ABSTRACT

Title of dissertation: COMPRESSED SENSING BEYOND THE IID AND STATIC DOMAINS: THEORY, ALGORITHMS AND APPLICATIONS

Abbas Kazemipour
Doctor of Philosophy, 2017

Dissertation directed by: Professors Min Wu and Behtash Babadi
Department of Electrical and Computer Engineering

Sparsity is a ubiquitous feature of many real world signals such as natural images and neural spiking activities. Conventional compressed sensing utilizes sparsity to recover low dimensional signal structures in high ambient dimensions using few measurements, where i.i.d measurements are at disposal. However real world scenarios typically exhibit non i.i.d and dynamic structures and are confined by physical constraints, preventing applicability of the theoretical guarantees of compressed sensing and limiting its applications. In this thesis we develop new theory, algorithms and applications for non i.i.d and dynamic compressed sensing by considering such constraints.

In the first part of this thesis we derive new optimal sampling-complexity tradeoffs for two commonly used processes used to model dependent temporal structures: the autoregressive processes and self-exciting generalized linear models. Our theoretical results successfully recovered the temporal dependencies in neural activities, financial data and traffic data.

Next, we develop a new framework for studying temporal dynamics by introducing compressible state-space models, which simultaneously utilize spatial and temporal sparsity. We develop a fast algorithm for optimal inference on such models and prove its optimal recovery guarantees. Our algorithm shows significant improvement in detecting sparse events in biological applications such as spindle detection and calcium deconvolution.

Finally, we develop a sparse Poisson image reconstruction technique and the first compressive two-photon microscope which uses lines of excitation across the sample at multiple angles. We recovered diffraction-limited images from relatively few incoherently multiplexed measurements, at a rate of 1.5 billion voxels per second.

COMPRESSED SENSING BEYOND THE I.I.D. AND STATIC DOMAINS: THEORY, ALGORITHMS AND APPLICATIONS

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Min Wu, Chair/Advisor
Behtash Babadi, Co-Advisor
Shaul Druckmann
Radu Balan
Prakash Narayan
Johnathan Fritz

Preface

Sparsity, coherence and dynamics are among the ubiquitous features of many real world signals and systems. Examples include natural images, sounds and neural spiking activities. These signals exhibit sparsity, that is there exists a basis for which the effective dimension of the signal is much smaller than its ambient dimension. Moreover natural signals entail rich temporal dynamics. Theory of compressed sensing [40, 45, 49, 50, 71, 145] is concerned with reconstruction of such signals by utilizing their sparse structures, using very few measurements . Compressed sensing provides sharp trade-offs between the number of measurement, sparsity, and estimation accuracy when random i.i.d measurements are at disposal. However, in most practical applications of interest, the measured signals and the corresponding covariates are highly interdependent and follow specific temporal dynamics. Although, the theory of compressed sensing has not considered non-i.i.d and dynamic domains, the recovery algorithms suggested by it show remarkable performance once the structure of the underlying signal is taken into account.

On a high level, research in compressive sensing is conducted in three main branches: theory, algorithms and applications. In this thesis we revisit all three branches for nonlinear models with interdependent covariates as well as dynamic compressive sensing with applications in neural signal processing, traffic modeling, financial data and media forensics. Three main proposed problems, our ongoing research and future work are included in details in the following chapters. The rest of this thesis is organized as follows: In Chapter 2 we introduce the problem of stable

estimation of high-order AR processes, generalizing results from i.i.d compressive sensing to general class of stable AR processes with sub-Gaussian innovations. In doing so we will show that spectral properties of stationary processes determine the interdependence of the covariates and in general the sampling-complexity tradeoffs of AR processes. In Chapter 3 we introduce the problem of robust estimation of generalized linear models. We will prove theoretical guarantees of compressive sensing for such models, characterizing the sampling-complexity tradeoffs between the interdependence of the covariates and the number of measurement. We further corroborate our theoretical guarantees with simulated data as well as an application to retinal ganglion cell spiking activities. In Chapter 4 we provide theoretical guarantees of stable state estimation where the underlying dynamics follow a compressible state-space model. Moreover we show application of our results in denoising and spike deconvolution from calcium imaging recordings of neural spiking activities and show promising recovery results from compressed imaging data. We propose an application of this method in designing compressive calcium imaging devices. Finally, in Chapter 6 we use ideas from projection microscopy to develop a two-photon imaging technique that scans lines of excitation across the sample at multiple angles, recovering high-resolution images from relatively few incoherently multiplexed measurements. By using a static image of the sample as a prior, we recover neural activity from megapixel fields of view at rates exceeding 1kHz.

Acknowledgment

I feel very fortunate to have the chance to learn from and collaborate with incredibly talented individuals in the past 5 years. First and foremost, I would like to thank my academic advisor, Professor Min Wu, who supported and believed in me throughout my PhD. I feel extremely indebted to her, for her patience, critical thinking and giving me the opportunity to follow my research interests. Second, I would like to thank my co-advisor, Professor Behtash Babadi. His unique blend of creativity, generosity and energy makes him a role model for my future career. I would also like to thank my mentors at Janelia Research Campus. I would like to thank Professor Shaul Druckmann and Dr. Kaspar Podgorski for their incredible support of my work and great contributions to my thesis.

I would like to thank my committee members and group leaders at Janelia for their encouragements and my collaborators at University of Maryland and Janelia who provided critical feedback on my work. Many of the interesting research problems and ideas came from fruitful discussions with them. I would also like to thank Professor Ali Olfat at University of Tehran for his incredible support and believing in me so early in my career.

I would like to thank my friends at UMD and Janelia, who made the past 5 years a plausible experience for me.

Last but not least, I would like to thank my family, especially my parents who always chose my happiness over their own convenience. Thank you for filling my life with love.

Notations

Throughout this thesis we use bold lower and upper case letters for denoting vectors and matrices, respectively. Parameter vectors are denoted by bold-face Greek letters. For example, $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]'$ denotes a p-dimensional parameter vector, with $[\cdot]'$ denoting the transpose operator. For a vector $\boldsymbol{\theta}$, we define its decomposition into positive and negative parts given by:

$$\boldsymbol{\theta} = \boldsymbol{\theta}^+ - \boldsymbol{\theta}^-,$$

where $\boldsymbol{\theta}^\pm = \max\{\pm\boldsymbol{\theta}, \mathbf{0}\}$. It can be shown that

$$\|\boldsymbol{\theta}^\pm\|_1 = \mathbf{1}'\boldsymbol{\theta}^\pm = \frac{\|\boldsymbol{\theta}\|_1 \pm \mathbf{1}'\boldsymbol{\theta}}{2}$$

are convex in $\boldsymbol{\theta}$. Similarly, for a summation

$$L = \sum_{i=1}^{n} l_i = L^+ - L^-,$$

where $L^+ = \sum_{i=1}^n \max\{l_i, 0\}$, $L^- = -\sum_{i=1}^n \max\{-l_i, 0\}$. We will use the notation $\mathbf{x}_i^j$ to denote the vector $[x_i, \cdots, x_j]^T$ for any $i, j \in \mathbb{Z}$ with $i \leq j$. We will denote the estimated values by $\widehat{(.)}$ and the biased estimates with the superscript $(.)^b$. Throughout the proofs, c_i's express absolute constants which may change from line to line where there is no ambiguity. By c_η we mean an absolute constant which only depends on a positive constant η. We denote the support of a vector $\mathbf{x}_t \in \mathbb{R}^p$ by $\text{supp}(\mathbf{x}_t)$ and its jth element by $(\mathbf{x}_t)_j$.

Given a sparsity level s and a vector $\mathbf{x}$, we denote the set of its s largest magnitude entries by S, and its best s-term approximation error by $\sigma_s(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_S\|_1$. When $\sigma_s(\mathbf{x}) \sim \mathcal{O}^{(1/2-\xi)}$ for some $\xi \geq 0$, we refer to $\mathbf{x}$ as (s, ξ)–compressible.

For simplicity of notation, we define $\mathbf{x}_0$ to be the all-zero vector in $\mathbb{R}^p$. For a matrix $\mathbf{A}$, we denote restriction of $\mathbf{A}$ to its first n rows by $(\mathbf{A})_n$.

We use the convention $[T] = \{1, \cdots, T\}$ and $\mathbf{W}_{[T]} = [\mathbf{w}_1, \cdots, \mathbf{w}_T]$, i.e. $\mathbf{w}_k$ represents the kth column of $\mathbf{W}_{[T]}$. $\odot$ and $\oslash$ denote elementwise multiplication and division respectively. Throughout the chapter we will use the terms innovations and spikes interchangeably. Unless otherwise stated, a function acts on a vector elementwise. For a matrix $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m \times n}$ its mixed p, q-norm is denoted by $\|\mathbf{A}\|_{p,q}$, i.e.

$$\|\mathbf{A}\|_{p,q} = \left[\sum_{i=1}^{m} \left(\sum_{j=1}^{n} |a_{ij}|^p \right)^{q/p} \right]^{1/q},$$

and $\|\mathbf{x}\|_{\mathbf{\Sigma}}^2 = \mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}$.

Contents

Chapter 1: Introduction

Data compression and its fundamental limits are one of the main questions in communication theory [64]. Classical communication theory treats data acquisition and data compression as two separate problems and has studied each of these modules extensively, for example in information theory it is well known that the fundamental limit of data compression is given by the entropy of source. With the emergence of big data applications and the prohibitive cost of sensing mechanisms for high resolution data, smarter sensing mechanisms seem to be inevitable.

The theory of compressive sensing [81] takes its name from the premise that in many applications data acquisition and compression can be performed simultaneously. This is done by utilizing the sparse structure of many natural signals such as images, sound etc. For example, natural images are known to be sparse in Fourier bases. By taking advantage of the sparse structure one could go beyond the fundamental limits imposed by physical constraints and uncertainty principles.

Research in compressive sensing can be categorized in three major branches: mathematical theory, algorithm design and applications. In this thesis we focus on all three aspects. From a mathematical perspective, we use tools from empirical process theory and statistical signal processing in order to study sampling complexity tradeoffs of compressive sensing for dynamic and non i.i.d compressive sensing. This is motivated by the fact that interdependence and dynamic structures are ubiquitous features of natural signals. From the algorithmic perspective, we provide fast algorithms for signal recovery under these conditions, and finally we will apply our theory in several applications of interest.

In order to facilitate reading of this of thesis we will provide a very brief introduction to compressive sensing and a few key results which will be used recurrently throughout the thesis. A more detailed treatment can be found in [81].

1.1 Sparse Solutions of Underdetermined Systems

A vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is s-sparse if it has at most s nonzero entries, i.e. if

$$\|\boldsymbol{\theta}\|_0 := \operatorname{card}\left(\operatorname{supp}(\boldsymbol{\theta})\right) \leq s.$$

We define

$$\sigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|_1 \tag{1.1}$$

and

$$\varsigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|_2 \tag{1.2}$$

which are scalar functions of $\boldsymbol{\theta}$ and s, and capture the compressibility of the parameter vector $\boldsymbol{\theta}$ in the ℓ_1 and ℓ_2 sense, respectively. Note that by definition $\varsigma_s(\boldsymbol{\theta}) \leq \sigma_s(\boldsymbol{\theta})$. For a fixed $\xi \in (0,1)$, we say that $\boldsymbol{\theta}$ is (s,ξ)-*compressible* if $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$ [145]. Note that when $\xi = 0$, the parameter vector $\boldsymbol{\theta}$ is exactly s-sparse.

Let

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{v}, \tag{1.3}$$

where $\mathbf{A} \in \mathbf{R}^{n \times p}$ is a known measurement matrix and $\mathbf{y} \in \mathbb{R}^n$ consists of n linear measurements of $\boldsymbol{\theta}$ and $\mathbf{v}$ is the bounded measurement noise satisfying

$$\|\mathbf{v}\|_2 \leq \epsilon.$$

Compressive sensing problem is concerned with solving (1.3) in the underdetermined setting, i.e. when $n < p$. In this setup, the key assumption of $\boldsymbol{\theta}$ being (close to) an s-sparse vector can be used to recover $\boldsymbol{\theta}$. In the absence of observation noise $\boldsymbol{\theta}$ can be *exactly* recovered from $\mathbf{y}$ if $\text{rank}(A) \geq 2s$, which requires at least $n \geq 2s$ measurements. Special designs of measurement matrices such as Vandermonde matrices or Fourier matrices have resulted in elegant reconstruction algorithms such as Prony's method [157].

In the presence of noise one would need higher number of measurements . Ideally one would like to solve the optimization of finding the *sparsest* $\boldsymbol{\theta}$ which satisfies the bounded noise constraints, i.e.

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta} \in \mathbb{R}^p} \quad \|\boldsymbol{\theta}\|_0, \\ &\text{subject to} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon. \end{aligned} \tag{1.4}$$

Unfortunately (1.4) is an NP-hard problem and due to the combinatorial nature of $\ell_0$norm. The convex relaxation of (1.4) is therefore used most often, which is achieved by replacing the ℓ_0 norm with an ℓ_1 norm, i.e.

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}\in\mathbb{R}^p} \quad \|\boldsymbol{\theta}\|_1, \\ &\text{subject to} \quad \|\mathbf{A}\boldsymbol{\theta}-\mathbf{y}\|_2 \leq \epsilon. \end{aligned} \tag{1.5}$$

1.2 Restricted Isometry Property

For the convex optimization problem (1.5) the following property is *sufficient* for stable recovery of $\boldsymbol{\theta}$:

Definition 1 (Restricted Isometry Property [50])**.** *The matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$ satisfies the restricted isometry property (RIP) [45] of order s, if for all s-sparse $\boldsymbol{\theta} \in \mathbb{R}^p$, we have*

$$(1-\delta_s)\|\boldsymbol{\theta}\|_2^2 \leq \|\mathbf{A}\boldsymbol{\theta}\|_2^2 \leq (1+\delta_s)\|\boldsymbol{\theta}\|_2^2, \tag{1.6}$$

where $\delta_s \in (0,1)$ is the smallest constant for which Eq. (1.6) holds.

Intuitively speaking, RIP requires that the linear measurements acts as an almost isometry on s-sparse vectors, resulting in *invertibility* of the underdetermined system of equations (1.3). The following formalizes this idea:

Theorem 1 (Implications of the RIP [47])**.** *Suppose that $\mathbf{A}$ satisfies the RIP of order $2S$ with $\delta_{2s} < \sqrt{2}-1$. Then any solution $\widehat{\boldsymbol{\theta}}$ to (1.5) satisfies*

$$\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_1 \leq c_1\sigma_s(\boldsymbol{\theta}) + c_2\sqrt{s}\epsilon, \tag{1.7}$$

and

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \leq c_1' \frac{\sigma_s(\boldsymbol{\theta})}{\sqrt{s}} + c_2' \epsilon. \tag{1.8}$$

For an arbitrary value $\delta_s \in (0,1)$ one can prove existence of a measurement matrix satisfying the RIP as long as $n > c_{\delta_s} s \log(p/s)$ namely:

Theorem 2 (RIP for Random Matrices [81])**.** *Let* $\mathbf{A}$ *be an* $n \times p$ *subgaussian random matrix. Then there exits a constant* $c > 0$ *depending only on subgaussian parameters, such that the restricted isometry constant of* $\frac{1}{\sqrt{n}\mathbf{A}}$ *satisfies* $\delta_s \leq \delta$ *with probability at least* $1 - \epsilon$ *provided*

$$n \geq \frac{c}{\delta^2} \left(s \log\left(ep/s \right) + \log(2/\epsilon) \right).$$

Setting $\epsilon = 2\exp(-\delta^2 n/2c)$ yields the condition

$$n \geq 2c/\delta^2 s \log\left(ep/s \right),$$

which guarantees that $\delta_s \leq \delta$ with probability at least $1 - 2\exp(-\delta^2 n/2c)$.

We will next discuss the commonly used reconstruction algorithms for sparse recovery.

1.3 Sparse Recovery Algorithms

Popular sparse recovery algorithms in compressive sensing can be divided into three main categories: optimization methods, greedy methods, and thresholding-based methods [81]. In this thesis we focus on developing algorithms for the optimization methods and greedy methods.

1.3.1 Optimization Problems

Three of the most popular optimization problems for sparse recovery are the quadratically constrained basis pursuit, lasso [203] and basis pursuit denoising. The quadratically constrained basis pursuit algorithm solves the optimization problem

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}\in\mathbb{R}^p} && \|\boldsymbol{\theta}\|_1, \\ &\text{subject to} && \|\mathbf{A}\boldsymbol{\theta}-\mathbf{y}\|_2 \leq \epsilon. \end{aligned} \tag{1.9}$$

Alternatively one can find the solution which is closest to the measurements while maintaining a controlled sparsity level. Such formulation is known as the lasso given by

$$\begin{aligned} &\text{minimize}_{\boldsymbol{\theta}\in\mathbb{R}^p} && \|\mathbf{A}\boldsymbol{\theta}-\mathbf{y}\|_2, \\ &\text{subject to} && \|\boldsymbol{\theta}\|_1 \leq \tau. \end{aligned} \tag{1.10}$$

A closely related optimization problem via the lagrangian formulation of both problems is known as basis pursuit denoising, given by

$$\text{minimize}_{\boldsymbol{\theta}\in\mathbb{R}^p} \quad \|\mathbf{A}\boldsymbol{\theta}-\mathbf{y}\|_2+\lambda\|\boldsymbol{\theta}|\|_1. \tag{1.11}$$

```
Input: $\mathbf{A}, \mathbf{y}$
Output: $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$
Initialization: $\begin{cases} \text{Start with the index set } S^{(0)} = \emptyset \\ \text{and the initial estimate } \widehat{\boldsymbol{\theta}}^{(0)}_{\mathsf{OMP}} = \mathbf{0} \end{cases}$
for $k = 1, 2, \cdots, s^\star$
  $j = \arg\max_i \left| \left( \mathbf{A}' \left( \mathbf{y} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{(k-1)}_{\mathsf{OMP}} \right) \right)_i \right|$
  $S^{(k)} = S^{(k-1)} \cup \{j\}$
  $\widehat{\boldsymbol{\theta}}^{(k)}_{\mathsf{OMP}} = \underset{\mathrm{supp}(\boldsymbol{\theta}) \subset S^{(k)}}{\arg\min} \|\mathbf{y} - \mathbf{A}\boldsymbol{\theta}\|_2$
end
```

Table 1.1: Orthogonal Matching Pursuit (OMP)

1.3.2 Greedy Algorithms

Although there exist fast solvers to convex problems of the type given by Eq. (1.9), these algorithms are polynomial time in n and p, and may not scale well with high-dimensional data. This motivates us to consider greedy solutions for the estimation of sparse parameters. In particular, in this thesis we will consider generalizations of the Orthogonal Matching Pursuit (OMP) [158] for general convex cost functions. The main idea behind the OMP is in the greedy selection stage, where the absolute value of the gradient of the cost function at the current solution is considered as the selection metric. The OMP algorithm adds one index to the estimated support of $\boldsymbol{\theta}$ at each step. A flowchart of the algorithm is given in Table (1.1).

