$ and a lower bound for $\log\det(I-2t\Psi)$ that can be used to further simplify the upper bound of $\log \mathrm{E}[e^{Z_{\mathcal{F}}t}]$. The main ingredient in the proof of Lemma 7 is the eigenvalue properties of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ that were established in Lemma 5. Substituting the bounds obtained in Lemma 7 in equation (2.8) we have:

$$\log \mathrm{E}[e^{Z_{\mathcal{F}}t}] \leq \left[\frac{2t^2}{1-2t} - t\right]\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 - \frac{d}{2}\log(1-4t^2), \tag{2.9}$$

Finally, to prove Theorem 1, we take the infimum of $\frac{2t^2}{1-2t} - t$ over $|t| < 1/2$ which is equal to $\sqrt{2} - 3/2$ at $t^\star = 1/2(1 - \sqrt{2}/2)$ and obtain the desired bound:

$$\begin{aligned}\inf_{|t|<1/2} \log \mathrm{E}[e^{Z_{\mathcal{F}} t}] &\leq -\frac{3-2\sqrt{2}}{2}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 - \frac{d}{2}\log(\sqrt{2}-1/2) \\ &\leq -\frac{3-2\sqrt{2}}{2}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2 + \frac{d}{2}.\end{aligned}$$

□

Now we prove the remaining lemmas.

Proof of Lemma 5: Before we prove the result let us introduce some notations:

- For any $\mathcal{F} \in \{1, 2, \cdots, p\}$, $V_{\mathcal{F}}$ is defined as the linear subspace spanned by the columns of $X_{\mathcal{F}}$,
- $V_{\mathcal{F}}^{\perp}$ stands for the subspace orthogonal to $V_{\mathcal{F}}$,
- $\tilde{V}_{\mathcal{F}}$ and $\tilde{V}_{\mathcal{T}}$ stand for $V_{\mathcal{F}} \cap (V_{\mathcal{F}} \cap V_{\mathcal{T}})^{\perp}$ and $V_{\mathcal{T}} \cap (V_{\mathcal{T}} \cap V_{\mathcal{F}})^{\perp}$, respectively,
- and finally for any subspace V, Π_V designates the orthogonal projection onto V. (With a slight abuse of notation, for any sparsity pattern $\mathcal{F}$, we use $\Pi_{\mathcal{F}}$ instead of $\Pi_{V_{\mathcal{F}}}$).

It is worthwhile noting that $\tilde{V}_{\mathcal{F}} \cap \tilde{V}_{\mathcal{T}}$ is empty. From Lemma 4.1 in (Bjorstad and Mandel, 1991), for any $V_{\mathcal{T}}$ and $V_{\mathcal{F}}$ in $\mathbb{R}^n$, it holds that,

$$\begin{aligned} V_{\mathcal{T}} &= \tilde{V}_{\mathcal{T}} \oplus (V_{\mathcal{T}} \cap V_{\mathcal{F}}), \\ V_{\mathcal{F}} \cup V_{\mathcal{T}} &= \tilde{V}_{\mathcal{T}} \oplus \tilde{V}_{\mathcal{F}} \oplus (V_{\mathcal{T}} \cap V_{\mathcal{F}}), \qquad (2.10) \\ V_{\mathcal{T}} \cap V_{\mathcal{F}} &\perp \tilde{V}_{\mathcal{F}} \oplus \tilde{V}_{\mathcal{T}}, \qquad (2.11)\end{aligned}$$

which yields

$$\begin{aligned} \Pi_{\mathcal{F}} &= \Pi_{\tilde{V}_{\mathcal{F}}} + \Pi_{V_{\mathcal{F}} \cap V_{\mathcal{T}}}, \\ \Pi_{\mathcal{T}} &= \Pi_{\tilde{V}_{\mathcal{T}}} + \Pi_{V_{\mathcal{F}} \cap V_{\mathcal{T}}},\end{aligned}$$

Consequently,

$$\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}} = \Pi_{\tilde{V}_{\mathcal{F}}} - \Pi_{\tilde{V}_{\mathcal{T}}}. \tag{2.12}$$

Since any set of columns of X with size less or equal to $2k$ are independent, for any $\mathcal{T}$ and $\mathcal{F}$ such that $|\mathcal{T}| = |\mathcal{F}| = k$ and $|\mathcal{F} - \mathcal{T}| = d$, we have:

$$\begin{aligned} V_{\mathcal{F}} \cap V_{\mathcal{T}} &= V_{\mathcal{F}\cap\mathcal{T}}, \\ V_{\mathcal{F}} \cup V_{\mathcal{T}} &= V_{\mathcal{F}\cup\mathcal{T}}, \end{aligned}$$

and

$$\dim(V_{\mathcal{F}\cap\mathcal{T}}) = |\mathcal{F} \cap \mathcal{T}| = k - d, \tag{2.13}$$

$$\dim(V_{\mathcal{F}\cup\mathcal{T}}) = |\mathcal{F} \cup \mathcal{T}| = k + d, \tag{2.14}$$

therefore,

$$\begin{aligned} \dim(\tilde{V}_{\mathcal{F}}) &= \dim(V_{\mathcal{F}}) - \dim(V_{\mathcal{F}} \cap V_{\mathcal{T}}) \\ &= \dim(V_{\mathcal{F}}) - \dim(V_{\mathcal{F}\cap\mathcal{T}}) = k - (k-d) = d = \dim(\tilde{V}_{\mathcal{T}}). \end{aligned}$$

The dimension of $(\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}})^{\perp}$ which is the null space of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ is equal to:

$$\dim\left((\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}})^{\perp}\right) = n - \dim(\tilde{V}_{\mathcal{F}}) - \dim(\tilde{V}_{\mathcal{T}}) = n - 2d.$$

We just proved that $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $n - 2d$ eigenvalues with eigenvalue zero. The range of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ is the $2d$ dimensional space $\tilde{V}_{\mathcal{F}} \cup \tilde{V}_{\mathcal{T}}$. Therefore, $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has $2d$ nonzero eigenvalues with absolute value less or equal to one (The eigenvalues of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ are equal to one only if $\tilde{V}_{\mathcal{F}} \perp \tilde{V}_{\mathcal{T}}$.)

If $v_{(\lambda)}$ is an eigenvector of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ with eigenvalue λ then we have

$$(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} = \lambda v_{(\lambda)}.$$

Next, we prove that the vector

$$v_{(-\lambda)} = v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)},$$

is an eigenvector of $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ with eigenvalue $-\lambda$. The proof presented in the following exploits the definition of the eigenvector $v_{(\lambda)}$:

$$\begin{aligned}
(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(-\lambda)} &= (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})\left(v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)}\right) \\
&= (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} - (\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\
&= \lambda v_{(\lambda)} - (\Pi_{\mathcal{F}} + \Pi_{\mathcal{F}}\Pi_{\mathcal{T}} - \Pi_{\mathcal{T}}\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})\, v_{(\lambda)} \\
&= -\Pi_{\mathcal{F}}\Pi_{\mathcal{T}}v_{(\lambda)} + \Pi_{\mathcal{T}}\Pi_{\mathcal{F}}v_{(\lambda)} \\
&= -\Pi_{\mathcal{F}}(\Pi_{\mathcal{F}} - \lambda)v_{(\lambda)} + \Pi_{\mathcal{T}}(\Pi_{\mathcal{T}} + \lambda)v_{(\lambda)} \\
&= -\Pi_{\mathcal{F}}(1 - \lambda)v_{(\lambda)} + \Pi_{\mathcal{T}}(1 + \lambda)v_{(\lambda)} \\
&= -(\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}})v_{(\lambda)} + \lambda(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\
&= -\lambda v_{(\lambda)} + \lambda(\Pi_{\mathcal{F}} + \Pi_{\mathcal{T}})v_{(\lambda)} \\
&= -\lambda v_{(-\lambda)}.
\end{aligned}$$

This means that for every eigenvector $v_{(\lambda)}$ with eigenvalue λ there exist another eigenvector $v_{(-\lambda)}$ with eigenvalue $-\lambda$.

Proof of Lemma 7: From Lemma 5, we know that $\Pi_{\mathcal{F}} - \Pi_{\mathcal{T}}$ has d pairs of nonzero positive and negative eigenvalues, whose magnitudes are equal. Let the positive eigenvalues be denoted by $\lambda_1, \cdots, \lambda_d$, then,

$$\begin{aligned}
\log\det(I - 2t\Psi) &= \sum_{i=1}^{d}\{\log(1 - 2t\lambda_i) + \log(1 + 2t\lambda_i)\} \\
&= \sum_{i=1}^{d}\log(1 - 4t^2\lambda_i^2)
\end{aligned}$$

Since, the eigenvalues are bounded by one, again by Lemma 5, $\log(1 - 4t^2\lambda_i^2)$ is lower bounded by $\log(1 - 4t^2)$; consequently,

$$\log\det(I - 2t\Psi) \leq d\log(1 - 4t^2).$$

To prove $\|(I - 2t\Psi)^{-1/2}\|_{2,2}^2 \leq (1 - 2t)^{-1}$, note that $(I - 2t\Psi)^{-1/2}$ has

- d eigenvalues equal to $(1-2t\lambda_1)^{-1/2}, \cdots, (1-2t\lambda_d)^{-1/2}$,
- d eigenvalues equal to $(1+2t\lambda_1)^{-1/2}, \cdots, (1+2t\lambda_d)^{-1/2}$,
- and $n-2d$ eigenvalues equal to one.

It is not hard to see that because $2t < 1$ and $\lambda_i < 1$ the top eigenvalue of $(I-2t\Psi)^{-1/2}$ is bounded above by $(I-2t)^{-1/2}$ and hence,

$$\|(I-2t\Psi)^{-1/2}\|_{2,2}^2 \;\leq\; (1-2t)^{-1}.$$

2.4 Proof of Theorem 2

We state two simple lemmas used to prove Theorem 2.

Lemma 8. *For Gaussian measurement matrices, with $X_{ij} \sim \mathcal{N}(0,1)$ the average error probability that the optimum decoder declares $\mathcal{F}$ is bounded by*

$$\Pr[\hat{\mathcal{T}}(y,X) = \mathcal{F}|\beta,\mathcal{T}] \leq \exp\left\{-\frac{n-k}{2}\log\left(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2\right) + \frac{|\mathcal{T}-\mathcal{F}|}{2}\right\}$$

where $C = 3-2\sqrt{2}$.

Lemma 9. *For the function*

$$g(r) := r\left[\frac{5}{2} + \log\left(\frac{k(p-k)}{r^2}\right)\right] - \frac{n-k}{2}\log(1+r\gamma)$$

defined on positive integers if

$$n-k > \max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2, \left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\}, \tag{2.15}$$

then

$$\max_{r=d,\cdots,k} g(r) \;\leq\; \max\{g(d), g(k)\}.$$

Before we prove the two lemmas, let us see how they imply Theorem 2.

2.4.1 Proof of Theorem 2

In order to find conditions under which $\Pr[|\mathcal{S}(\hat{\beta}) - \mathcal{T}| \geq d]$ asymptotically goes to zero, we exploit the union bound in conjunction with counting arguments and the previously stated two lemmas.

First, note that the event $|\mathcal{S}(\hat{\beta}) - \mathcal{T}| \geq d$ can be written as the union of the events $\mathcal{S}(\hat{\beta}) = \mathcal{F}$ for all sparsity patterns $\mathcal{F}$ such that $|\mathcal{F} - \mathcal{T}| \geq d$. The union bound allows us to bound the probability of the event $\cup_{|\mathcal{F}-\mathcal{T}|\geq d}\{\mathcal{S}(\hat{\beta}) = \mathcal{F}\}$ by the sum of probabilities of events like $\mathcal{S}(\hat{\beta}) = \mathcal{F}$. In mathematical terms,

$$\begin{aligned}\Pr[|\mathcal{S}(\hat{\beta}) - \mathcal{T}| \geq d] &= \Pr\left[\cup_{|\mathcal{F}-\mathcal{T}|\geq d}\{\mathcal{S}(\hat{\beta}) = \mathcal{F}\}\right] \\ &\leq \sum_{r=d}^{k} \sum_{|\mathcal{F}-\mathcal{T}|=r} \Pr\left[\mathcal{S}(\hat{\beta}) = \mathcal{F}\right].\end{aligned}$$

Lemma 8 which is based on generating functions of chi-square distributions introduces an upper bound for the event $\mathcal{S}(\hat{\beta}) = \mathcal{F}$; namely,

$$\Pr[\hat{\mathcal{T}}(y, X) = \mathcal{F}|\beta, \mathcal{T}] \leq e^{-\frac{n-k}{2}\log(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2)+\frac{|\mathcal{T}-\mathcal{F}|}{2}}$$

with $C = 3 - 2\sqrt{2}$. If we replace $C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$ with the lower bound $C|\mathcal{T} - \mathcal{F}|\beta_{\min}^2$ which follows the definition of $\beta_{\min}$ we obtain an upper bound for the event $\mathcal{S}(\hat{\beta}) = \mathcal{F}$ that does not depend on $\mathcal{F}$ as long as $|\mathcal{F} - \mathcal{T}|$ is fixed. The number of sparsity patterns $\mathcal{F}$ that are different from $\mathcal{T}$ in exactly r elements is $\binom{k}{r}\binom{p-k}{r}$. Therefore, we can bound $\sum_{|\mathcal{F}-\mathcal{T}|=r} \Pr\left[\mathcal{S}(\hat{\beta}) = \mathcal{F}\right]$ by $\binom{k}{r}\binom{p-k}{r}e^{-\frac{n-k}{2}\log(1+Cr\beta_{\min}^2)+\frac{r}{2}}$. To summarize, exploiting inequality $\log\binom{a}{b} < b\log(\frac{ae}{b})$ we have:

$$\Pr[|\mathcal{S}(\hat{\beta}) - \mathcal{T}| \geq d] \leq \sum_{r=d}^{k} e^{r\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{r^2}\right)\right]-\frac{n-k}{2}\log\left(1+Cr\beta_{\min}^2\right)}. \tag{2.16}$$

Let $g(r)$ stand for the exponent in the previous equation

$$g(r) := r\left[\frac{5}{2} + \log\left(\frac{k(p-k)}{r^2}\right)\right] - \frac{n-k}{2}\log(1 + r\gamma)$$

where we defined

$$\gamma := C\beta_{\min}^2.$$

From Lemma 9 we know that if

$$n-k > \max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2, \left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\} \tag{2.17}$$

then $\max_{r=d,\cdots,k} g(r) \leq \max\{g(d), g(k)\}$ and therefore

$$\begin{aligned}\Pr[|\mathcal{S}(\hat{\beta})-\mathcal{T}| \geq d] &\leq \sum_{r=d}^{k} e^{g(r)} \\ &\leq ke^{\max\{g(d),g(k)\}}. \end{aligned} \tag{2.18}$$

For $\Pr[|\mathcal{S}(\hat{\beta})-\mathcal{T}| \geq d] \to 0$ it suffices that $g(d)$ and $g(k)$ go to $-\infty$ fast enough. In the statement of Theorem 2 we have the following condition

$$n-k > B\frac{d\log(\frac{k(p-k)}{d^2})+d}{\log(1+d\gamma)}$$

that results in the following upper bound

$$\begin{aligned} g(d) &= d\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{d^2}\right)\right]-\frac{n-k}{2}\log(1+d\gamma) && (2.19)\\ &\leq d\left[\frac{5}{2}+\log\left(\frac{k(p-k)}{d^2}\right)\right]-\frac{B}{2}\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right] \\ &\leq -\frac{B-5}{2}\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right]. && (2.20)\end{aligned}$$

Hence, if

$$n-k > B\max\left\{\frac{d\log(\frac{k(p-k)}{d^2})+d}{\log(1+d\gamma)}, \frac{k\log(\frac{k(p-k)}{k^2})+k}{\log(1+k\gamma)}\right\} \tag{2.21}$$

then

$$\max\{g(d),g(k)\} \geq \frac{B-5}{2}\max\left\{\left[d\log\left(\frac{k(p-k)}{d^2}\right)+d\right], \left[k\log\left(\frac{k(p-k)}{k^2}\right)+k\right]\right\}.$$

Therefore, inequalities (2.17) and (2.21) which are the main conditions in Theorem 1, imply that

$$\Pr[|\mathcal{S}(\hat{\beta})-\mathcal{T}| \geq d] < k\max\left\{\left[\frac{ek(p-k)}{d^2}\right]^{-dB^\star}, \left[\frac{e(p-k)}{k}\right]^{-kB^\star}\right\}$$

where $B^\star = \frac{B-5}{2}$.

□

Now we prove the remaining lemmas.

Proof of Lemma 8: The columns of $X_{\mathcal{F}}$ and $X_{\mathcal{T}-\mathcal{F}}$ are, by definition, disjoint and therefore independent Gaussian random matrices with column spaces spanning random independent $|\mathcal{F}|$- and $|\mathcal{T}-\mathcal{F}|$-dimensional subspaces, respectively. The Gaussian random vector $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ has i.i.d. Gaussian entries with variance $\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$. Therefore, we conclude that, since the random Gaussian vector $X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}$ is projected onto the subspace orthogonal to the random column space of $X_{\mathcal{F}}$, the quantity $\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2/\|\beta_{\mathcal{T}-\mathcal{F}}\|^2$ is a chi-square random variable with $n-k$ degrees of freedom. Thus,

$$\begin{aligned}
\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|\beta,\mathcal{T}] &= \mathrm{E}_X\left\{\Pr[\hat{\mathcal{T}}(y,X)=\mathcal{F}|X,\beta,\mathcal{T}]\right\} \\
&\overset{1}{\leq} \mathrm{E}_X\left\{e^{-\frac{C}{2}\|(I-\Pi_{\mathcal{F}})X_{\mathcal{T}-\mathcal{F}}\beta_{\mathcal{T}-\mathcal{F}}\|^2+\frac{d}{2}}\right\} \\
&= \mathrm{E}_{W\sim\chi^2_{n-k}} e^{-\frac{C}{2}W\|\beta_{\mathcal{T}-\mathcal{F}}\|^2+\frac{d}{2}} \\
&\overset{2}{=} e^{-\frac{n-k}{2}\log(1+C\|\beta_{\mathcal{T}-\mathcal{F}}\|^2)+\frac{d}{2}}.
\end{aligned}$$

The first inequality follows from Theorem 1 and the second equality comes from the well-known formula (see for example (Severini, 2005)) for the moment-generating function of a chi-square random variable; that is, $\mathrm{E}_{W\sim\chi^2_{n-k}} e^{tW} = (1-2t)^{-\frac{n-k}{2}}$ for $2t<1$.

Proof of Lemma 9: Let us first explain the idea behind this Lemma. We aim to prove that under certain conditions, for some $r_0 \in [1,k]$, $g(r)$ is a decreasing function for $r \in [1,r_0]$ and an increasing function for $r \in [r_0,k]$. This yields the desired upper bound,

$$\max_{r\in[d,k]} g(r) \leq \max\{g(d), g(k)\}. \tag{2.22}$$

We begin by taking derivatives of $g(r)$ to prove the aforementioned claim:

$$\begin{aligned} g'(r) &= \frac{1}{2} + \log\left(\frac{k(p-k)}{r^2}\right) - \frac{\gamma(n-k)}{2(1+r\gamma)} \\ g''(r) &= \frac{-4(1+r\gamma)^2 + r\gamma^2(n-k)}{2r(1+r\gamma)^2}. \end{aligned}$$

Note that in the following steps we use inequality (2.15), i.e.

$$n-k > \max\left\{4k\left(1+\frac{1}{k\gamma}\right)^2, \left(1+\frac{1}{\gamma}\right)[1+2\log(k(p-k))]\right\},$$

to prove inequality (2.22):

1. $g''(r) = 0$ has two solutions $r_1^\star$ and $r_2^\star$ such that $r_1^\star < r_2^\star$. Due to the positivity of the denominator and the quadratic and concave nature of the numerator of $g''(r)$, we have:
   (a) $g''(r) < 0$ for $r < r_1^\star$,
   (b) $g''(r) > 0$ for $r_1^\star < r < r_2^\star$,
   (c) $g''(r) < 0$ for $r_2^\star < r$.
2. From inequality (2.15) we have $n-k > 4k\left(1+\frac{1}{k\gamma}\right)^2$ which ensures that $g''(k) > 0$. Therefore, we have $r_1^\star < k < r_2^\star$. This implies the convexity of $g(r)$ for $r \in [r_1^\star, k]$ and the negativity of $g''(r)$ for $r < r_1^\star$. We have two situations depending on whether $1 < r_1^\star$ or not:
   (a) $1 < r_1^\star$: From inequality (2.15) we have $n-k > \frac{1+\gamma}{\gamma}[1+2\log(k(p-k))]$ which implies that $g'(1) < 0$. This, in conjunction with $g''(r) < 0$ for $r < r_1^\star$, implies that $g(r)$ is decreasing for $r \in [1, r_1^\star]$.
   (b) $r_1^\star \leq 1$: $g(r)$ is convex for $r \in [1, k]$.
3. Either case, i.e. $g(r)$ is convex for $r \in [1, k]$ or decreasing for all $r \in [1, r_1^\star]$ and convex for $r \in [r_1^\star, k]$, proves the desired inequality (2.22).

2.5 Conclusion

In this paper, we examined the probability that the optimal decoder declares an incorrect sparsity pattern. We obtained an upper bound for any generic measurement matrix, and this allowed us to calculate the error probability in the case of random measurement matrices. In the special case when the entries of the measurement matrix are i.i.d. normal random variables, we computed an upper bound on the expected error probability. Sufficient conditions on exact sparsity pattern recovery were obtained, and they were shown to improve the previous results (Wainwright, 2007; Akcakaya and Tarokh, 2008; Fletcher et al., 2008; Karbasi et al., 2009). Moreover, these results asymptotically match (in terms of growth rate) the corresponding necessary condition presented in (Wang et al., 2008). An interesting open problem is to extend the sufficient conditions derived in this work to non-Gaussian and sparse measurement matrices.

Chapter 3

Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods

This chapter is based on the paper "Efficient estimation of two-dimensional firing rate surfaces via Gaussian process methods" (Rahnama Rad and Paninski, 2010).

3.1 Introduction

A common problem in statistical neural data analysis is to estimate the firing rate of a neuron given some two-dimensional variable. Spatial examples include the estimation of "place fields" in the hippocampus (Brown et al., 1998), "grid fields" in entorhinal cortex (Hafting et al., 2005), and position- or velocity-fields in motor cortex (Gao et al., 2002; Paninski et al., 2004a). Spatiotemporal examples include

the estimation of tuning curves that change as a function of time (Frank et al., 2002; Rokni et al., 2007); purely temporal examples include models of spike-history effects (Kass and Ventura, 2001) or the tracking of firing rates that change as a function of both intra- and inter-trial times during a behavioral task (Czanner et al., 2008). Finally, more abstract examples arise in the context of spike-triggered covariance analyses (Rust et al., 2005; Aguera y Arcas and Fairhall, 2003). More generally, the estimation of the intensity function of two-dimensional point processes is a central problem in a variety of other scientific fields, including forestry and astronomy (Moeller and Waagepetersen, 2004).

A number of methods have appeared in the literature to address this problem. It is worth briefly reviewing some of these approaches here, in order to illustrate some of the computational and statistical aspects of this two-dimensional point-process smoothing problem. Perhaps the most direct (and common) approach is to write $p(spike|\vec{x})$, the conditional probability of observing a spike in a small time bin given the two-dimensional signal $\vec{x}$, as

$$p(spike|\vec{x}) = \frac{p(spike, \vec{x})}{p(\vec{x})},$$

and then to estimate the probability densities in the numerator and denominator via standard nonparametric methods, either via histogram or kernel smoothing methods (Devroye and Lugosi, 2001); thus our estimate of the conditional firing rate is obtained as a ratio of estimated densities $\hat{p}(spike, \vec{x})/\hat{p}(\vec{x})$. The advantages of this method include its conceptual simplicity and its computational speed; in particular, linear smoothing methods for obtaining $\hat{p}(spike, \vec{x})$ and $\hat{p}(\vec{x})$ essentially involve a standard spatial convolution operation, which may be computed efficiently via the fast Fourier transform. Also, uncertainty in the estimated firing rates can be quantified via standard bootstrap methods (though this may be computationally expensive). However, the disadvantages of this approach are quite well-known (Kass et al., 2003; Kass et al., 2005): if the kernel width (or histogram bin) in the

density estimate is chosen to be too large, then the estimated firing rate surface is oversmoothed; on the other hand, if the kernel width is too small, then the division by the small, noisy estimated density $\hat{p}(\vec{x})$ can lead to large, noisy fluctuations which can mask the underlying structure in the estimated firing rate.

Another important but somewhat more subtle disadvantage of this direct ratio approach has to do with the "adaptivity" of the estimator. Speaking roughly, we would like our estimator to smooth out the data more in areas where fewer observations are available (and where the estimate is bound to be noisier), while letting the data "speak for itself" and applying minimal smoothing in regions where many $\vec{x}$ observations are available (where reliable estimates can be made without too much spatial averaging). The ratio estimator as described above does not have this important adaptive property. It is of course possible to make this smoother adaptive: one method is to let the kernel width scale roughly inversely with the number of samples $\vec{x}$ observed in a local region. However, this method is somewhat ad hoc; more importantly, since this adaptive smoothing can not be computed via a simple convolution, the fast Fourier methods no longer apply, making the method much slower and therefore obviating one of the main advantages of this ratio approach. Finally, it is well-known that the firing rate typically depends not just on a single location variable $\vec{x}$, but also on additional covariates, e.g., the time since the last spike (Berry and Meister, 1998; Kass and Ventura, 2001; Frank et al., 2002; Paninski, 2004), or the local activity of other cells in the network (Harris et al., 2003; Paninski et al., 2004b; Truccolo et al., 2005; Paninski et al., 2007; Pillow et al., 2008); it is difficult to systematically incorporate these covariate effects in the simple nonparametric ratio approach.