The choice of the maximum number of iterations $s^\star$ will be discussed in detail later. One can repeat the process until a stopping criterion is met. The advantage of the OMP algorithm is breaking the high (p)-dimensional optimization problem into several low ($\leq s^\star$)-dimensional problem which can usually be solved much faster.

1.4 Theoretical Guarantees for Convex Cost Functions

In many applications of interest the measurements are not linear, or the noise is not additive or Gaussian. In these scenarios the objective function is not in the form of squared errors. We will now introduce the theoretical requirements which generalize RIP to general convex cost functions, and the corresponding theoretical guarantees.

1.4.1 Restricted Strong Convexity

General convex cost functions are usually in the form of a loss function or a negative log-likelihood. Estimation problems for such cost functions are known as M-estimation problems. The general form of an M-estimation problem is given by

$$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{L}(\boldsymbol{\theta}), \tag{1.12}$$

and its ℓ_1-regularized counterpart is given by

$$\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{L}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1. \tag{1.13}$$

where, $\mathfrak{L}(\boldsymbol{\theta})$ is a cost function which is convex in $\boldsymbol{\theta}$, $\gamma_n > 0$ is a regularization parameter and $\boldsymbol{\Theta}$ is the convex set of admissible solutions.

The main sufficient condition for theoretical guarantees of general convex cost functions is the notion of Restricted Strong Convexity (RSC) [148]. By the convexity of the cost function, it is clear that a small change in $\boldsymbol{\theta}$ results in a small change in the cost function. However, the converse is not necessarily true. Intuitively speaking, the RSC condition guarantees that the converse holds: a small change in the cost implies a small change in the parameter vector, i.e., the cost function is not too *flat* around the true parameter vector. A depiction of the RSC condition for $p = 2$, adopted from [148], is given in Figure 1.1. In Figure 1.1(a), the RSC does not hold since a change along θ_2 does not change the log-likelihood, whereas the log-likelihood in Figure 1.1(b) satisfies the RSC.

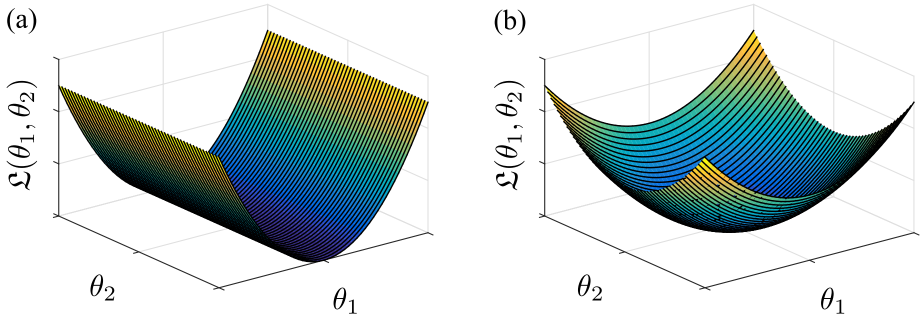

(a)
$\mathfrak{L}(\theta_1, \theta_2)$
θ_2
θ_1
(b)
$\mathfrak{L}(\theta_1, \theta_2)$
θ_2
θ_1

Figure 1.1: Illustration of RSC (a) RSC does not hold (b) RSC does hold.

More formally, if the log-likelihood is twice differentiable at $\boldsymbol{\theta}$, the RSC is equivalent to existence of a lower quadratic bound on the negative log-likelihood:

$$\mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\psi}, \boldsymbol{\theta}) := \mathfrak{L}(\boldsymbol{\theta} + \boldsymbol{\psi}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\psi}' \nabla \mathfrak{L}(\boldsymbol{\theta}) \geq \kappa \|\boldsymbol{\psi}\|_2^2, \tag{1.14}$$

for a positive constant $\kappa > 0$ and all $\boldsymbol{\psi} \in \mathbb{R}^p$ in a carefully-chosen neighborhood of $\boldsymbol{\theta}$ depending on s and ξ. Based on the results of [148], when the RSC is satisfied, sufficient conditions akin to that in Theorem 1 can be obtained by estimating the Euclidean extent of the solution set around the true parameter vector. Here we restate the main result of [148] concerning RSC and its implications in controlling the estimation error for general convex cost functions:

Proposition 1 (Implications of RSC (Theorem 1 of [148]))**.** *For a negative log-likelihood $\mathfrak{L}(\boldsymbol{\theta})$ which satisfies the RSC with parameter κ, every solution to the convex optimization problem (1.13) satisfies*

$$\left\| \widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta} \right\|_2 \leq \frac{2\gamma_n \sqrt{s}}{\kappa} + \sqrt{\frac{2\gamma_n \sigma_s(\boldsymbol{\theta})}{\kappa}} \tag{1.15}$$

with a choice of the regularization parameter

$$\gamma_n \geq 2 \left\| \nabla \mathfrak{L}(\boldsymbol{\theta}) \right\|_\infty . \tag{1.16}$$

The first term in the bound (1.16) is increasing in s and corresponds to the estimation error of the s largest components of $\boldsymbol{\theta}$ in magnitude, whereas the second term is decreasing in s and represents the cost of replacing $\boldsymbol{\theta}$ with its best s-sparse approximation.

Similarly the counterpart of OMP for general cost functions was introduced in [235] and is summarized in Table 1.2.

The main theoretical result regarding the generalized OMP is given by the following Proposition stating that the greedy procedure is successful in obtaining a reasonable $s^\star$-sparse approximation, if the cost function satisfies the RSC:

```
Input: $\mathfrak{L}(\boldsymbol{\theta}), s^\star$
Output: $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)}$
Initialization: $\begin{cases} \text{Start with the index set } S^{(0)} = \emptyset \\ \text{and the initial estimate } \widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(0)} = 0 \end{cases}$
for $k = 1, 2, \cdots, s^\star$
  $j = \arg\max_i \left| \left( \nabla \mathfrak{L} \left( \widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(k-1)} \right) \right)_i \right|$
  $S^{(k)} = S^{(k-1)} \cup \{j\}$
  $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(k)} = \underset{\text{supp}(\boldsymbol{\theta}) \subset S^{(k)}}{\arg\min} \ \mathfrak{L}(\boldsymbol{\theta})$
end
```

Table 1.2: Generalized Orthogonal Matching Pursuit

Proposition 2 (Guarantees of OMP (Theorem 2.1 of [235]))**.** *Suppose that $\mathfrak{L}(\boldsymbol{\theta})$ satisfies RSC with a constant $\kappa > 0$. Let $s^\star$ be a constant such that*

$$s^\star = \mathcal{O}(s \log s), \tag{1.17}$$

Then, we have

$$\left\| \widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)} - \boldsymbol{\theta}_s \right\|_2 \leq \frac{\sqrt{6}\epsilon_{s^\star}}{\kappa},$$

where $\epsilon_{s^\star}$ satisfies

$$\epsilon_{s^\star} \leq \sqrt{s^\star + s} \| \nabla \mathfrak{L}(\boldsymbol{\theta}_s) \|_\infty. \tag{1.18}$$

1.5 Roadmap of the Thesis

As noted earlier most of the theoretical guarantees of compressive sensing provide sharp trade-offs between the number of measurement, sparsity, and estimation accuracy when random i.i.d measurements are at disposal. However, in most practical applications of interest, the measured signals and the corresponding covariates are highly interdependent and follow specific dynamics. In Chapters 2 and 3 we will generalize these theoretical guarantees to two large classes of stationary processes, namely the autoregressive (AR) processes and generalized linear models (GLM), where the covariates are highly nonlinear and non-i.i.d. Our theoretical results provide insight into the tradeoffs in sampling requirements and a measure of dependence in these models.

In Chapters 4 and 6 we provide fast iterative algorithms for compressible state-space models as well as general convex optimization problems with positivity constraints. We will use ideas from projection microscopy to develop a two-photon imaging technique that scans lines of excitation across the sample at multiple angles, recovering high-resolution images from relatively few incoherently multiplexed measurements. By using a static image of the sample as a prior, we recover neural activity from megapixel fields of view at rates exceeding 1kHz.

Chapter 2: Sampling Requirements for Stable Autoregressive Estimation

Autoregressive (AR) models are among the most fundamental tools in analyzing time series. They have been useful for modeling signals in many applications including financial time series analysis [182] and traffic modeling [8, 9, 25, 60, 78, 178]. Due to their well-known approximation property, these models are commonly used to represent stationary processes in a parametric fashion and thereby preserve the underlying structure of these processes [11]. In general, the ubiquitous long-range dependencies in real-world time series, such as financial data, results in AR model fits with large orders [182].

In this chapter, we close the gap in theory of compressed sensing for non-i.i.d. data by providing theoretical guarantees on stable estimation of autoregressive, where the history of the process takes the role of the interdependent covariates. In doing so, we relax the assumptions of i.i.d. covariates and exact sparsity common in CS. Our results indicate that utilizing sparsity recovers important information about the intrinsic frequencies of such processes.

We consider the problem of estimating the parameters of a linear univariate autoregressive model with sub-Gaussian innovations from a limited sequence of consecutive observations. Assuming that the parameters are compressible, we analyze the performance of the ℓ_1-regularized least squares as well as a greedy estimator of the parameters and characterize the sampling trade-offs required for stable recovery in the non-asymptotic regime. In particular, we show that for a fixed sparsity level, stable recovery of AR parameters is possible when the number of samples scale *sublinearly* with the AR order. Our results improve over existing sampling complexity requirements in AR estimation using the LASSO, when the sparsity level scales faster than the square root of the model order. We further derive sufficient conditions on the sparsity level that guarantee the minimax optimality of the ℓ_1-regularized least squares estimate. Applying these techniques to simulated data as well as real-world datasets from crude oil prices and traffic speed data confirm our predicted theoretical performance gains in terms of estimation accuracy and model selection.

2.1 Introduction

Autoregressive (AR) models are among the most fundamental tools in analyzing time series. Applications include financial time series analysis [182] and traffic modeling [8, 9, 25, 60, 78, 178]. Due to their well-known approximation property, these models are commonly used to represent stationary processes in a parametric fashion and thereby preserve the underlying structure of these processes [11]. In order to leverage the approximation property of AR models, often times parameter sets of very large order are required [169]. For instance, any autoregressive moving average (ARMA) process can be represented by an AR process of infinite order. Statistical inference using these models is usually performed through fitting a long-order AR model to the data, which can be viewed as a truncation of the infinite-order representation [85, 86, 105, 188]. In general, the ubiquitous long-range dependencies in real-world time series, such as financial data, results in AR model fits with large orders [182].

In various applications of interest, the AR parameters fit to the data exhibit sparsity, that is, only a small number of the parameters are non-zero. Examples include autoregressive communication channel models, quasi-oscillatory data tuned around specific frequencies and financial time series [21, 136, 178]. The non-zero AR parameters in these models correspond to significant time lags at which the underlying dynamics operate. Traditional AR order selection criteria such as the Final Prediction Error (FPE) [13], Akaike Information Criterion (AIC) [12] and Bayesian Information Criterion (BIC) [185], are based on asymptotic lower bounds on the mean squared prediction error. Although there exist several improvements over these traditional results aiming at exploiting sparsity [105,188,221], the resulting criteria pertain to the asymptotic regimes and their finite sample behavior is not well understood [89]. Non-asymptotic results for AR estimation, such as [89,144], do not fully exploit the sparsity of the underlying parameters in favor of reducing the sample complexity. In particular, for an AR process of order p, sufficient sampling requirements of $n \sim \mathcal{O}(p^4) \gg p$ and $n \sim \mathcal{O}(p^5) \gg p$ are established in [89] and [144], respectively.

A relatively recent line of research employs the theory of compressed sensing (CS) for studying non-asymptotic sampling-complexity trade-offs for regularized M-estimators. In recent years, the CS theory has become the standard framework for measuring and estimating sparse statistical models [45, 50, 71]. The theoretical guarantees of CS imply that when the number of incoherent measurements are roughly proportional to the sparsity level, then stable recovery of the model parameters is possible. A key underlying assumption in many existing theoretical analyses of linear models is the independence and identical distribution (i.i.d.) of the covariates' structure. The matrix of covariates is either formed by fully i.i.d. elements [22,180], is based on row-i.i.d. correlated designs [173,237], is Toeplitz-i.i.d. [99], or circulant i.i.d. [175], where the design is extrinsic, fixed in advance and is independent of the underlying sparse signal. The matrix of covariates formed from the observations of an AR process does not fit into any of these categories, as the intrinsic history of the process plays the role of the covariates. Hence the underlying interdependence in the model hinders a straightforward application of existing CS results to AR estimation. Recent non-asymptotic results on the estimation of multivariate AR (MVAR) processes have been relatively successful in utilizing sparsity for such dependent structures. For Gaussian and low-rank MVAR models, respectively, sub-linear sampling requirements have been established in [96,130] and [147], using regularized LS estimators, under bounded operator norm assumptions on the transition matrix. These assumptions are shown to be restrictive for MVAR processes with lags larger than 1 [224]. By relaxing these boundedness assumptions for Gaussian, sub-Gaussian and heavy-tailed MVAR processes, respectively, sampling requirements of $n \sim \mathcal{O}(s \log p)$ and $\mathcal{O}((s \log p)^2)$ have been established in [28] and [224,227]. However, the quadratic scaling requirement in the sparsity level for the case of sub-Gaussian and heavy-tailed innovations incurs a significant gap with respect to the optimal guarantees of CS (with linear scaling in sparsity), particularly when the sparsity level s is allowed to scale with p.

In this chapter, we consider two of the widely-used estimators in CS, namely the ℓ_1-regularized Least Squares (LS) or the LASSO and the Orthogonal Matching Pursuit (OMP) estimator, and extend the non-asymptotic recovery guarantees of the CS theory to the estimation of univariate AR processes with compressible parameters using these estimators. In particular, we improve the aforementioned gap between non-asymptotic sampling requirements for AR estimation and those promised by compressed sensing by providing sharper sampling-complexity trade-offs which improve over existing results when the sparsity grows faster than the square root of p. Our focus on the analysis of univariate AR processes is motivated by the application areas of interest in this chapter which correspond to one-dimensional time series. Existing results in the literature [28, 96, 130, 147, 224, 227], however, consider the MVAR case and thus are broader in scope. We will therefore compare our results to the univariate specialization of the aforementioned results. Our main contributions can be summarized as follows:

First, we establish that for a univariate AR process with sub-Gaussian innovations when the number of measurements scales *sub-linearly* with the product of the ambient dimension p and the sparsity level s, i.e., $n \sim \mathcal{O}(s(p \log p)^{1/2}) \ll p$, then stable recovery of the underlying AR parameters is possible using the LASSO and the OMP estimators, even though the covariates are highly interdependent and solely based on the history of the process. In particular, when $s \propto p^{\frac{1}{2}+\delta}$ for some $\delta \geq 0$ and the LASSO is used, our results improve upon those of [224, 227], when specialized to the univariate AR case, by a factor of $p^{\delta}(\log p)^{3/2}$. For the special case of Gaussian AR processes, stronger results are available which require a scaling of $n \sim \mathcal{O}(s \log p)$

[28]. Moreover, our results provide a theory-driven choice of the number of iterations for stable estimation using the OMP algorithm, which has a significantly lower computational complexity than the LASSO.

Second, in the course of our analysis, we establish the Restricted Eigenvalue (RE) condition [31] for $n \times p$ design matrices formed from a realization of an AR process in a Toeplitz fashion, when $n \sim \mathcal{O}(s(p \log p)^{1/2}) \ll p$. To this end, we invoke appropriate concentration inequalities for sums of *dependent* random variables in order to capture and control the high interdependence of the design matrix. In the special case of a white noise sub-Gaussian process, i.e., a sub-Gaussian i.i.d. Toeplitz measurement matrix, we show that our result can be strengthened from $n \sim \mathcal{O}(s(p \log p)^{1/2})$ to $n \sim \mathcal{O}(s(\log p)^2)$, which improves by a factor of $s/\log p$ over the results of [99] requiring $n \sim \mathcal{O}(s^2 \log p)$.

Third, we establish sufficient conditions on the sparsity level which result in the minimax optimality of the ℓ_1-regularized LS estimator. Finally, we provide simulation results as well as application to oil price and traffic data which reveal that the sparse estimates significantly outperform traditional techniques such as the Yule-Walker based estimators [196]. We have employed statistical tests in time and frequency domains to compare the performance of these estimators.

The rest of the chapter is organized as follows. In Section 2.2, we will introduce the notations and problem formulation. In Section 2.3, we will describe several methods for the estimation of the parameters of an AR process, present the main theoretical results of this chapter on robust estimation of AR parameters, and establish the minimax optimality of the ℓ_1-regularized LS estimator. Section 2.4 includes our simulation results on simulated data as well as the real-world financial and traffic data, followed by concluding remarks in Section 2.5.

2.2 Problem Formulation

Consider a univariate AR(p) process defined by

$$x_k = \theta_1 x_{k-1} + \theta_2 x_{k-2} + \cdots + \theta_p x_{k-p} + w_k = \boldsymbol{\theta}' \mathbf{x}_{k-p}^{k-1} + w_k, \tag{2.1}$$

where $\{w_k\}_{k=-\infty}^{\infty}$ is an i.i.d sub-Gaussian innovation sequence with zero mean and variance σ_{w}^2. This process can be considered as the output of an LTI system with transfer function

$$H(z) = \frac{\sigma_{\mathsf{w}}^2}{1 - \sum_{\ell=1}^{p} \theta_\ell z^{-\ell}}. \tag{2.2}$$

Throughout the chapter we will assume $\|\boldsymbol{\theta}\|_1 \leq 1-\eta < 1$ to enforce the stability of the filter. We will refer to this assumption as *the sufficient stability assumption*, since an AR process with poles within the unit circle does not necessarily satisfy $\|\boldsymbol{\theta}\|_1 < 1$. However, beyond second-order AR processes, it is not straightforward to state the stability of the process in terms of its parameters in a closed algebraic form, which in turn makes both the analysis and optimization procedures intractable. As we will show later, the only major requirement of our results is the boundedness of the spectral spread (also referred to as condition number) of the AR process. Although the sufficient stability condition is more restrictive, it will significantly simplify the spectral constants appearing in the analysis and clarifies the various trade-offs in the sampling bounds (See, for example, Corollary 3).

The AR(p) process given by $\{x_k\}_{k=-\infty}^{\infty}$ in (2.1) is stationary in the strict sense. Also by (2.2) the power spectral density of the process equals

$$S(\omega) = \frac{\sigma_{\mathsf{w}}^2}{|1 - \sum_{\ell=1}^{p} \theta_\ell e^{-j\ell\omega}|^2}. \tag{2.3}$$

The sufficient stability assumption implies boundedness of the spectral spread of the process defined as

$$\rho = \sup_{\omega} S(\omega) \Big/ \inf_{\omega} S(\omega).$$

We will discuss how this assumption can be further relaxed in Appendix A.1.2. The spectral spread of stationary processes in general is a measure of how quickly the process reaches its ergodic state [89]. An important property that we will use later in this chapter is that the spectral spread is an upper bound on the eigenvalue spread of the covariance matrix of the process of arbitrary size [101].

We will also assume that the parameter vector $\boldsymbol{\theta}$ is compressible (to be defined more precisely later), and can be well approximated by an s-sparse vector where $s \ll p$. We observe n consecutive snapshots of length p (a total of $n+p-1$ samples) from this process given by $\{x_k\}_{k=-p+1}^{n}$ and aim to estimate $\boldsymbol{\theta}$ by exploiting its sparsity; to this end, we aim at addressing the following questions in the non-asymptotic regime:

- Are the conventional LASSO-type and greedy techniques suitable for estimating $\boldsymbol{\theta}$?

- What are the sufficient conditions on n in terms of p and s, to guarantee stable recovery?

- Given these sufficient conditions, how do these estimators perform compared to conventional AR estimation techniques?

Traditionally, the Yule-Walker (YW) equations or least squares formulations are used to fit AR models. Since these methods do not utilize the sparse structure of the parameters, they usually require $n \gg p$ samples in order to achieve satisfactory performance. The YW equations can be expressed as

$$\mathbf{R}\boldsymbol{\theta} = \mathbf{r}_{-p}^{-1}, \quad r_0 = \boldsymbol{\theta}'\mathbf{r}_{-p}^{-1} + \sigma_{\mathsf{w}}^2, \tag{2.4}$$

where $\mathbf{R} := \mathbf{R}_{p\times p} = \mathbb{E}[\mathbf{x}_1^p \mathbf{x}_1^{p'}]$ is the $p \times p$ covariance matrix of the process and $r_k = \mathbb{E}[x_i x_{i+k}]$ is the autocorrelation of the process at lag k. The covariance matrix R and autocorrelation vector $\mathbf{r}_{-p}^{-1}$ are typically replaced by their sample counterparts. Estimation of the AR(p) parameters from the YW equations can be efficiently carried out using the Burg's method [43]. Other estimation techniques include LS regression and maximum likelihood (ML) estimation. In this chapter, we will consider the Burg's method and LS solutions as comparison benchmarks. When n is comparable to p, these two methods are known to exhibit substantial performance differences [137].

When fitted to the real-world data, the parameter vector $\boldsymbol{\theta}$ usually exhibits a degree of sparsity. That is, only certain lags in the history have a significant contribution in determining the statistics of the process. These lags can be thought of as the intrinsic delays in the underlying dynamics. To be more precise, for a sparsity level $s < p$, we denote by $S \subset \{1, 2, \cdots, p\}$ the support of the s largest elements of $\boldsymbol{\theta}$ in absolute value, and by $\boldsymbol{\theta}_s$ the best s-term approximation to $\boldsymbol{\theta}$. We also define

$$\sigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|_1 \text{ and } \varsigma_s(\boldsymbol{\theta}) := \|\boldsymbol{\theta} - \boldsymbol{\theta}_s\|_2, \tag{2.5}$$

which capture the compressibility of the parameter vector $\boldsymbol{\theta}$ in the ℓ_1 and ℓ_2 sense, respectively. Note that by definition $\varsigma_s(\boldsymbol{\theta}) \leq \sigma_s(\boldsymbol{\theta})$. For a fixed $\xi \in (0, 1)$, we say that $\boldsymbol{\theta}$ is (s, ξ)-*compressible* if $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$ [145] and $(s, \xi, 2)$-*compressible* if $\varsigma_s(\boldsymbol{\theta}) = \mathcal{O}(s^{1-\frac{1}{\xi}})$. Note that $(s, \xi, 2)$-*compressibility* is a weaker condition than (s, ξ)-*compressibility* and when $\xi = 0$, the parameter vector $\boldsymbol{\theta}$ is exactly s-sparse.

Finally, in this chapter, we are concerned with the compressed sensing regime where $n \ll p$, i.e., the observed data has a much smaller length than the ambient dimension of the parameter vector. The main estimation problem of this chapter can be summarized as follows: *given observations* $\mathbf{x}_{-p+1}^n$ *from an AR process with sub-Gaussian innovations and bounded spectral spread, the goal is to estimate the unknown p-dimensional* $(s, \xi, 2)$-*compressible AR parameters* $\boldsymbol{\theta}$ *in a stable fashion (where the estimation error is controlled) when* $n \ll p$.

2.3 Theoretical Results

In this section, we will describe the estimation procedures and present the main theoretical results of this chapter.

2.3.1 ℓ_1-regularized least squares estimation

Given the sequence of observations $\mathbf{x}_{-p+1}^n$ and an estimate $\widehat{\boldsymbol{\theta}}$, the normalized estimation error can be expressed as:

$$\mathfrak{L}\left(\widehat{\boldsymbol{\theta}}\right) := \frac{1}{n} \left\| \mathbf{x}_1^n - \mathbf{X}\widehat{\boldsymbol{\theta}} \right\|_2^2, \tag{2.6}$$

where

$$\mathbf{X} = \begin{bmatrix} x_{n-1} & x_{n-2} & \cdots & x_{n-p} \\ x_{n-2} & x_{n-3} & \cdots & x_{n-p-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_0 & x_{-1} & \cdots & x_{-p+1} \end{bmatrix}. \tag{2.7}$$

Note that the matrix of covariates $\mathbf{X}$ is Toeplitz with highly interdependent elements. The LS solution is thus given by:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{LS}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min}\, \mathfrak{L}(\boldsymbol{\theta}), \tag{2.8}$$

where

$$\boldsymbol{\Theta} := \{\boldsymbol{\theta} \in \mathbb{R}^p | \ \|\boldsymbol{\theta}\|_1 < 1 - \eta\}$$

is the convex feasible region for which the stability of the process is guaranteed. Note that the sufficient constraint of $\|\boldsymbol{\theta}\|_1 < 1 - \eta$ is by no means necessary for stability. However, the set of all $\boldsymbol{\theta}$ resulting in stability is in general not convex. We have thus chosen to cast the LS estimator of Eq. (2.8) –as well as its ℓ_1-regularized version that follows– over a convex subset $\boldsymbol{\Theta}$, for which fast solvers exist. In addition, as we will show later, this assumption significantly clarifies the various constants appearing in our theoretical analysis. In practice, the Yule-Walker estimate is obtained without this constraint, and is guaranteed to result in a stable AR process. Similarly, for the LS estimate, this condition is relaxed by obtaining the unconstrained LS estimate and checking *post hoc* for stability [160].

Consistency of the LS estimator given by (2.8) was shown in [182] when $n \to \infty$ for Gaussian innovations. In the case of Gaussian innovations the LS estimates correspond to conditional ML estimation and are asymptotically unbiased under mild conditions, and with p fixed, the solution converges to the true parameter vector as $n \to \infty$. For fixed p, the estimation error is of the order $\mathcal{O}(\sqrt{p/n})$ in general [99]. However, when p is allowed to scale with n, the convergence rate of the estimation error is not known in general.

In the regime of interest in this thesis, where $n \ll p$, the LS estimator is ill-posed and is typically regularized with a smooth norm. In order to capture the compressibility of the parameters, we consider the ℓ_1-regularized LS estimator:

$$\widehat{\boldsymbol{\theta}}_{\ell_1} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{L}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1, \tag{2.9}$$

where $\gamma_n > 0$ is a regularization parameter. This estimator, deemed as the Lagrangian form of the LASSO [203], has been comprehensively studied in the sparse recovery literature [118, 139, 220] as well as AR estimation [118, 144, 221, 237]. A general asymptotic consistency result for LASSO-type estimators was established in [118]. Asymptotic consistency of LASSO-type estimators for AR estimation was shown in [221, 237]. For sparse models, non-asymptotic analysis of the LASSO with covariate matrices from row-i.i.d. correlated design has been established in [220, 237].

In many applications of interest, the data correlations are exponentially decaying and negligible beyond a certain lag, and hence for large enough p, autoregressive models fit the data very well in the prediction error sense. An important question is thus how many measurements are required for estimation stability? In the *overdetermined* regime of $n \gg p$, the non-asymptotic properties of LASSO for model selection of AR processes has been studied in [144], where a sampling requirement of $n \sim \mathcal{O}(p^5)$ is established. Recovery guarantees for LASSO-type estimators of multivariate AR parameters in the *compressive* regime of $n \ll p$ are studied in [28,96,130,147,224,227]. In particular, sub-linear scaling of n with respect to the ambient dimension is established in [96,130] for Gaussian MVAR processes and in [147] for low-rank MVAR processes, respectively, under the assumption of bounded operator norm of the transition matrix. In [28] and [224,227], the latter assumption is relaxed for Gaussian, sub-Gaussian, and heavy-tailed MVAR processes, respectively. These results have significant practical implications as they will reveal sufficient conditions on n with respect to p as well as a criterion to choose γ_n, which result in stable estimation of $\boldsymbol{\theta}$ from a considerably short sequence of observations. The latter is indeed the setting that we consider in this chapter, where the ambient dimension p is fixed and the goal is to derive sufficient conditions on $n \ll p$ resulting in stable estimation.

It is easy to verify that the objective function and constraints in Eq. (2.9) are convex in $\boldsymbol{\theta}$ and hence $\widehat{\boldsymbol{\theta}}_{\ell_1}$ can be obtained using standard numerical solvers. Note that the solution to (2.9) might not be unique. However, we will provide error bounds that hold for all possible solutions of (2.9), with high probability.

Recall that, the Yule-Walker solution is given by

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw}} := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{J}(\boldsymbol{\theta}) = \widehat{\mathbf{R}}^{-1}\widehat{\mathbf{r}}_{-p}^{-1}, \tag{2.10}$$

where $\mathfrak{J}(\boldsymbol{\theta}) := \|\widehat{\mathbf{R}}\boldsymbol{\theta} - \widehat{\mathbf{r}}_{-p}^{-1}\|_2$. We further consider two other sparse estimators for $\boldsymbol{\theta}$ by penalizing the Yule-Walker equations. The ℓ_1-regularized Yule-Walker estimator is defined as:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{2,1}} := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{J}(\boldsymbol{\theta}) + \gamma_n\|\boldsymbol{\theta}\|_1, \tag{2.11}$$

where $\gamma_n > 0$ is a regularization parameter. Similarly, using the robust statistics instead of the Gaussian statistics, the estimation error can be re-defined as:

$$\mathfrak{J}_1(\boldsymbol{\theta}) := \|\widehat{\mathbf{R}}\boldsymbol{\theta} - \widehat{\mathbf{r}}_{-p}^{-1}\|_1,$$

we define the ℓ_1-regularized estimates as

$$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{1,1}} := \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\min} \quad \mathfrak{J}_1(\boldsymbol{\theta}) + \gamma_n\|\boldsymbol{\theta}\|_1. \tag{2.12}$$

2.3.2 Greedy estimation

Although there exist fast solvers for the convex problems of the type given by (2.9), (2.11) and (2.12), these algorithms are polynomial time in n and p, and may not scale well with the dimension of data. This motivates us to consider greedy solutions for the estimation of $\boldsymbol{\theta}$. In particular, we will consider and study the performance of a generalized Orthogonal Matching Pursuit (OMP) algorithm [158,235]. A flowchart of this algorithm is given in Table 2.1 for completeness. At each iteration, a new component of $\boldsymbol{\theta}$ for which the gradient of the error metric $\mathfrak{f}(\boldsymbol{\theta})$ is the largest in absolute value is chosen and added to the current support. The algorithm proceeds for a total of $s^\star = \mathcal{O}(s \log s)$ steps, resulting in an estimate with $s^\star$ components. When the error metric $\mathfrak{L}(\boldsymbol{\theta})$ is chosen, the generalized OMP corresponds to the original OMP algorithm. For the choice of the YW error metric $\mathfrak{J}(\boldsymbol{\theta})$, we denote the resulting greedy algorithm by ywOMP.

```
Input: $\mathfrak{f}(\boldsymbol{\theta}), s^\star$
Output: $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}} = \widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{OMP}}$
Initialization: { Start with the index set $S^{(0)} = \emptyset$
                { and the initial estimate $\widehat{\boldsymbol{\theta}}^{(0)}_{\mathsf{OMP}} = 0$
for $k = 1, 2, \cdots, s^\star$
  $j = \arg\max_i \left| \left( \nabla \mathfrak{f} \left( \widehat{\boldsymbol{\theta}}^{(k-1)}_{\mathsf{OMP}} \right) \right)_i \right|$
  $S^{(k)} = S^{(k-1)} \cup \{j\}$
  $\widehat{\boldsymbol{\theta}}^{(k)}_{\mathsf{OMP}} = \underset{\mathrm{supp}(\boldsymbol{\theta}) \subset S^{(k)}}{\arg\min} \; \mathfrak{f}(\boldsymbol{\theta})$
end
```

Table 2.1: Generalized Orthogonal Matching Pursuit (OMP)

2.3.3 Estimation performance guarantees

The main theoretical result regarding the estimation performance of the ℓ_1-regularized LS estimator is given by the following theorem:

Theorem 3. *If $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, there exist positive constants d_0, d_1, d_2, d_3 and d_4 such that for $n > s \max\{d_0(\log p)^2, d_1(p \log p)^{1/2}\}$ and a choice of regularization parameter $\gamma_n = d_2\sqrt{\frac{\log p}{n}}$, any solution $\widehat{\boldsymbol{\theta}}_{\ell_1}$ to (2.9) satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\right\|_2 \leq d_3\sqrt{\frac{s \log p}{n}} + \sqrt{d_3\sigma_s(\boldsymbol{\theta})}\sqrt[4]{\frac{\log p}{n}}, \tag{2.13}$$

with probability greater than $1 - \mathcal{O}(\frac{1}{n^{d_4}})$. The constants depend on the spectral spread of the process and are explicitly given in the proof.

Similarly, the following theorem characterizes the estimation performance bounds for the OMP algorithm:

Theorem 4. *If $\boldsymbol{\theta}$ is $(s, \xi, 2)$-compressible for some $\xi < 1/2$, there exist positive constants d'_0, d'_1, d'_2, d'_3 and d'_4 such that for $n > s \log s \max\{d'_0(\log p)^2, d'_1(p \log p)^{1/2}\}$, the OMP estimate satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}} - \boldsymbol{\theta}\right\|_2 \leq d'_2\sqrt{\frac{s \log s \log p}{n}} + d'_3\frac{\log s}{s^{\frac{1}{\xi}-2}} \tag{2.14}$$

after $s^\star = 4\rho s \log 20\rho s$ iterations with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d'_4}}\right)$. The constants depend on the spectral spread of the process and are explicitly given in the proof.

The results of Theorems 3 and 6 suggest that under suitable compressibility assumptions on the AR parameters, one can estimate the parameters reliably using the ℓ_1-regularized LS and OMP estimators with much fewer measurements compared to those required by the Yule-Walker/LS based methods. To illustrate the significance of these results further, several remarks are in order:

Remark 1. The sufficient stability assumption of $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$ is restrictive compared to the class of stable AR models. In general, the set of parameters $\boldsymbol{\theta}$ which admit a stable AR process is not necessarily convex. This condition ensures that the resulting estimates of (2.9)-(2.12) pertain to stable AR processes and at the same time can be obtained by convex optimization techniques, for which fast solvers exist. A common practice in AR estimation, however, is to solve for the unconstrained problem and check for the stability of the resulting AR process *post hoc.* In our numerical studies in Section 2.4, this procedure resulted in a stable AR process in all cases. Nevertheless, the stability guarantees of Theorems 3 and 6 hold for the larger class of stable AR processes, even though they may not necessarily be obtained using convex optimization techniques. We further discuss this generalization in Appendix A.1.2.