Parametric statistical models lie at the other end of the spectrum. We may model the firing rate $p(spike|\vec{x}) \approx p(spike|\vec{x}, \theta)$, where θ is a finite-dimensional parameter, and then fit θ directly to the observed data via standard likelihood-based

methods (Brown et al., 1998; Kass et al., 2005). Confidence intervals on the estimated firing rates may again be obtained by bootstrapping, or by the standard likelihood asymptotic methods based on the observed Fisher information (though this approach is only effective when many data observations are available and, more importantly, when the model is known to provide a good explanation of the data); in addition, it is easy to incorporate covariates (e.g., the effect of the local spike history). Parametric methods can be very powerful when a good model is available, but the results are highly dependent on the model family chosen. For example, two-dimensional unimodal Gaussian surface models for place fields can be effective for some hippocampal cells, but fail badly when modeling grid cells, which display many bumps in their firing rate surfaces. Computation in parametric models involves optimization over the parameter θ, and therefore typically scales like $O(\dim(\theta)^3)$; this adverse scaling encourages researchers to reduce the dimensionality of θ, at the expense of model flexibility[1]. Finally, local maxima in the model's objective function surface can be a significant concern in some cases.

State-space methods for estimating time-varying tuning curves represent something of a compromise between these two approaches (Brown et al., 2001; Frank et al., 2002; Czanner et al., 2008; Paninski et al., 2010). These methods are quite effective in the spatiotemporal cases cited above, but do not apply directly to the purely spatial setting. The idea is to fit a parametric model to the tuning curve, but then to track the changes in this tuning curve as a function of time using what amounts to a temporal smoothing method. These methods can be cast in a fully Bayesian setting that permits the calculation of various measures of uncertainty of the estimated firing rates and the incorporation of our prior knowledge about the smoothness of the firing rate in time and space. Computation in these methods

[1]The typical $O(\dim(\theta)^3)$ complexity is due to the matrix-solve step involved in Newton-Raphson optimization over the parameter θ. Solving a linear equation involving a $N \times N$ matrix leads to the $O(N^3)$ computational complexity.

scales roughly as $O(\dim(\theta)^3 T)$ (Paninski et al., 2010), where θ again represents the model parameter and T represents the number of time points at which we are estimating the firing rate.

In this paper we discuss a Bayesian nonparametric approach to the two-dimensional point-process smoothing problem which is applicable in both the spatial and spatiotemporal settings. Our methods are in a sense a generalization of the temporal state-space methods for point process smoothing; we will see that very similar local computational properties can be exploited in the spatial case. The resulting estimator is adaptive in the sense described above, comes equipped with confidence intervals, and can be computed efficiently. Our methods may also be considered a generalization of the techniques described by (Gao et al., 2002), and as a computationally efficient relative of the techniques considered in (Cunningham et al., 2007; Cunningham et al., 2008). We will discuss the relationship to this work below in more depth, after describing our methods in more detail.

3.2 Methods

3.2.1 The doubly stochastic point process model

We model neural activity as a point process with rate λ, with λ depending smoothly on some two-dimensional variable $\vec{x}$. For technical reasons which we will discuss further below, we will model the firing rate in terms of a smooth nonnegative function $f(.)$ applied to a two-dimensional surface $z(\vec{x})$; this surface $z(\vec{x})$, in turn, is assumed to be a smooth function which we will estimate from the observed point process data. Thus, the firing rate map $\lambda(\vec{x}) = f(z(\vec{x}))$ will itself be a smooth nonnegative function of $\vec{x}$. We will study several somewhat distinct experimental settings and show that they all can be conveniently cast in these basic terms. The

experimental settings we have in mind are:

1. We observe a spatial point process whose rate is given by $\lambda(\vec{x}) = f[z(\vec{x})]$.

2. We observe a temporal point process whose rate is given by $\lambda_t = f\left[z(\vec{x}_t)\right]$, where $\vec{x}_t$ is some known time-varying path through space (e.g., the time-varying position of a rat in a maze (Brown et al., 1998) or the hand position in a motor experiment (Paninski et al., 2004b)).

3. We make repeated observations of a temporal point process whose mean rate function may change somewhat from trial to trial[2]; in this case we may model the rate as $\lambda_t^{(i)}$, where t denotes the time within a trial and i denotes the trial number (Frank et al., 2002; Czanner et al., 2008).

4. We observe a temporal process whose rate is given by $\lambda(t) = f\left[z(x(t), t)\right]$, where $x(t)$ is some known time-varying path through a one-dimensional space (e.g., the time-varying position of a rat in a linear maze), and the one-dimensional tuning curve $f[z(x, t)]$ changes as a function of time (Frank et al., 2002; Rokni et al., 2007).

5. We observe a temporal process whose rate is given by $\lambda(t) = f\left[z(t, \tau)\right]$, where $z(t, \tau)$ depends on absolute time t and the time since the last spike τ. Models of this general form are discussed in (Kass and Ventura, 2001), who termed these "inhomogeneous Markov interval" models.

We provide detailed formulations for each of the mentioned applications in appendix A. Each of these formulations may be elaborated by the inclusion of additional terms, as we discuss in section 3.2.6 below. In all cases, the nonnegative function $f(.)$ is assumed to be convex and log-concave (Paninski, 2004). We further model

[2]Thanks to C. Shalizi for pointing out this example.

z as a sample from a Gaussian process with covariance function $C(\vec{x}, \vec{x}')$ (Cressie, 1993; Rasmussen and Williams, 2006); as we discuss below, this allows us to encode our a priori assumptions about the smoothness of z in a convenient, flexible fashion[3]. In this setting, the resulting point process is doubly-stochastic and is known as a Cox process (Snyder and Miller, 1991; Moeller and Waagepetersen, 2004); in the special case that $f(.) = \exp(.)$, the process is called a log-Gaussian Cox process (Moeller et al., 1998). Related models have seen several applications in the fields of neural information processing and neural data analysis (Smith and Brown, 2003; Jackson, 2004; Brockwell et al., 2004; Sahani, 1999; Wu et al., 2004; Wu et al., 2006; Yu et al., 2006); for example, the temporal point-process smoothing methods developed by Brown and colleagues (Frank et al., 2002; Smith and Brown, 2003; Czanner et al., 2008) may be interpreted in this framework. In particular, as mentioned above, (Gao et al., 2002) and (Cunningham et al., 2007; Cunningham et al., 2008) applied similar techniques to the problem of estimating spatial receptive fields; we will discuss the relationship to this work in more depth in the discussion section below.

3.2.2 Smoothing priors

To set the stage for our main development over the next two sections, it is helpful to review some concepts in Bayesian smoothing and estimation. There is a very large statistical literature on smoothing in one and more dimensions (Wahba, 1990; Green and Silverman, 1994). For conceptual simplicity, let's begin by reviewing the one-dimensional smoothing problem from a Bayesian point of view. In this setting we have a univariate sequential ordered series $y_1, y_2, \cdots, y_n$ observed at locations $\tau_1, \tau_2, \cdots, \tau_n$ on a one-dimensional grid. The goal is to approximate this series by

[3]We will discuss the advantages of placing the prior on z, instead of directly on the firing rate $f(z)$, in the next sections.

a smooth continuous function $z(\tau)$, i.e.

$$y_i = z(\tau_i) + \epsilon_i, \tag{3.1}$$

where ϵ_i is measurement error. Assuming that equation (3.1) is the true model and the measurement error is Gaussian with zero mean and variance σ^2, the probability of the observed data $D = \{y_i\}_{i=1,\cdots,n}$ given the smooth continuous function $z(\tau)$ is given by:

$$p(D|z) = (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} e^{-\frac{(y_i - z(\tau_i))^2}{2\sigma^2}}. \tag{3.2}$$

The maximum likelihood estimate $z_{ML}(\tau_i) = y_i$ is obtained by maximizing the logarithm of equation (3.2) over $z(\tau_i)$ for $i = 1, \cdots, n$. Unfortunately, the resulting estimate is not necessarily a smooth continuous function; in fact, in this case the MLE is not even well-defined except at the observed points $\tau = \tau_i$, since the likelihood does not depend on $z(\tau)$ for $\tau \neq \tau_i$.

The standard approach for remedying this problem is to introduce a functional of $z(.)$ to penalize non-smooth functions:

$$\hat{z}_\gamma = \arg\max_z \log P(D|z) - \gamma \mathcal{F}(z). \tag{3.3}$$

The first term accounts for the fit of the data, while the second penalizes the roughness of $z(\tau)$. The functional $\mathcal{F}(z)$ must be bigger for non-smooth functions $z(.)$ compared to smooth functions. One common choice for $\mathcal{F}(.)$ is $\int [\frac{\partial z}{\partial \tau}]^2 d\tau$, the total power of the first derivative (Wahba, 1990). Another common choice for $\mathcal{F}(.)$ is the integrated length of the square of the second derivative which gives greater smoothness than the first derivative constraint. The tuning parameter $\gamma > 0$ exists to balance fitness (in terms of log-likelihood) versus smoothness (in terms of integrated square of the first derivative). In the limit of small γ the estimate $\hat{z}_\gamma(.)$ better fits the data and is less smooth. As we increase the tuning parameter, $\hat{z}_\gamma(.)$ becomes smoother but fits the data less well. The tuning parameter may be

chosen by hand (using our a priori knowledge of the smoothness of the function z), or by any of several existing data-based methods, such as generalized cross validation, expectation maximization, generalized maximum likelihood or empirical Bayes (Wahba, 1990; Hastie et al., 2001). From a Bayesian point of view, the penalizing term can be interpreted as a prior on $z(.)$:

$$\log P(z|\gamma) \propto -\gamma \mathcal{F}(z).$$

Since $P(z|D,\gamma) \propto P(D|z)P(z|\gamma)$ as a function $z(.)$, the maximum a posteriori (MAP) estimate of $z(.)$ is given by

$$\begin{aligned} \hat{z}_\gamma &= \arg\max_z P(z|D,\gamma) \\ &= \arg\max_z \log P(D|z) + \log P(z|\gamma), \end{aligned}$$

which is equivalent to (3.3) when $\log P(z|\gamma) \propto -\gamma \mathcal{F}(z)$.

Likewise, priors based on smoothing constraints can be used to estimate two-dimensional surfaces given point process observations. The motivation is as in the one-dimensional Gaussian case: without any prior on smoothness, estimating the rate map (by maximizing the log-likelihood) results in a jagged, discontinuous rate surface with singularities on the observed spikes. One simple and convenient prior penalizes large differences in the two-dimensional surface z, exactly as we discussed above in the one-dimensional case:

$$\log P(z(.)|\gamma) \propto -\gamma \mathcal{F}(z) = -\gamma \int \left[(\frac{\partial}{\partial x} z)^2 + (\frac{\partial}{\partial y} z)^2 \right] dxdy, \tag{3.4}$$

where $\vec{x} = (x, y)$ and $z = z(\vec{x})$. If $\mathcal{F}(z)$ is close to zero, then z will be smooth (Wahba, 1990).

Three brief technical points are worth noting here. First, this prior is not a proper probability distribution because it can not be normalized to one (i.e., $\int \exp[-\gamma \mathcal{F}(z)]dz = \infty$.). However, in most cases the posterior distribution $P(z(.)|D,\gamma)$ will still be integrable even if we use such an improper prior (Gelman et al., 2003). Note that to

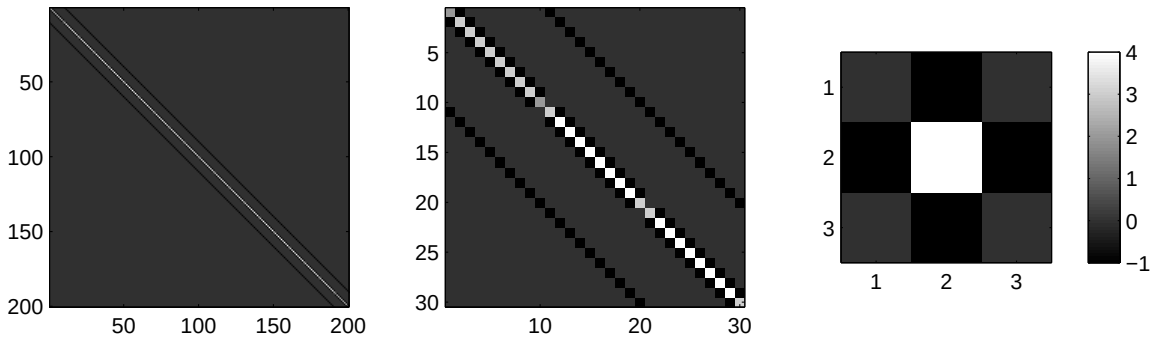

Figure 3.1: Example of the inverse prior covariance matrix C^{-1}, with $\gamma = 1$ and $\epsilon = 0$. The penalty functional $\mathcal{F}(z)$ is implemented via a quadratic form $z^T C^{-1} z$; choosing the inverse prior covariance C^{-1} to be sparse banded allows us to efficiently compute the posterior expected firing rate $\hat{\lambda}$. Left: The banded structure of C^{-1}, in an example setting where z is represented as a 10×20 grid. Middle: The first 30×30 entries of C^{-1}. Right: The five-point "stencil" implemented in C^{-1}; note that only nearest spatial neighbors are involved in the computation of the penalty.

take into account the a priori boundedness of $z(\vec{x})$ we may augment the roughness penalty in a simple way:

$$\tilde{\mathcal{F}}(z) = \int \left[(\frac{\partial}{\partial x} z)^2 + (\frac{\partial}{\partial y} z)^2 \right] dxdy + \epsilon \int z^2 dxdy; \tag{3.5}$$

this small extra term makes $P(z|\gamma) \propto \exp[-\gamma \tilde{\mathcal{F}}(z)]$ a proper prior. The scalar ϵ here sets the inverse scale of z: smaller values of ϵ correspond to larger prior variance in the Gaussian prior specified by $\log P(z|\gamma)$. Second, it is possible to tune the smoothness along the horizontal and vertical directions independently. This is useful when the two dimensions are measured in different units (e.g., time and location). This is easily done by introducing two roughness tuning parameters as follows:

$$\mathcal{F}(z) = \int \left[\gamma_x (\frac{\partial}{\partial x} z)^2 + \gamma_y (\frac{\partial}{\partial y} z)^2 \right] dxdy. \tag{3.6}$$

Third, it is possible to consider penalties based on higher derivatives, as in the one dimensional case (Wahba, 1990). For example, the penalty of equation (3.6) based

on the second derivative is as follows:

$$\mathcal{F}_2(z) = \int \left[\gamma_x (\frac{\partial^2 z}{\partial x^2})^2 + \gamma_y (\frac{\partial^2 z}{\partial y^2})^2 \right] dxdy.$$

As before, the MAP estimate of z is defined as:

$$\hat{z}_\gamma = \arg\max_z \{\log P(D|z) + \gamma \tilde{\mathcal{F}}(z)\}.$$

(We will introduce the likelihood $P(D|z)$ for the point process case in the next section.)

To implement this estimator numerically, we must discretize space and represent the function z in vector form. We may simply discretize the spatial variable $\vec{x}$ and then concatenate the resulting matrix z by appending its columns to construct a vector[4]. In the discrete domain, any of the aforementioned penalties can be written in terms of quadratic forms in the vector z, i.e. as $z^T C^{-1} z$ for an appropriate sparse positive semi-definite matrix C^{-1}. See Figure 3.1 for an example of C^{-1} such that $z^T C^{-1} z$ implements $\mathcal{F}(z)$ for a surface z which is represented in the discrete domain on a 10×20 grid. Note that the exact form of C in the case where the prior $\log P(z|\gamma) \propto -z^T C^{-1} z$ is improper may not exist. This is due to the fact that some of the eigenvalues of C^{-1} may be zero in which case C does not exist.

It is important to remember that the Gaussian prior corresponding to the exponent of the negative quadratic penalty $z^T C^{-1} z$ only acts as a regularizer, and does not imply that we are modeling the random surface $z(\vec{x})$ as a two-dimensional unimodal Gaussian surface as a function of the two-dimensional spatial variable $\vec{x}$; instead, $p(z)$ is a Gaussian function of the much higher-dimensional vector z, and therefore in general samples z from this prior may have quite arbitrary multimodal shapes as a function of $\vec{x}$, as illustrated in Fig. 3.2.

[4]With a slight abuse of notation we interchangeably use z for both the vector representation and the grid representation. The difference should be clear from the context.

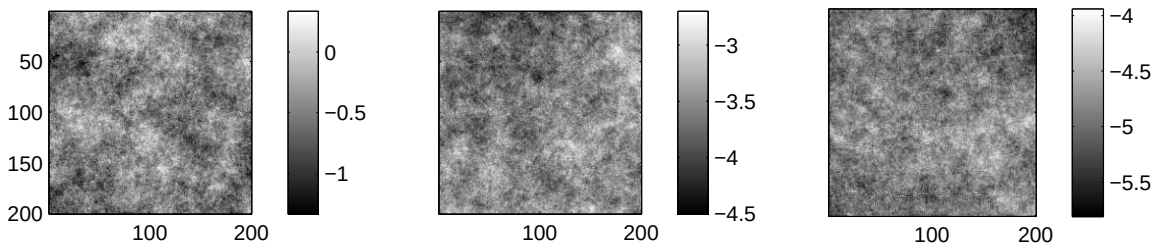

Figure 3.2: Three independent samples z drawn from the Gaussian prior with covariance matrix C and mean zero, with $\gamma = 20$ and $\epsilon = 10^{-6}$. Note that samples can take a fairly arbitrary shape, though slowly-varying (correlated) structure is visible in each case.

3.2.3 Computing the posterior

Now our main goal is to efficiently perform computations with the posterior distribution $p(z|D)$ of the random surface z given the observed spike train data D[5]. For example, given $p(z|D)$ we can estimate the firing rate by taking the conditional expectation

$$\hat{\lambda}(\vec{x}) = \mathrm{E}\big(f[z(\vec{x})]|D\big) = \int f(u)p(z(\vec{x}) = u|D)du.$$

It is well-known that for convex and log-concave $f(.)$ the log-posterior

$$\log p(z|D) = \log p(D|z) + \log p(z) + const.$$

is concave as a function of z (Paninski, 2004; Paninski, 2005; Cunningham et al., 2007), since both the prior $p(z)$ and the point-process likelihood $p(D|z)$ are log-concave in z[6], and log-concavity is preserved under multiplication. As a result

[5]Note that the posterior depends on the prior which itself is function of the hyper-parameter γ discussed in the previous section. All the posterior probabilities for the rest of the paper are for a fixed γ unless otherwise mentioned and to simplify notation we discard the dependence on γ.

[6]The point-process log-likelihood is given generically as (Snyder and Miller, 1991)

$$\log p(D|z) = \sum_i \log \lambda(t_i) - \int_0^T \lambda(t)dt,$$

where t_i are the observed spike times and $[0, T]$ is the time interval over which the spike train is observed; for details on how to compute the likelihood in each of the settings mentioned in section

$\log p(z|D)$ has no non-global local maxima in z, and therefore standard gradient ascent algorithms are guaranteed to converge to a global maximum if one exists. Furthermore, this log-concavity allows the development of efficient approximation and sampling algorithms for the posterior $p(z|D)$ using the Laplace approximation (Ahmadian et al., 2009; Kass and Raftery, 1995), as we discuss below.

As mentioned earlier, we assume a Gaussian prior on z:

$$\log p(z) = -z^T C^{-1} z + const. \tag{3.7}$$

The inverse covariance matrix C^{-1} encodes both the smoothness and boundedness of z, as discussed in the previous section. Now our basic approximation is a standard Laplace approximation (Fahrmeir and Kaufmann, 1991; Kass and Raftery, 1995; Paninski et al., 2007) for the posterior:

$$p(z|D) \approx \frac{1}{(2\pi)^{d/2}|C_D|^{1/2}} \exp\left(-\frac{1}{2}(z-\hat{z}_D)^T C_D^{-1}(z-\hat{z}_D)\right), \tag{3.8}$$

where $d = \dim(z)$,

$$\hat{z}_D = \arg\max_z p(z|D)$$

and

$$C_D^{-1} = C^{-1} + H_D, \tag{3.9}$$

with

$$H_D = -\nabla\nabla_z \log p(D|z)_{z=\hat{z}_D}.$$

In words, this is just a second-order approximation of the concave function $\log p(z|D)$ about its peak $\hat{z}_D$. We have found that this approximation is acceptably accurate when the log-prior and log-likelihood are smooth and concave, as is the case here; see e.g. (Paninski et al., 2010; Pillow et al., 2011; Ahmadian et al., 2009) for further discussion.

3.2.1, see Appendix A.

Two items are worth noting. First, $\hat{z}_D$ may be found via ascending the objective function $\log p(z|D)$ by the Newton-Raphson algorithm. Since this function is concave and is therefore unimodal, as emphasized above, we don't need to worry about local maxima. In principle finding the maximum of a concave function is straightforward (Boyd and Vandenberghe, 2004). The difficulty arises when the dimensionality of z is large which in our applications might be as large as $\sim 10^5$. Second, C_D^{-1} is quite easy to compute once we have $\hat{z}_D$, since H_D is a diagonal matrix (as can be demonstrated by explicit computation; see equation (3.11) below). The key to computing the posterior distribution in equation (3.8) is to develop efficient methods for computing $\hat{z}_D$. The standard Newton-Raphson ascent method requires that we solve the linear equation

$$\left(C^{-1} - \nabla\nabla_z \log p(D|z)_{z=\hat{z}^{(i)}}\right) w = \nabla_z \log p(z|D)_{z=\hat{z}^{(i)}} \tag{3.10}$$

for the search direction w, where $\hat{z}^{(i)}$ denotes our estimate of z after i iterations of Newton-Raphson. For example, in the simplest setting (case 1 described in section 3.2.1), we have the standard point-process log-likelihood (Snyder and Miller, 1991)

$$\log p(D|z) = \sum_j \log f[z(\vec{x}_j)] - \int f[z(\vec{x})]d\vec{x} + const.$$

where j indexes the location $\vec{x}_j$ where the j-th spike was observed, and so

$$\frac{\partial \log p(D|z)}{\partial z(\vec{x})} = -f'[z(\vec{x})]d\vec{x} + \sum_j \frac{f'}{f}[z(\vec{x})]\delta(\vec{x} - \vec{x}_j)$$

and

$$\frac{\partial^2 \log p(D|z)}{\partial z(\vec{x})\partial z(\vec{x}')} = \begin{cases} -f''[z(\vec{x})]d\vec{x} + \sum_j \frac{f''f-(f')^2}{f^2}[z(\vec{x})]\delta(\vec{x} - \vec{x}_j) & \text{if } \vec{x} = \vec{x}' \\ 0 & \text{otherwise,} \end{cases} \tag{3.11}$$

where $f'(.)$ and $f''(.)$ denote the first and second (scalar) derivatives of the function $f(.)$. Note that if $f(.) = \exp(.)$, the second term of the first line in equation (3.11) is zero.

So the feasibility of this smoothing method rests primarily on the tractability of the Newton step (3.10), which in turn rests on our ability to solve equations of the form

$$\left(C^{-1} + H\right) w = b$$

as a function of the unknown vector w, for diagonal matrices H. For general $d \times d$ matrices C^{-1}, this will require $O(d^3)$ time[7], which is intractable for reasonably-sized z. (Cunningham et al., 2007; Cunningham et al., 2008) introduced techniques for speeding up the computations in this general case to find an approximate MAP estimate of the rate map which behaved reasonably well in numerical examples; we take a different approach here and restrict our attention to a special subclass of covariance functions C which is flexible enough for our needs but at the same time allows us to perform the necessary computations much more efficiently than in the general $O(d^3)$ case.

Before we discuss these computational issues, though, it is worth mentioning a few important statistical properties of the estimator $\hat{\lambda}$ for the firing rate. First, the Bayesian approach allows to systematically calculate various measures of the uncertainty of the estimator $\hat{\lambda}$ (as we will discuss at more length below), and it is straightforward to incorporate our prior knowledge about the smoothness of z in the definition of the covariance function C. In addition, the Bayesian estimator, by construction, functions as an adaptive smoother: because the Bayesian estimator represents a balance between the data and our prior beliefs about z, the estimator will smooth less in regions where the data are highly informative, and vice versa. Quantitatively, this balance of data versus prior is determined by the size of the "observed Fisher information matrix" H_D compared to the inverse prior covariance C^{-1}, and therefore depends both on the observed data and the nonlinear func-

[7]Note that computing H_D requires just $O(T)$ time, where T is the length of the experiment, and therefore this step is not rate-limiting.

tion $f(.)$; for $f(.) = \exp(.)$, for example, H_D increases with the firing rate, since $\lambda(u) = f''(u)$ is monotonically increasing in u, and therefore the effective smoothing width decreases in regions of high firing rate, as desired. More concretely, when the observed information matrix H_D is large compared to the prior covariance C, the posterior uncertainty (measured by the posterior covariance C_D, equation (3.9)) is approximately H_D^{-1}, whereas the posterior uncertainty reverts to the prior uncertainty (i.e., C) when the observed data D are less informative. This adaptive behavior can also be understood by examining the matrix equation (3.10) in the Fourier domain: since the inverse covariance C^{-1} is typically chosen to penalize high-frequency fluctuations (Theunissen et al., 2001), larger values of the diagonal term H_D correspond to a local spatial filter $(C^{-1} + H_D)^{-1}$ which passes higher spatial frequencies, and which is therefore more spatially localized (Paninski, 2005).