Remark 2. When $\boldsymbol{\theta} = \mathbf{0}$, i.e., the process is a sub-Gaussian white noise and hence the matrix $\mathbf{X}$ is i.i.d. Toeplitz with sub-Gaussian elements, the constants d_1 and d'_1 in Theorems 1 and 2 vanish, and the measurement requirements strengthen to $n > d_0 s (\log p)^2$ and $n > d'_0 s \log s (\log p)^2$, respectively. Comparing this sufficient condition with that of [99] given by $n \sim \mathcal{O}(s^2 \log p)$ reveals an improvement of order $s(\log p)^{-1}$ by our results.

Remark 3. When $\boldsymbol{\theta} \neq \mathbf{0}$, the dominant measurement requirements are $n > d_1 s (p \log p)^{1/2}$ and $n > d'_1 s \log s (p \log p)^{1/2}$. Comparing the sufficient condition $n \sim \mathcal{O}(s (p \log p)^{1/2})$ of Theorem 3 with those of [45, 50, 71, 220] for linear models with i.i.d. measurement matrices or row-i.i.d. correlated designs [173, 237] given by $n \sim \mathcal{O}(s \log p)$ a loss of order $\mathcal{O}((p/\log p)^{1/2})$ is incurred, although all these conditions require $n \ll p$. However, the loss seems to be natural as it stems from a major difference of our setting as compared to traditional CS: each row of the measurement matrix $\mathbf{X}$ highly depends on the entire observation sequence $\mathbf{x}_1^n$, whereas in traditional CS, each row of the measurement matrix is only related to the corresponding measurement. Hence, the aforementioned loss can be viewed as the price of self-averaging of the process accounting for the low-dimensional nature of the covariate sample space and the high inter-dependence of the covariates to the observation sequence. Recent results on M-estimation of sparse MVAR processes with sub-Gaussian and heavy-tailed innovations [224, 227] require $n \sim \mathcal{O}(s^2(\log p)^2)$ when specialized to the univariate case, which compared to our results improve the loss of $\mathcal{O}((p/\log p)^{1/2})$ to $(\log p)^2$ with the additional cost of quadratic requirement in the sparsity s. However, in the over-determined regime of $s \propto p^{\frac{1}{2}+\delta}$ for some $\delta \geq 0$, our results imply $n \sim \mathcal{O}(p^{1+\delta}(\log p)^{1/2})$, providing a saving of order $p^{\delta}(\log p)^{3/2}$ over those of [224, 227].

Remark 4. It can be shown that the estimation error for the LS method in general scales as $\sqrt{p/n}$ [99] which is not desirable when $n \ll p$. Our result, however, guarantees a much smaller error rate of the order $\sqrt{s \log p/n}$. Also, the sufficiency conditions of Theorem 6 require high compressibility of the parameter vector $\boldsymbol{\theta}$ ($\xi < 1/2$), whereas Theorem 3 does not impose any extra restrictions on $\xi \in (0, 1)$. Intuitively speaking, these two comparisons reveal the trade-off between computational complexity and measurement/compressibility requirements for convex optimization vs. greedy techniques, which are well-known for linear models [41].

Remark 5. The condition $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$ in Theorem 3 is not restricting for the processes of interest in this chapter. This is due to the fact that the boundedness assumption on the spectral spread implies an exponential decay of the parameters (See Lemma 1 of [89]). Finally, the constants d_1, d_1' are increasing with respect to the spectral spread of the process ρ. Intuitively speaking, the closer the roots of the filter given by (2.2) get to the unit circle (corresponding to larger ρ and smaller η), the slower the convergence of the process will be to its ergodic state, and hence more measurements are required. A similar dependence to the spectral spread has appeared in the results of [89] for ℓ_2-regularized least squares estimation of AR processes.

Remark 6. The main ingredient in the proofs of Theorems 3 and 6 is to establish the restricted eigenvalue (RE) condition introduced in [31] for the covariates matrix $\mathbf{X}$. Establishing the RE condition for the covariates matrix $\mathbf{X}$ is a nontrivial problem due to the high interdependence of the matrix entries. We will indeed show that if the sufficient stability assumption holds, then with $n \sim \mathcal{O}\left(s \max\{d_0(\log p)^2, d_1(p\log p)^{1/2}\}\right)$ the sample covariance matrix is sharply concentrated around the true covariance matrix and hence the RE condition can be guaranteed. All constants appearing in Theorems 3 and 6 are explicitly given in Appendix A.1.2. As a typical numerical example, for $\eta = 0.9$ and $\sigma_w^2 = 0.1$, the constants of Theorem 3 can be chosen as

$d_0 \approx 1000, d_1 \approx 3 \times 10^8, d_2 \approx 0.15, d_3 \approx 140$, and $d_4 = 1$. The full proofs are given in Appendix A.1.2.

2.3.4 Minimax optimality

In this section, we establish the minimax optimality of the ℓ_1-regularized LS estimator for AR processes with sparse parameters. To this end, we will focus on the class $\mathcal{H}$ of stationary processes which admit an AR(p) representation with s-sparse parameter $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$. The theoretical results of this section are inspired by the results of [89] on non-asymptotic order selection via ℓ_2-regularized LS estimation in the absence of sparsity, and extend them by studying the ℓ_1-regularized LS estimator of (2.9).

We define the maximal *estimation* risk over $\mathcal{H}$ to be

$$\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}) := \sup_{\mathcal{H}} \left(\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \right] \right)^{1/2}. \tag{2.15}$$

The minimax estimator is the one minimizing the maximal estimation risk, i.e.,

$$\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}} := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\min} \quad \mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}). \tag{2.16}$$

Minimax estimator $\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}$, in general, cannot be constructed explicitly [89], and the common practice in non-parametric estimation is to construct an estimator $\widehat{\boldsymbol{\theta}}$ which is *order optimal* as compared to the minimax estimator:

$$\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}) \leq L \mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}). \tag{2.17}$$

with $L \geq 1$ being a constant. One can also define the minimax *prediction* risk by the maximal prediction error over all possible realizations of the process:

$$\mathcal{R}^2_{\mathsf{pre}}(\widehat{\boldsymbol{\theta}}) := \sup_{\mathcal{H}} \mathbb{E}\left[\left(x_k - \widehat{\boldsymbol{\theta}}'\mathbf{x}_{k-p}^{k-1}\right)^2\right]. \tag{2.18}$$

In [89], it is shown that an ℓ_2-regularized LS estimator with an order $p^\star = \mathcal{O}(\log n)$ is minimax optimal. This order pertains to the denoising regime where $n \gg p$. Hence, in order to capture long order lags of the process, one requires a sample size exponentially large in p, which may make the estimation problem computationally infeasible. For instance, consider a 2-sparse parameter with only θ_1 and θ_p being non-zero. Then, in order to achieve minimax optimality, $n \sim \mathcal{O}(2^p)$ measurements are required. In contrast, in the compressive regime where $s, n \ll p$, the goal, instead of selecting p, is to find conditions on the sparsity level s, so that for a given n and large enough p, the ℓ_1-regularized estimator is minimax optimal without explicit knowledge of the value of s (See for example, [46]).

In the following proposition, we establish the minimax optimality of the ℓ_1-regularized estimator over the class of sparse AR processes with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

Proposition 3. *Let $\mathbf{x}_1^n$ be samples of an AR process with s-sparse parameters satisfying $\|\boldsymbol{\theta}\|_1 \leq 1-\eta$ and $s \leq \min\left\{\frac{1-\eta}{\sqrt{8\pi}\eta}\sqrt{\frac{n}{\log p}}, \frac{n}{d_1(p\log p)^{1/2}}, \frac{n}{d_0(\log p)^2}\right\}$. Then, we have:*

$$\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\ell_1}) \leq L\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}).$$

where L is a constant and is only a function of η and σ^2_{w} and is explicitly given in the proof.

Remark 5. Proposition 3 implies that ℓ_1-regularized LS is minimax optimal in estimating the s-sparse parameter vector $\boldsymbol{\theta}$, for small enough s. The proof of the Proposition 3 is given in Appendix A.1.4. This result can be extended to compressible $\boldsymbol{\theta}$ in a natural way with a bit more work, but we only present the proof for the case of s-sparse $\boldsymbol{\theta}$ for brevity. We also state the following proposition on the prediction performance of the ℓ_1-regularized LS estimator:

Proposition 4. *Let $\mathbf{x}_{-p+1}^{n}$ be samples of an AR process with s-sparse parameters and Gaussian innovations, then there exists a positive constant d_5 such that for large enough n, p and s satisfying $n > d_1 s(p \log p)^{1/2}$, we have:*

$$\mathcal{R}_{\mathsf{pre}}^2(\widehat{\boldsymbol{\theta}}_{\ell_1}) \leq d_5 \frac{s \log p}{n} + \sigma_{\mathsf{w}}^2. \qquad (2.19)$$

It can be readily observed that for $n \gg s \log p$ the prediction error variance is very close to the variance of the innovations. The proof is similar to Theorem 3 of [89] and is skipped in this chapter for brevity.

2.4 Application to Simulated and Real Data

In this section, we study and compare the performance of Yule-Walker based estimation methods with those of the ℓ_1-regularized and greedy estimators given in Section 2.3. These methods are applied to simulated data as well as real data from crude oil price and traffic speed.

2.4.1 Simulation studies

In order to simulate an AR process, we filtered a Gaussian white noise process using an IIR filter with sparse parameters. Figure 2.1 shows a typical sample path of the simulated AR process used in our analysis. For the parameter vector $\boldsymbol{\theta}$, we chose a length of $p = 300$, and employed $n = 1500$ generated samples of the corresponding process for estimation. The parameter vector $\boldsymbol{\theta}$ is of sparsity level $s = 3$ and $\eta = 1 - \|\boldsymbol{\theta}\|_1 = 0.5$. A value of $\gamma_n = 0.1$ is used, which is slightly tuned around the theoretical estimate given by Theorem 3. The order of the process is assumed to be known. We compare the performance of seven estimators: 1) $\widehat{\boldsymbol{\theta}}_{\mathsf{LS}}$ using LS, 2) $\widehat{\boldsymbol{\theta}}_{\mathsf{yw}}$ using the Yule-Walker equations, 3) $\widehat{\boldsymbol{\theta}}_{\ell_1}$ from ℓ_1-regularized LS, 4) $\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$ using OMP, 5) $\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{2,1}}$ using Eq. (2.11), 6) $\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{1,1}}$ using Eq. (2.12), and 7) $\widehat{\boldsymbol{\theta}}_{\mathsf{ywOMP}}$ using the cost function $\mathfrak{J}(\boldsymbol{\theta})$ in the generalized OMP. Note that for the LS and Yule-Walker estimates, we have relaxed the condition of $\|\boldsymbol{\theta}\|_1 < 1$, to be consistent with the common usage of these methods. The Yule-Walker estimate is guaranteed to result in a stable AR process, whereas the LS estimate is not [160]. Figure 2.2 shows the estimated parameter vectors using these algorithms. It can be visually observed that ℓ_1-regularized and greedy estimators (shown in purple) significantly outperform the Yule-Walker-based estimates (shown in orange).

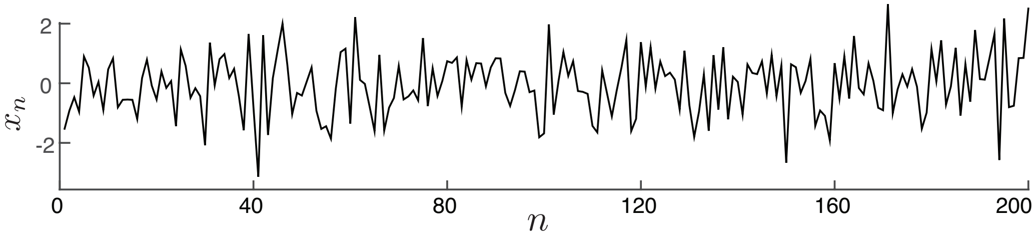


Figure 2.1: Samples of the simulated AR process.

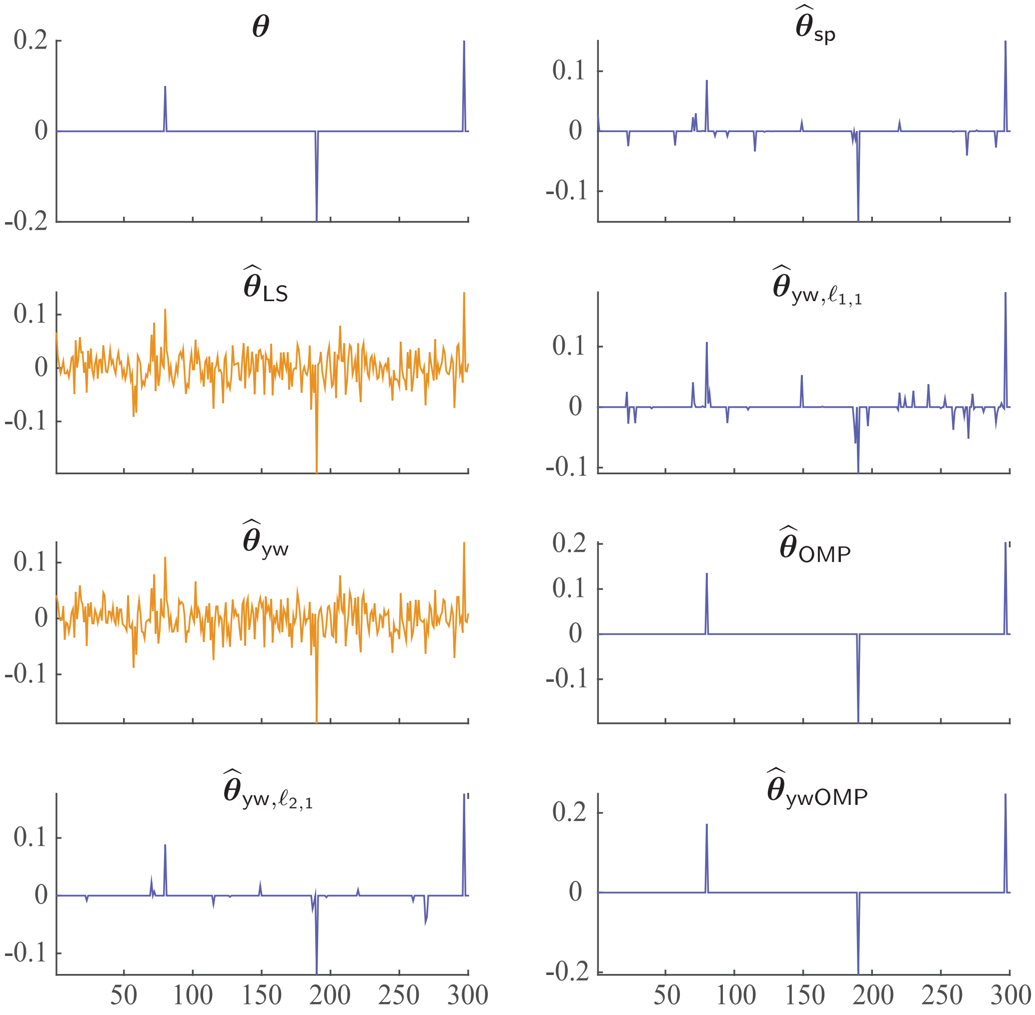


Figure 2.2: Estimates of $\boldsymbol{\theta}$ for $n = 1500$, $p = 300$, and $s = 3$ (These results are best viewed in the color version).

In order to quantify the latter observation precisely, we repeated the same experiment for $p = 300, s = 3$ and $10 \leq n \leq 10^5$. A comparison of the normalized MSE of the estimators vs. n is shown in Figure 2.3. As it can be inferred from Figure 2.3, in the region where n is comparable to or less than p (shaded in light purple), the sparse estimators have a systematic performance gain over the Yule-Walker based estimates, with the ℓ_1-regularized LS and ywOMP estimates outperforming the rest.

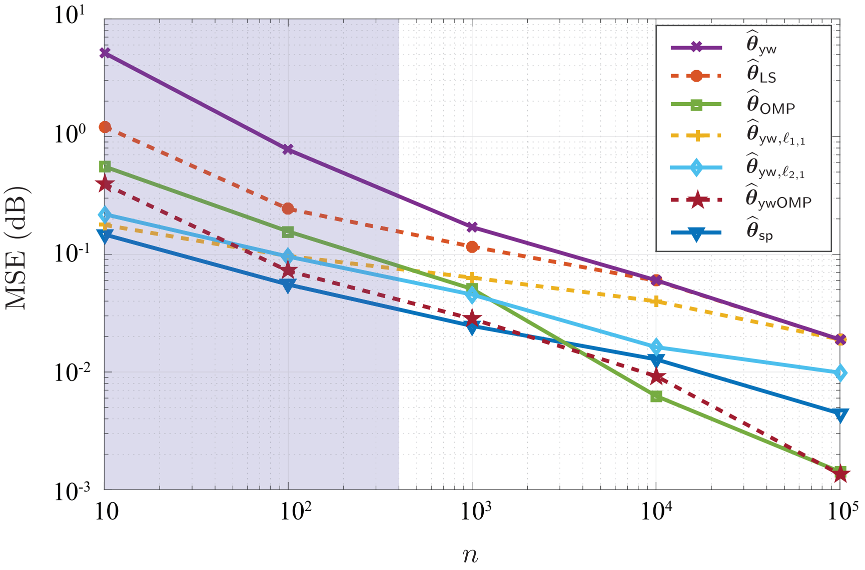


Figure 2.3: MSE comparison of the estimators vs. the number of measurements n. The shaded region corresponds to the compressive regime of $n < p$.

The MSE comparison in Figure 2.3 requires one to know the true parameters. In practice, the true parameters are not available for comparison purposes. In order to quantify the performance gain of these methods, we use statistical tests to assess the goodness-of-fit of the estimates. The common chi-square type statistical tests, such as the F-test, are useful when the hypothesized distribution to be tested against is discrete or categorical. For our problem setup with sub-Gaussian innovations, we will use a number of statistical tests appropriate for AR processes, namely, the Kolmogorov-Smirnov (KS) test, the Cramér-von Mises (CvM) criterion, the spectral Cramér-von Mises (SCvM) test and the Anderson-Darling (AD) [15, 65, 107]. A summary of these tests is given in Appendix B.1. Table 2.2 summarizes the test statistics for different estimation methods. Cells colored in orange (darker shade in grayscale) correspond to traditional AR estimation methods and those colored in blue (lighter shade in grayscale) correspond to the sparse estimator with the best performance among those considered in this work. These tests are based on the known results on limiting distributions of error residuals. As noted from Table 2.2, our simulations suggest that the OMP estimate achieves the best test statistics for the CvM, AD and KS tests, whereas the ℓ_1-regularized estimate achieves the best SCvM statistic.

Table 2.2: Goodness-of-fit tests for the simulated data

Estimate \ Test	CvM	AD	KS	SCvM
$\boldsymbol{\theta}$	0.31	1.54	0.031	0.009
$\widehat{\boldsymbol{\theta}}_{\mathsf{LS}}$	0.68	5.12	0.037	0.017
$\widehat{\boldsymbol{\theta}}_{\mathsf{yw}}$	0.65	4.87	0.034	0.025
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	0.34	1.72	0.030	0.009
$\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$	0.29	1.45	0.028	0.009
$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{2,1}}$	0.35	1.80	0.032	0.009
$\widehat{\boldsymbol{\theta}}_{\mathsf{yw},\ell_{1,1}}$	0.42	2.33	0.040	0.008
$\widehat{\boldsymbol{\theta}}_{\mathsf{ywOMP}}$	0.29	1.46	0.030	0.009

2.4.2 Application to the analysis of crude oil prices

In this and the following subsection, we consider applications with real-world data. As for the first application, we apply the sparse AR estimation techniques to analyze the crude oil price of the Cushing, OK WTI Spot Price FOB dataset [1]. This dataset consists of 7429 daily values of oil prices in dollars per barrel. In order to avoid outliers, usually the dataset is filtered with a moving average filter of high order. We have skipped this procedure by visual inspection of the data and selecting $n = 4000$ samples free of outliers. Such financial data sets are known for their non-stationarity and long order history dependence. In order to remove the deterministic trends in the data, one-step or two-step time differencing is typically used. We refer to [178] for a full discussion of this detrending method. We have used a first-order time differencing which resulted in a sufficient detrending of the data. Figure 2.4 shows the data used in our analysis. We have chosen $p = 150$ by inspection. The histogram of first-order differences as well the estimates are shown in Figure 2.5.

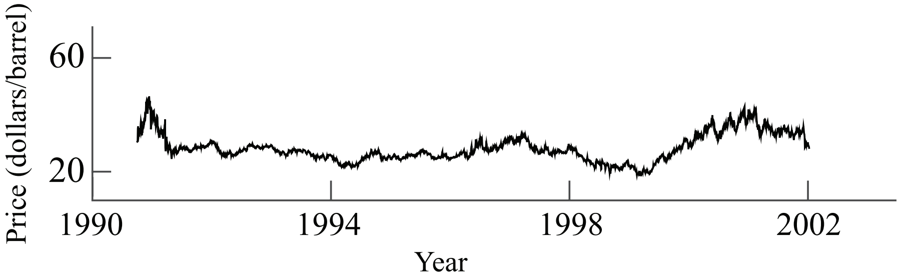


Figure 2.4: A sample segment of the Cushing, OK WTI Spot Price FOB data.

A visual inspection of the estimates in Figure 2.5 shows that the ℓ_1-regularized LS ($\widehat{\boldsymbol{\theta}}_{\ell_1}$) and OMP ($\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$) estimates consistently select specific time lags in the AR

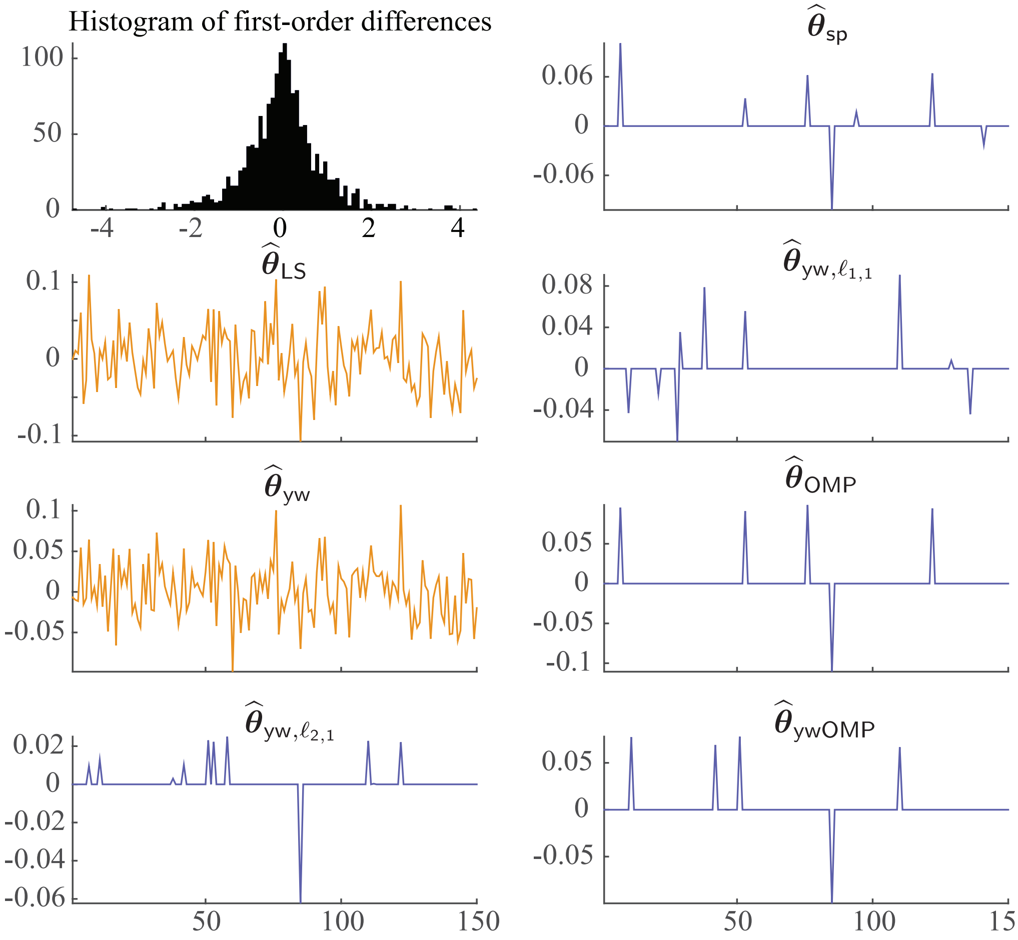


Figure 2.5: Estimates of $\boldsymbol{\theta}$ for the second-order differences of the oil price data.

parameters, whereas the Yule-Walker and LS estimates seemingly overfit the data by populating the entire parameter space. In order to perform goodness-of-fit tests, we use an even/odd two-fold cross-validation. Table 2.3 shows the corresponding test statistics, which reveal that indeed the ℓ_1-regularized and OMP estimates outperform the traditional estimation techniques.

Table 2.3: Goodness-of-fit tests for the crude oil price data

Test Estimate	CvM	AD	KS	SCvM
$\widehat{\boldsymbol{\theta}}_{\text{LS}}$	0.88	5.55	0.055	0.046
$\widehat{\boldsymbol{\theta}}_{\text{yw}}$	0.58	3.60	0.043	0.037
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	0.27	1.33	0.031	0.020
$\widehat{\boldsymbol{\theta}}_{\text{OMP}}$	0.22	1.18	0.025	0.022
$\widehat{\boldsymbol{\theta}}_{\text{yw},\ell_{2,1}}$	0.28	1.40	0.027	0.021
$\widehat{\boldsymbol{\theta}}_{\text{yw},\ell_{1,1}}$	0.24	1.26	0.027	0.022
$\widehat{\boldsymbol{\theta}}_{\text{ywOMP}}$	0.23	1.18	0.026	0.022

2.4.3 Application to the analysis of traffic data

Our second real data application concerns traffic speed data. The data used in our simulations is the INRIX ® speed data for I-495 Maryland inner loop freeway (clockwise) between US-1/Baltimore Ave/Exit 25 and Greenbelt Metro Dr/Exit 24 from 1 Jul, 2015 to 31 Oct, 2015 [2, 3]. The reference speed of 65 mph is reported. Our aim is to analyze the long-term, large-scale periodicities manifested in these data by fitting high-order sparse AR models. Given the huge length of the data and its high variability, the following pre-processing was made on the original data:

1. The data was downsampled by a factor of 4 and averaged by the hour in order to reduce its daily variability, that is each lag corresponds to one hour.

2. The logarithm of speed was used for analysis and the mean was subtracted. This reduces the high variability of speed due to rush hours and lower traffic during weekends and holidays.

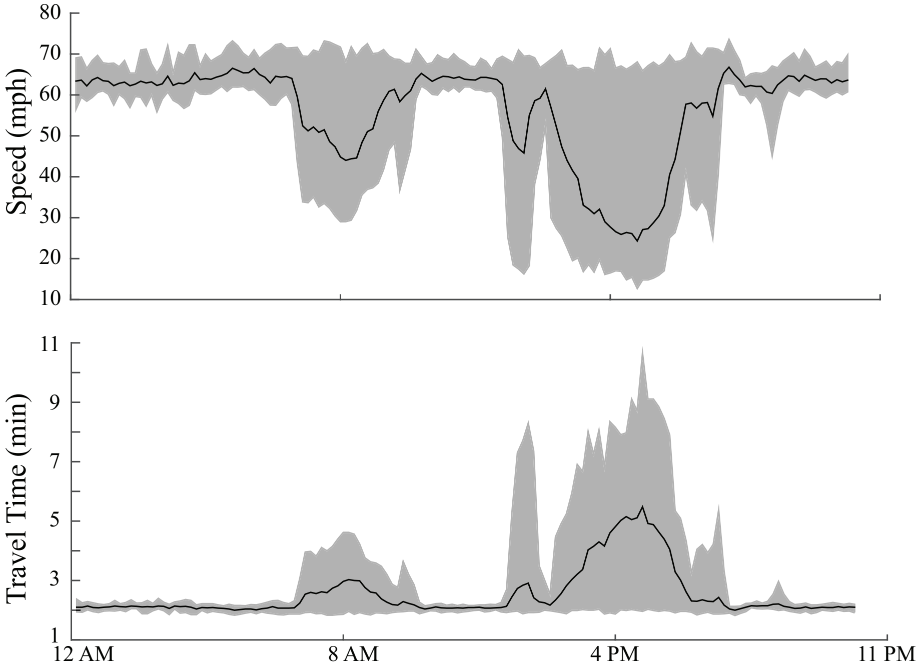


Figure 2.6: A sample of the speed and travel time data for I-495.

Figure 2.6 shows a typical average weekly speed and travel time in this dataset and the corresponding 25-75-th percentiles. As can be seen the data shows high variability around the rush hours of 8 am and 4 pm. In our analysis, we used the first half of the data ($n = 1500$) for fitting, from which the AR parameters and the distribution and variance of the innovations were estimated. The statistical tests were designed based on the estimated distributions, and the statistics were computed accordingly using the second half of the data. We selected an order of $p = 200$ by inspection and noting that the data seems to have a periodicity of order 170 samples.

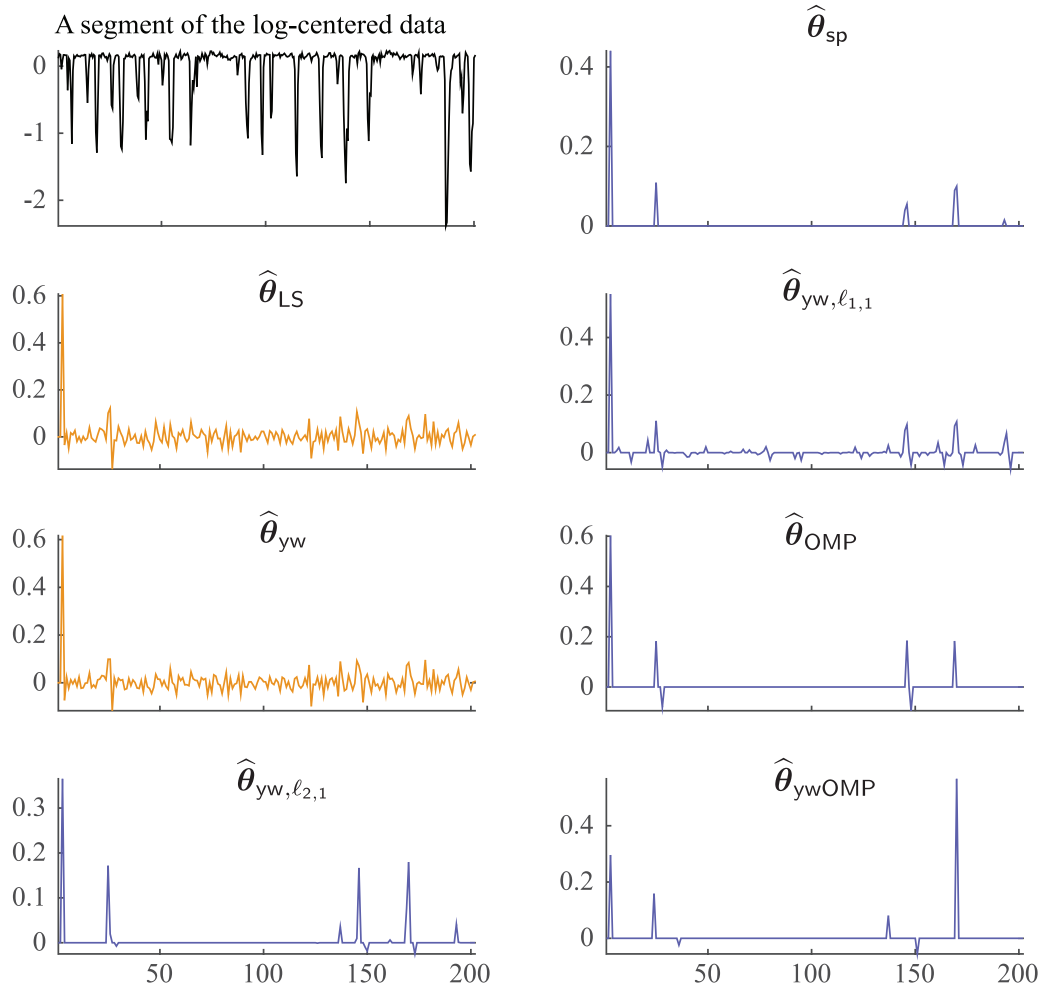


Figure 2.7: Estimates of $\boldsymbol{\theta}$ for the traffic speed data.

Figure 2.7 shows part of the data used in our analysis as well as the estimated parameters. The ℓ_1-regularized LS ($\widehat{\boldsymbol{\theta}}_{\ell_1}$) and OMP ($\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$) are consistent in selecting the same components of $\boldsymbol{\theta}$. These estimators pick up two major lags around which $\boldsymbol{\theta}$ has its largest components. The first lag corresponds to about 24 hours which is mainly due to the rush hour periodicity on a daily basis. The second lag is around

150 − 170 hours which corresponds to weekly changes in the speed due to lower traffic over the weekend. In contrast, the Yule-Walker and LS estimates do not recover these significant time lags.

Table 2.4: Goodness-of-fit tests for the traffic speed data

Test / Estimate	CvM	AD	KS	SCvM
$\widehat{\boldsymbol{\theta}}_{\mathsf{yw}}$	0.012	0.066	0.220	0.05
$\widehat{\boldsymbol{\theta}}_{\ell_1}$	1.4×10^{-7}	2.1×10^{-6}	6.7×10^{-4}	0.25
$\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}$	0.017	0.082	0.220	1.49
$\widehat{\boldsymbol{\theta}}_{\mathsf{ywOMP}}$	0.025	0.122	0.270	0.14

Statistical tests for a selected subset of the estimators are shown in Table 2.4. Interestingly, the ℓ_1-regularized LS estimator significantly outperforms the other estimators in three of the tests. The Yule-Walker estimator, however, achieves the best SCvM test statistic.

2.5 Concluding Remarks

In this chapter, we investigated sufficient sampling requirements for stable estimation of AR models in the non-asymptotic regime using the ℓ_1-regularized LS and greedy estimation (OMP) techniques. We have further established the minimax optimality of the ℓ_1-regularized LS estimator. Compared to the existing literature, our results provide several major contributions. First, when $s \sim p^{\frac{1}{2}+\delta}$ for some $\delta \geq 0$, our results suggest an improvement of order $\mathcal{O}(p^{\delta}(\log p)^{3/2})$ in the sampling requirements for the estimation of univariate AR models with sub-Gaussian innovations using the LASSO, over those of [224] and [227] which require $n \sim \mathcal{O}(p^2(\log p)^2)$ for stable AR estimation.

When specialized to a sub-Gaussian white noise process, i.e., establishing the RE condition of i.i.d. Toeplitz matrices, our results provide an improvement of order $\mathcal{O}(s/\log p)$ over those of [99]. Second, although OMP is widely used in practice, the choice of the number of greedy iterations is often ad-hoc. In contrast, our theoretical results prescribe an analytical choices of the number of iterations required for stable estimation, thereby promoting the usage of OMP as a low-complexity algorithm for AR estimation. Third, we established the minimax optimality of the ℓ_1-regularized LS estimator for the estimation of sparse AR parameters.

We further verified the validity of our theoretical results through simulation studies as well as application to real financial and traffic data. These results show that the sparse estimation methods significantly outperform the widely-used Yule-Walker based estimators in fitting AR models to the data. Although we did not theoretically analyze the performance of sparse Yule-Walker based estimators, they seem to perform on par with the ℓ_1-regularized LS and OMP estimators based on our numerical studies. Finally, as we will see in the next chapter our results provide a striking connection to estimation of sparse self-exciting discrete point process models. These models regress an observed binary spike train with respect to its history via Bernoulli or Poisson statistics, and are often used in describing spontaneous activity of sensory neurons. Our results have shown that in order to estimate a sparse history-dependence parameter vector of length p and sparsity s in a stable fashion, a spike train of length $n \sim \mathcal{O}(s^{2/3}p^{2/3}\log p)$ is required. This leads us to conjecture that these sub-linear sampling requirements are sufficient for a larger class of autoregressive processes, beyond those characterized by linear models. Finally, our minimax optimality result requires the sparsity level s to grow at most as fast as $\mathcal{O}(n/(p \log p)^{1/2})$. We consider further relaxation of this condition, as well as the generalization of our results to sparse MVAR processes as future work.