3.2.4 Priors based on nearest-neighbor penalties lead to fast computation

In this section we will describe how to choose the inverse prior covariance C^{-1} so that we can solve the Newton step in a computationally efficient manner while retaining the statistical efficiency and biological plausibility[8] of our estimator. The basic insight here is that if $[C^{-1}]_{\vec{x},\vec{y}} = 0$ whenever $\vec{x}$ and $\vec{y}$ are not neighbors on the discrete two-dimensional grid[9], then C^{-1} may be written in block-tridiagonal form with tridiagonal blocks, and our equation resembles a discrete Poisson equation, for which highly efficient multigrid solvers are available which require just $O(d)$ time (Press et al., 1992). Even standard methods for solving the equation (as imple-

[8]Biological plausibility refers here to the exclusion of sharp discontinuities and singularities in the rate map.

[9]In examples considered in this paper we will focus mainly on the nearest-neighbor case, but the methods may be applied more generally when $[C^{-1}]_{\vec{x},\vec{y}} = 0$ if $\vec{x}$ and $\vec{y}$ are separated by a distance of more than n pixels, where n is small ($n = 1$ in the nearest-neighbor case).

mented, e.g., in Matlab's $A\backslash b$ call) are quite efficient here, requiring just $O(d^{3/2})$ time[10]. We have found that a very simple Newton-Raphson algorithm exploiting these efficient linear algebra techniques (and a simple backtracking method to ensure that the objective increases with each Newton step) converges in just a few iterations, therefore providing a rapid and stable algorithm for computing $\hat{z}_D$; the optimization takes just a few seconds on a laptop computer for $d \sim 10^4$. In contrast, if we represent z in some finite-dimensional basis set B in which fast matrix solving methods are not available, then each Newton step generically requires $O(\dim(B)^3)$ time. Thus we see that the fast sparse matrix techniques allow us to investigate spatial receptive fields of much higher resolution, since d may be made much larger than $\dim(B)$ at the same computational cost.

Once $\hat{z}_D$ is obtained using these efficient methods, we estimate the firing rate map by:

$$\mathrm{E}[f(z(\vec{x}))|D] = \int f(u)p(z(\vec{x}) = u|D)du \approx \int G_{\hat{z}_D(\vec{x}),\mathrm{Var}[z(\vec{x})|D]}(u)f(u)du,$$

where we have applied the Laplace approximation in the second step and abbreviated the Gaussian density in u with mean μ and variance σ^2 as $G_{\mu,\sigma^2}(u)$. In the case that $f(.) = \exp(.)$, $f[z(\vec{x})]$ has a lognormal distribution and we may read off the conditional mean $\mathrm{E}[f(z(\vec{x}))|D]$ and variance $\mathrm{Var}[f(z(\vec{x}))|D]$ using standard results about this distribution; for example,

$$\mathrm{E}[\exp(z(\vec{x}))|D] = \exp\left(\hat{z}_D(\vec{x}) + \frac{1}{2}\mathrm{Var}[z(\vec{x})|D]\right). \tag{3.12}$$

In any case, we need to compute the conditional variances $\mathrm{Var}[z(\vec{x})|D]$, which under the Laplace approximation are given by the diagonal elements of C_D. These terms

[10]This $O(d^{3/2})$ scaling requires that a good ordering is found to minimize fill-in during the forward sweep of the Gaussian elimination algorithm; code to find such a good ordering (via "approximate minimum degree" algorithms (Davis, 2006)) is built into the Matlab call $A\backslash b$ when A is represented as a sparse banded matrix. See also (Sanches et al., 2008) for an approach based on a related Sylvester equation; this approach is quite different but turns out to have the same computational complexity.

may be computed without ever computing the full matrix C_D (which would require $O(d^2)$ storage) by making use of the banded structure of C_D^{-1}, via standard algorithms such as the forward-backward Kalman smoother or the method described in (Asif and Moura, 2005). This step requires $O(d^2)$ time. However, in many cases (as we will see below), $\mathrm{Var}[z(\vec{x})|D]$ is somewhat smoother than $\mathrm{E}[z(\vec{x})|D]$, and may therefore be computed on a coarser scale (making d smaller) and then interpolated to a finer scale; thus this $O(d^2)$ time scaling is not a major limitation.

Another important application of the Laplace approximation is to compute the marginal likelihood $p(D|\theta) = \int p(z, D|\theta)dz$, where θ denotes parameters we might want to optimize over in the context of model selection (e.g., the hyperparameters setting the spatial scale and variance of the prior covariance C; see Results section below), or in constructing hierarchical models of the observed rate maps over multiple neurons (Behseta et al., 2005; Geffen et al., 2009). The Laplace approximation for this marginal likelihood is

$$\log p(D|\theta) = \log \int p(D, z|\theta)dz \approx \log p(\hat{z}_D|\theta) + \log p(D|\hat{z}_D, \theta) - \frac{1}{2}\log |C_D^{-1}| + const., \tag{3.13}$$

where "const." is constant in θ and the dependence of $\hat{z}_D$ (and therefore H_D) on θ has been left implicit to avoid cluttering the notation. The first two terms here are easy to compute once $\hat{z}_D$ has been obtained: the first (Gaussian) term is a sparse banded quadratic form, and the second is the usual point-process loglikelihood. The third term requires the determinant of a sparse banded matrix, which again may be computed efficiently and stably via a Cholesky decomposition; the "chol" function in Matlab again automatically takes advantage of the sparse banded nature of C_D^{-1} here, and requires just $O(d^{3/2})$ time.

It is also worth noting that the generalization to non-Gaussian priors of the form

$$p(z) \propto \exp\left(\sum_{ij} h_{ij}[z(\vec{x}_i) - z(\vec{x}_j)]\right),$$

for some collection of smooth, symmetric, concave functions $h_{ij}(.)$, is straightforward, since the log-posterior remains smooth and concave. For example, by using sub-quadratic penalty functions $h_{ij}(.)$ we can capture sharper edge effects than in the Gaussian prior (Gao et al., 2002); conversely, if we use penalty functions of the form

$$h_{ij}(u) = \begin{cases} 0 & |u| < K \\ -\infty & otherwise \end{cases},$$

then we may directly impose Lipschitz constraints on z (Coleman and Sarma, 2007) (the resulting concave objective function is non-smooth, but may be optimized stably via interior-point methods (Boyd and Vandenberghe, 2004; Cunningham et al., 2007; Cunningham et al., 2008; Vogelstein et al., 2008; Koyama and Paninski, 2009; Paninski et al., 2010)). Once again, to maintain the sparse banded structure of the Hessian $C^{-1} + H$ we choose $h_{ij}(.)$ to be uniformly zero for non-neighboring $(\vec{x}_i, \vec{x}_j)$. We recover the Gaussian case if we choose $h_{ij}(.)$ to be quadratic with negative curvature; in this case, the nonzero elements of the inverse prior covariance matrix $(C^{-1})_{ij}$ correspond to the pairs (i, j) for which the functions $h_{ij}(.)$ are not uniformly zero.

3.2.5 MCMC methods

While we have found the Laplace approximation to be quite effective in the applications we have studied (Ahmadian et al., 2009), in many cases it may be useful to draw samples from the posterior distribution $p(z|D)$ directly, either for Monte Carlo computation of the firing rate estimator $\hat{\lambda}$ or for visualization and model checking purposes. The prewhitened Metropolis-adjusted Langevin or hybrid Monte Carlo algorithms (Robert and Casella, 2005; Ahmadian et al., 2009) are standard MCMC algorithms that are well-suited for sampling from the near-Gaussian pos-

terior $p(z|D)$. To implement this algorithm here, we only need an efficient method for solving the prewhitening equation

$$Rw = \eta$$

for w, where η is a standard normal sample vector and R is the Cholesky decomposition of C_D^{-1}. As noted above, R may be computed in $O(d^{3/2})$ time here, and each call to the solver for $Rw = \eta$ again requires just $O(d^{3/2})$ time, although due to the high dimensionality of z here the chain may require many steps to mix properly (Robert and Casella, 2005; Ahmadian et al., 2009). We have not explored this approach extensively.

3.2.6 Using the Schur complement to handle non-banded cases

There are two important settings where our sparse banded matrix methods need to be modified slightly. First, in some cases we would like to impose periodic boundary conditions on our estimate $\hat{\lambda}$. For example, (Rokni et al., 2007) analyzed the dynamics of tuning curves as a function of arm direction; since direction is a periodic variable, our estimate $\hat{\lambda}$ should also be periodic in the direction, for all values of time (i.e., we have to impose "cylindrical" — periodic in direction but not in time — boundary conditions on $\hat{\lambda}$). We can ensure periodicity in our estimate $\hat{\lambda}$ simply by choosing our inverse prior covariance matrix C^{-1} to have the desired cylindrical boundary conditions. However, this means that C_D^{-1} no longer has a banded form; the upper-right and lower-left corner blocks of this matrix are nonzero.

Our second example involves the inclusion of covariate information. Each of the experimental settings introduced in section 3.2.1 may be elaborated by including

additional covariate information (e.g., spike history effects (Paninski, 2004; Truccolo et al., 2005)). For example, instead of modeling the rate of our observed temporal point process as $\lambda(t) = f[z(\vec{x})]$, we could use the model $\lambda(t) = f[z(\vec{x}) + W_t\theta]$ instead, where W denotes a matrix of known (fixed) covariates and θ is a set of weights we would like to fit simultaneously with z. We can proceed by directly optimizing $\log p(z, \theta|D)$ (assuming θ has a log-concave prior which is independent of z); this joint optimization in (θ, z) is tractable, again, due to the special structure of the Hessian matrix of the objective function $\log p(z, \theta|D)$ here. If we order the parameter vector as $\{z, \theta\}$, the Hessian may be written in block form $H = \begin{pmatrix} H_{zz} & H_{\theta z}^T \\ H_{\theta z} & H_{\theta\theta} \end{pmatrix}$, where H_{zz} has the special sparse banded form discussed above. Thus, in both of these examples we have to solve a linear equation involving a block matrix $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{pmatrix}$, where the size of the block H_{11} is much larger than that of the block H_{22}, and the large block H_{11} is sparse banded. These systems may be solved easily in $O(d^{3/2})$ time via Schur complement methods; for details see appendix B and for an illustration see Figure 3.9 below.

3.3 Results

In this section we will describe several applications of the methods described above, to both simulated and real spike train data. In all examples we find the posterior expectation of the firing rate map by equation (3.12). We assume that $f(.) = \exp(.)$ and use $\log p(z) \propto -\gamma\mathcal{F}(z)$, as defined in equation (3.4). Recall that any convex and log-concave $f(.)$, e.g. $\log(1+e^x)$, could be used instead of the exponential non-linearity. All hyper-parameters γ are estimated by the empirical Bayes method described in appendix D.

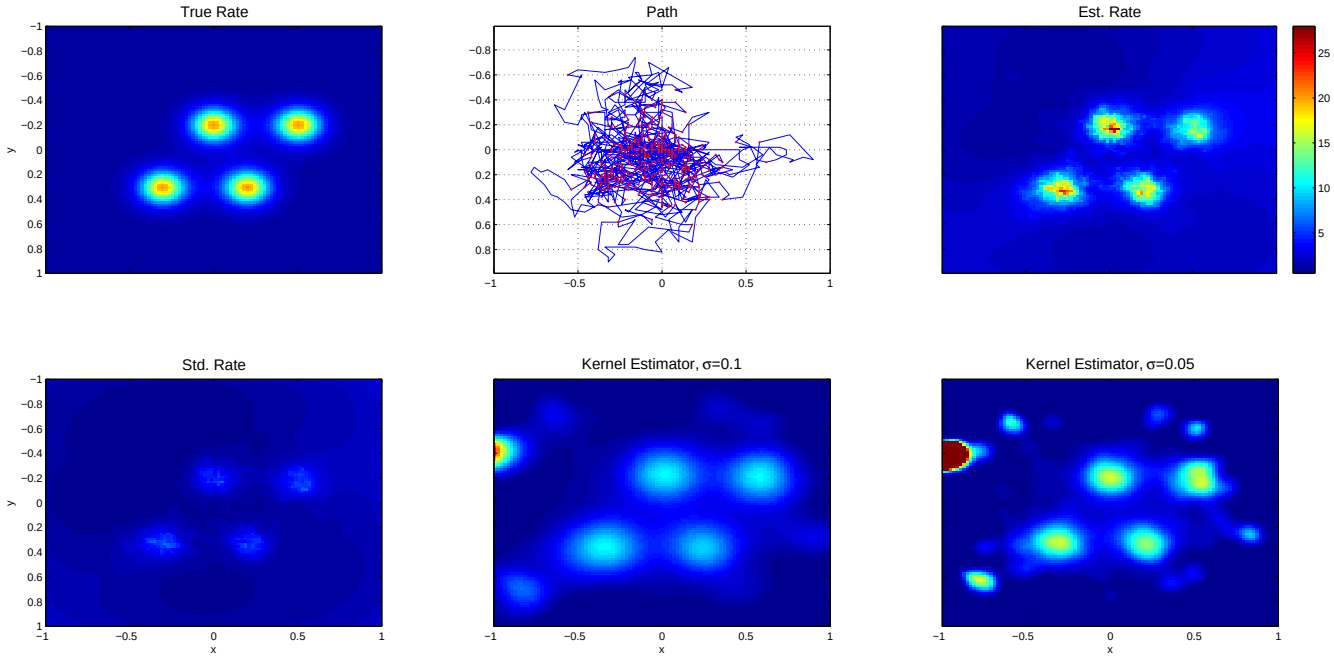

Figure 3.3: Estimating the two-dimensional firing rate given a spike train observed simultaneously with a time-varying path $\vec{x}(t)$. The simulation was done with 10ms time bins and a total of ~ 2800 spikes over 600 seconds. Top left: The true firing rate surface $\lambda(\vec{x})$. Top middle: The trace of the path $\vec{x}(t)$ through the two dimensional space for the first 10 seconds with red dots indicating the spikes. Top right: The posterior expectation of the firing rate surface, i.e. $\mathrm{E}\left[f(z)|D\right]$. Bottom left: The posterior standard deviation of the firing rate surface. Bottom middle and right: The kernel estimator with isotropic bandwidth $\sigma = 0.1$ and $\sigma = 0.05$, respectively. For small bandwidth the kernel estimator is very noisy, especially at the corners, where no samples are available.

3.3.1 Synthetic data

3.3.1.1 A two-dimensional spatial place field

We begin with the second example introduced in section 3.2.1: we observe a temporal point process whose rate is given by $\lambda(t) = \exp[z(\vec{x}(t))]$, where $\vec{x}(t)$ is a (known) time-varying path through space. $\vec{x}(t)$ was sampled from two dimensional random

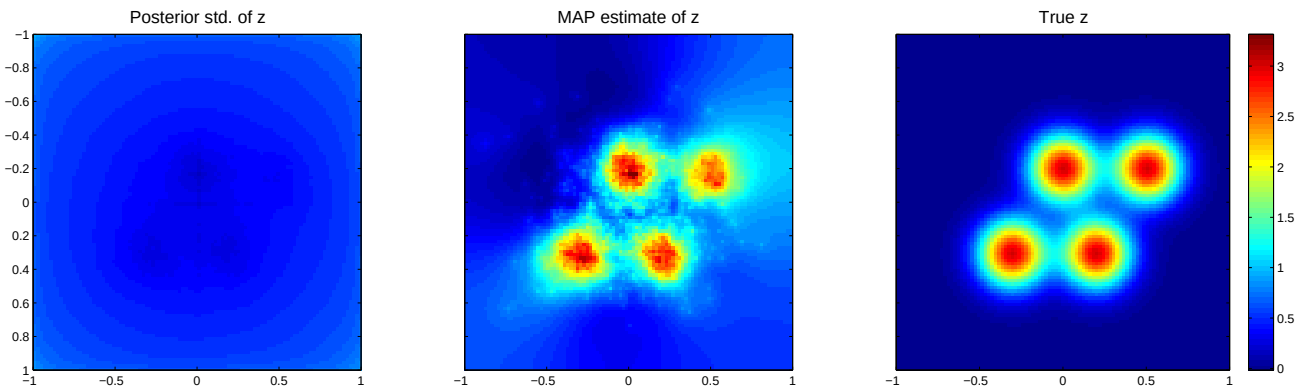

Figure 3.4: The spatial fields $z(x, y)$ corresponding to the estimated firing rates $\lambda(x, y)$ shown in Fig. 3.3; units are dimensionless and correspond to variations in the log-firing rate. Note that in this figure the posterior standard deviation and the MAP estimate of the latent surface z is presented as opposed to figure 3.3 which presents the posterior expectations of firing rate surface, i.e. $\mathrm{E}\left[f(z)|D\right]$. Left: The posterior standard deviation of $z(x, y)$, i.e. $\mathrm{std}\left[z|D\right]$, which is smaller around the center because more samples are available from that region (c.f. the top middle panel of Figure (3.3)). Middle: The MAP estimate of $z(x, y)$. Right: The true $z(x, y)$.

walk. The unknown underlying random two-dimensional surface $z(\vec{x})$ is assumed to be constant in time. The experiment we are trying to simulate is the estimation of the firing rate surface of a single grid cell from the recorded spike train of the corresponding cell. Figure 3.4 illustrates an estimated place field: we see that the method provides a very accurate estimate of the true $z(x, y)$ for centrally-located points $\vec{x}$, where the space has been sampled densely (see the top middle panel of Figure 3.3). For more peripheral points $\vec{x}$, on the other hand, less data are available. Here the estimated firing rate relies more heavily on the prior, and reverts to a flat surface (since a gradient-penalizing prior, as discussed in section 3.2.2, was used here). As emphasized above, these estimates are computationally efficient, requiring just a few seconds on a laptop computer to recover surfaces $\hat{z}$ described by $\sim 10^4$ parameters. The middle bottom and right bottom panel of Figure 3.3

shows the results of the popular Gaussian kernel estimator (see appendix C for details). As we increase the bandwidth of the Gaussian kernel the estimate becomes smoother. The kernel estimator performs poorly in this example mainly because it is not able to adapt its bandwidth according to the local informativeness of the observations and only a sparse sample is available. The performance of the kernel estimator is much better in the case that observations of all points in the domain of interest are available, as we will see in section 3.3.1.3.

3.3.1.2 A temporally-varying one-dimensional spatial field

Next we examine setting 4 from section 3.2.1. Here we have a one dimensional tuning curve which changes with time. In this example z is placed on a 100×200 grid. The experiment we are trying to simulate is the estimation of temporally varying tuning curve from a single cell recording. This can correspond to the estimation of the one dimensional spatial receptive field of cell which changes over time while the rat is running back and forth on a one dimensional track. We observe a sample path, x_t (which was sampled from a one dimensional random walk), along with a point process of rate $\lambda(t) = f[z(x_t, t)]$; the resulting estimate of the underlying spatiotemporal firing rate is shown in Figure 3.5. Similar results as in Figure 3.3 are obtained: as long as the path samples the space enough, we obtain a reasonable estimate of the changing tuning curve but where insufficient data are available the estimator reverts to the prior. As before, we assume that $f(.) = \exp(.)$ and use $\log p(z) \propto -\gamma_x \int \left(\frac{\partial z}{\partial x}\right)^2 dxdt - \gamma_t \int \left(\frac{\partial z}{\partial t}\right)^2 dxdt$. The hyper-parameters γ_x and γ_t were estimated using the empirical Bayes method explained in appendix D. Figure 3.6 shows the result of the linear Gaussian kernel smoothing with different combinations of temporal and spatial bandwidths. The kernel estimator is more problematic in this application; the output of this estimator depends heavily on the sample path x_t. Of course the Bayesian estimate also depends on the path, but

this estimator is better able to balance the information gained from the data with our prior information about the smoothness of the rate map.

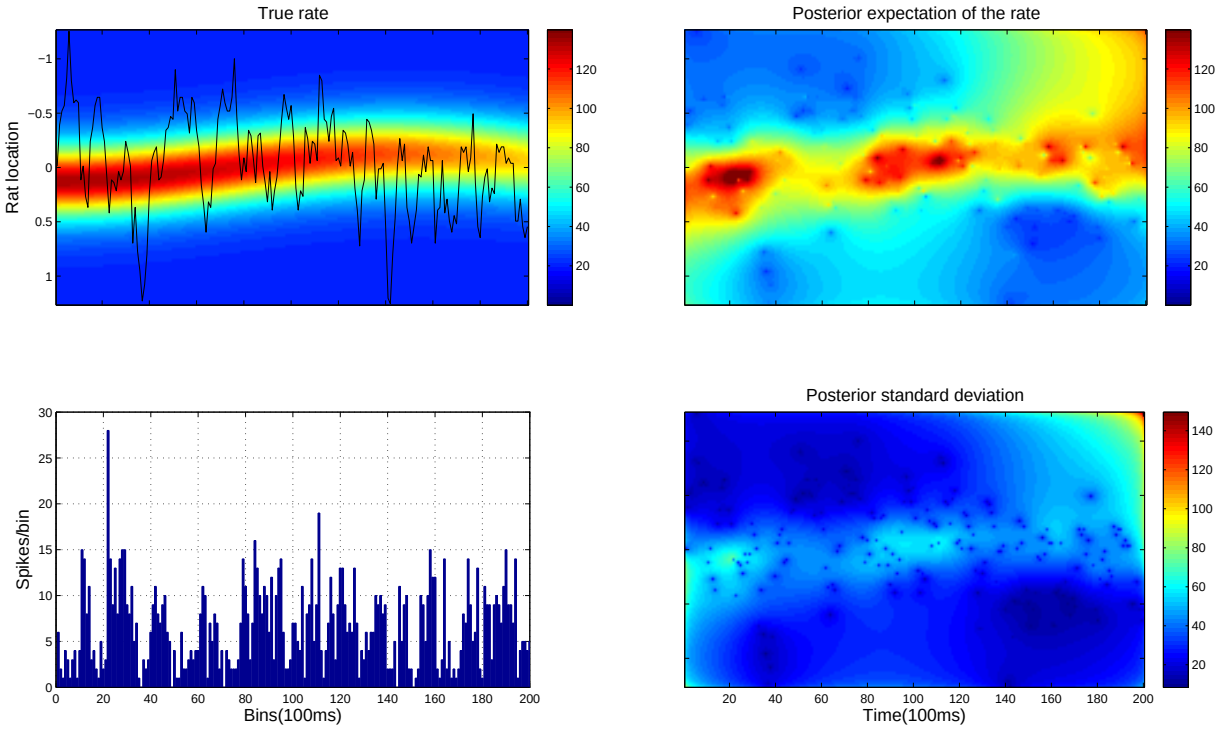

Figure 3.5: Estimating a one-dimensional time varying spatial tuning curve. Top left: The actual color map of the rate surface as a function of location and time (color) and the observed one-dimensional path of the animal as a function of time (black trace). Top right: The posterior expectation of the rate for a 20s period with a total of ~ 1300 spikes. The rate map as a function of location and time is observed very sparsely and for areas like the top right or the bottom middle of the rate map no observations are available as is clear from the path of the animal. The posterior expectation of the rate map at unobserved parts is effectively smoothed based on observations from other parts. Note in particular that in the upper right, where no data are available, the estimate reverts to the prior, which forces the inferred rate to be a flat extrapolation of the observed data from the right middle of the rate map. Bottom left: Observed spike count. Bottom right: The posterior standard deviation of the firing rate surface. The standard deviation increases with the firing rate (c.f. Eq. 3.12) and is higher at the lower half and top right part where limited data are available; c.f. the black path shown in the top left panel.

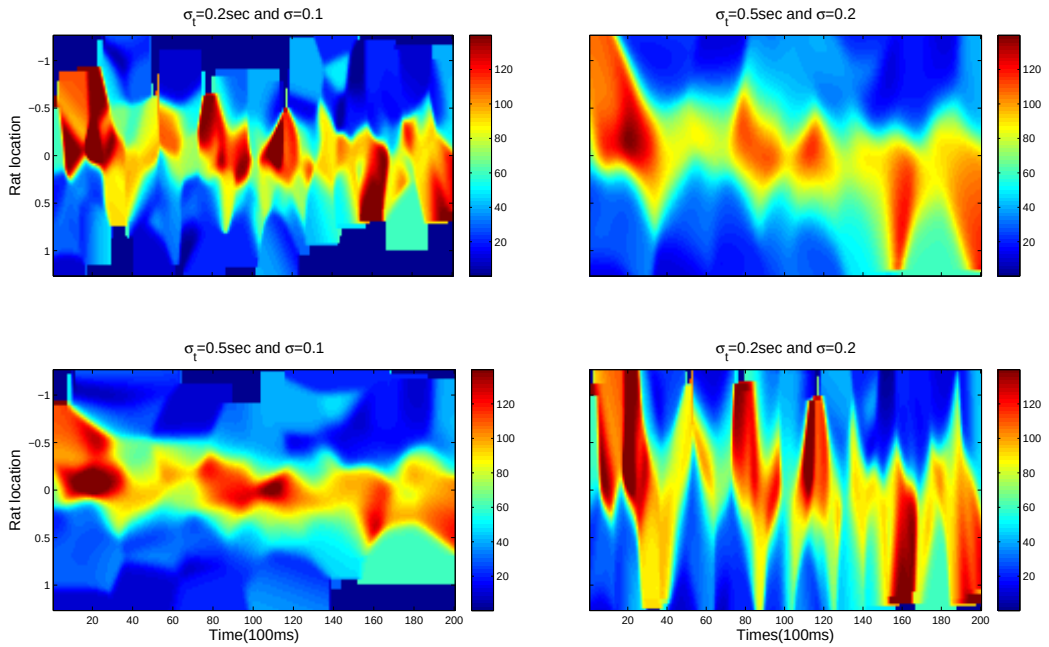

Figure 3.6: Kernel estimator of the one-dimensional time varying spatial tuning curve of Figure 3.5. Each panel corresponds to a different combination of spatial and temporal bandwidth. For small bandwidths (e.g., top left panel), the estimate is quite noisy. The dark blue blocks seen in the top and bottom left figures are due to the fact that we don't have enough samples from those regions and that the bandwidth is small (i.e., the estimate is undefined at these locations). As we increase the bandwidth this problem seems to disappear but still the map is heavily influenced by the path of the animal; c.f. the black path shown in in the top left panel of Figure 3.5.

3.3.1.3 Trial-by-trial firing rate modulations

Finally, we analyze the simulated example data involving the between-trial and within-trial neural spiking dynamics (as in setting 3 in section 3.2.1) from (Czanner et al., 2008); this data set was simulated to emulate recorded data from a monkey performing a location-scene association task. See the top left panel of Figure 3.7 for the simulated spike train over 50 trials: this model neuron displays strong non-stationarity both within and between trials.

For N trials each having duration $T\delta$, we use the following model:

$$P(D|z) = \prod_{i=1}^{N}\prod_{t=1}^{T} \frac{e^{-f(z_{i,t}+H_{i,t})\delta}(f(z_{i,t}+H_{i,t})\delta)^{n_{i,t}}}{n_{i,t}!},$$

where $n_{i,t}$ stands for the number of spikes within the the interval $(t\delta, (t+1)\delta]$ in the ith trial, δ for bin size, and T the number of bins in one trial. The history effect is defined by:

$$H_{i,t} = \sum_{t'=1}^{\tau} h_{t'} n_{i,t-t'},$$

where h_t stands for the spike history term and τ for its duration; $\tau = 30$ ms here, following (Czanner et al., 2008). Note that $z(i,t)$ lies on a $N \times T$ grid which is a 10^4 dimensional space because of $N = 50$ and $T = 200$. The estimate of (z, h) is found by the joint optimization

$$(\hat{z}, \hat{h}) = \arg\max_{z,h} \left\{ \log P(D|z,h) + \sum_{t}\sum_{n=1}^{N-1} \left[\gamma_n \big[z(i+1,t) - z(i,t)\big]^2 + \frac{\gamma_t}{\delta}\big[z(i,t+\delta) - z(i,t)\big]^2 \right] \right\}, \tag{3.14}$$

where the hyper-parameters γ_n and γ_t determine how strongly the estimate is smoothed across trials and within trials, respectively. The hyper-parameters are estimated using the empirical Bayes method described in appendix D. The joint optimization is performed using the methods discussed in section 3.2.6; once $\hat{z}$ is found it is straightforward to calculate the posterior expectation and standard deviation of the rate map as described in section 3.2.4.

See Figure 3.7 for data, results and comparisons. The latent surface z and the history term h_t were estimated simultaneously. However, for the estimated firing rate maps shown in the middle panel of figure 3.7 the effect of the spiking history, which varies much more sharply as a function of time t than does the latent surface z, is excluded from this plot, to emphasize what was specifically referred to as the "stimulus component" (in our terminology, the z-dependent component) of the firing rate surface in (Czanner et al., 2008). The posterior expected rate map (middle

left panel of Figure 3.7) and the simple kernel estimator (bottom right panel of Figure 3.7) provide qualitatively similar results, since in this case we have an observation for every point on the rate grid (unlike the case analyzed in Figs. 3.5-3.6). The smooth posterior expectation here illustrates the power of sharing statistical information both within and between trials. The state-space method implemented in (Czanner et al., 2008) (middle right panel of Figure 3.7), on the other hand, only smooths across trials, not across neighboring time bins, and therefore leads to a much noisier estimate[11].

The marginal loglikelihood $\log P(D|\gamma_n, \gamma_t)$ of the hyper-parameters (γ_n, γ_t) is shown in the top right panel of Figure 3.7. In appendix D we describe how to efficiently estimate this marginal likelihood. By plotting the marginal log likelihood of the hyper-parameters over a grid it is possible to choose the best values of these smoothing hyper-parameters, as is shown in the top right panel of Figure 3.7. The influence of the hyper-parameters on $\hat{z}$ is clear from Figure 3.8; as we increase γ_n (top rows to bottom rows), smoothing across trials becomes stronger. Likewise the right columns are smoother over time compared to the left ones.

[11]More concretely, the state-space method discussed in (Czanner et al., 2008) may be understood as a version of the Gaussian process method discussed here (Paninski et al., 2010): the state-space term encodes a Gaussian prior on a latent variable which modulates the firing rate in an exponential manner, exactly as in our model if we take $f(.) = \exp(.)$. The major difference is that (Czanner et al., 2008) choose their state-space model parameters such that the the corresponding prior inverse covariance matrix C^{-1} lacks the second term in the r.h.s. of equation (3.14), which enforces smoothness across neighboring time bins. (Specifically, in their implementation the state-space covariance matrix $\Sigma = cov(\epsilon_k)$ is diagonal, so the firing rate is estimated independently in each time bin; see (Czanner et al., 2008) for notation and details.) There are also more minor differences in the computation of the posterior expectation of the firing rate; see (Paninski et al., 2010) for further discussion.

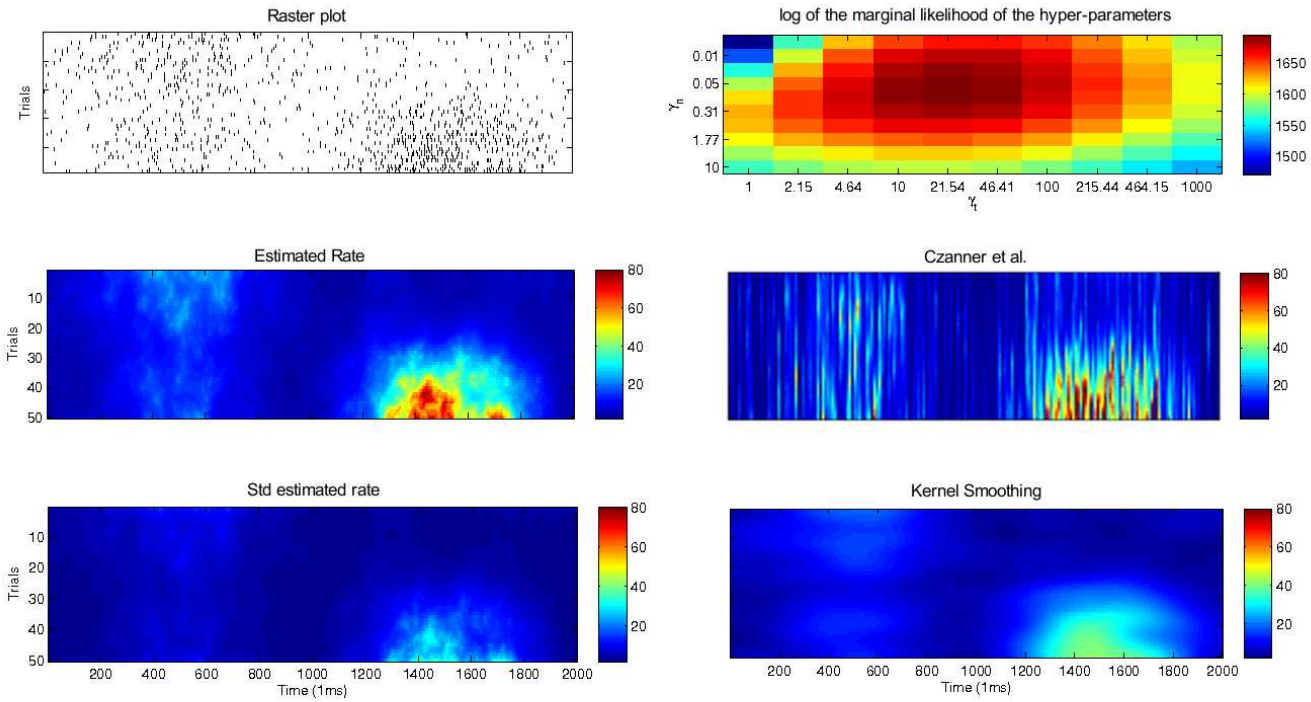

Figure 3.7: Estimating the firing rate in the context of significant trial-to-trial nonstationarity. Top left: The observed spike trains for different trials; see (Czanner et al., 2008) for simulation details. Top right: The log of the marginal likelihood of the hyper-parameters γ_t and γ_n; the empirical Bayes method discussed in appendix D chooses the "best" smoothing parameters by maximizing this function. Middle left: Posterior expectation of the firing rate, i.e. $\mathrm{E}[f(z)|D]$ which in (Czanner et al., 2008) was specifically mentioned as the stimulus component of the firing rate, computed using hyper-parameters (γ_t, γ_n) chosen via empirical Bayes (i.e., maximizing the surface shown in the top right panel). This estimated model (including the estimated history effects $H_{i,t}$, not shown here) passed the Kolmogorov-Smirnov goodness of fit test described in (Brown et al., 2002) at the 99% level. Middle right: Smoothed estimate using the method discussed in (Czanner et al., 2008). This estimate of the firing rate surface was referred to as the "stimulus component" in (Czanner et al., 2008). Again, for clarity, the latent variable z and the history term were estimated jointly but we only show the "stimulus component" (excluding the discontinuous spike-history effect) here. Bottom left: The posterior standard deviation of the estimated firing rate surface, using the same hyperparameters as in the middle left panel. Bottom right: output of kernel smoother. The time bandwidth and trial bandwidth are 100ms and 3 trials, respectively. Note that the kernel and Bayesian methods seem to perform well here; the state-space method of (Czanner et al., 2008) seems to undersmooth the data in the t direction.

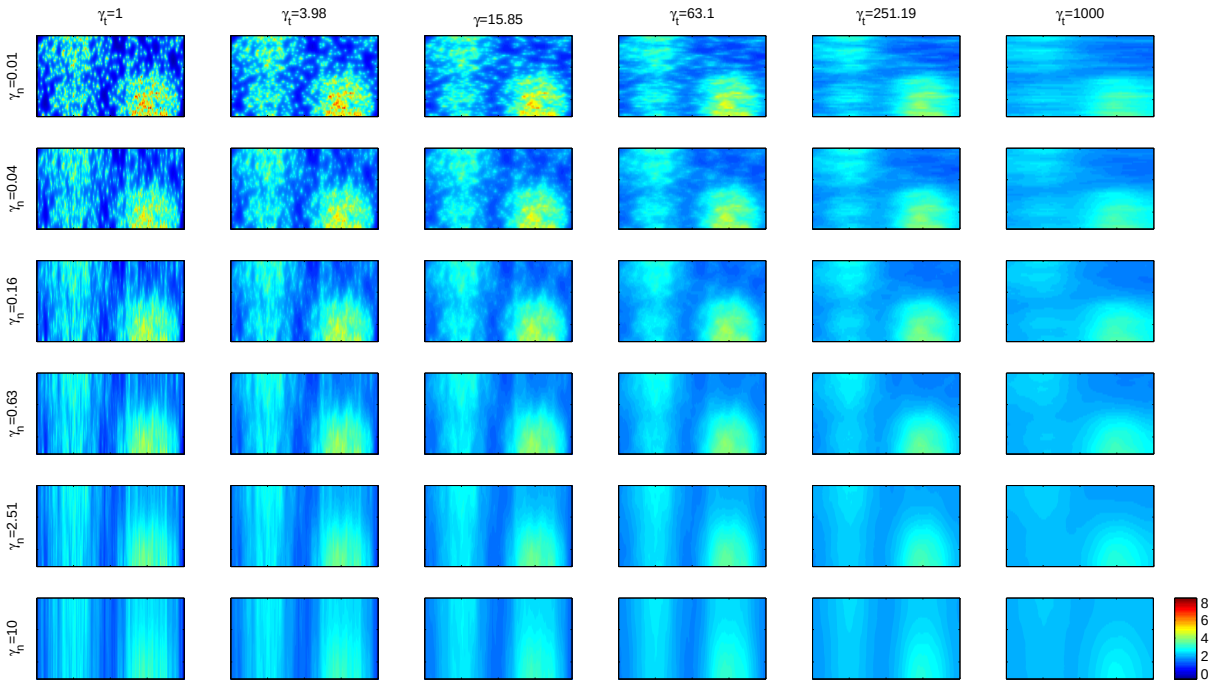

Figure 3.8: Estimating the latent surface $z(t, n)$ in the context of significant trial-to-trial non-stationarity, for different settings of the hyper-parameters (γ_t, γ_n), as in eq. (3.14). The top rows correspond to smaller γ_n and bottom ones to bigger γ_n; the left columns correspond to smaller γ_t and right columns to bigger γ_t. To be concrete, γ_t increases logarithmically from 1 to 1000 from the left columns to the right columns. Similarly, γ_n increases logarithmically from 0.01 to 10 from the top rows to the bottom rows. As we increase γ_n the smoothing across trials becomes stronger. Similarly, by increasing γ_t the temporal smoothing becomes stronger. Figure 3.7, upper right, displays the corresponding marginal log-likelihood for each of these hyperparameter settings.

3.3.2 Real data

3.3.2.1 Two dimensional spatial firing map

Now we apply our methods to data previously analyzed in (Paninski et al., 2004a; Paninski et al., 2004b). In these experiments, a monkey was trained to manually track a moving target on a two-dimensional plane, guided by visual feedback on a

computer monitor. The hand position $\vec{x}(t)$ was recorded simultaneously with the spike trains of several neurons in the primary motor cortex (MI).

As is well-known, MI neurons are tuned to a variety of kinematic variables, including the hand position, velocity, and acceleration. To explore the nonlinear properties of MI tuning to these variables, we fit a model of the form $\lambda(t) = f[W_t\theta + z(\vec{x}(t))]$, where the six-column covariate matrix

$$W_t = \begin{bmatrix} \vec{x}(t) & \dfrac{\partial \vec{x}(t)}{\partial t} & \dfrac{\partial^2 \vec{x}(t)}{\partial t^2} \end{bmatrix}$$

contains the observed time-varying horizontal and vertical position, velocity, and acceleration, and θ denotes a six-dimensional set of linear weights acting on W_t. By including the linear terms $W_t\theta$ in the model and by using the gradient penalty in the prior for z, we ensure that the estimated $z(\vec{x})$ contains only nonlinear effects as a function of $\vec{x}$, since any linear trend will be accounted for by the $W_t\theta$ term. More precisely, any linear trend in z has non-zero gradient and will therefore be penalized, whereas this linear dependence is parametrically included in $W_t\theta$, which is unpenalized here. Therefore, the estimate of z will not show any linear dependence, allowing us to isolate any non-linear dependence in z. Note that z lies on a 100×100 grid. The results are shown in Figure 3.9; we see that, as reported in (Paninski et al., 2004b), the tuning here is largely linear in W_t and the non-linear dependence which is captured by z is less significant compared to the linear dependence. (Similar results were observed in other cells and when z was allowed to depend on velocity instead of position; data not shown.)

3.4 Discussion

We have introduced Gaussian process methods for estimating the conditional intensity function of two-dimensional point processes, and demonstrated the application

of these methods in a variety of neural coding settings. Our basic approach was to approximate the posterior distribution of the rate map using the Laplace approximation constructed by finding the MAP estimate and the Hessian at that point. The prior was chosen to enforce local smoothness while retaining the computational efficiency of the Newton-Raphson ascent method used to find the MAP estimate.

Our work is closest to that of (Gao et al., 2002), (Czanner et al., 2008), and (Cunningham et al., 2007; Cunningham et al., 2008). We presented an explicit comparison of our method with that of (Czanner et al., 2008) in section 3.3.1.3 above. (Gao et al., 2002) discussed the estimation of two-dimensional firing rates in the context of motor cortical data recorded in the same experiments as the data shown in Figure 3.9; this previous work emphasized the importance of nearest-neighbor smoothing penalties to obtain valid estimates of the firing rate, and also discussed the relative benefits of quadratic vs. sub-quadratic penalty functions for recovering sharper features in the estimated rate surfaces. We have extended this work here by casting these methods in a Gaussian process setting, which allows us to provide estimates of the posterior uncertainty and of the marginal likelihood of the observed data. This framework allowed us to approach a number of additional applications, going beyond the estimation of a single spatial rate map. Our work focused especially on the computational efficiency of these techniques: we emphasized the log-concavity of the posterior and the use of efficient linear algebra methods for optimization. We also developed methods to include additional covariate information in the estimates, and discussed the use of non-quadratic penalizers (as introduced by (Gao et al., 2002)) within the same computationally-efficient paradigm.

The work of (Cunningham et al., 2007; Cunningham et al., 2008) is even closer in spirit to ours[12]; the Bayesian viewpoint is emphasized throughout in that paper.

[12]We should also mention (Cressie and Johannesson, 2008; Macke et al., 2010) here, who discuss yet another major alternative method for speeding computation in spatial Gaussian process models, in this case via imposing a low-rank structure on the prior covariance which may then be

The major difference is that (Cunningham et al., 2007; Cunningham et al., 2008) tackled the case of fairly general covariance functions, whereas we have limited our attention to covariance functions whose inverses contain only local potentials; this restriction allows us to exploit efficient computational linear algebra methods and makes our estimator significantly faster. (The beneficial computational properties of the banded matrices that result from these local potentials are of course well-known and exploited extensively in the spline literature (Wahba, 1990).) One additional technical difference is that (Cunningham et al., 2007; Cunningham et al., 2008) imposed nonnegativity constraints directly on the Gaussian process, instead of mapping the Gaussian process through a rectifying function $f(.)$ as we have done here. This direct positivity-conditioning approach makes inference of the conditional mean and variance of the firing rate somewhat more difficult, since the marginal distribution of the multidimensional truncated Gaussian distribution is difficult to approximate (whereas in the case treated in this paper we can compute the mean and conditional variance of the estimated firing rate $\hat{\lambda}$ analytically, under the Laplace approximation). (Cunningham et al., 2007; Cunningham et al., 2008) used the MAP estimate to approximate the conditional expectation of the firing rate; this approximation is valid in the "high-information" limit, where the data likelihood dominates the variability of the prior. In cases where less data are available, MCMC techniques such as the hit-and-run algorithm (Lovasz and Vempala, 2003) can be employed to sample efficiently from the log-concave posterior distribution, though the "corners" due to the positivity prior enforced in (Cunningham et al., 2007; Cunningham et al., 2008) cause the MCMC chain to mix more slowly than in the case of the smooth posterior in the current work (Ahmadian et al., 2009).

Recently, a fast nonparametric rate estimation method (Brown et al., 2009a; Brown

exploited computationally via the Woodbury matrix lemma. See (Cressie, 1993; Rasmussen and Williams, 2006) for further background and discussion.

et al., 2009b) based on variance stabilization transforms was introduced with desirable theoretical optimality properties. The variance stabilization transform turns the relatively complicated problem into a standard homoscedastic Gaussian regression problem and then any good nonparametric Gaussian regression procedure (e.g., wavelet smoothing) can be applied. One interesting direction for future work would be to combine the favorable properties enjoyed by this completely nonparametric method with those enjoyed by our Bayesian method; for example, it is not clear how to incorporate inhomogeneous observations (as described, for example, in Figs. 3.5-3.6 here) or additional covariate effects into the variance-stabilization method.

We should also note that a number of fully-Bayesian methods have been developed to perform point-process smoothing in the one-dimensional case; the Bayesian adaptive regression splines (BARS) method described in (DiMatteo et al., 2001; Kass et al., 2003) is perhaps the most popular in the neuroscience community. These methods are based on MCMC integration over a suitable posterior and often provide state-of-the-art estimation accuracy, but at significantly greater computational cost than the optimization approach pursued here. Extensions of the BARS method to the two-dimensional case are feasible but have not yet been pursued, to our knowledge; we would expect that the fast two-dimensional Laplace approximation methods we have developed here would be useful in this extended BARS setting.

Finally, we should note that there are a number of well-known connections between the point-process and density estimation problems. Gaussian process methods for density estimation have been explored intensively in the statistics and physics literature (Good and Gaskins, 1971; Thorburn, 1986; Bialek et al., 1996; Holy, 1997; Schmidt, 2000; Paninski, 2005). One interesting avenue for future work would be to explore the application of the computational methods developed here to problems in two-dimensional density and conditional density estimation.

3.5 Appendix I: Detailed formulations of the different experimental setups

Here we discuss the experimental settings introduced in section 3.2.1 in somewhat more detail.

1. We observe a spatial point process on a grid whose rate is given by $\lambda(\vec{x}) = f[z(\vec{x})]$. The likelihood of the observed spike train is given by:

$$P(D|z) = \prod_i \frac{e^{-\lambda(\vec{x}_i)\delta x}(\lambda(\vec{x}_i)\delta x)^{n_i}}{n_i!},$$

   where the product is over all points of the grid and δx is the spatial binwidth and n_i is the number of spikes observed in the i-th bin.

2. We observe a temporal point process whose rate is given by $\lambda_t = f\left[z(\vec{x}_t)\right]$, where $\vec{x}_t$ is some known time-varying path through space (e.g., the time-varying position of a rat in a maze (Brown et al., 1998) or the hand position in a motor experiment (Paninski et al., 2004b)). Here the likelihood is given by

$$P(D|z) = \prod_{t=0}^{T} \frac{e^{-\lambda_t \delta t}(\lambda_t \delta t)^{n_t}}{n_t!}, \tag{3.15}$$

   where the path of the animal during $[0, T\delta t]$ does not necessarily cover all points of the grid. This setting is different from the first one in two ways. First, the time-varying path through space might not cover the whole space, and therefore we typically will not have observations for every point in space. Second, given the observed temporal spiking activity we are able to include the spiking history in the model (Paninski, 2004; Truccolo et al., 2005). In this case we have:

$$\lambda_t = f[z(\vec{x}_t) + H_t], \tag{3.16}$$

where $H_t = \sum_i h_{t-t_i}$; here t_i is time of the ith spike and h_t designates the spike-history waveform. Note that we can generalize this model by adding other time-varying covariates, as discussed in section 3.2.6.

3. We make repeated observations of a temporal point process whose mean rate function may change from trial to trial; in this case we may model the rate as $\lambda_t^{(i)}$, where t denotes the time within a trial and i denotes the trial number. For N trials each having duration $\delta t T_i$, we have:

$$P(D|z) = \prod_{i=1}^{N} \prod_{t=0}^{T_i} \frac{e^{-\lambda_t^{(i)} \delta t} (\lambda_t^{(i)} \delta t)^{n_t^{(i)}}}{n_t^{(i)}!},$$

where $n_t^{(i)}$ stands for the number of spikes within the $(t\delta_t, (t+1)\delta_t]$ timebin of the ith trial.

4. We observe a temporal process whose rate is given by $\lambda(t) = f\left[z(x(t), t)\right]$, where $x(t)$ is some known time-varying path through a one-dimensional space. $P(D|z)$ is given by equation (3.15). Here the two dimensions correspond to time and the one dimensional position, i.e. $\lambda(x, t) = f(x, t)$. However, since the path x is changing over time we represented the firing rate as $\lambda(t) = f(x_t, t)$.

5. We observe a temporal process whose rate is given by $\lambda(t) = f\left[z(t, \tau)\right]$, where $z(t, \tau)$ depends on absolute time t and the time since the last spike τ (Kass and Ventura, 2001). Imagine we observe the spike train $\{t_i\}_{i=1,\cdots,l}$ over a period of $[0\ \delta t T]$ seconds. The likelihood is given by

$$P(D|z) \propto e^{-\int_0^{\delta t T} f[z(t, t-\tau(t))]dt} \left(\prod_{i=2}^{l} f\left[z(t_i, t_i - t_{i-1})\right] \right) f\left[z(t_1, \infty)\right],$$

where $\tau(t)$ is the time since last spike from time t and l the total number of spikes over a period of $[0\ \delta t T]$. In the discrete domain, for small enough δt

such that the number of spikes n_t in $(t\delta t, (t+1)\delta t]$ is either zero or one, we have:

$$\log P(D|z) \approx \sum_{l=0}^{T} n_t \log f\left[z(l\delta t, \tau(l\delta t))\right] + (1-n_t)\log\left(1 - f\left[z(l\delta t, \tau(l\delta t))\delta t\right]\right) + const.$$

3.6 Appendix II: Schur complement to handle non-banded matrices

As discussed in section 3.2.6, in some cases we have to solve the linear equation

$$Hx = b,$$

involving a block matrix $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{pmatrix}$, where the size of the block H_{11} is much larger than that of the block H_{22}, and the large block H_{11} is sparse banded. We have

$$H^{-1} = \begin{pmatrix} H_{11}^{-1} + H_{11}^{-1}H_{12}(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} H_{12}^T H_{11}^{-1} & -H_{11}^{-1}H_{12}(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} \\ -(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} H_{12}^T H_{11}^{-1} & (H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} \end{pmatrix}.$$

Write x and b as $(x_1^T \;\; x_2^T)^T$ and $(b_1^T \;\; b_2^T)^T$, respectively. We have

$$\begin{aligned} x_1 &= H_{11}^{-1}b_1 + H_{11}^{-1}H_{12}(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} H_{12}^T H_{11}^{-1} b_1 - H_{11}^{-1}H_{12}(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} b_2, \\ x_2 &= -(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} H_{12}^T H_{11}^{-1} b_1 + (H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} b_2. \end{aligned}$$

Assume H_{11} is $d \times d$ and H_{12} is $d \times k$, where $d >> k$ or more specifically $k = O(1)$. Finding $H_{11}^{-1}y$ for any y takes $O(d^{3/2})$, therefore by writing

$$(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} = H_{22}^{-1} - H_{22}^{-1} H_{12}^T (H_{11} + H_{12} H_{22}^{-1} H_{12}^T)^{-1} H_{12} H_{22}^{-1},$$

one can find $(H_{22} - H_{12}^T H_{11}^{-1} H_{12})^{-1} y$ for any y in $O(d^{3/2})$ and therefore x_1 and x_2 can be found in $O(d^{3/2})$. Note that we want to avoid calculating H_{11}^{-1} because of the storage cost.