Chapter 3: Robust Estimation of Self-Exciting Generalized Linear Models

In this chapter, we close the gap in theory of compressed sensing for non-i.i.d. data by providing theoretical guarantees on stable estimation of self-exciting generalized linear models. We consider the problem of estimating self-exciting generalized linear models from limited binary observations, where the history of the process serves as the covariate. In doing so, we relax the assumptions of i.i.d. covariates and exact sparsity common in CS. Our results indicate that utilizing sparsity recovers important information about the intrinsic frequencies of such processes. We analyze the performance of two classes of estimators, namely the ℓ_1-regularized maximum likelihood and greedy estimators, for a canonical self-exciting process and characterize the sampling tradeoffs required for stable recovery in the non-asymptotic regime. Our results extend those of compressed sensing for linear and generalized linear models with i.i.d. covariates to those with highly inter-dependent covariates. We further provide simulation studies as well as application to real spiking data from the mouse's lateral geniculate nucleus and the ferret's retinal ganglion cells under different nonlinear forward models which agree with our theoretical predictions.

3.1 Introduction

The theory of compressed sensing (CS) has provided a novel framework for measuring and estimating statistical models governed by sparse underlying parameters [40, 45, 49, 50, 71, 145]. In particular, for linear models with random covariates and sparsity of the parameters, the CS theory provides sharp trade-offs between the number of measurement, sparsity, and estimation accuracy. Typical theoretical guarantees imply that when the number of random measurements are roughly proportional to sparsity, then stable recovery of these sparse models is possible.

Beyond those described by linear models, observations from binary phenomena form a large class of data in natural and social sciences. Their ubiquity in disciplines such as neuroscience, physiology, seismology, criminology, and finance has urged researchers to develop formal frameworks to model and analyze these data. In particular, the theory of point processes provides a statistical machinery for modeling and prediction of such phenomena. Traditionally, these models have been employed to predict the likelihood of self-exciting processes such as earthquake occurrences [153, 217], but have recently found applications in several other areas. For instance, these models have been used to characterize heart-beat dynamics [23, 210] and violence among gangs [75]. Self-exciting point process models have also found significant applications in analysis of neuronal data [37, 38, 155, 156, 163, 191, 207].

In particular, point process models provide a principled way to regress binary spiking data with respect to extrinsic stimuli and neural covariates, and thereby forming predictive statistical models for neural spiking activity. Examples include place cells in the hippocampus [38], spectro-temporally tuned cells in the primary auditory cortex [44], and spontaneous retinal or thalamic neurons spiking under tuned intrinsic frequencies [33, 129]. Self-exciting point processes have also been utilized in assessing the functional connectivity of neuronal ensembles [59, 115]. When fitted to neuronal data, these models exhibit three main features: first, the underlying parameters are nearly sparse or compressible [59, 206]; second, the covariates are often highly structured and correlated; and third, the input-output relation is highly nonlinear. Therefore, the theoretical guarantees of compressed sensing do not readily translate to prescriptions for point process estimation.

Estimation of these models is typically carried out by Maximum Likelihood (ML) or regularized ML estimation in discrete time, where the process is viewed as a Generalized Linear Model (GLM). In order to adjust the regularization level, empirical methods such as cross-validation are typically employed [59]. In the signal processing and information theory literature, sparse signal recovery under Poisson statistics has been considered in [142] with application to the analysis of ranking data. In [172], a similar setting has been studied, with motivation from imaging by photon-counting devices. Finally, in theoretical statistics, high-dimensional M-estimators with decomposable regularizers, such as the ℓ_1-norm, have been studied for GLMs [148].

A key underlying assumption in the existing theoretical analysis of estimating GLMs is the independence and identical distribution (i.i.d.) of covariates. This assumption does not hold for self-exciting processes, since the history of the process takes the role of the covariates. Nevertheless, regularized ML estimators show remarkable performance in fitting GLMs to neuronal data with history dependence and highly non-i.i.d. covariates. In this chapter, we close this gap by presenting new results on robust estimation of compressible GLMs, relaxing the assumptions of i.i.d. covariates and exact sparsity common in CS.

In particular, we will consider a canonical GLM and will analyze two classes of estimators for its underlying parameters: the ℓ_1-regularized maximum likelihood and greedy estimators. We will present theoretical guarantees that extend those of CS theory and characterize fundamental trade-offs between the number of measurements, model compressibility, and estimation error of GLMs in the non-asymptotic regime. Our results reveal that when the number of measurements scale sub-linearly with the product of the ambient dimension and a generalized measure of sparsity (modulo logarithmic factors), then stable recovery of the underlying models is possible, even though the covariates solely depend on the history of the process. We will further discuss the extensions of these results to more general classes of GLMs. Finally, we will present applications to simulated as well as real data from two classes of neurons exhibiting spontaneous activity, namely the mouse's lateral geniculate nucleus and the ferret's retinal ganglion cells, which agree with our theoretical predictions. Aside from their theoretical significance, our results are particularly important in light of the technological advances in neural prostheses, which require robust neuronal system identification based on compressed data acquisition.

The rest of the chapter is organized as follows: In Section 3.2, we present our notational conventions, preliminaries and problem formulation. In Section 3.3, we discuss the estimation procedures and state the main theoretical results of this chapter. Section 3.4 provides numerical simulations as well as application to real data. In Section 3.5, we discuss the implications of our results and outline future research directions. Finally, we present the proofs of the main theoretical results and give a brief background on relevant statistical tests in Appendices A.2–B.2.

3.2 Preliminaries and Problem Formulation

We first give a brief introduction to self-exciting GLMs (see [66] for a detailed treatment).

We consider a sequence of observations in the form of binary spike trains obtained by discretizing continuous-time observations (e.g. electrophysiology recordings), using bins of length Δ. We assume that not more than one event fall into any given bin. In practice, this can always be achieved by choosing Δ small enough. The binary observation at bin i is denoted by x_i. The observation sequence can be modeled as the outcome of conditionally independent Poisson or Bernoulli trials, with a spiking probability given by $\mathbb{P}(x_i = 1) =: \lambda_{i|H_i}$, where $\lambda_{i|H_i}$ is the spiking probability at bin i given the history of the process H_i up to bin i.

These models are widely-used in neural data analysis and are motivated by the continuous time point processes with history dependent conditional intensity functions [66]. For instance, given the history of a continuous-time point process H_t up to time t, a conditional intensity of $\lambda(t|H_t) = \lambda$ corresponds to the homogeneous Poisson process. As another example, a conditional intensity of $\lambda(t|H_t) = \mu + \int_{-\infty}^{t} \theta(t-\tau) dN(\tau)$ corresponds to a process known as the Hawkes process [100] with base-line rate μ and history dependence kernel $\theta(\cdot)$. Under the assumption of the orderliness of a continuous-time point process, a discretized approximation to these processes can be obtained by binning the process by bins of length Δ, and defining the spiking probability by $\lambda_i := \lambda(i\Delta|H_{i\Delta})\Delta + o(\Delta)$. In this chapter, we consider discrete random processes characterized by the spiking probability $\lambda_{i|H_i}$, which are either inherently discrete or employed as an approximation to continuous-time point process models.

Throughout the rest of the chapter, we drop the dependence of $\lambda_{i|H_i}$ on H_i to simplify notation, denote it by λ_i and refer to it as spiking probability. Given the sequence of binary observed data $\mathbf{x}_1^n$, the negative log-likelihood function under the Bernoulli statistics can be expressed as:

$$\mathfrak{L}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \left\{x_i \log \lambda_i + (1-x_i)\log(1-\lambda_i)\right\}. \tag{3.1}$$

Another common likelihood model used in the analysis of neuronal spiking data corresponds to Poisson statistics [206], for which the negative log-likelihood takes the following form:

$$\mathfrak{L}(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^{n} \left\{x_i \log \lambda_i - \lambda_i\right\}. \tag{3.2}$$

Throughout the chapter, we will focus on binary observations governed by Bernoulli statistics, whose negative log-likelihood is given in Eq. (3.1). In applications such as electrophysiology in which neuronal spiking activities are recorded at a high sampling rate, the binning size Δ is very small and the Bernoulli and Poisson statistics coincide.

When the discrete process is viewed as an approximation to a continuous-time process, these log-likelihood functions are known as the Jacod log-likelihood approximations [66]. We will present our analysis for the negative log-likelihood given by (3.1), but our results can be extended to other statistics including (3.2) (See the remarks of Section 3.3).

Throughout this chapter $\mathbf{x}_{-p+1}^{n}$ will be considered as the observed spiking sequence which will be used for estimation purposes. A popular class of models for λ_i is given by GLMs. In its general form, a GLM consists of two main components: an observation model and an equation expressing some (possibly nonlinear) function of the observation mean as a *linear* combination of the covariates. In neural systems, the covariates consist of external stimuli as well as the history of the process. Inspired by spontaneous neuronal activity, we consider *fully* self-exciting processes, in which the covariates are only functions of the process history. As for a canonical GLM inspired by the Hawkes process, we consider a process for which the spiking probability is a *linear* function of the process history:

$$\lambda_i := \mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1}, \tag{3.3}$$

where μ is a positive constant representing the base-line rate, and $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]'$ is a parameter vector denoting the history dependence of the process. We further assume that the process is non-degenerate, i.e., it will not terminate in an infinite sequence of zeros. We refer to this GLM, viewed as a random process, as the *canonical self-exciting process*. Other popular models in the computational neuroscience literature include the log-link model where $\lambda_i = \exp(\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1})$ and the logistic-link model where $\lambda_i = \frac{\exp(\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1})}{1+\exp(\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1})}$. The parameter vector $\boldsymbol{\theta}$ can be thought of as the binary equivalent of autoregressive (AR) parameters in linear AR models.

When fitted to neuronal spiking data, the parameter vector $\boldsymbol{\theta}$ exhibits a degree of sparsity [59,206], that is, only certain lags in the history have a significant contribution in determining the statistics of the process. These lags can be thought of as the preferred or intrinsic delays in the spontaneous response of a neuron. To be more precise, for a sparsity level $s < p$, we denote by $\boldsymbol{\theta}_s$ the best s-term approximation to $\boldsymbol{\theta}$.

Finally, in this chapter, we are concerned with the compressed sensing regime where $n \ll p$, i.e., the observed data has a much smaller length than the ambient dimension of the parameter vector. The main estimation problem of this chapter is the following: *given observations $\mathbf{x}_{-p+1}^{n}$ from the canonical self-exciting process, the goal is to estimate the unknown baseline rate μ and the p-dimensional (s, ξ)-compressible history dependence parameter vector $\boldsymbol{\theta}$ in a stable fashion (where the estimation error is controlled) when $n \ll p$.*

3.3 Theoretical Results

In this section, we consider two estimators for $\boldsymbol{\theta}$, namely, the ℓ_1-regularized ML estimator and a greedy estimator, and present the main theoretical results of this chapter on the estimation error of these estimators. Note that when μ is not known, the following results can be applied to the augmented parameter vector $[\mu, \boldsymbol{\theta}']'$. We analyze the case of known μ for simplicity of presentation.

3.3.1 ℓ_1-Regularized ML Estimation

The natural estimator for the parameter vector is the ML estimator, which is widely used in neuroscience [206], which by virtue of (3.1) is given by:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{ML}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathfrak{L}(\boldsymbol{\theta}), \tag{3.4}$$

where $\boldsymbol{\Theta}$ is the relaxed closed convex feasible region for which $0 \leq \lambda_i \leq 1$ given by the conditions:

$$\begin{aligned} &0 < \pi_{\min} \leq \mu - \mathbf{1}'\boldsymbol{\theta}^-, \\ &\mu + \mathbf{1}'\boldsymbol{\theta}^+ \leq \pi_{\max} < 1/2, \end{aligned} \tag{$\star$}$$

for some constants $\pi_{\min}$ and $\pi_{\max}$. This first inequality incurs minimal loss of generality, as $\pi_{\min}$ can be chosen to be arbitrarily small. The restriction of $\pi_{\max} < 1/2$ ensures that the process is fast mixing and has mainly been adopted for technical convenience. This assumption incurs some loss of generality, as it excludes processes for which the maximum spiking probability exceeds 1/2. However, due to the low spiking probability of typical neuronal activity, this loss is tolerable for the applications of interest in this chapter (see Section 3.4).

In the regime of interest when $n \ll p$, the ML estimator is ill-posed and is typically regularized with a smooth norm. In order to capture the compressibility of the parameters, we consider the ℓ_1-regularized ML estimator:

$$\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \mathfrak{L}(\boldsymbol{\theta}) + \gamma_n \|\boldsymbol{\theta}\|_1. \tag{3.5}$$

where $\gamma_n > 0$ is a regularization parameter. It is easy to verify that the objective function and constraints in Eq. (3.5) are convex in $\boldsymbol{\theta}$ and hence $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$ can be obtained using standard numerical solvers. Note that the solution to (3.5) might not be unique. However, we will provide error bounds that hold for all possible solutions of (3.5), with high probability.

It is known that ML estimates are asymptotically unbiased under mild conditions, and with p fixed, the solution converges to the true parameter vector as $n \to \infty$. However, it is not clear how fast the convergence rate is for finite n or when p is not fixed and is allowed to scale with n. This makes the analysis of ML estimators, and in general regularized M-estimators, very challenging [148]. Nevertheless, such an analysis has significant practical implications, as it will reveal sufficient conditions on n with respect to p as well as a criterion to choose γ_n, which result in a stable estimation of $\boldsymbol{\theta}$. Finally, note that we are fixing the ambient dimension p throughout the analysis. In practice, the history dependence is typically negligible beyond a certain lag and hence for a large enough p, GLMs fit the data very well.

3.3.2 Greedy Estimation

Although there exist fast solvers to convex problems of the type given by Eq. (3.5), these algorithms are polynomial time in n and p, and may not scale well with high-dimensional data. This motivates us to consider greedy solutions for the estimation of $\boldsymbol{\theta}$. In particular, we will consider a generalization of the Orthogonal Matching Pursuit (OMP) [158, 235] for general convex cost functions. A flowchart of this algorithm is given in Table 3.1, which we denote by the Point Process Orthogonal

```
Input: $\mathfrak{L}(\boldsymbol{\theta}), s^\star$
Output: $\widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}}$
Initialization: $\begin{cases} \text{Start with the index set } S^{(0)} = \emptyset \\ \text{and the initial estimate } \widehat{\boldsymbol{\theta}}^{(0)}_{\mathsf{POMP}} = 0 \end{cases}$
for $k = 1, 2, \cdots, s^\star$
  $j = \arg\max_{i} \left| \left( \nabla \mathfrak{L} \left( \widehat{\boldsymbol{\theta}}^{(k-1)}_{\mathsf{POMP}} \right) \right)_i \right|$
  $S^{(k)} = S^{(k-1)} \cup \{j\}$
  $\widehat{\boldsymbol{\theta}}^{(k)}_{\mathsf{POMP}} = \underset{\text{supp}(\boldsymbol{\theta}) \subset S^{(k)}}{\arg\min} \mathfrak{L}(\boldsymbol{\theta})$
end
```

Table 3.1: Point Process Orthogonal Matching Pursuit (POMP)

Matching Pursuit (POMP) algorithm. At each iteration, the component in which the objective function has the largest deviation is chosen and added to the current support. The algorithm proceeds for a total of $s^\star$ steps, resulting in an estimate with $s^\star$ components.

The main idea behind the generalized OMP is in the greedy selection stage, where the absolute value of the gradient of the cost function at the current solution is considered as the selection metric. Consider an estimate $\widehat{\boldsymbol{\theta}}^{(k-1)}$ at the $(k-1)$-st stage of the generalized OMP for a quadratic cost function of the form $\|\mathbf{b} - \mathbf{A}\boldsymbol{\theta}\|_2^2$, with $\mathbf{b}$ and $\mathbf{A}$ denoting the observation vector and covariates matrix, respectively. Then, the gradient takes the form $\mathbf{A}'(\mathbf{b} - \mathbf{A}\widehat{\boldsymbol{\theta}}^{(k-1)})$ which is exactly the correlation vector between the residual error and the columns of $\mathbf{A}$ as in the original OMP algorithm.

3.3.3 Theoretical Guarantees

Recall that the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ is assumed to be (s, ξ)-compressible, so that $\sigma_s(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_S\|_1 = \mathcal{O}(s^{1-\frac{1}{\xi}})$, and the observed data are given by the vector $\mathbf{x}_{-p+1}^n \in \{0,1\}^{n+p-1}$, all in the regime of $s, n \ll p$. In the remainder of this chapter, we assume that $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The main theoretical result regarding the performance of the ℓ_1-regularized ML estimator is given by the following theorem:

Theorem 5. *If $\sigma_s(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, there exist constants d_1, d_2, d_3 and d_4 such that for $n > d_1 s^{2/3} p^{2/3} \log p$ and a choice of $\gamma_n = d_2\sqrt{\frac{\log p}{n}}$, any solution $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$ to (3.5) satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta}\right\|_2 \leq d_3\sqrt{\frac{s \log p}{n}} + \sqrt{d_3 \sigma_s(\boldsymbol{\theta})}\sqrt[4]{\frac{\log p}{n}}, \tag{3.6}$$

with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d_4}}\right)$.