3.7 Appendix III: Kernel Estimator

For every point $\vec{x}$, let the empirical frequency of visits to that point and the relative frequency of observed spikes be designated by $m_{\vec{x}}$ and $r_{\vec{x}}$, respectively; thus the total number of observed spikes at $\vec{x}$ is $m_{\vec{x}} r_{\vec{x}}$. Further, let M stand for $\sum_{\vec{r}'} m_{\vec{r}'}$ where the summation is over all points in the domain. Write $p(n_{\vec{x}} = 1|\vec{x})$, for $n_{\vec{x}} \in \{0, 1\}$ as the indicator of a spike, as $p(n_{\vec{x}} = 1, \vec{x})/p(\vec{x})$. One method to estimate $p(\vec{x})$ and $p(n_{\vec{x}} = 1|\vec{x})$ is to build a histogram of the relative frequency of appearance of the empirical data. Instead the kernel estimator uses a smoothed estimate of the histogram as follows:

$$\begin{aligned} \hat{p}(\vec{x}) &= \sum_{\vec{x}'} \frac{m_{\vec{x}'}}{M} k(\vec{x}, \vec{x}') \\ \hat{p}(n_{\vec{x}} = 1|\vec{x}) &= \sum_{\vec{x}'} \frac{r_{\vec{x}} m_{\vec{x}'}}{M} k(\vec{x}, \vec{x}'), \end{aligned}$$

where $k(\vec{x}, \vec{x}')$ is called the kernel. For example the Gaussian kernel is defined as

$$k(\vec{x}, \vec{x}') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|\vec{x}-\vec{x}'\|^2}{2\sigma^2}},$$

and σ is the bandwidth parameter to control the smoothness of the estimate. The kernel estimator of the conditional probability simplifies to:

$$\hat{p}(n_{\vec{x}} = 1|\vec{x}) = \frac{\sum_{\vec{x}'} r_{\vec{x}'} m_{\vec{x}'} k(\vec{x}, \vec{x}')}{\sum_{\vec{x}'} m_{\vec{x}'} k(\vec{x}, \vec{x}')}.$$

3.8 Appendix IV: Empirical Bayes method to estimate the hyper-parameters

Here we illustrate how to exploit the Laplace approximation to obtain an Empirical Bayes (maximum marginal likelihood) estimate of the smoothing parameters. We

work with the following generative model:

$$P[D, z|\psi] = P[D|z]P[z|\psi],$$

where ψ denotes a possible vector of hyper-parameters and $\log P[z|\psi] \propto -\frac{1}{2}z^T C_\psi^{-1} z$. Thus

$$\begin{aligned} P[D|\psi] &= \int P[D|z]P[z|\psi]dz & (3.17)\\ &\approx P[D|\hat{z}_D(\psi)]P[\hat{z}_D(\psi)|\psi]\int e^{-\frac{1}{2}(z-\hat{z}_D(\psi))^T C_D^{-1}(\psi)(z-\hat{z}_D(\psi))}dz & (3.18)\\ &= (2\pi)^{d/2}|C_D(\psi)|^{1/2}P[D|\hat{z}_D(\psi)]P[\hat{z}_D(\psi)|\psi]. & (3.19)\end{aligned}$$

For any ψ, the MAP estimate $\hat{z}_D(\psi)$ (and therefore $\log P[D|\psi]$ and the following choice for ψ as the maximizer of $\log P[D|\psi]$) is available in $O(d^{3/2})$ time:

$$\begin{aligned} \hat{\psi} &= \arg\max_\psi \Big\{\frac{1}{2}\log|C_D(\psi)| + \log P[D|\hat{z}_D(\psi)] + \log P[\hat{z}_D(\psi)|\psi]\Big\}\\ &= \arg\max_\psi \Big\{-\frac{1}{2}\log|C_\psi^{-1} - \nabla\nabla_z \log p(D|z)_{z=\hat{z}_D(\psi)}| + \log P[D|\hat{z}_D(\psi)] + \log P[\hat{z}_D(\psi)|\psi]\Big\}.\end{aligned}$$

Specifically, to calculate $(1/2)\log|H|$ stably we use $\sum(\log(\mathrm{diag}(\mathrm{chol}(H))))$ in Matlab, which runs in $O(d^{3/2})$.

Acknowledgements

We thank Y. Ahmadian and M. Sahani for helpful conversations and G. Czanner for providing her code discussed in section (3.3.1.3). We are grateful to the referees and X. Pitkow for critical comments and carefully reading the manuscript. LP is supported by an NSF CAREER award, an Alfred P. Sloan Research Fellowship, a Gatsby Foundation Pilot Grant, and a McKnight Scholar award.

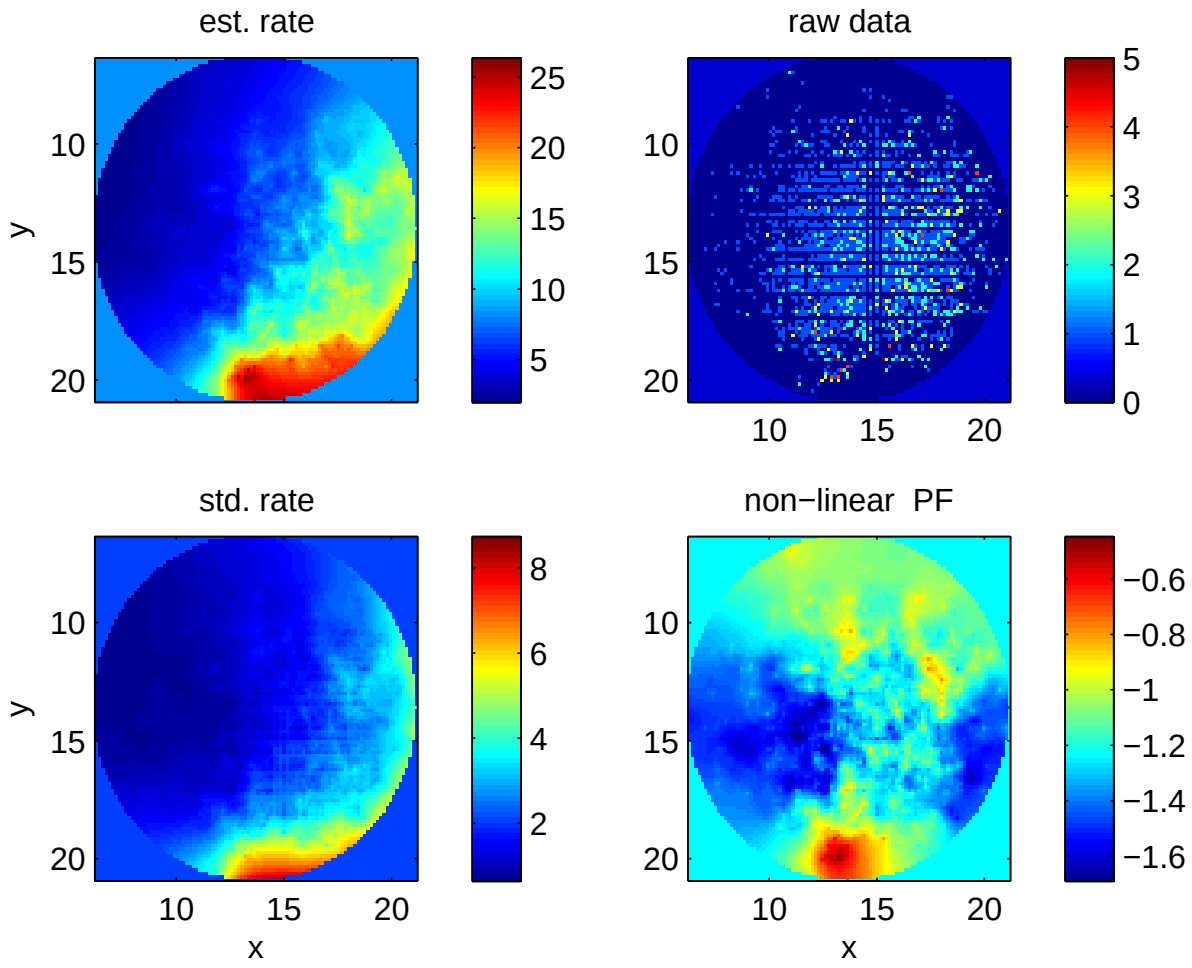

Figure 3.9: Estimating the nonlinearity in the position-dependent firing rate of an MI neuron. The data and predictions are confined to the indicated circles. Top left: Predicted firing rate of a single neuron as a function of position at zero velocity and acceleration, estimated via the Bayesian methods discussed here. Top right: The number of spikes in 50ms windows at different points in the position space. (The striped appearance here is due to aliasing effects, and should be ignored.) Bottom left: The standard deviation of the predicted firing rate. Note that the posterior uncertainty increases towards the more sparsely-sampled perimeter. Bottom right: The nonlinear part ($z(\vec{x})$) of the estimated spatial receptive-field. Note the very small scale of the nonlinear effect compared to the linear trend shown in the top left panel, consistent with the results of (Paninski et al., 2004b).

Chapter 4

Information Rates and Optimal Decoding in Large Neural Populations

This chapter is based on the paper "Information Rates and Optimal Decoding in Large Neural Populations" (Rahnama Rad and Paninski, 2011).

4.1 Introduction

It has long been argued that many key questions in neuroscience can best be posed in information-theoretic terms; the efficient coding hypothesis discussed in (Attneave, 1954; Barlow, 1961; Barlow et al., 1989; Atick, 1992), represents perhaps the best-known example. Answering these questions quantitatively requires us to compute the Shannon information rate of neural channels, whether numerically using experimental data or analytically in mathematical models. In many cases it is useful to exploit connections with "ideal observer" analysis, in which the performance

of an optimal Bayesian decoder places fundamental bounds on the performance of any biological system given access to the same neural information. However, the non-linear, non-Gaussian, and correlated nature of neural responses has hampered the development of this theory, particularly in the case of high-dimensional and/or time-varying stimuli.

The neural decoding literature is far too large to review systematically here; instead, we will focus our attention on work which has attempted to develop an analytical theory to simplify these complex decoding and information-rate problems. Two limiting regimes have received significant analytical attention in the neuroscience literature. In the "high-SNR" regime, $n \to \infty$, where n is the number of neurons encoding the signal of interest; if the information rate of each neuron is bounded away from zero and neurons respond in a conditionally weakly-dependent manner given the stimulus, then the total information provided by the neural population becomes infinite, and the error rate of any reasonable neural decoder tends to zero. For discrete stimuli, the Shannon information is effectively determined in this asymptotic limit by a simpler quantity known as the Chernoff information (Cover and Thomas, 1991; Kang and Sompolinsky, 2001); for continuous stimuli, maximum likelihood estimation is asymptotically optimal, and the asymptotic Shannon information is controlled by the Fisher information (Clarke and Barron, 1990; Brunel and Nadal, 1998). On the other hand we can consider the "low-SNR" limit, where only a few neurons are observed and each neuron is asymptotically weakly tuned to the stimulus. In this limit, the Shannon information tends to zero, and under certain conditions the optimal Bayesian estimator (which can be strongly nonlinear in general) can be approximated by a simpler linear estimator; see (Bialek and Zee, 1990) and more recently (Pillow et al., 2011) for details.

In this paper, we study information transmission and optimal decoding in what we would argue is a more biologically-relevant "intermediate" regime, where n is large

but the total amount of information provided by the population remains finite, and the problem of decoding the stimulus given the population neural activity remains nontrivial.

4.2 Likelihood in the intermediate regime: the inhomogeneous Poisson case

For clarity, we begin by analyzing the information in a simple population of neurons, represented as inhomogenous Poisson processes that are conditionally independent given the stimulus. We will extend our analysis to more general neural populations in the next section. In response to the stimulus, at each time step t neuron i fires with probability $\lambda_i(t)dt$, where the rate is given by

$$\lambda_i(t) = f\left[b_i(t) + \epsilon \ell_{i,t}(\theta)\right], \tag{4.1}$$

where $f(.)$ is a smooth rectifying non-linearity and ϵ is a gain factor controlling each neuron's sensitivity. The baseline firing rate is determined by $b_i(t)$ and is independent of the input signal. The true stimulus at time t is defined by θ_t, and θ abbreviates the time varying stimulus $\theta_{0:T}$ in the time interval $[0, Tdt]$. The term $\ell_{i,t}(\theta)$ summarizes the dependence of the neuron's firing rate on θ; depending on the setting, this term may represent e.g. a tuning curve or a spatiotemporal filter applied to the stimulus (see examples below).

The likelihood includes all the information about the stimulus encoded in the population's spiking response. Neuron i's response at time step t is designated by by the binary variable $r_i(t)$. The log-likelihood at the parameter value ϑ (which may be different from the true parameter θ) is given by the standard point-process formula

(Snyder and Miller, 1991):

$$L_\vartheta(r) := \log p(r|\vartheta) \quad = \quad \sum_{i=1}^{n}\sum_{t=0}^{T} r_i(t)\log\lambda_i(t) - \lambda_i(t)dt. \tag{4.2}$$

This expression can be expanded around $\epsilon = 0$:

$$L_\vartheta(r) \quad = \quad L_\vartheta(r)|_{\epsilon=0} + \epsilon\frac{\partial L_\vartheta(r)}{\partial\epsilon}|_{\epsilon=0} + \frac{1}{2}\epsilon^2\frac{\partial^2 L_\vartheta(r)}{\partial\epsilon^2}|_{\epsilon=0} + O(n\epsilon^3),$$

where

$$\begin{aligned}
\frac{\partial L_\vartheta(r)}{\partial\epsilon}|_{\epsilon=0} &= \sum_{i,t}\ell_{i,t}(\vartheta)\Big\{r_i(t)\frac{f'}{f}\big(b_i(t)\big) - f'(b_i(t))dt\Big\} \\
\frac{\partial^2 L_\vartheta(r)}{\partial\epsilon^2}|_{\epsilon=0} &= \sum_{i,t}\ell_{i,t}^2(\vartheta)\Big\{r_i(t)\big(\frac{f'}{f}\big)'\big(b_i(t)\big) - f''(b_i(t))dt\Big\}.
\end{aligned}$$

Let r_i denote the vector representation of the ith neuron's spike train and let[1]

$$\begin{aligned}
g_i(r_i) &:= \Big[r_i(1)\ \frac{f'}{f}(b_i(1)) - f'(b_i(1))dt \quad \cdots \quad r_i(T)\frac{f'}{f}(b_i(T)) - f'(b_i(T))dt\Big]^T \\
h_i(r_i) &:= \Big[r_i(1)\big(\frac{f'}{f}\big)'(b_i(1)) - f''(b_i(1))dt \quad \cdots \quad r_i(T)\big(\frac{f'}{f}\big)'(b_i(T)) - f''(b_i(T))dt\Big]^T \\
\ell_i(\vartheta) &:= \Big[\ell_{i,1}(\vartheta)\ \ \ell_{i,2}(\vartheta) \quad \cdots \quad \ell_{i,T}(\vartheta)\Big]^T;
\end{aligned}$$

then

$$L_\vartheta(r) \quad = \quad L_\vartheta(r)|_{\epsilon=0} + \epsilon\sum_{i=1}^{n}\ell_i(\vartheta)^T g_i(r_i) + \frac{1}{2}\epsilon^2\sum_{i=1}^{n}\ell_i(\vartheta)^T\mathrm{diag}[h_i(r_i)]\ell_i(\vartheta) + O(n\epsilon^3).$$

This second-order loglikelihood expansion is standard in likelihood theory (van der Vaart, 1998); as usual, the first term is constant in ϑ and can therefore be ignored, while the third (quadratic) term controls the curvature of the loglikelihood at $\epsilon = 0$, and scales as ϵn^2. In the high-SNR regime discussed above, where $n \to \infty$ and ϵ is fixed, the likelihood becomes sharply peaked at θ (and therefore the Fisher information, which may be understood as the curvature of the log-likelihood at θ,

[1]With a slight abuse of notation, we use T for both the total number of time steps and the transpose operation; the difference is clear from the context.

controls the asymptotics of the estimation error in the case of continuous stimuli), and estimation of θ becomes easy; in the low-SNR regime, we fix n and consider the $\epsilon \to 0$ limit.

Now, finally, we can more precisely define the "intermediate" SNR regime: we will focus on the case of large populations ($n \to \infty$), but in order to keep the total information in a finite range we need to scale the sensitivity ϵ as $\epsilon \sim n^{-1/2}$. In this setting, the error term $O(n\epsilon^3) = O(n^{-\frac{1}{2}}) = o(1)$ and can therefore be neglected, and the law of large numbers (LLN) implies that

$$\epsilon^2 \frac{\partial^2 L_\vartheta(r)}{\partial \epsilon^2}|_{\epsilon=0} \quad = \quad \mathrm{E}_{r|\theta}\left[\frac{1}{n}\sum_i \ell_i(\vartheta)^T \mathrm{diag}[h_i(r_i)]\ell_i(\vartheta)\right];$$

consequently, the quadratic term $\epsilon^2 \frac{\partial^2 L_\vartheta(r)}{\partial \epsilon^2}|_{\epsilon=0}$ will be independent of the observed spike train and therefore void of information about θ. So the first derivative term is the only part of the likelihood that depends both on the neural activity and ϑ, and may therefore be considered a sufficient statistic in this asymptotic regime: all the information about the stimulus is summarized in

$$\epsilon \frac{\partial L_\vartheta(r)}{\partial \epsilon}|_{\epsilon=0} = \frac{1}{\sqrt{n}}\sum_i \ell_i(\vartheta)^T g_i(r_i). \tag{4.3}$$

We may further apply the central limit theorem (CLT) to this sum of independent random vectors to conclude that this term converges to a Gaussian process indexed by ϑ (under mild technical conditions that we will ignore here, for clarity). Thus this model enjoys the local asymptotic normality property observed in many parametric statistical models (van der Vaart, 1998): all of the information in the data can be summarized asymptotically by a sufficient statistic with a sampling distribution that turns out to be Gaussian.

4.2.1 Example: Linearly filtered stimuli and state-space models

In many cases neurons are modeled in terms of simple rectified linear filters responding to the stimulus. We can handle this case easily using the language introduced above, if we let K_i denote the matrix implementing the transformation $(K_i\theta)_t = \ell_{i,t}(\theta)$, the projection of the stimulus onto the i-th neuron's stimulus filter. Then,

$$\epsilon \frac{\partial L_\vartheta(r)}{\partial \epsilon}|_{\epsilon=0} = \vartheta^T \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_i^T \left(\mathrm{diag}\left[\frac{f_i'}{f_i}\right] r_i - f_i' dt \right) \right] := \vartheta^T \Delta(r),$$

where f_i stands for the vector version of $f[b_i(t)]$. Thus all the information in the population spike train can be summarized in the random vector $\Delta(r)$, which is a simple linear function of the observed spike train data. This vector has an asymptotic Gaussian distribution, with mean and covariance

$$\begin{aligned}
\mathrm{E}_{r|\theta}\left(\Delta(r)\right) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_i^T \left(\mathrm{diag}\left[\frac{f_i'}{f_i}\right] \left(f_i dt + f_i' dt \frac{K_i\theta}{\sqrt{n}} + O(\frac{1}{n}) \right) - f_i' dt \right) \\
&= \left[\frac{1}{n} \sum_{i=1}^{n} K_i^T \mathrm{diag}\Big[\frac{f_i'^2}{f_i} dt\Big] K_i \right] \theta + O(\frac{1}{\sqrt{n}}) \\
J := \mathrm{cov}_{r|\theta}\left(\Delta(r)\right) &= \frac{1}{n} \sum_{i=1}^{n} K_i^T \mathrm{diag}\left[\frac{f_i'}{f_i}\right] \mathrm{cov}_{r|\theta}\Big[r_i\Big] \mathrm{diag}\left[\frac{f_i'}{f_i}\right] K_i \\
&= \frac{1}{n} \sum_{i=1}^{n} K_i^T \mathrm{diag}\Big[\frac{f_i'^2}{f_i} dt\Big] K_i + O(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Thus, the neural population's non-linear and temporally dynamic response to the stimulus is as informative in this intermediate regime as a single observation from a standard Gaussian experiment, in which the parameter θ is filtered linearly by J and corrupted by Gaussian noise. All of the filtering properties of the population are summarized by the matrix J. (Note that if we consider each K_i as a random sample from some distribution of filters, then J will converge by the law of large numbers to a matrix we can compute explicitly.)

Thus in many cases we can perform optimal Bayesian decoding of θ given the spike trains quite easily. For example, if θ has a zero mean Gaussian prior distribution with covariance C_θ, then the posterior mean and the maximum-a-posteriori (MAP) estimate is well-known and coincides with the optimal linear estimate (OLE):

$$\hat{\theta}_{OLE}(r) = E(\theta|r) = (J + C_\theta^{-1})^{-1}\Delta(r). \tag{4.4}$$

We may compute the Shannon information $I(\theta : r)$ between r and θ in a similarly direct fashion. We know that, asymptotically, the sufficient statistic $\Delta(r)$ is as informative as the full population response r

$$I(\theta : r) = I(\theta : \Delta(r)).$$

In the case that the prior of θ is Gaussian, as above, then the information can therefore be computed quite explicitly via standard formulas for the linear-Gaussian channel (Cover and Thomas, 1991):

$$I(\theta : \Delta(r)) = \frac{1}{2}\log\det(I + JC_\theta). \tag{4.5}$$

To summarize, when the encodings $\ell_{i,t}(\theta)$ are linear in θ, and we are in the intermediate-SNR regime, and the parameter θ has a Gaussian prior distribution, then the optimal Bayesian estimate is obtained by applying a linear transformation to the sufficient statistic $\Delta(r)$ which itself is linear in the spike train, and the mutual information between the stimulus and full population response has a particularly simple form. These results help to extend previous theoretical studies (Bialek and Zee, 1990; Salinas and Abbott, 1994; Snippe, 1996; Pillow et al., 2011) demonstrating that in some cases linear decoding can be optimal, and also shed some light on recent experimental studies indicating that optimal linear and nonlinear Bayesian estimators often have similar performance in practice (Macke et al., 2011; Lawhern et al., 2011).

To work through a concrete example, consider the case that the temporal sequence of parameter values θ_t is generated by an autoregressive process:

$$\theta_{t+1} \quad = \quad A\theta_t + \eta_t \qquad \eta_t \sim \mathcal{N}(0, R),$$

for a stable dynamics matrix A and positive-semidefinite covariance matrix R. Further assume that the observation matrices K_i act instantaneously, i.e., K_i is block-diagonal with blocks $K_{i,t}$, and therefore the responses are modeled as

$$r_i(t) \sim Poiss[f(b_i(t) + \epsilon K_{i,t}\theta_t)dt].$$

Thus θ and the responses r together represent a state-space model. This framework has been shown to lead to state-of-the-art performance in a wide variety of neural data analysis settings (Eden et al., 2004; Paninski et al., 2010; Koyama et al., 2010). To understand optimal inference in this class of models in the intermediate SNR regime, we may follow the recipe outlined above: we see that the asymptotic sufficient statistic in this model can be represented as

$$\Delta_t = J_t\theta_t + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, J_t),$$

where the effective filter matrix J defined above is block-diagonal (due to the block-diagonal structure of the filter matrices K_i), with blocks we have denoted J_t. Thus Δ_t represents observations from a linear-Gaussian state-space model, i.e., a Kalman filter model (Roweis and Ghahramani, 1999). Optimal decoding of θ given the observation sequence $\Delta_{1:T}$ can therefore be accomplished via the standard forward-backward Kalman filter-smoother (Durbin and Koopman, 2001; Shumway and Stoffer, 2006); see Fig. 4.1 for an illustration. The information rate $\lim_{T\to\infty} I(\theta_{0:T} : r_{0:T}) = \lim_{T\to\infty} I(\theta_{0:T} : \Delta(r)_{0:T})$ may be computed via similar recursions in the stationary case (i.e., when J_t is constant in time). The result may be expressed most explicitly in terms of a matrix which is the solution of a Riccati equation involving the effective Kalman model parameters; the details are provided in the appendix.

4.2.2 Nonlinear examples: orientation coding, place fields, and small-time expansions

While the linear setting discussed above can handle many examples of interest, it does not seem general enough to cover two well-studied decoding problems: inferring the orientation of a visual stimulus from a population of cortical neurons (Seung and Sompolinsky, 1993; Berens et al., 2011), or inferring position from a population of hippocampal or entorhinal neurons (Brown et al., 1998). In the former case, the stimulus is a phase variable, and therefore does not fit gracefully into the linear setting described above; in the latter case, place fields and grid fields are not well-approximated as linear functions of position. If we apply our general theory in these settings, the interpretation of the encoding function $\ell_i(\theta)$ does not change significantly: $\ell_i(\theta)$ could represent the tuning curve of neuron i as a function of the orientation of the visual stimulus, or of the animal's location in space. However, without further assumptions the limiting sufficient statistic, which is a weighted sum of these encoding functions $\ell_i(\theta)$ (recall eq. 4.3) may result in an infinite-dimensional Gaussian process, which may be computationally inconvenient.

To simplify matters somewhat, we can introduce a mild assumption on the tuning functions $\ell_i(\theta)$. Let's assume that these functions may be expressed in some low-dimensional basis: $\ell_i(\theta) = K_i\Phi(\theta)$, for some vectors K_i, and $\Phi(\theta)$ is defined to map θ into an mT-dimensional space which is usually smaller than $\dim(\theta) = \dim(\theta_t)T$. This finite-basis assumption is very natural: in the orientation example, tuning curves are periodic in the angle θ_t and are therefore typically expressed as sums of a few Fourier functions; similarly, two-dimensional finite Fourier or Zernike bases are often used to represent grid or place fields (Brown et al., 1998). The key point here is that we may now simply follow the derivation of the last section with $\Phi(\theta)$ in place of θ; we find that the sufficient statistic may be represented asymptotically

as an mT-dimensional Gaussian vector with mean J and covariance $J\Phi(\theta)$, with J defined as in the preceding section.

We should note that this nonlinear case does remain slightly more complicated than the linear case in one respect: while the likelihood with respect to $\Phi(\theta)$ reduces to something very simple and tractable, the prior (which is typically defined as a function of θ) might be some complicated function of the remapped variable $\Phi(\theta)$. So in most interesting nonlinear cases we can no longer compute the optimal Bayesian decoder or the Shannon information rate analytically. However, our approach does lead to a major simplification in numerical investigations into theoretical coding issues. For example, to examine the coding efficiency of a population of neurons encoding an orientation variable in this intermediate SNR regime we do not need to simulate the responses of the entire population (which would involve drawing nT random variables, for some large population size n); instead, we only need to draw a single equivalent mT-dimensional Gaussian vector $\Delta(r)$, and quantify the decoding performance based on the approximate loglikelihood

$$L_\vartheta(r) = L_\vartheta(r)|_{\epsilon=0} + \Phi(\vartheta)^T\Delta(r) + \frac{1}{2}\Phi(\vartheta)^T J\Phi(\vartheta) + O(\frac{1}{\sqrt{n}}),$$

which as emphasized above has a simple quadratic form as a function of $\Phi(\vartheta)$. Since m can typically be chosen to be much smaller than n, this approach can result in significant computational savings.

We now switch gears slightly and examine another related intermediate regime in which nonlinear encoding plays a key role: instead of letting the sensitivity ϵ of each neuron become small (in order to keep the total information in the population finite), we could instead keep the sensitivity constant and let the time period over which we are observing the population scale inversely with the population size n. This short-time limit is sensible in some physiological and psychophysical contexts (Thorpe et al., 1996) and was examined analytically in (Panzeri et al., 1999) to

study the impact of inter-neuron dependencies on information transmission. Our methods can also be applied to this short-time limit. We begin by writing the loglikelihood of the observed spike count vector r in a single time-bin of length dt:

$$L_\vartheta(r) := \log p(r|\theta) \quad = \quad \sum_i r_i \log f\left[b_i + \ell_i(\vartheta)\right] - f\left[b_i + \ell_i(\vartheta)\right] dt.$$

The second term does not depend on r; therefore, all information in r about θ resides in the sufficient statistic

$$\Delta_\vartheta(r) := \sum_i r_i \log f\left[b_i + \ell_i(\vartheta)\right].$$

Since the i-th neuron fires with probability $f\left[b_i + \ell_i(\theta)\right] dt$, the mean of $\Delta_\vartheta(r)$ scales with ndt, and it is clear that $dt = 1/n$ is a natural scaling of the time bin. With this scaling $\Delta_\vartheta(r)$ converges to a Gaussian stochastic process with mean

$$\mathrm{E}_{r|\theta}[\Delta_\vartheta(r)] = \frac{1}{n} \sum_i f\left[b_i + \ell_i(\theta)\right] \log f\left[b_i + \ell_i(\vartheta)\right]$$

and covariance

$$\mathrm{cov}_{r|\theta}[\Delta_\vartheta(r), \Delta_{\vartheta'}(r)] = \frac{1}{n} \sum_i f\left[b_i + \ell_i(\theta)\right] \Big(\log f\left[b_i + \ell_i(\vartheta)\right]\Big) \Big(\log f\left[b_i + \ell_i(\vartheta')\right]\Big),$$

where we have used the fact that the variance of a Poisson random variable coincides with its mean.

In general, this limiting Gaussian process will be infinite-dimensional. However, if we choose the exponential nonlinearity ($f(.) = \exp(.)$) and the encoding functions $\ell_i(\theta)$ are of the finite-dimensional form considered above, $\ell_i(\theta) = K_i^T \Phi(\theta)$, then the $\log f[b_i + \ell_i(\vartheta)]$ term in the definition of $\Delta_\vartheta(r)$ simplifies: in this case, all information about θ is captured by the sufficient statistic

$$\Delta(r) = \sum_i r_i K_i.$$

If we again let $dt = 1/n$, then we find that $\Delta(r)$ converges to a finite-dimensional Gaussian random vector with mean and covariance

$$\mathrm{E}_{r|\theta}[\Delta(r)] = \frac{1}{n}\sum_i f\left[b_i + K_i^T\Phi(\theta)\right]K_i; \qquad \mathrm{cov}_{r|\theta}[\Delta(r)] = \frac{1}{n}\sum_i f\left[b_i + K_i^T\Phi(\theta)\right]K_iK_i^T;$$

again, if the filters K_i are modeled as independent draws from some fixed distribution, then the above normalized sums converge to their expectations, by the LLN. Thus, as in the intermediate-SNR regime, we see that inference can be dramatically simplified in this short-time setting.

4.3 Likelihood in the intermediate regime: non-Poisson effects

We conclude by discussing the generalization to non-Poisson networks with interneuronal dependencies and nontrivial correlation structure. We generalize the rate equation (4.1) to

$$\lambda_i(t) = f_i\left[b_i(t) + \epsilon\ell_{i,t}(\theta)\middle|\mathcal{H}_t\right],$$

where $\mathcal{H}_t$ stands for the spiking activity of all neurons prior to time t: $\mathcal{H}_t = \{r_i(t')\}_{t'<t, 1\le i\le n}$. Note that the influence of spiking history may be different for each neuron: refractory periods, self-inhibition and coupling between neurons can be formulated by appropriately defining the dependence of $f_i(.)$ on $\mathcal{H}_t$.

We begin, as usual, by expanding the log-likelihood. The basic point-process likelihood (eq. 4.2) remains valid. Let $g_i(r)$ and $h_i(r)$ denote the vector versions of

$$r_i(t)\frac{f'}{f}\Big[b_i(t)\big|\mathcal{H}_t\Big] - f_i'\Big[b_i(t)\big|\mathcal{H}_t\Big]dt \quad \text{and} \quad r_i(t)\big(\frac{f'}{f}\big)'\Big[b_i(t)\big|\mathcal{H}_t\Big] - f_i''\Big[b_i(t)\big|\mathcal{H}_t\Big]dt,$$

respectively, analogously to the Poisson case. Then, the first and second terms in

the expansion of the loglikelihood may be written as

$$\epsilon \frac{\partial L_\vartheta(r)}{\partial \epsilon}|_{\epsilon=0} = \epsilon \sum_i \ell_i^T(\vartheta) g_i(r) \qquad \text{and} \qquad \frac{1}{2}\epsilon^2 \frac{\partial^2 L_\vartheta(r)}{\partial \epsilon^2}|_{\epsilon=0} = \frac{1}{2}\epsilon^2 \sum_i \ell_i^T(\vartheta) \mathrm{diag}[h_i(r)] \ell_i(\vartheta),$$

as before. For independent neurons, the log-likelihood was composed of normalized sums of independent random variables that converged to a Gaussian process, by the CLT. In the history-dependent, coupled case, $g_i(r)$ and $h_i(r)$ depend not only on the i-th neuron's activity r_i, but rather on the whole network history. Nonetheless, under technical conditions on the network's dependence structure (to ensure that the firing rates and correlations in the network remain bounded), we may still exploit versions of the LLN and CLT. Thus, under conditions ensuring the validity of the LLN we may conclude that, as before, the second-order term $\epsilon^2 \frac{\partial^2 L_\vartheta(r)}{\partial \epsilon^2}|_{\epsilon=0}$ converges to its expectation under the intermediate $\epsilon \sim n^{-\frac{1}{2}}$ scaling, and therefore carries no information about θ. When we discard this second-order term, along with higher-order terms that are negligible in the intermediate-SNR, large-n limit, we are left once again with the gradient term $\epsilon \frac{\partial L_\vartheta(r)}{\partial \epsilon}|_{\epsilon=0} = \frac{1}{\sqrt{n}} \sum_i \ell_i(\vartheta)^T g_i(r)$, which under appropriate conditions (ensuring the validity of a CLT) will converge to a Gaussian process limit whose mean and covariance we can often compute analytically.

Let's turn to a specific example, in order to make these claims somewhat more concrete. Consider a network with weak couplings and possibly strong self-inhibition and history dependence; more precisely, we assume that interneuronal conditional cross-covariances are weak, given the stimulus:

$$\mathrm{cov}[r_i(t), r_j(t+\tau)|\theta] = O(n^{-1}) \qquad \text{for} \quad i \neq j.$$

See, e.g., (Ginzburg and Sompolinsky, 1994; Toyoizumi et al., 2009) for further discussion of this condition, which is satisfied for many spiking networks in which the synaptic weights scale uniformly as $O(n^{-1})$. For simplicity, we will also restrict our attention to linear encoding functions, though generalizations to the nonlinear

case are straightforward. Thus, as before, let K_i denote the matrix implementing the transformation $(K_i\theta)_t = \ell_{i,t}(\theta)$, the projection of the stimulus onto the i-th neuron's stimulus filter. Then

$$\epsilon \frac{\partial L_\vartheta(r)}{\partial \epsilon}|_{\epsilon=0} = \vartheta^T \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_i^T \left(\text{diag} \left[\frac{f_i'}{f_i} \right] r_i - f_i' dt \right) \right],$$

where f_i stands for the vector version of $f_i\Big[b_i(t)\Big|\mathcal{H}_t\Big]$; in other words, the t-th entry of $f_i dt$ is the probability of observing a spike in the interval $[t, t+dt]$, given the network spiking history $\mathcal{H}_t$ in the absence of input. Our sufficient statistic is therefore exactly as in the Poisson setting,

$$\Delta(r) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_i^T \left(\text{diag} \left[\frac{f_i'}{f_i} \right] r_i - f_i' dt \right), \tag{4.6}$$

except for the history-dependence induced through the redefinition of f_i.

Computing the necessary means and covariances in this case requires more work than in the Poisson case; see the appendix for details. It is helpful (though not necessary) to make the stationarity assumption $b_i(t) \equiv b_i$, which implies in this setting that $\mathrm{E}(\frac{f_i'^2}{f_i})$ can also be chosen to be time-invariant; in this case the limiting covariance and mean of the sufficient statistic are given by

$$J := \text{cov}_{r|\theta}\left[\Delta(r)\right] = \frac{1}{n} \sum_{i=1}^{n} K_i \text{diag}\Big[\mathrm{E}_{r|\theta=0}\big(\frac{f_i'^2}{f_i} dt\big)\Big] K_i; \qquad \mathrm{E}_{r|\theta}\left[\Delta(r)\right] = J\theta,$$

where the expectations are over the spontaneous network activity in the absence of any input. In short, once again, we have $\Delta(r) \to_D \mathcal{N}(J\theta, J)$. Analytically, the only challenge here is to compute the expectations in the definition of J. In many cases this can be done analytically (e.g., in any population of uncoupled renewal-process neurons), or by using mean-field theory (Toyoizumi et al., 2009), or numerically by simply calculating the mean firing rate of the network in the undriven state $\theta = 0$.

We examine this convergence quantitatively in Fig. 4.1. In this case the stimulus θ_t was a sample path from a one-dimensional autoregressive (AR(1)) process. Spikes

were generated according to

$$\lambda_i(t) = \lambda_o \exp\left(\frac{\theta_t}{\sqrt{n}} + \sum_{j=1}^{n} w_{ji} I_j(t)\right) 1_{\tau_i(t)>\tau_{\mathrm{ref}}},$$

where $I_j(t)$ is the synaptic input from the j-th cell (generated by convolving the spike train r_j with an exponential of time constant 20 ms), w_{ji} is the synaptic weight matrix coupling the output of neuron j to the input of neuron i, $\tau_i(t)$ is the time since the last spike; therefore, $1_{\tau_i(t)>\tau_{\mathrm{ref}}}$ enforces the absolute refractory period τ_{ref}, which was set to be 2 ms here. Since the encoding filters K_i act instantaneously in this model (K_i can be represented as a delta function, weighted by $n^{-1/2}$), the observed spike trains can be considered observations from a state-space model, as described above. The weights w_{ji} were generated randomly from a uniform distribution on the interval $-[5/n, 5/n]$, with self-weights $w_{ii} = 0$, and $\sum_j w_{ji} = 0$ to enforce detailed balance in the network. Note that, while the interneuronal coupling is weak in this example, the autocorrelation in these spike trains is quite strong on short time scales, due to the absolute refractory effect.

We compared two estimators of θ: the full (nonlinear) MAP estimate $\hat{\theta}_{MAP} = \arg\max_\theta p(\theta|r)$, which we computed using the fast direct optimization methods described in (Koyama and Paninski, 2009; Paninski et al., 2010), and the limiting optimal estimator $\hat{\theta}_\Delta := (J + C_\theta^{-1})^{-1}\Delta(r)$. Note that J is diagonal; we computed the expectations in the definition of J using the numerical approach described above in this simulation, though in other simulations (with uncoupled renewal-model populations) we checked that the fully-analytical approach gave the correct solution. In addition, C_θ^{-1} is tridiagonal in this state-space setting; thus the linear matrix equation in eq. (4.4) can be solved efficiently in $O(T)$ time using standard tridiagonal matrix solvers. We find that, as predicted, the full nonlinear Bayesian estimator $\hat{\theta}_{MAP}$ approaches the limiting optimal estimator $\hat{\theta}_\Delta$ as n becomes large; $n = 10$ is basically sufficient in this case, although of course the convergence will

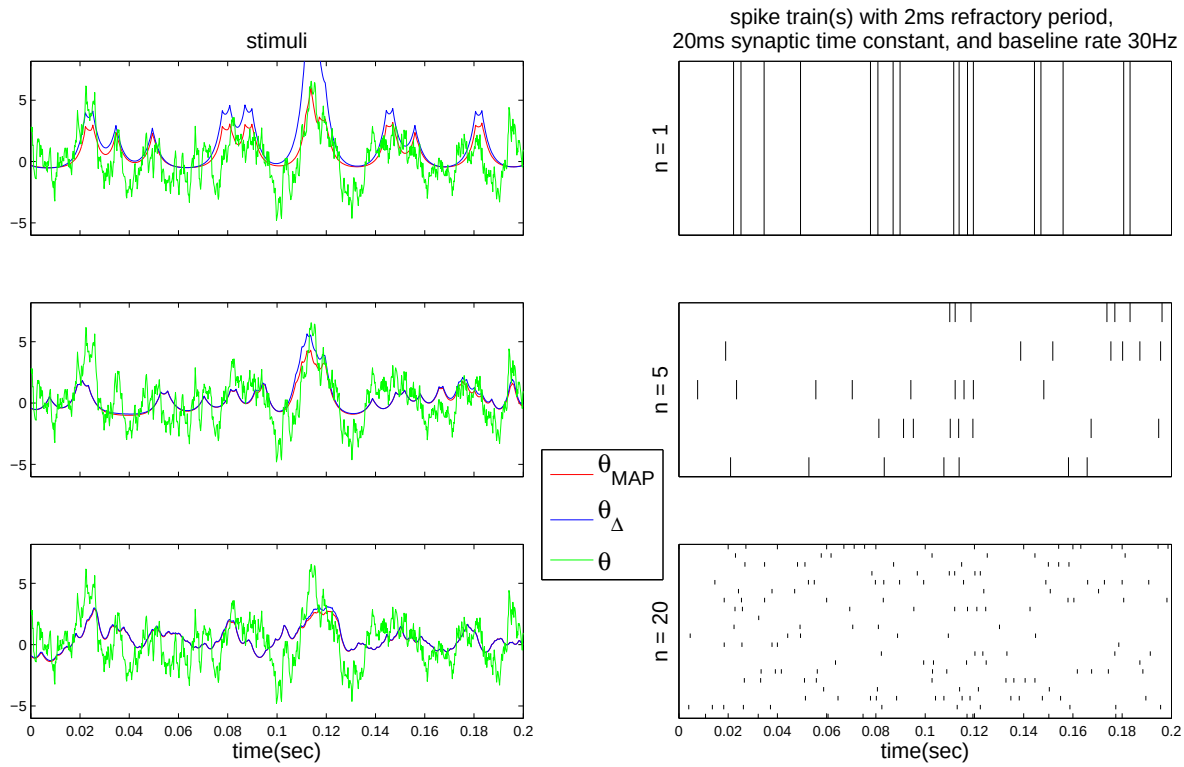

Figure 4.1: The left panels show the true stimulus (green), MAP estimate (red) and the limiting optimal estimator $\hat{\theta}_\Delta := (J + C_\theta^{-1})^{-1}\Delta(r)$ (blue) for various population sizes n. The right panels show the spike trains used to compute these estimates. Note that the same true stimulus was used in all three simulations. As n increases, the linear decoder converges to the MAP estimate, despite the nonlinear and correlated nature of the network model generating the spike trains (see main text for details).

be slower for larger values of the gain factor ϵ (or, equivalently, larger filters K_i or larger values of the variance of θ_t).

We conclude with a few comments about these results. First, note that the covariance matrix J we have computed here coincides almost exactly with what we computed previously in the Poisson case. Indeed, we can make this connection much more precise: we can always choose an equivalent Poisson network with rates defined so that the $\mathrm{E}_{r|\theta=0}[(f_i')^2/f_i]$ term in the non-Poisson network matches the $(f_i')^2/f_i$ term in the Poisson network. Since J determines the information rate

completely, we conclude that for any weakly-coupled network there is an equivalent Poisson network which conveys exactly the same information in the intermediate regime. However, note that the the sufficient statistic $\Delta(r)$ is different in the Poisson and non-Poisson settings, since the f'/f term linearly reweights the observed spikes, depending on how likely they were given the history; thus the optimal Bayesian decoder incorporates non-Poisson effects explicitly.

A number of interesting questions remain open. For example, while we expect a LLN and CLT to continue to hold in many cases of strong, structured interneuronal coupling, computing the asymptotic mean and covariance of the sufficient statistic $\Delta(r)$ may be more challenging in such cases, and new phenomena may arise. We also hope in the future to examine the effect of latent correlated variability (as discussed, e.g., in the recent work of (Yu et al., 2009; Vidne et al., 2009)) on the results presented here.

4.4 Appendix: Information rates in the Kalman model

For completeness, in this appendix we provide the details of the computation of the information rate in the Kalman model. The information rate is the difference between the prior entropy rate and the posterior entropy rate of the stimulus. The former can be calculated using the Markov property (Cover and Thomas, 1991); namely,

$$\lim_{T\to\infty}\frac{1}{T}H(\theta_{1:T}) \;=\; \lim_{T\to\infty}\frac{1}{T}\left[H(\theta_1)+\sum_{t=2}^{T}H(\theta_t|\theta_{t-1})\right]=\frac{1}{2}\log\det R+\text{constant},$$

where R is the dynamics noise covariance defined in the state-space section of the main text, and constant denotes a term that will cancel with the same term in the posterior entropy rate and can therefore be ignored.

We provide three methods of increasingly explicit form for computing the posterior entropy rate. The posterior distribution of the stimulus given data is a Gaussian distribution; therefore, the posterior entropy depends on the determinant of the posterior covariance matrix $\mathrm{cov}[\theta_{1:T}\big|\Delta_{1:T}]$. This matrix is of size $Td\times Td$, where $d=dim(\theta_t)$. The inverse of this matrix is block-tridiagonal (Paninski et al., 2010), with blocks of size $d\times d$, and we may therefore compute the determinant of this matrix in $O(T)$ time using standard block-tridiagonal determinant recursions. Examining these recursions leads to a Riccati-like equation that determines the posterior entropy rate, $\lim_{T\to\infty}(1/T)\log\det\mathrm{cov}[\theta_{1:T}\big|\Delta_{1:T}]+\text{constant}$.

Alternatively, we can use a method described in (Huggins and Paninski, 2011), based on the Gaussian integral identity

$$\log p(\Delta)=\log\int p(\theta,\Delta)d\theta=\log p(\hat\theta)+\log p(\Delta|\hat\theta)+\frac{1}{2}\log\det\mathrm{cov}(\theta|\Delta)+\text{constant},$$

where $\hat\theta=\mathrm{E}(\theta_{1:T}|\Delta_{1:T})$ can be computed via the standard forward-backward Kalman

recursions. Since this formula is valid for any value of Δ, e.g., $\Delta = 0$, we can compute the marginal log-probability $\log p(\Delta)$ via the standard forward recursion for the Kalman filter, and $\log p(\hat{\theta})$ and $\log p(\Delta|\hat{\theta})$ by plugging $\hat{\theta}$ into the log-prior $\log p(\theta)$ and the log-likkelihood $\log p(\Delta|\theta)$, which are both computable explicitly in this model. This leaves us with the $\frac{1}{2}\log\det\mathrm{cov}(\theta|\Delta)$ term; taking limits of the result divided by T provides the posterior entropy rate.

Finally, a third, explicit method to compute the posterior entropy rate may be derived as follows:

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{T}H(\theta_{1:T}|\Delta_{1:T}) &= \lim_{T\to\infty}\frac{1}{T}\Big[\mathrm{E}_{\Delta_{1:T}}H(\theta_1|\Delta_{1:T}) + \sum_{t=2}^{T}\mathrm{E}_{\Delta_{1:T}}H(\theta_t|\theta_{t-1},\Delta_{1:T})\Big] \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\frac{1}{2}\log\det\mathrm{cov}[\theta_t|\theta_{t-1},\Delta_{1:T}].\end{aligned}$$

The covariance $\mathrm{cov}[\theta_t|\theta_{t-1},\Delta_{1:T}]$ can be expressed in terms of the forward covariance matrix $C_t^f = \mathrm{cov}[\theta_t|\Delta_{1:t}]$ and the backward covariance matrix $C_t^s = \mathrm{cov}[\theta_t|\Delta_{1:T}]$; the joint covariance of θ_t and θ_{t+1} given the full observation Δ can be expressed as (Durbin and Koopman, 2001; Shumway and Stoffer, 2006):

$$\begin{pmatrix} C_t^s & C_{t+1}^s K_t^T \\ K_t C_{t+1}^s & C_{t+1}^s, \end{pmatrix},$$

where

$$\mathrm{cov}(\theta_t|\Delta_{1:t-1}) = AC_{t-1}^f A^T + R$$

and

$$K_t = C_t^f J^T[\mathrm{cov}(\theta_t|\Delta_{1:t-1})]^{-1}.$$

Using the standard formula for computing the conditional covariance of a Gaussian we have:

$$\mathrm{cov}[\theta_t|\theta_{t-1},\Delta_{1:T}] = C_t^s - K_{t-1}C_t^s K_{t-1}^T.$$

Finally, we have:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=2}^{T}\frac{1}{2}\log\det\operatorname{cov}[\theta_t|\theta_{t-1},\Delta_{1:T}] \;=\; \frac{1}{2}\log|C^s-KC^sK^T|$$

where

$$C^s=\lim_{T\to\infty}C^s_{T/2} \qquad \text{and} \qquad K \;=\; \lim_{T\to\infty}K_{T/2}.$$

These matrices can be found using the Riccatti equations:

$$C^f \;=\; \Big[(AC^fA^T+R)^{-1}+J\Big]^{-1} \qquad \text{and} \qquad C^s-KC^sK^T=C^f-K_t(AC^fA^T+R)K_t^T.$$

Appendix: Mean and Covariance of sufficient statistic with History Dependence

The expectation and covariance of $\Delta(r)$ should be calculated over the distribution of network activity r in response to input θ. The expectation of

$$\Delta(r) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_i^T\left(\operatorname{diag}\left[\frac{f_i'}{f_i}\right]r_i-f_i'dt\right)$$

depends on the expectation of $\frac{f_i'}{f_i}(t)r_i(t)-f_i'(t)dt$. Note that by conditioning on the history $\mathcal{H}_t$

$$\begin{aligned}
\mathrm{E}_{r|\theta}\left\{\frac{f_i'}{f_i}(t)r_i(t)-f_i'(t)dt\right\} &= \mathrm{E}_{r|\theta}\left\{\mathrm{E}_{r|\theta}\Big[\frac{f_i'}{f_i}(t)r_i(t)-f_i'(t)dt\Big|\mathcal{H}_t\Big]\right\}\\
&= \mathrm{E}_{r|\theta}\left\{\frac{f_i'}{f_i}(t)\mathrm{E}_{r|\theta}\Big[r_i(t)\Big|\mathcal{H}_t\Big]-f_i'(t)dt\right\}\\
&= \mathrm{E}_{r|\theta}\left\{\frac{f_i'}{f_i}(t)\Big[f_i(t)dt+f_i'dt\frac{(K_i\theta)_t}{\sqrt{n}}+O(\frac{1}{n})\Big]-f_i'(t)dt\right\}\\
&= \mathrm{E}_{r|\theta}\Big(\frac{f_i'^2}{f_i}(t)dt\Big)\frac{(K_i\theta)_t}{\sqrt{n}}+O(\frac{1}{n}) \qquad (4.7)
\end{aligned}$$

therefore the expectation of $\Delta(r)$ is simplified to

$$\begin{aligned}
\mathrm{E}_{r|\theta}\left(\Delta(r)\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} K_i^T\left\{\mathrm{E}_{r|\theta}\Big(\mathrm{diag}\left[\frac{f_i'^2}{f_i}dt\right]\Big)\frac{K_i\theta}{\sqrt{n}} + O(\frac{1}{n})\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} K_i^T\mathrm{E}_{r|\theta}\Big(\mathrm{diag}\left[\frac{f_i'^2}{f_i}dt\right]\Big)K_i\theta + O(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n}\sum_{i=1}^{n} K_i^T\mathrm{E}_{r|\theta=0}\Big(\mathrm{diag}\left[\frac{f_i'^2}{f_i}dt\right]\Big)K_i\theta + O(\frac{1}{\sqrt{n}})
\end{aligned}$$