Similarly, the following theorem characterizes the performance bounds for the POMP estimate:

Theorem 6. *If $\boldsymbol{\theta}$ is (s, ξ)-compressible for some $\xi < 1/2$, there exist constants d_1', d_2', d_3' and d_4' such that for $n > d_1' s^{2/3} p^{2/3} (\log s)^{2/3} \log p$, the POMP estimate satisfies the bound*

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}} - \boldsymbol{\theta}\right\|_2 \leq d_2'\sqrt{\frac{s \log s \log p}{n}} + d_3'\frac{\log s}{s^{\frac{1}{\xi}-2}} \tag{3.7}$$

after $s^\star = \mathcal{O}(s \log s)$ iterations with probability greater than $1 - \mathcal{O}\left(\frac{1}{n^{d_4'}}\right)$.

Full proofs of Theorems 5 and 6 are given in Appendix A.2.

Remarks. An immediate comparison of the sufficient condition $n = \mathcal{O}(s^{2/3}p^{2/3}\log p)$ of Theorem 5 with those of [148] for GLM models with i.i.d. covariates given by $n = \mathcal{O}(s \log p)$ reveals that a loss of order $\mathcal{O}(p^{2/3}s^{-1/3})$ is incurred due to the interdependence of the covariates. However, the sample space of n i.i.d. covariates is np-dimensional, whereas in our problem the sample space is only $(n+p)$-dimensional. Hence, the aforementioned loss can be viewed as the price of self-averaging of the process accounting for the low-dimensional nature of the covariate sample space. To the best of our knowledge, the dominant loss of $\mathcal{O}(p^{2/3})$ in both theorems does not seem to be significantly improvable, as self-exciting processes are known to converge quite slowly to their ergodic state [176]. On a related note, the analysis of the sampling requirements of linear AR models reveals a loss of $\mathcal{O}(p^{1/2})$ in the number of measurements [112].

The sufficient condition of Theorem 6 given by $n = \mathcal{O}(s^{2/3}p^{2/3}\left(\log s\right)^{2/3}\log p)$ implies an extra loss of $\left(\log s\right)^{2/3}$ due to the greedy nature of the solution. Theorem 6 also requires a high compressibility level of the parameter vector $\boldsymbol{\theta}$ ($\xi < 1/2$), whereas Theorem 5 does not impose any extra restrictions on $\xi \in (0, 1)$. Intuitively speaking, this comparison reveals the trade-off between computational complexity and compressibility requirements for convex optimization vs. greedy techniques, which is well-known for linear models [40].

The constants d_i, d'_i, $i = 1, \cdots, 4$, α and β are explicitly given in the proof of the theorems in Appendix A.2. As for a typical numerical example, for $\pi_{\min} = 0.01$ and $\pi_{\max} = 0.49$, the constants of Theorem 5 can be chosen as $d_1 \approx 10^3, d_2 = 50, d_3 \approx 10^4$ and $d_4 = 4$.

The main ingredient in the proofs of Theorems 5 and 6 is inspired by the beautiful treatment of Negahban et al. in [148] in establishing the notion of Restricted Strong Convexity (RSC). The major technical challenge for the canonical self-exciting process, as opposed to the GLM models with i.i.d. covariates in [148], lies in the fact that the covariates are highly inter-dependent as they are formed by the history of the process. Hence, it is not straightforward to establish RSC with high probability, as the large deviation techniques used for i.i.d. random vectors do not hold. We establish the RSC for the canonical self-exciting process in two steps (see Lemma 12 in Appendix A.2). First, we show that RSC holds for the expected value of the negative log-likelihood $\mathbb{E}[\mathfrak{L}(\boldsymbol{\theta})]$, and then by invoking results on concentration of dependent random variables show that the negative log-likelihood $\mathfrak{L}(\boldsymbol{\theta})$ resides in a sufficiently small neighborhood of $\mathbb{E}[\mathfrak{L}(\boldsymbol{\theta})]$ with high probability, and hence satisfies the RSC.

The remainder of the proof of Theorem 5 establishes that upon satisfying the RSC, the estimation error can be suitably bounded (Proposition 1, Appendix A.2). Similarly, Theorem 2 is proven using the RSC of the canonical self-exciting process together with the results adopted from [235] on the performance of OMP for convex cost functions (Proposition 6, Appendix A.2).

Extensions. For simplicity and clarity of presentation, we have opted to present the proofs for the case of known μ and for the canonical self-exciting process as a canonical GLM. The following corollary extends our results to the case of unknown μ.

Corollary 1. *The claims of Theorems 5 and 6 hold when μ is not known, except for possibly slightly different constants.*

Proof. The proof is given in Appendix A.2. □

The canonical self-exciting process can be generalized to a larger class of GLMs by generalization of its spiking probability function. In a more general form we can consider a spiking probability function given by

$$\lambda_i = \phi\left(\mu + \boldsymbol{\theta}'\mathbf{x}_{i-p}^{i-1}\right),$$

where $\phi(\cdot)$ is a possibly nonlinear function for which $0 < \lambda_i < 1$. In their continuous form, such processes are referred to as the *nonlinear Hawkes process* [238]. Two of the commonly-used models in neural data analysis are the log-link and logistic-link models. Our prior numerical studies in [110] revealed a similar performance improvement of the ℓ_1-regularized ML and the greedy solution over the ML estimate for the log-link model. Stationarity of these discrete processes can be proved similar to the canonical self-exciting process (see Appendix A.3). The latter fact is key to extending our proofs to other models and is summarized by the following corollary:

Corollary 2. *Theorems 5 and 6 hold when the spiking probability is given by* $\lambda_i = \phi\left(\mu + \boldsymbol{\theta}'\mathbf{x}_{i-p}^{i-1}\right)$ *for some continuous, bounded, convex and twice-differentiable function* $\phi(\cdot)$ *(e.g.,* $\phi(x) = \exp(x)$ *or* $\phi(x) = \mathsf{logit}^{-1}(x)$*) for which* $0 < \lambda_i < 1/2$*, except for different constants.*

Proof. The proof is given in Appendix A.2. □

3.4 Application to Simulated and Real Data

In this section, we study the performance of the conventional ML estimator, the ℓ_1-regularized ML estimator, and the POMP estimator on simulated data as well as real spiking data recorded from the mouse's lateral geniculate nucleus (LGN) neurons and the ferret's retinal ganglion cells (RGC). We have archived a MATLAB implementation of the estimators used in this chapter using the CVX package [91] on the open source repository GitHub and made it publicly available [4].

3.4.1 Simulation Studies

In order to simulate spiking data governed by the canonical self-exciting process, we sequentially generate spikes using (3.3). We have used $\mu = 0.1$, $\pi_{\min} = 0.01$, $\pi_{\max} = 0.49$, $p = 1000$, $s = 3$ and $n = 950$ for simulation purposes. Figure 3.1 shows 500 samples of the canonical self-exciting process generated using a history dependence parameter vector shown in Figure 3.2(a). The parameter vector $\boldsymbol{\theta}$ is compressible with a sparsity level of $s = 3$ and $\sigma_3(\boldsymbol{\theta}) = 0.05$. A value of $\gamma_n = 0.1$ is used to obtain the ℓ_1-regularized ML estimate, which is slightly tuned around the theoretical estimate given by Theorem 5. Figures 3.2(b), 3.2(c), and 3.2(d) show the estimated history dependence parameter vectors using ML, ℓ_1-regularized ML, and POMP, respectively. It can be readily visually observed that regularized ML and POMP significantly outperform the ML estimate in finding the correct values of $\boldsymbol{\theta}$. More specifically, the components at lags 405 and 800 (indicated by the gray arrows) are underestimated by the ML estimator, and their contribution is distributed among several falsely identified smaller lag components.

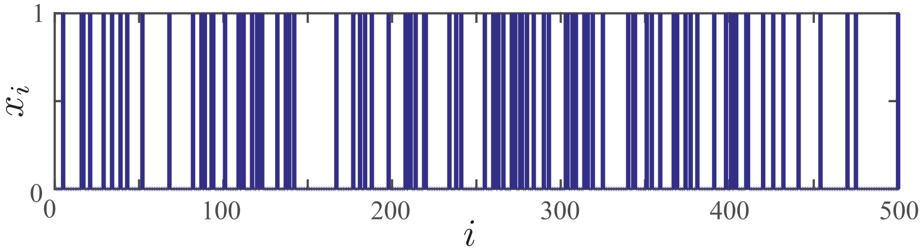
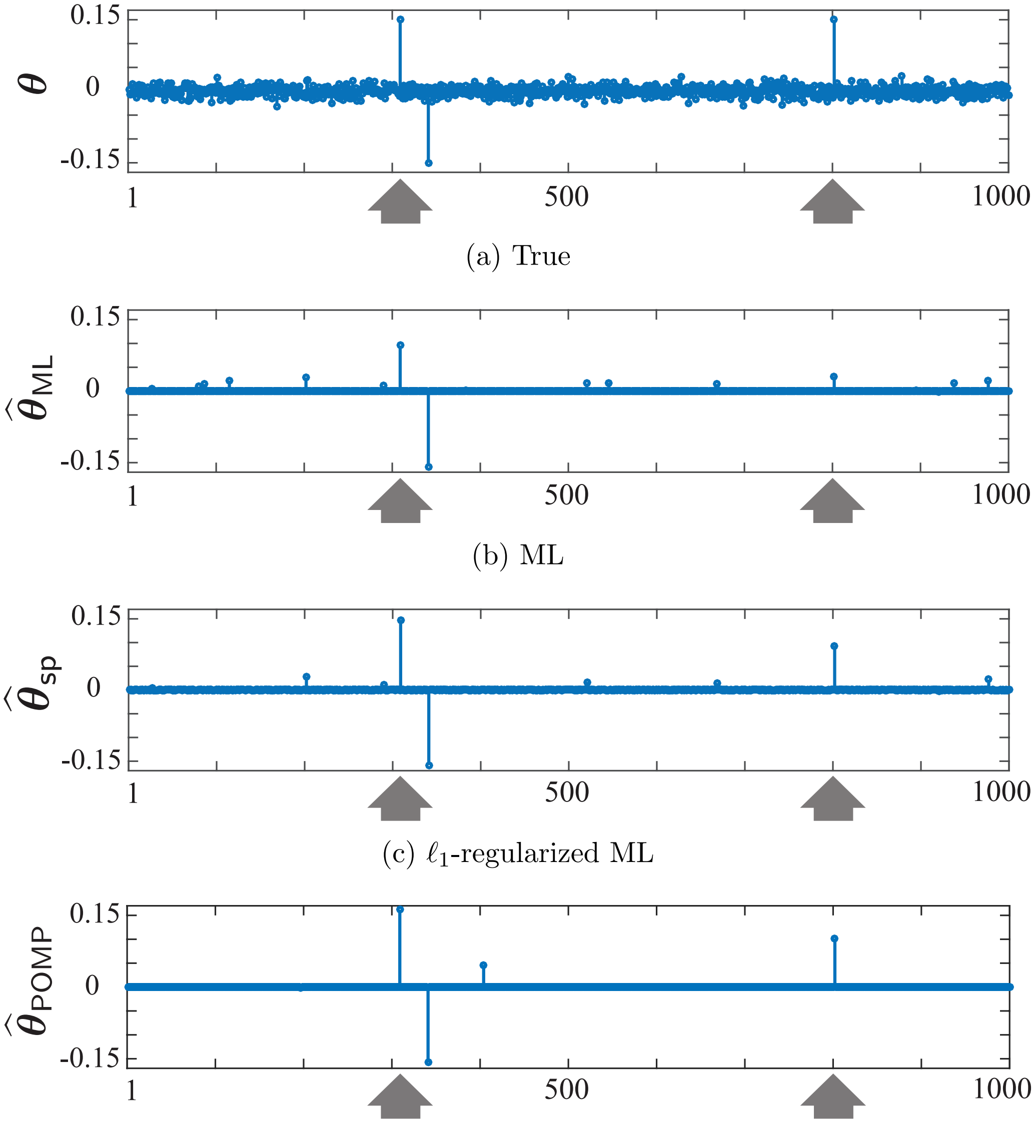


Figure 3.1: A sample of the simulated canonical self-exciting process.



(a) True

(b) ML

(c) ℓ_1-regularized ML

(d) POMP

Figure 3.2: (a) True parameters vs. (b) ML, (c) ℓ_1-regularized ML, and (d) POMP estimates.

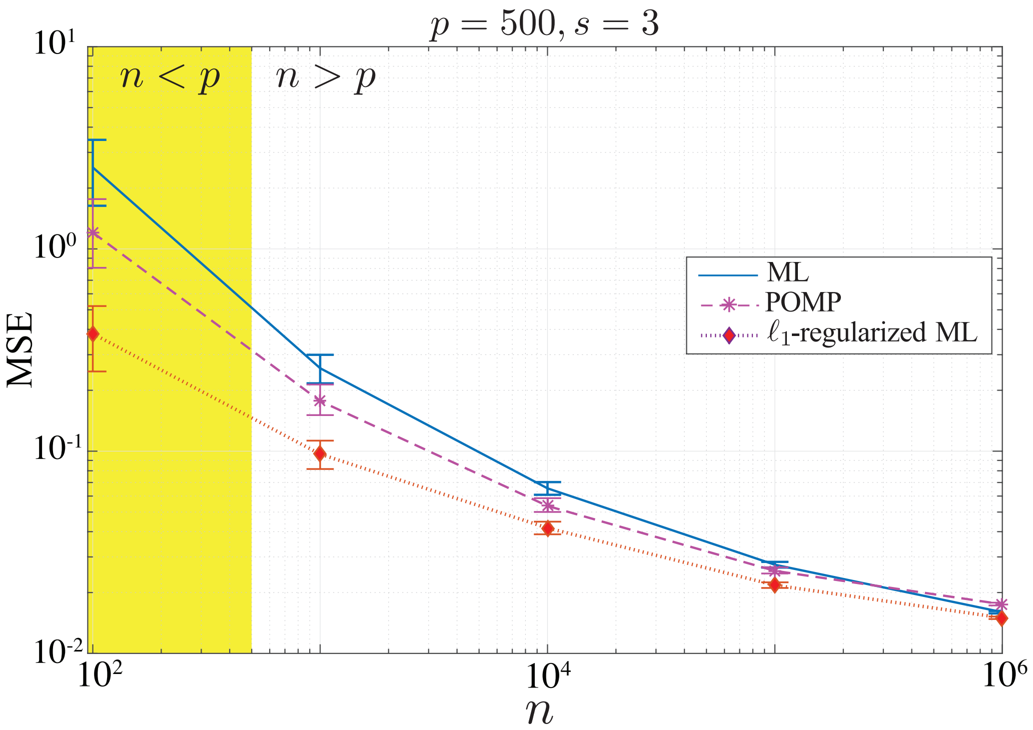
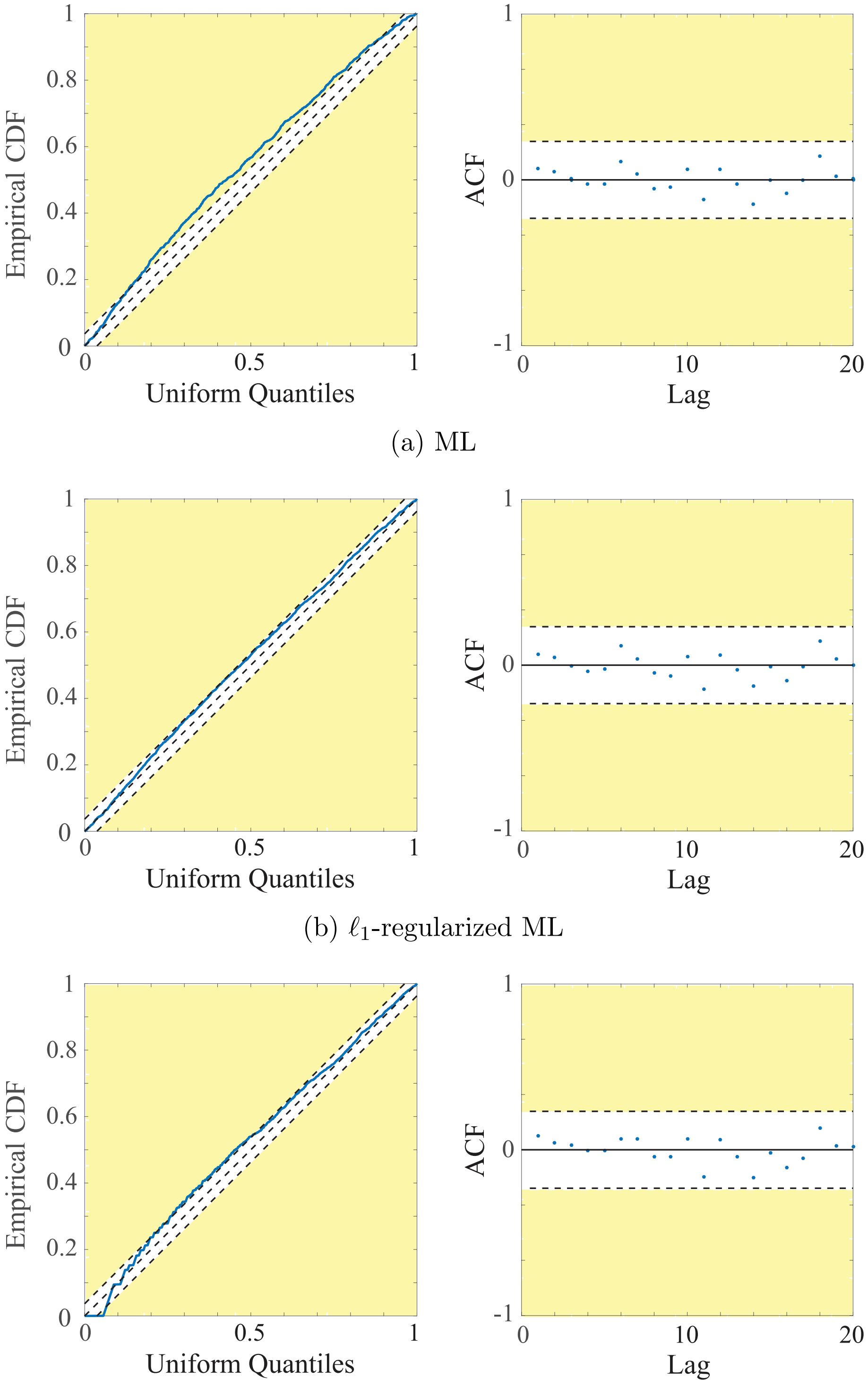


Figure 3.3: MSE performance of the ML, ℓ_1-regularized ML and POMP estimators.