The covariance can be written as

$$\mathrm{cov}_{r|\theta}\left(\Delta(r)\right) = \frac{1}{n}\sum_{i=1}^{n} K_i^T\mathrm{cov}_{r|\theta}\left[\mathrm{diag}\left[\frac{f_i'}{f_i}\right]r_i - f_i'dt\right]K_i \tag{4.8}$$

$$+ \frac{1}{n}\sum_{i\neq j} K_i^T\mathrm{cov}_{r|\theta}\left[\mathrm{diag}\left[\frac{f_i'}{f_i}\right]r_i - f_i'dt\ ,\ \mathrm{diag}\left[\frac{f_j'}{f_j}\right]r_j - f_j'dt\right]K_j \tag{4.9}$$

First, we calculate the sum in equation (4.8); second, we show that for weak coupling the sum in equation (4.9) is $O(\frac{1}{\sqrt{n}})$. For simplicity of presentation, let us define

$$Z_{i,t} := r_i(t)\frac{f_i'}{f_i}(t) - f_i'(t)dt.$$

The terms in the sum of equation (4.8) are auto-covariances that can be written as ($\tau \geq 0$)

$$\begin{aligned}
&\mathrm{cov}_{r|\theta}\left[\mathrm{diag}\left[\frac{f_i'}{f_i}\right]r_i - f_i'dt\right]_{t,t+\tau} = \mathrm{cov}_{r|\theta}\left[Z_{i,t}, Z_{i,t+\tau}\right] \\
&= \mathrm{E}\left[\mathrm{cov}\left[Z_{i,t}, Z_{i,t+\tau}\middle|\mathcal{H}_{t+\tau}\right]\right] + \mathrm{cov}\left[\mathrm{E}[Z_{i,t}|\mathcal{H}_{t+\tau}], \mathrm{E}[Z_{i,t+\tau}|\mathcal{H}_{t+\tau}]\right] \\
&= \delta(\tau)\mathrm{var}_{r|\theta}\Big[Z_{i,t}\Big] + \mathrm{cov}_{r|\theta}\left[Z_{i,t}, \frac{f_i'^2}{f_i}(t+\tau)dt\right]\frac{(K_i\theta)_{t+\tau}}{\sqrt{n}} + O(\frac{1}{n}) \\
&= \delta(\tau)\left[\mathrm{E}\big[\mathrm{var}(Z_{i,t}|\mathcal{H}_t)\big] + \mathrm{var}\big[\mathrm{E}(Z_{i,t}|\mathcal{H}_t)\big]\right] + O(\frac{1}{\sqrt{n}}) \\
&= \delta(\tau)\mathrm{E}_{r|\theta}\big[\frac{f_i'^2}{f_i}(t)dt\big] + O(\frac{1}{\sqrt{n}}) \\
&= \delta(\tau)\mathrm{E}_{r|\theta=0}\big[\frac{f_i'^2}{f_i}(t)dt\big] + O(\frac{1}{\sqrt{n}})
\end{aligned}$$

Next, we show that if the cross-correlation in the network activity is small

$$\mathrm{cov}[r_i(t), r_j(t+\tau)] \sim \frac{1}{n}$$

then, the sum of cross-covariance terms in equation (4.9) is negligible because

$$\begin{aligned}
& \mathrm{cov}_{r|\theta}\left[\mathrm{diag}\left[\frac{f_i'}{f_i}\right] r_i - f_i' dt, \mathrm{diag}\left[\frac{f_j'}{f_j}\right] r_j - f_j' dt\right]_{t,t+\tau} = \mathrm{cov}_{r|\theta}\left[Z_{i,t}, Z_{j,t+\tau}\right] \\
= \; & \mathrm{E}\left[\mathrm{cov}\left[Z_{i,t}, Z_{j,t+\tau}\middle|\mathcal{H}_{t+\tau}\right]\right] + \mathrm{cov}\left[\mathrm{E}[Z_{i,t}|\mathcal{H}_{t+\tau}], \mathrm{E}[Z_{j,t+\tau}|\mathcal{H}_{t+\tau}]\right] \\
= \; & \mathrm{cov}\left[Z_{i,t}, \mathrm{E}[Z_{j,t+\tau}|\mathcal{H}_{t+\tau}]\right] \\
= \; & \mathrm{cov}\left[r_i(t)\frac{f_i'}{f_i}(t) - f_i'(t)dt, \frac{f_j'^2}{f_j}(t+\tau)\right]\left\{\frac{(K_j\theta)_{t+\tau}}{\sqrt{n}} + O(\frac{1}{n})\right\} = O(\frac{1}{n^{3/2}})
\end{aligned}$$

; thus,

$$\frac{1}{n}\sum_{i\neq j} K_i^T \mathrm{cov}_{r|\theta}[r_i \cdot \frac{f_i'}{f_i} - f_i' dt, r_j \cdot \frac{f_j'}{f_j} - f_j' dt] K_j \;=\; O\left(\frac{1}{\sqrt{n}}\right).$$

Chapter 5

Distributed Parameter Estimation in Networks

This chapter is based on the paper "Distributed Parameter Estimation in Networks" (Rahnama Rad and Tahbaz-Salehi, 2010)

5.1 Introduction

Information aggregation is a fundamental problem in multi-agent systems. In many scenarios, observations are distributed throughout the network in such a way that no agent has access to enough data to learn a relevant parameter in isolation, and therefore, agents face the task of recovering the truth by engaging in communication with one another. Such problems are ubiquitous in social and economic networks, as well as networks engineered for specific applications. For example, Kotler (Kotler, 1986) and Ioannides and Loury (Ioannides and Loury, 2004) document how people base their decisions on their neighbors' information when purchasing consumer products or adopting new technologies, respectively. Similarly, the main

goal of distributed sensor and robotic networks is to aggregate relevant decentralized information, so that a pre-specified task can be performed properly (see e.g., Jadbabaie, Lin, and Morse (Jadbabaie et al., 2003) and Bullo, Cortés, and Martínez (Bullo et al., 2009)).

The goal of this paper is to develop a recursive model for aggregation of dispersed information over networks, where the measurements of each agent are only partially informative about the unknown parameter. In order to resolve the local identification problems they face,[1] agents in our model update their estimates as a function of their neighbors' beliefs. More specifically, we assume that at discrete time intervals, each agent sets its belief as the geometric mean of the likelihood of its observation and its neighbors' beliefs, and uses the mode of the updated belief function as the estimate for the unknown parameter.

We show that despite the absence of local identifiability across the network, agents' estimates are weakly consistent (i.e., converge to the truth in probability), provided that there exists a directed information path connecting any two agents in the network. In other words, we prove that as long as the underlying network is *strongly connected*, information is properly aggregated over the network and the local identification problems are resolved. We also show that as observations accumulate, the distribution of agents' estimates converge to a normal distribution. The consistency and asymptotic normality of agents' estimates hold regardless of the distribution of their measurements and the structure of the network (beyond of course, the strong connectivity requirement). Furthermore, we characterize the asymptotic covariance matrix of the distributed estimates in terms of agents' signal structures, as well as the network topology. Using this characterization, we show that in bidirectional

[1]Throughout the paper, by local (global) identifiability, we mean the possibility of consistently estimating the parameter through an agent's private data (the data observed by all agents). The terminology should not be mistaken by the concepts of local and global indistinguishably in a neighborhood of the true parameter in the parameter space.

networks, distributed estimators are as efficient as any centralized estimator with access to the collection of signals observed across the network. This efficiency is achieved even if the communication network is highly sparse.

Our work is related to the collection of works on learning in networks in economics, as well as distributed estimation and consensus algorithms in the control literature. The consensus literature (such as DeGroot (DeGroot, 1974), Jadbabaie, Lin, and Morse (Jadbabaie et al., 2003), and Golub and Jackson (Golub and Jackson, 2010)) studies models in which a collection of agents asymptotically agree on the same value. Golub and Jackson provide conditions under which the asymptotic consensus value coincides with the true underlying parameter in *large* networks. In the same spirit is Xiao, Boyd, and Lall (Xia et al., 2005), which uses the consensus update to compute the maximum-likelihood estimate of the underlying parameter in a distributed fashion. These papers, however, do not address the problem of local identifiability, as they assume that all agents' observations are equally informative. As a main point of departure from previous studies, we consider agents who face local identification problems due to their different signal structures. Moreover, we show that as time progresses, not only the agents agree on their estimates, but also their consensus estimate converges to the true underlying parameter.

More relevant to our paper is (Jadbabaie et al., 2010), which studies distributed non-Bayesian learning in social networks. However, unlike (Jadbabaie et al., 2010), we study the problem of estimating a parameter in a continuum and in presence of continuous observations. Furthermore, we characterize the rate of convergence and the efficiency of the estimates. Finally, our work is also relevant to (Kar et al., 2008), who focus on a non-stationary update with time-decaying weight sequences associated with consensus and innovation updates. In contrast to (Kar et al., 2008), in this paper, we address general non-linear observation models and present a stationary update for the beliefs.

The rest of the paper is organized as follows. In the next section, we describe the model and present the dynamics according to which agents update their estimates of the true parameter. In Section III, we prove that all agents' estimates are consistent. Asymptotic normality is proved in Section IV, where we also compute the asymptotic variance of agents' estimates. In Section V, we investigate the efficiency of the distributed estimators and compare our results with centralized maximum likelihood estimation. Section VI concludes.

5.2 The Model

5.2.1 Agents and Observations

Let $N = \{1, 2 \ldots, n\}$ denote a group of agents, located on a network, who are assigned the task of estimating an unknown parameter $\theta^* \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$ is a convex parameter space. At discrete time steps $t \in \mathbb{N}$, each agent observes noisy and partially informative signals that can be used in estimating the parameter. More specifically, at any given time period t, agent i observes a random signal $s_t^i \in \mathbb{R}^p$, drawn from a distribution with conditional probability density $\ell_i(\cdot|\theta)$. We assume that agents' signals are i.i.d. over time and independent from the observations of all other agents.

The signals observed by a single agent, although potentially informative, do not reveal the parameter completely; i.e., each agent faces an identification problem. Two parameters are said to be observationally equivalent from the point of view of an agent if the conditional distributions of the signals coincide. We denote the set of parameters that are observationally equivalent to θ^* from the point of view of agent i by $\bar{\Theta}_i \triangleq \{\theta \in \Theta : \mathbb{P}[\ell_i(s^i|\theta) = \ell_i(s^i|\theta^*)] = 1\}$.[2]

[2]Throughout the paper, $\mathbb{P}$ refers to the probability distribution induced by the true parameter

Despite the local identification problems faced by the agents, we assume that the true parameter is identifiable if one has access to the signals observed by all agents.

Assumption (GI): The true parameter is *globally identifiable*; that is, $\bigcap_{i=1}^{n} \bar{\Theta}_i = \{\theta^*\}$.

Clearly, in the absence of the above assumption, even an agent with access to all the data collected across the network over time would not be able to consistently estimate θ^*.

In addition to Assumption (GI), we impose the following regularity conditions on the observation models of the agents:

(A1) $\ell_i(\cdot|\theta)$ is twice continuously differentiable in θ for all realizations of data.

(A2) $\log \ell_i(\cdot|\theta)$ is concave in θ for all observations.

(A3) $\ell_i(s^i|\theta)$ is a measurable function of s^i for all $\theta \in \Theta$.

(A4) $\mathbb{E}[\log^2 \ell_i(s_1^i|\theta)] < \infty$ for all i.

(A5) $\mathbb{E}\left[\sup_{\theta\in\mathcal{B}} \|\nabla_{\theta\theta} \log \ell_i(s_1^i|\theta)\|\right] < \infty$, for some neighborhood $\mathcal{B}$ of θ^*, where $\nabla_{\theta\theta}$ denotes the Hessian with respect to the parameter vector θ.

The above assumptions are quite mild and many of the usual distribution families, such as normals and exponentials, satisfy them. We have made these assumptions for simplicity, and our results hold under much weaker restrictions as well.

Finally, we define the *Fisher information matrix* corresponding to agent i's observation model as the covariance of its score function; that is,

$$\mathcal{I}_i(\theta) = \mathbb{E}\left[\nabla_\theta \, \psi_\theta^i(s_1^i) \nabla_\theta \, \psi_\theta^i(s_1^i)'\right] \tag{5.1}$$

θ^*, and $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$.

where $\psi^i_\theta(s^i_t) \triangleq \log \ell_i(s^i_t|\theta)$ and ∇_θ denotes the gradient with respect to the parameter vector θ. As the definition suggests $\mathcal{I}_i$ is a $d \times d$ symmetric and positive semi-definite matrix.

5.2.2 Network Structure

In addition to signals $\{s^i_t\}_{t=1}^\infty$ observed privately over time, each agent can communicate with a subset of other agents known as its *neighbors*. We capture this neighborhood relation with a directed graph $G = (V, E)$, where each vertex in V corresponds to an agent $i \in N$, and there exists a directed edge $(j, i) \in E$ from vertex j to vertex i if agent i has access to the belief function of agent j. We denote the set of neighbors of agent i with N_i, and impose the following restriction on the network:

Assumption (C): The communication graph G is strongly connected; that is, there exists a directed path from any vertex to any other vertex in G.

Intuitively, Assumption (C) guarantees the possibility of information flow between any two agents (either directly or indirectly) in the network. The next sections will highlight the role played by this assumption in guaranteeing consistency and asymptotic normality of agents' estimates.

5.2.3 Belief Dynamics and Estimates

In order to aggregate the information provided to them over time – either through observations or communication with neighbors – agents hold and update beliefs over the parameter space Θ. More specifically, we denote the belief of agent i at time t with $\mu_{i,t} : \Theta \longrightarrow \mathbb{R}^+$, a probability measure over Θ. As for the dynamics, we assume that each agent updates its belief function as a geometric mean of its

neighbors' beliefs and its own observation likelihood function; or equivalently, the log-posterior beliefs of each agent is a linear combination of its neighbors' log-beliefs and its log-likelihood function:

$$\nu_{i,t+1}(\theta) = \lambda_i \log \ell_i(s^i_{t+1}|\theta) + \sum_{j \in N_i \cup \{i\}} w_{ij}\nu_{j,t}(\theta) + c_{i,t} \tag{5.2}$$

where $\nu_{i,t}(\theta) \triangleq \log \mu_{i,t}(\theta)$ is the logarithm of the belief function, $\lambda_i > 0$ is the weight that agent i assigns to its private observations, $w_{ij} > 0$ is the weight assigned to the beliefs of agent j in its neighborhood, and $c_{i,t}$ is a normalization constant which ensures that $\mu_{i,t+1}(\theta)$ is a well-defined probability density over Θ. Note that constants $c_{i,t}$ do not depend on the parameter θ. Throughout the paper, we assume that $\lambda_i = \lambda$, and $\sum_{j \in N_i \cup \{i\}} w_{ij} = 1$, for all $i \in N$.

Given its beliefs at any given time period, agent i's estimate of the true parameter is defined as a maximizer of its belief function; that is,[3]

$$\hat{\theta}_{i,t} \in \arg\max_{\theta \in \Theta} \nu_{i,t}(\theta). \tag{5.3}$$

Note that $\hat{\theta}_{i,t}$ is a random variable that depends on the data observed by agents up to time t. In the next section we show that this point estimator always exists and is a measurable function of the data. Moreover, note that due to the identification problem faced by each agent, the maximizer is not necessarily unique at all times. In that case, $\hat{\theta}_{i,t}$ can correspond to any solution of (5.3).

In order to simplify notation, we write update (5.2) in matrix form as

$$\nu_{t+1}(\theta) = W\nu_t(\theta) + \lambda\psi_\theta(s_{t+1}) + c_t \qquad \forall \theta \in \Theta$$

where $W = [w_{ij}]$ is a stochastic matrix with $w_{ij} = 0$ if $j \notin N_i \cup \{i\}$, and c_t is a vector of constants independent of θ. Thus, at any time t, we have

$$\nu_t(\theta) = W^t\nu_0(\theta) + \lambda \sum_{\tau=1}^{t} W^{t-\tau}\psi_\theta(s_\tau) + c'_t,$$

[3]Given the fact that log is a monotone function, defining the estimate as the mode of the log-belief function is equivalent to defining it as the maximizer of the belief function itself.

where c'_t is a vector that depends on past observations of all agents, but not θ. Finally, we define

$$\Phi_{i,t}(\theta) \triangleq \frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}[W^{t-\tau}]_{ij}\psi_\theta^j(s_\tau^j)$$

which is a function of agents' observations as well as the parameter. Therefore,

$$\nu_{i,t}(\theta) = \lambda t\Phi_{i,t}(\theta) + \sum_{j=1}^{n} W_{ij}^t \nu_{j,0}(\theta) + c'_{i,t} \tag{5.4}$$

where the second term only depends on the priors and the last term is a constant not depending on θ. This immediately implies that for large enough t, the point estimator $\hat{\theta}_{i,t}$ coincides with the maximizer of $\Phi_{i,t}(\theta)$ over Θ.

5.3 Consistency

In this section, we prove that under relatively mild assumptions, all agents' estimates of the true parameter are asymptotically consistent in probability; that is, $\hat{\theta}_{i,t} \xrightarrow{p} \theta^*$ for all i as $t \to \infty$. Before presenting our results on consistency, we state a few lemmas. The proofs can be found in the Appendix.

Our first lemma establishes that the point estimator of each agent is well-defined.

Lemma 10. *Suppose that $\theta^* \in \operatorname{int}\Theta$. Then, there exists a measurable function of the data $\hat{\theta}_{i,t}$ that solves (5.3).*

The next lemma shows that the beliefs of all agents converge asymptotically to a limit independent of their priors.

Lemma 11. *Suppose that Assumption (C) holds. Then,*

$$\Phi_{i,t}(\theta) \xrightarrow{p} \Phi_\infty(\theta) \triangleq \sum_{j=1}^{n} z_j \mathbb{E}[\log \ell_j(s_1^j|\theta)] \tag{5.5}$$

for all $\theta \in \Theta$, where $z = [z_i]$ is the stationary distribution of a Markov chain with W as its probability transition matrix.

Note that under Assumption (C), matrix W corresponds to an aperiodic and irreducible Markov chain, and therefore, has a unique stationary distribution z, with all elements strictly positive. Moreover, the limiting normalized log-posterior belief function $\Phi_\infty(\theta)$ is independent of i for all values of θ, and as a result, for large enough t, the beliefs of all agents get arbitrarily close. This implies that, as observations accumulate, the agents' estimates get closer to one another.

The next lemma establishes that the limiting log-posterior belief function $\Phi_\infty(\theta)$ is uniquely maximized at the true parameter θ^*, if the truth is globally identifiable and the network of agents is strongly connected.

Lemma 12. *Suppose that Assumptions (C) and (GI) hold. Then,*

$$\arg\max_{\theta \in \Theta} \ \Phi_\infty(\theta) = \{\theta^*\},$$

where $\Phi_\infty(\theta)$ is defined in (5.5).

Both Assumptions (C) and (GI) are required for the above lemma to hold. Clearly, in the presence of a global identification problem in the network, there exists a $\theta \neq \theta^*$ for which $\Phi_\infty(\theta) = \Phi_\infty(\theta^*)$ on almost all sample paths, and therefore, the limiting log-posterior belief function is not uniquely maximized. On the other hand, a network which is not strongly connected corresponds to a random walk with some transient states which implies that vector z will have at least one element, say z_k, equal to zero. As a result, the identification problem of agent k persists and leads to a non-unique solution to the maximization problem.

We now present the main result of this section.

Theorem 13. *Suppose that $\theta^* \in \operatorname{int}\Theta$ and that Assumptions (C) and (GI) hold. Then, the point estimators of all agents are weakly consistent; that is*

$$\hat{\theta}_{i,t} \xrightarrow{p} \theta^* \qquad \forall i.$$

Proof. First, note that for large enough t, the estimate $\hat{\theta}_{i,t}$ coincides with the maximizer of $\Phi_{i,t}(\theta)$ over Θ. On the other hand, by Lemma 11, the convex function $\Phi_{i,t}(\theta)$ converges to $\Phi_\infty(\theta)$ in probability for all θ. As established by Lemma 12, $\Phi_\infty(\theta)$ is uniquely maximized at θ^*, and therefore, by Theorem 2.7 of Newey and McFadden (Newey and McFadden, 1994), the maximizer of $\Phi_{i,t}(\theta)$ converges in probability to θ^* for all $i \in N$. Thus, the estimator of ever agent is weakly consistent. □

Theorem 13 establishes that as the number of observations grows, the estimate of each agent converges to the parameter corresponding to the true data generating process. The importance of this result lies in the fact that asymptotic consistency is achieved despite the fact that all agents face some identification problem – in the sense that no agent can consistently estimate the true parameter in isolation. However, if agents have access to the information held by their neighbors and the communication graph is strongly connected, then information is properly aggregated over the network, and the estimate of every agent converges to the true parameter.

The other notable fact about Theorem 13 is that consistency is achieved regardless of the network's structure. More specifically, as long as the network is strongly connected, its topology and the weights w_{ij} assigned by the agents to their neighbors do not affect convergence of the estimates to the truth. However, in the next sections, we show that the network structure determines the efficiency of the distributed estimators.

5.4 Asymptotic Normality

In this section, we prove that the agents' estimates are asymptotically normally distributed and characterize their asymptotic covariance matrices.

We start by stating two auxiliary lemmas, which are proved in the Appendix. Lemma 14 is simply a weak law of large numbers for the Hessian of the log-likelihood of the observations, whereas Lemma 15 is a central limit theorem for the gradients.

Lemma 14. *Suppose that* $\{\bar{\theta}_{i,t}\}_{i\in N}$ *are consistent estimators of* θ^**, and suppose Assumption (C) holds. Then,*

$$-\nabla_{\theta\theta}\Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*) \qquad \forall i.$$

Lemma 15. *Suppose that Assumption (C) holds. Then, for all* $i \in N$

$$\sqrt{t}\nabla_\theta\, \Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, \sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*)\big).$$

We are now ready to state and prove the main result of this section.

Theorem 16. *Suppose that Assumptions (C) and (GI) hold. Then,*

$$\sqrt{t}(\hat{\theta}_{i,t} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathrm{Avar}) \tag{5.6}$$

where the asymptotic covariance matrix is given by

$$\mathrm{Avar} = \left[\sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*)\right]^{-1} \sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*) \left[\sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*)\right]^{-1}. \tag{5.7}$$

Proof. By definition, $\hat{\theta}_{i,t}$ is a maximizer of $\Phi_{i,t}(\theta)$, and therefore, it must be the case that $\nabla_\theta\, \Phi_{i,t}(\hat{\theta}_{i,t}) = 0$. On the other hand, by the mean value theorem, we have

$$\nabla_\theta\, \Phi_{i,t}(\hat{\theta}_{i,t}) = \nabla_\theta\, \Phi_{i,t}(\theta^*) + \nabla_{\theta\theta}\, \Phi_{i,t}(\bar{\theta}_{i,t})(\hat{\theta}_{i,t} - \theta^*),$$

where $\bar{\theta}_{i,t}$ is a mean value between θ^* and $\hat{\theta}_{i,t}$. Thus, we can solve for $(\hat{\theta}_{i,t} - \theta^*)$ and get

$$\sqrt{t}(\hat{\theta}_{i,t} - \theta^*) = -\sqrt{t}\left[\nabla_{\theta\theta}\,\Phi_{i,t}(\bar{\theta}_{i,t})\right]^{-1}\nabla_{\theta}\,\Phi_{i,t}(\theta^*).$$

Since $\bar{\theta}_{i,t}$ lies between θ^* and $\hat{\theta}_{i,t}$, it is a consistent estimator for θ^*,[4] and therefore, Lemma 14 implies that $\nabla_{\theta\theta}\,\Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} -\sum_{j=1}^{n} z_j^2\mathcal{I}_j(\theta^*)$. Note that the global identifiability assumption guarantees that $\sum_j z_j^2\mathcal{I}_j(\theta^*)$ is non-singular. On the other hand, Lemma 15 guarantees that $\sqrt{t}\nabla_{\theta}\,\Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}(0, \sum_{j=1}^{n} z_j\mathcal{I}_j(\theta^*))$. At this point, the theorem trivially follows by Slutsky's theorem.[5] □

Theorem 16 states that the agents' estimates are normally distributed as the sample size grows. As the proof suggests, the key idea behind asymptotic normality is that in large samples, estimators are approximately equal to linear combinations of sample averages (a consequence of applying the mean value theorem), so that the central limit theorem can be applied (Newey and McFadden, 1994). The theorem also states that distributed estimators, like the centralized maximum likelihood estimator, are $\sqrt{t}$-consistent. Finally, expression (5.7) provides the asymptotic covariance matrix of the estimates in terms of the network structure and information matrices corresponding to agents' observation models.

5.5 Estimator Efficiency and Network Topology

In the previous section, we derived asymptotic variance of the distributed estimators. In this section, we investigate their efficiency in terms of the network structure, as well as the observation model of each agent. Our next theorem compares the

[4]Note that in Theorem 13 we established that $\hat{\theta}_{i,t}$ is consistent.

[5]Slutsky's theorem states that if $x_t \xrightarrow{d} x$ and $y_t \xrightarrow{p} c$ where c is a constant, then, $x_t y_t \xrightarrow{d} cY$.

distributed estimator with a centralized estimator, and provides a bound for its performance.

Theorem 17. *Suppose that Assumptions (GI) and (C) hold. Then, asymptotic variance of the distributed estimator satisfies*

$$\text{Avar} \succeq [\mathcal{I}_c(\theta^*)]^{-1} \tag{5.8}$$

where $\mathcal{I}_c(\theta)$ denotes the Fisher information matrix of a centralized estimator with access to the observations of all agents. Moreover, the above bound is tight if W is doubly stochastic.

Before presenting the proof, a few remarks are in order. First note that $[\mathcal{I}_c(\theta^*)]^{-1}$ is equal to asymptotic variance of the maximum-likelihood estimator of a centralized entity with access to the measurements of all agents. In other words, equation (5.8) simply means that the distributed estimators are never more efficient (in the Cramér-Rao sense) than a centralized maximum likelihood estimator. This is not surprising, as one expects that decentralization can never lead to a more efficient estimation.

The second part of the theorem, however, is more striking. It basically states if the weight matrix W is doubly stochastic, then the distributed estimator is as efficient as any centralized estimator. For example, if all communication links are bidirectional and the weights that each pair of agents assign to one another are equal (i.e., $w_{ij} = w_{ji}$), then decentralization does not sacrifice efficiency, regardless of how sparse the network is.

Proof of Theorem 17. We first compute $\mathcal{I}_c(\theta)$ in terms of the Fisher information matrices corresponding to agents' observation models. By independence of observations across agents, we have

$$\ell(s_t|\theta) = \ell_1(s_t^1|\theta)\ell_2(s_t^2|\theta)\cdots\ell_n(s_t^n|\theta),$$

which implies

$$\begin{aligned}\mathcal{I}_c(\theta^*) &= \mathbb{E}\left[\sum_{j=1}^{n}\nabla_\theta\psi^j_{\theta^*}(s^j_1)\sum_{i=1}^{n}\nabla'_\theta\psi^i_{\theta^*}(s^i_1)\right]\\ &= \sum_{j=1}^{n}\mathbb{E}\left[\nabla_\theta\psi^j_{\theta^*}(s^j_1)\nabla_\theta\psi^j_{\theta^*}(s^j_1)'\right]\\ &= \sum_{j=1}^{n}\mathcal{I}_j(\theta^*),\end{aligned}$$

where we have used the fact that $\mathbb{E}\left[\nabla_\theta \log \ell_i(s^i_1|\theta^*)\right] = 0$ (see proof of Lemma 15). Therefore, in order to prove (5.8), we need to show that

$$Q = \sum_{j=1}^{n}\mathcal{I}_j(\theta^*) - \sum_{j=1}^{n} z_j\mathcal{I}_j(\theta^*)\left[\sum_{j=1}^{n} z_j^2\mathcal{I}_j(\theta^*)\right]^{-1}\sum_{j=1}^{n}\mathcal{I}_j(\theta^*)$$

is positive semi-definite. Note that Q is the Schur complement of

$$X = \begin{bmatrix}\sum_j z_j^2\mathcal{I}_j(\theta^*) & \sum_j z_j\mathcal{I}_j(\theta^*)\\ \\ \sum_j z_j\mathcal{I}_j(\theta^*) & \sum_j \mathcal{I}_j(\theta^*)\end{bmatrix}$$

which can be easily verified to be positive semi-definite.[6] Thus, Q is also positive semi-definite, which proves the first part of the theorem.[7]

To prove the second part, we use the fact that if W is doubly stochastic, then its corresponding Markov chain has a uniform stationary distribution, that is, $z_i = \frac{1}{n}$. Therefore, expression (5.7) reduces to

$$\text{Avar} = \left[\sum_{j=1}^{n}\mathcal{I}_j(\theta^*)\right]^{-1} = \left[\mathcal{I}_c(\theta^*)\right]^{-1}$$

which is the asymptotic covariance matrix of the centralized maximum likelihood estimator. This proves that the bound is tight. ■ □

[6]Note that $u'Xu = \sum_j (z_j u'_1 + u'_2)\mathcal{I}_j(\theta^*)(z_j u_1 + u_2) \geq 0$ for all $u' = [u'_1\ u'_2]$.

[7]For more on Schur complement and its properties, see for example, Boyd and Vandenberghe (Boyd and Vandenberghe, 2004), page 650.

As a final remark, we emphasize that although sufficient, double stochasticity of W is not necessary for efficiency of the distributed estimator. For example, it is possible to achieve efficiency by assigning a zero weight on an agent whose signals are non-informative, and have the rest of the weights equally shared among the rest of the agents. A complete characterization of efficiency conditions is part of our ongoing research.

5.6 Conclusions

In this paper, we studied a model of distributed estimation over a network, where each agent faces a local identification problem – in the sense that it cannot consistently estimate a parameter of interest in isolation. The agents engage in communication with their neighbors in order to resolve their identification problems. We showed that as long as the true parameter is globally identifiable (i.e., there is enough information across the network for it to be uniquely identified) and the communication network is strongly connected (i.e., there exists a direct or indirect information path connecting any two agents), then all agents can consistently estimate the true parameter as observations accumulate. Moreover, we proved that under some regularity assumptions on the observation models, the agents' estimates are asymptotically normally distributed. Finally, we computed the asymptotic variance of the distributed estimators, and showed that in bidirectional networks, the agents' estimators are as efficient as any centralized estimator, regardless of the sparsity of the network.

Appendix: Omitted Proofs

Proof of Lemma 10. The proof is along the lines of the proof of Lemma 7.1 in Hayashi (Hayashi,), and therefore, is omitted. ■

□

Proof of Lemma 11. We first show that variance of $\Phi_{i,t}(\theta)$ converges to zero, for all i and θ:

$$\begin{aligned}\mathrm{var}[\Phi_{i,t}(\theta)] &= \frac{1}{t^2}\sum_{\tau=1}^{t}\sum_{j=1}^{n}[W_{ij}^{t-\tau}]^2\mathrm{var}[\psi_\theta^j(s_1^j)] \\ &\leq \frac{1}{t}\sum_{j=1}^{n}\mathrm{var}[\psi_\theta^j(s_1^j)] \longrightarrow 0,\end{aligned}$$

and therefore, $\Phi_{i,t}(\theta) - \mathbb{E}[\Phi_{i,t}(\theta)] \xrightarrow{p} 0$. On the other hand, we have

$$\begin{aligned}\mathbb{E}[\Phi_{i,t}(\theta)] &= \sum_{j=1}^{n}\left[\frac{1}{t}\sum_{\tau=1}^{t}W^{t-\tau}\right]_{ij}\mathbb{E}[\psi_\theta^j(s_1^j)] \\ &\longrightarrow \sum_{j=1}^{n}[\mathbf{1}z']_{ij}\mathbb{E}[\psi_\theta^j(s_1^j)] \\ &= \sum_{j=1}^{n}z_j\mathbb{E}[\psi_\theta^j(s_1^j)],\end{aligned}$$

where we used the fact that W corresponds to an aperiodic and irreducible Markov chain with the unique stationary distribution z (guaranteed by Assumption (C)), and that Cesàro means preserve convergent sequences and their limits. Thus, we have

$$\Phi_{i,t}(\theta) \xrightarrow{p} \sum_{j=1}^{n}z_j\mathbb{E}[\psi_\theta^j(s_1^j)]$$

for all $i \in N$ and all $\theta \in \Theta$, which completes the proof. ■

□

Proof of Lemma 12. By Jensen's inequality,

$$\mathbb{E}\left[\log \frac{\ell_j(s_1^j|\theta)}{\ell_j(s_1^j|\theta^*)}\right] \leq \log \mathbb{E}\left[\frac{\ell_j(s_1^j|\theta)}{\ell_j(s_1^j|\theta^*)}\right] = 0,$$

implying

$$\mathbb{E}[\log \ell_j(s_1^j|\theta)] \leq \mathbb{E}[\log \ell_j(s_1^j|\theta^*)]$$

with equality holding if and only if $\theta \in \bar{\Theta}_j$. Therefore, the set of maximizers of $\mathbb{E}[\log \ell_j(s_1^j|\theta)]$ coincides with the set of parameters that are observationally equivalent to θ^*. Thus, by Assumption (GI), θ^* is the unique maximizer of their weighted sum. Notice that once again we are using the fact that all elements of vector z are strictly positive. ■

□

Proof of Lemma 14. First, notice that by a simple weak law of large numbers argument, $\nabla_{\theta\theta}\Phi_{i,t}(\theta) - \mathbb{E}\nabla_{\theta\theta}\Phi_{i,t}(\theta)$ converges to zero in probability, pointwise for all $\theta \in \Theta$. Moreover, we have

$$\mathbb{E}\nabla_{\theta\theta}\Phi_{i,t}(\theta) \longrightarrow \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)]$$

for all θ, where once again we have used Assumption (C) and the convergence of Cesàro means. Therefore,

$$\nabla_{\theta\theta}\Phi_{i,t}(\theta) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)] \qquad \forall \theta \in \Theta.$$

Now Corollary 2.2 of Newey (Newey, 1991) implies that under Assumptions (A1)–(A5), $\nabla_{\theta\theta}\Phi_{i,t}(\theta)$ converges uniformly in probability to $\sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)]$, and therefore, by Theorem 4.1.5 of Amemiya (Amemiya, 1985), for any consistent estimator $\bar{\theta}_{i,t} \xrightarrow{p} \theta^*$, we have

$$\nabla_{\theta\theta}\Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_{\theta^*}^j(s_1^j)].$$

Finally, the information matrix equality implies that

$$\mathbb{E}[\nabla_{\theta\theta}\psi^j_{\theta^*}(s^j_1)] = -\mathbb{E}\left[\nabla_\theta\,\psi^j_{\theta^*}(s^j_1)\nabla_\theta\,\psi^j_{\theta^*}(s^j_1)'\right]$$

which is equal to $-\mathcal{I}_j(\theta^*)$, by definition. This completes the proof. ■

□

Proof of Lemma 15. The proof of this lemma relies on the multivariate extension of the Lindeberg-Feller central limit theorem, which can be found in van der Vaart (van der Vaart, 1998), Proposition 2.27. But first, notice that by Lemma 3.6 of Newey and McFadden (Newey and McFadden, 1994), we have

$$\mathbb{E}\left[\nabla_\theta \log \ell_i(s^i_1|\theta^*)\right] = 0,$$

implying that $\mathbb{E}\nabla_\theta\,\Phi_{i,t}(\theta^*) = 0$.

In order to apply the Lindeberg-Feller CLT, we need to show that the Lindeberg condition is satisfied; that is

$$\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}(W^{t-\tau})^2_{ij}\mathbb{E}\left[\|\nabla_\theta\,\psi^j_{\theta^*}\|^2\mathbb{I}_{\left\{W^{t-\tau}_{ij}\|\nabla_\theta\,\psi^j_{\theta^*}\|>\epsilon\sqrt{t}\right\}}\right] \to 0$$

for all $\epsilon > 0$, as $t \to \infty$, where $\mathbb{I}$ denotes the indicator function, and for notational simplicity, we have dropped the dependence of $\nabla_\theta\,\psi^j_{\theta^*}$ on the observations s^j. Verifying that the Lindeberg condition is straightforward: the left hand-side is bounded above by expression

$$\max_{1\le j\le n}\mathbb{E}\left[\|\nabla_\theta\,\psi^j_{\theta^*}\|^2\mathbb{I}_{\left\{\|\nabla_\theta\,\psi^j_{\theta^*}\|>\epsilon\sqrt{t}\right\}}\right]$$

which converges to zero for all $\epsilon > 0$ as $t \to \infty$. Thus, by the Lindeberg-Feller CLT, $\sqrt{t}\Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}(0, S)$, where S is given by

$$\begin{aligned} S &= \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}(W^{t-\tau})^2_{ij}\mathbb{E}\left[\nabla_\theta\,\psi^j_{\theta^*}(s^j_1)\nabla_\theta\,\psi^j_{\theta^*}(s^j_1)'\right] \\ &= \sum_{j=1}^{n} z_j^2\mathcal{I}_j(\theta^*) \end{aligned}$$

where we have used the fact that $W^t \longrightarrow \mathbf{1}z'$, and the definition of the Fisher information matrix in (5.1). ■

□

Bibliography

Aguera y Arcas, B. and Fairhall, A. (2003). What causes a neuron to spike? *Neural Computation*, 15:1789–1807.

Ahmadian, Y., Pillow, J., and Paninski, L. (2009). Efficient Markov Chain Monte Carlo methods for decoding population spike trains. *Under review, Neural Computation*.

Akcakaya, M. and Tarokh, V. (2008). Noisy compressive sampling limits in linear and sublinear regimes. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 1–4.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Asif, A. and Moura, J. (2005). Block matrices with l-block banded inverse: Inversion algorithms. *IEEE Transactions on Signal Processing*, 53:630–642.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, pages 213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, pages 217–234.

Barlow, H. B., Kaushal, T. P., and Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural Computation*, 1:412–423.

Behseta, S., Kass, R., and Wallstrom, G. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, 92:419–434.

Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S., and Bethge, M. (2011). Reassessing optimal neural population codes with neurometric functions. *Proceedings of the National Academy of Sciences*, 108:4423–4428.

Berry, M. and Meister, M. (1998). Refractoriness and neural precision. *J. Neurosci.*, 18:2200–2211.

Bialek, W., Callan, C., and Strong, S. (1996). Field theories for learning probability distributions. *Physical Review Letters*, 77:4693–4697.

Bialek, W. and Zee, A. (1990). Coding and computation with neural spike trains. *Journal of Statistical Physics*, 59:103–115.

Bjorstad, P. and Mandel, J. (1991). On the spectra of sums of orthogonal projections with applications to parallel computing. *BIT Numerical Mathematics*, 31:76–88.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Oxford University Press.

Brockwell, A., Rojas, A., and Kass, R. (2004). Recursive Bayesian decoding of motor cortical signals by particle filtering. *Journal of Neurophysiology*, 91:1899–1907.

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

Brown, E., Frank, L., Tang, D., Quirk, M., and Wilson, M. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18:7411–7425.

Brown, E., Nguyen, D., Frank, L., Wilson, M., and Solo, V. (2001). An analysis of neural receptive field plasticity by point process adaptive filtering. *PNAS*, 98:12261–12266.

Brown, L., Cai, T., Zhang, R., Zhao, L., and Zhou, H. (2009a). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields. To appear.*

Brown, L., Cai, T., and Zhou, H. (2009b). Nonparametric regression in exponential familie. *Annals of Statistics. To appear.*

Brunel, N. and Nadal, J.-P. (1998). Mutual information, fisher information, and population coding. *Neural Comput.*, 10(7):1731–1757.

Bullo, F., Cortes, J., and Martinez, S. (2009). *Distributed Control of Robotic Networks*. Princeton University Press.

Clarke, B. and Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453 – 471.

Coleman, T. and Sarma, S. (2007). A computationally efficient method for modeling neural spiking activity with point processes nonparametrically. *IEEE Conference on Decision and Control*.

Cover, T. and Thomas, J. (1991). *Elements of information theory*. Wiley, New York.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal Of The Royal Statistical Society Series B*, 70(1):209–226.

Cunningham, J., Yu, B., Shenoy, K., and Sahani, M. (2007). Inferring neural firing rates from spike trains using Gaussian processes. *NIPS*.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008). Fast Gaussian process methods for point process intensity estimation. *ICML*, pages 192–199.

Czanner, G., Eden, U., Wirth, S., Yanike, M., Suzuki, W., and Brown, E. (2008). Analysis of between-trial and within-trial neural spiking dynamics. *Journal of Neurophysiology*, 99:2672–2693.

Davis, T. (2006). *Direct Methods for Sparse Linear Systems*. SIAM.

DeGroot, M. (1974). Reaching a Consensus. *Journal of American Statistical Association*, 69(345):118–121.

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

di Bernardo, D., Thompson, M. J., Gardner, T., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., Schaus, S., and Collins, J. J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotech*, 23(3):377–383.

DiMatteo, I., Genovese, C., and Kass, R. (2001). Bayesian curve fitting with free-knot splines. *Biometrika*, 88:1055–1073.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Eden, U. T., Frank, L. M., Barbieri, R., Solo, V., and Brown, E. N. (2004). Dynamic analyses of neural encoding by point process adaptive filtering. *Neural Computation*, 16:971–998.

Fahrmeir, L. and Kaufmann, H. (1991). On Kalman filtering, posterior mode estimation and fisher scoring in dynamic exponential family regression. *Metrika*, 38:37–60.

Fletcher, A., Rangan, S., and Goyal, V. (2008). Necessary and sufficient conditions on sparsity pattern recovery. *CoRR*, abs/0804.1839.

Frank, L., Eden, U., Solo, V., Wilson, M., and Brown, E. (2002). Contrasting patterns of receptive field plasticity in the hippocampus and the entorhinal cortex: An adaptive filtering approach. *J. Neurosci.*, 22(9):3817–3830.

Gao, Y., Black, M., Bienenstock, E., Shoham, S., and Donoghue, J. (2002). Probabilistic inference of arm motion from neural activity in motor cortex. *NIPS*, 14:221–228.

Geffen, M. N., Broome, B. M., Laurent, G., and Meister, M. (2009). Neural encoding of rapidly fluctuating odors. *Neuron*, 61:570–586.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. CRC Press.

Ginzburg, I. and Sompolinsky, H. (1994). Theory of correlations in stochastic neural networks. *Phys Rev E*, 50(4):3171–3191.

Golub, B. and Jackson, M. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149.

Good, I. and Gaskins, R. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. CRC Press.

Hafting, T., Fyhn, M., Molden, S., Moser, M., and Moser, E. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436:801–806.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hayashi, F. *Econometrics*. Princeton University Press.

Holy, T. (1997). The analysis of data from continuous probability distributions. *Physical Review Letters*, 79:3545–3548.

Huggins, J. and Paninski, L. (2011). Optimal experimental design for sampling voltage on dendritic trees in the low-snr regime. *Under review*.

Ioannides, M. and Loury, L. (2004). Job information networks, neighborhood effects, and inequality. *The Journal of Economic Literature*, 42(2):1056–1093.

Jackson, B. (2004). Including long-range dependence in integrate-and-fire models of the high interspike-interval variability of cortical neurons. *Neural Computation*, 16:2125–2195.

Jadbabaie, A., Lin, J., and Morse, A. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001.

Jadbabaie, A., Sandroni, A., and Tahbaz-Salehi, A. (2010). Non-bayesian social learning. *Submitted*.

Kang, K. and Sompolinsky, H. (2001). Mutual information of population codes and distance measures in probability space. *Phys. Rev. Lett.*, 86:4958–4961.

Kar, S., Moura, J., and Ramanan, K. (2008). Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication. *Unpublished manuscript*.

Karbasi, A., Hormati, A., Mohajer, S., and Vetterli, M. (2009). Support recovery in compressed sensing: An estimation theoretic approach. In *2009 IEEE International Symposium on Information Theory*.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kass, R. and Ventura, V. (2001). A spike-train probability model. *Neural Comp.*, 13:1713–1720.

Kass, R., Ventura, V., and Cai, C. (2003). Statistical smoothing of neuronal data. *Network: Computation in Neural Systems*, 14:5–15.

Kass, R. E., Ventura, V., and Brown, E. N. (2005). Statistical issues in the analysis of neuronal data. *J Neurophysiol*, 94:8–25.

Kotler, P. (1986). *The Principles of Marketing*. Prentice-Hall Inc., Englewood Cliffs, NJ.

Koyama, S., Bolde, L., Shalizi, C., and Kass, R. (2010). Approximate methods for state-space models. *Journal of ASA*, 105(489):170–180.

Koyama, S. and Paninski, L. (2009). Efficient computation of the maximum a posteriori path and parameter estimation in integrate-and-fire and more general state-space models. *Journal of Computational Neuroscience*, In press.

Lawhern, V., Wu, W., Hastopoulos, N., and Paninski, L. (2011). Population decoding of motor cortical activity using a generalized linear model with hidden states. *Journal of Neuroscience Methods*.

Lovasz, L. and Vempala, S. (2003). The geometry of logconcave functions and an $O^*(n^3)$ sampling algorithm. Technical Report 2003-04, Microsoft Research.

Macke, J., Sing, L., Cunningham, J.P. snd Yu, B., Shenoy, K., and Sahani, M. (2011). Modelling low-dimensional dynamics in recorded spiking populations. *COSYNE*.

Macke, J. H., Gerwinn, S., Kaschube, M., White, L. E., and Bethge, M. (2010). Bayesian estimation of orientation preference maps. *Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference (NIPS 2009)*, pages 1195–1203.

Moeller, J., Syversveen, A., and Waagepetersen, R. (1998). Log-Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25:451–482.

Moeller, J. and Waagepetersen, R. (2004). *Statistical inference and simulation for spatial point processes*. Chapman Hall.

Newey, W. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167.

Newey, W. and McFadden, D. (1994). *Large sample estimation and hypothesis Testing*, volume 4, pages 2111–2245. Elsevier.

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

Paninski, L. (2005). Log-concavity results on Gaussian process methods for supervised and unsupervised learning. *Advances in Neural Information Processing Systems*, 17.

Paninski, L., Ahmadian, Y., Ferreira, D., Koyama, S., Rahnama Rad, K., Vidne, M., Vogelstein, J., and Wu, W. (2010). A new look at state-space models for neural data. *Journal of Computational Neuroscience*, 29(1):107–126.

Paninski, L., Fellows, M., Hatsopoulos, N., and Donoghue, J. (2004a). Spatiotemporal tuning properties for hand position and velocity in motor cortical neurons. *Journal of Neurophysiology*, 91:515–532.

Paninski, L., Fellows, M., Shoham, S., Hatsopoulos, N., and Donoghue, J. (2004b). Superlinear population encoding of dynamic hand trajectory in primary motor cortex. *J. Neurosci.*, 24:8551–8561.

Paninski, L., Pillow, J., and Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. In Cisek, P., Drew, T., and Kalaska, J., editors, *Computational Neuroscience: Progress in Brain Research*. Elsevier.

Panzeri, S., Schultz, S., Treves, A., and Rolls, E. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society London B*, 266(1423):1001–1012.

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

Pillow, J., Shlens, J., Paninski, L., Sher, A., Litke, A., Chichilnisky, E., and Simoncelli, E. (2008). Spatiotemporal correlations and visual signaling in a complete neuronal population. *Nature*, 454:995–999.

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

Rahnama Rad, K. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Information Theory*, 57(7).

Rahnama Rad, K. and Paninski, L. (2010). Efficient, adaptive estimation of two-dimensional firing rate surfaces via gaussian process methods. *Network: Computation in Neural Systems*, 21:142–168.

Rahnama Rad, K. and Paninski, L. (2011). Information rates and optimal decodding in large neural populations. *Submitted to NIPS*.

Rahnama Rad, K. and Tahbaz-Salehi, A. (2010). Distributed parameter estimation in networks. In *49th IEEE Conference on Decision and Control*, pages 5050–5055.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rauhut, H., Schnass, K., and Vandergheynst, P. (2008). Compressed sensing and redundant dictionaries. *IEEE Transactions Information Theory*, 54(5):2210 – 2219.

Reeves, G. and Gastpar, M. (2008). Sampling bounds for sparse support recovery in the presence of noise. In *IEEE International Symposium on Information Theory*, pages 2187 – 2191.

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

Rokni, U., Richardson, A., Bizzi, E., and Seung, S. (2007). Motor learning with unstable neural representations. *Neuron*, 54:653–666.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345.

Rust, N., Schwartz, O., Movshon, A., and Simoncelli, E. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, 46:945–956.

Sahani, M. (1999). *Latent variable models for neural data analysis*. PhD thesis, California Institute of Technology.

Salinas, E. and Abbott, L. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, 1:89–107.

Sanches, J., Nascimento, J., and Marques, J. (2008). Medical image noise reduction using the sylvester-lyapunov equation. *Image Processing, IEEE Transactions on*, 17:1522–1539.

Schmidt, D. M. (2000). Continuous probability distributions from finite data. *Phys. Rev. E*, 61(2):1052–1055.

Seung, H. S. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90:10749–10753.

Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge University Press.

Shumway, R. and Stoffer, D. (2006). *Time Series Analysis and Its Applications*. Springer.

Smith, A. and Brown, E. (2003). Estimating a state-space model from point process observations. *Neural Computation*, 15:965–991.

Snippe, H. (1996). Parameter extraction from population codes: A critical assesment. *Neural Computation*, 8:511–529.

Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer-Verlag.

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

Thorburn, D. (1986). A Bayesian approach to density estimation. *Biometrika*, 73:65–75.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Toyoizumi, T., Rahnama Rad, K., and Paninski, L. (2009). Mean-field approximations for coupled populations of generalized linear model spiking neurons with Markov refractoriness. *Neural Computation*, 21:1203–1243.

Truccolo, W., Eden, U., Fellows, M., Donoghue, J., and Brown, E. (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.

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

Vidne, M., Kulkarni, J., Ahmadian, Y., Pillow, J., Shlens, J., Chichilnisky, E., Simoncelli, E., and Paninski, L. (2009). Inferring functional connectivity in an ensemble of retinal ganglion cells sharing a common input. *COSYNE*.

Vinje, W. and Gallant, J. (2000). Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science*, 287(5456):1273–1276.

Vogelstein, J., Babadi, B., Watson, B., Yuste, R., and Paninski, L. (2008). Fast nonnegative deconvolution via tridiagonal interior-point methods, applied to calcium fluorescence data. *Statistical analysis of neural data (SAND) conference*.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wainwright, M. (2007). Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, pages 961–965.

Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202.

Wang, W., Wainwright, M., and Ramchandran, K. (2008). Information-theoretic limits on sparse support recovery: Dense versus sparse measurements. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2197–2201.

Wu, W., Black, M. J., Mumford, D., Gao, Y., Bienenstock, E., and Donoghue, J. (2004). Modeling and decoding motor cortical activity using a switching Kalman filter. *IEEE Transactions on Biomedical Engineering*, 51:933–942.

Wu, W., Gao, Y., Bienenstock, E., Donoghue, J. P., and Black, M. J. (2006). Bayesian population coding of motor cortical activity using a Kalman filter. *Neural Computation*, 18:80–118.

Xia, L., Boyd, S., and Lall, S. (2005). A scheme for robust distributed sensor fusion based on average consensus. pages 63–70, Los Angeles, CA.

Yu, B., Afshar, A., Santhanam, G., Ryu, S., Shenoy, K., and Sahani, M. (2006). Extracting dynamical structure embedded in neural activity. *NIPS*.

Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102:614–635.

Zibulevsky, M. and Pearlmutter, B. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13:863–882.