In order to quantify this performance gain, we repeated the same experiment by generating realizations corresponding to randomly chosen supports of size $s = 3$ for $\boldsymbol{\theta}$ and spike trains of length $10^2 \le n \le 10^6$. In each case, the magnitudes of the components of $\boldsymbol{\theta}$ were chosen to satisfy the assumptions ($\star$). For a given $\boldsymbol{\theta}$, the mean-square-error (MSE) of the estimate $\widehat{\boldsymbol{\theta}}$ is defined as $\widehat{\mathbb{E}}\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2\}$, where $\widehat{\mathbb{E}}\{\cdot\}$ is the sample average over the realizations of the process. Figure 3.3 shows the results of this simulation, where a similar systematic performance gain is observed. The left segment of the plot (shaded in yellow) and the right segment correspond to the compressive ($n < p$) and denoising ($n > p$) regimes, respectively. Error bars on the plot indicate 90% quantiles of the MSE for this simulation obtained by multiple realizations. As it can be inferred from Figure 3.3, the ℓ_1-regularized ML and POMP have a systematic performance gain over the ML estimate in the compressive regime, where $n \ll p$, with the former outperforming the rest. In the denoising regime, the performance of the ℓ_1-regularized and ML become closer, while the POMP saturates to a higher MSE floor. The latter observation can be explained by the fact that the POMP can only estimate $s^\star$ components (including those of $\boldsymbol{\theta}_s$), and fails to capture the $(p - s^\star)$ compressible components. This results in an MSE floor above that obtained by ML, for large values of n.



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.4: KS and ACF tests at 95% confidence level, for the ML, ℓ_1-regularized ML and POMP estimates.

The MSE comparison in Figure 3.3 requires one to know the true parameters. In practice, the true parameters are unknown, and statistical tests are typically used to assess the goodness-of-fit of the estimates to the observed data. We use the Kolmogorov-Smirnov (KS) test and the autocorrelation function (ACF) test to assess the goodness-of-fit. These tests are based on the time-rescaling theorem for point processes [39], which states that if the time axis is rescaled using the estimated conditional intensity function of the inhomogeneous Poisson process, the resulting point process is a homogeneous Poisson process with unit rate. Thereby, one can test for the validity of the time-rescaling theorem via two statistical tests: the KS test reveals how close the empirical quantiles of the time-rescaled point process to the true quantiles of a unit rate Poisson process, and the ACF test reveals how close the ISI distribution of the time-rescaled process is to the true ISI distribution of a unit rate Poisson process. Details of these tests are given in Appendix B.2.

Figure 3.4 shows the KS and ACF tests at a 95% confidence level for the ML ℓ_1-regularized ML, and the POMP estimates from Figure 3.2. The yellow shades mark the regions below the specified confidence levels. The ML estimate fails to pass the KS test, while the regularized and POMP estimates pass both tests.

3.4.2 Application to Spontaneous Neuronal Spiking Activity

Background and motivation

Early studies of spontaneous neuronal activity from the cat's cochlear nucleus [87] marked a significant breakthrough in computational neuroscience by going beyond the so-called Poisson hypothesis, by which single neurons were assumed to be firing according to homogeneous Poisson statistics. The diversity of the ISIs deduced from the spontaneous activity of the cochlear neurons led to the development of more sophisticated statistical models based on renewal process theory, resulting in the Gamma and inverse Gaussian ISI descriptions of spontaneous neuronal activity [88, 209]. Due to the analytical difficulties involved in working with these models, their generalization to a broader range of spiking statistics is not straightforward.

In light of the more recent discoveries on the role of spontaneous neuronal activity in brain development [32, 228], its relation to functional architecture [208], and its functional significance in a variety of modalities including retinal [228], visual [132], auditory [205], hippocampal [194], cerebellar [10], and thalamic [164] function, the modeling and analysis of this phenomenon has sparked a renewed interest among researchers in recent years. In particular, models based on GLMs have shown to overcome the analytical difficulties of the abovementioned models based on renewal theory, and have been successfuly used in relating the spontaneous neuronal activity to instrinsic and extrinsic neural covariates [24, 39, 155, 156] as well as inferring the functional connectivity of neuronal ensembles [59, 115]. The above-mentioned results rely on the accuracy of the ML estimation of these models. In addition, the estimated parameters are typically sparse. Therefore, the ℓ_1-regularized ML and POMP estimators are expected to offer a more robust alternative than the ML, especially under the limited observation setting.

In order to evaluate the performance of these estimators on real data, in the remainder of this section we will compare the performance of the ML, ℓ_1-regularized ML, and POMP estimators in modeling the spontaneous spiking activity recorded from two different types of neurons, namely the mouse's lateral geniculate nucleus and the ferret's retinal ganglion cells.

In the following analysis, the regularization parameter γ_n was chosen using a two-fold cross-validation refinement around the value obtained from our theoretical results. The length of the history components p was chosen by first selecting a large enough p as an upper bound for the expected correlation length of neuronal spontaneous activity (estimated as ~ 1.5 s), followed by reducing p to the point where an increase in the history length does not result in significantly detected history components.

Application to LGN spiking activity

We first compare the performance of the estimators on the LGN neurons. The LGN is part of the thalamus in the brain, which acts as a relay from the retina to the primary visual cortex [108]. The data were recorded at $1ms$ resolution from the mouse LGN neurons using single-unit recording [184]. We used about 5 seconds of data from one neuron for the analysis. In order to capture the history dependence governing the spontaneous spiking activity of the LGN neuron, we model the spiking probability using the canonical self-exciting process model with $p = 100$ ($\Delta = 1ms$). Figure 3.5 shows the spiking data used in the analysis.

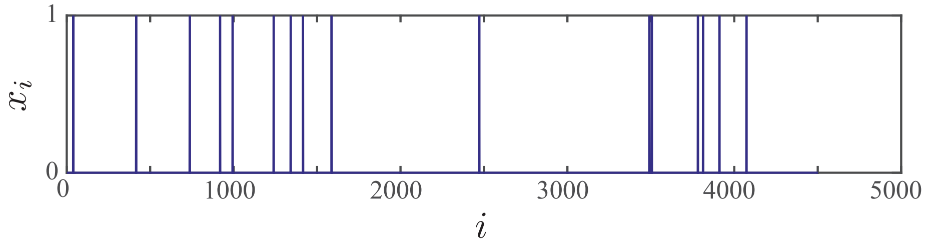
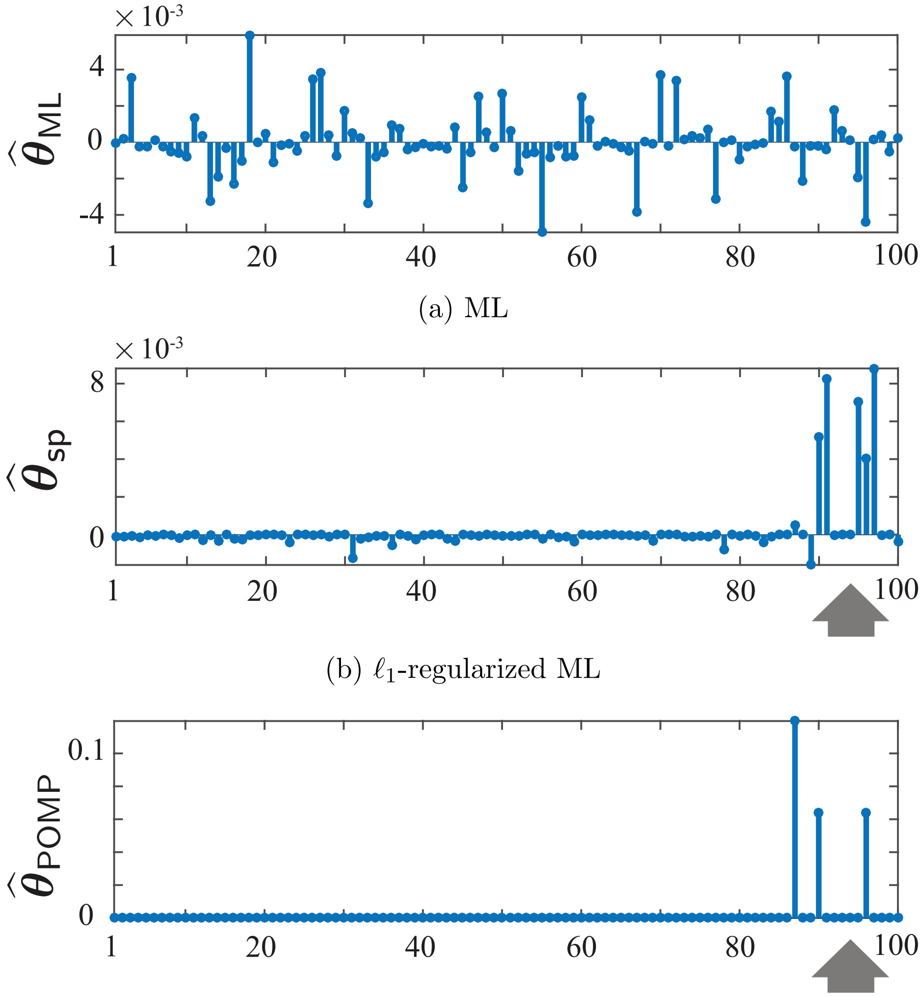


Figure 3.5: The LGN spiking data used in the analysis.

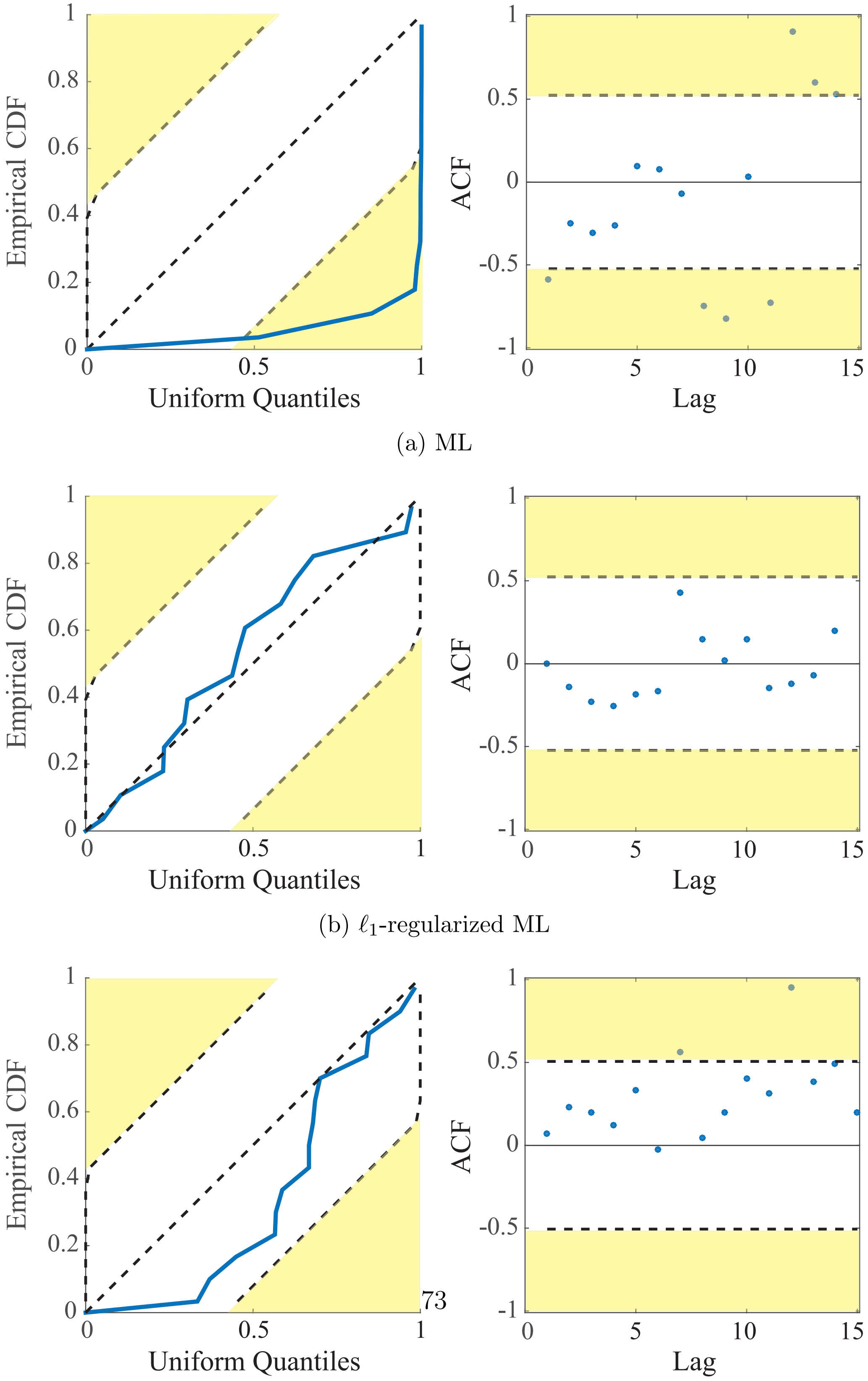


(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.6: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the LGN spiking parameters.



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.7: KS and ACF tests at 99% confidence level, for the ML, ℓ_1-regularized ML and POMP estimates.

Figure 3.6 shows the estimated history dependence parameter vectors using the three methods. Both the regularized ML (Figure 3.6(b)) and POMP (Figure 3.6(c)) estimates capture significant history dependence components around a lag of 90–95 ms (marked by the upward arrows). In [33], an intrinsic neuronal oscillation frequency of around 10 Hz has been reported in around 30% of all classes of mouse retinal cells under experiment, using combined two-photon imaging and patch-clamp recording. Our results are indeed consistent with the above mentioned findings about the intrinsic spiking frequency of retinal neurons. To see this, we consider the power spectral density of the canonical self-exciting process given by:

$$S(\omega) = \frac{1}{2\pi}\left(\pi_\star^2\delta(\omega) + \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2\,|1-\Theta(\omega)|^2}\right), \tag{3.8}$$

where $\Theta(\omega)$ is the discrete-time Fourier transform of $\boldsymbol{\theta}$ and $\pi_\star = \mu/(1-\mathbf{1}'\boldsymbol{\theta})$ denotes the stationary distribution probability of spiking. The derivation of the power spectral density is given in Appendix A.3. The power spectral density of the canonical self-exciting process resembles the Bartlett spectrum of the Hawkes process [26, 27, 100], whose peaks correspond to the significant oscillatory components of the underlying process. Our estimated parameter vectors $\boldsymbol{\theta}$ using the regularized ML and POMP have significant nonzero components around lags of $90 \leq k \leq 95$. As a result, $S(\omega)$ peaks at $\omega = \frac{2\pi}{k\Delta}$. Hence, $f = \frac{1}{k\Delta}$ is an estimate of the significant intrinsic frequency of the underlying self-exciting process. Using the estimated numerical values, the intrinsic frequency is around 10.5–11 Hz, which is consistent with experimental findings of [33]. Compared to the method in [33], our estimates are obtained using much shorter recordings of spiking activity and provide a principled framework to study the oscillatory behavior of LGN neurons using sparse GLM estimation.

Note that there is a difference in the orders of magnitudes of the POMP estimate compared to the ML and regularized ML estimates. This is due to the fact that the POMP estimate is exactly s-sparse, whereas the ML and regularized ML estimates consist of $p = 100$ non-zero values. In order to assess the goodness-of-fit of these estimates, we invoke the KS and ACF tests. Figure 3.7 shows the corresponding KS and ACF test plots. As it is implied from Figure 3.7(a), the ML estimate fails both tests due to overfitting, whereas the regularized ML (Figure 3.7(b)) passes both tests at the specified confidence levels. The POMP estimate (Figure 3.7(c)), however, passes the KS test while marginally failing the ACF test. The latter observation implies that the seemingly negligible components of the parameter vector captured by the regularized ML estimate seem to be important in explaining the statistics of the observed data.

Application to RGC spiking activity

We will next study the performance of the estimators on spiking data recorded from the RGCs of neonatal and adult ferrets [225]. The retinal ganglion cells are located in the innermost layer of the retina. They integrate information from photoreceptors and project them into the brain [29]. The data were recorded using a multi-electrode array from the ferret retina at 50 μs [225]. We used 5 seconds of data from one neuron for the analysis (neuron 2, session 1, adult data set, CARMEN data base

[76]). Figure 3.8 shows a segment of the spiking data used in our analysis. The RGC activity in the adult ferret is characterized by bursts of activity with a mean firing rate of 9 ± 7 Hz, which are separated by 0.5–1 s intervals [225].

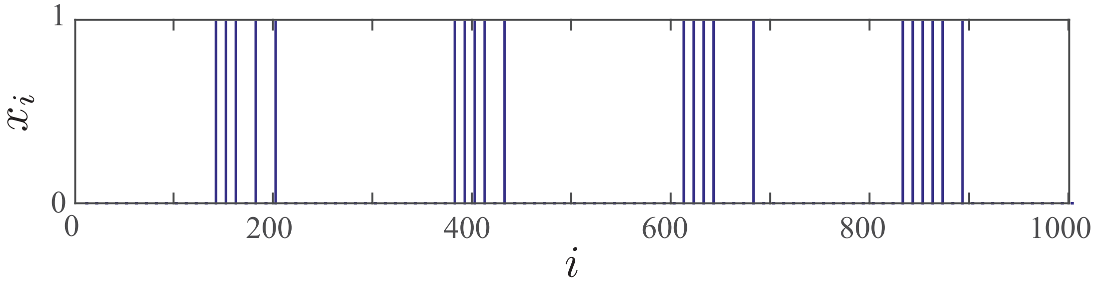
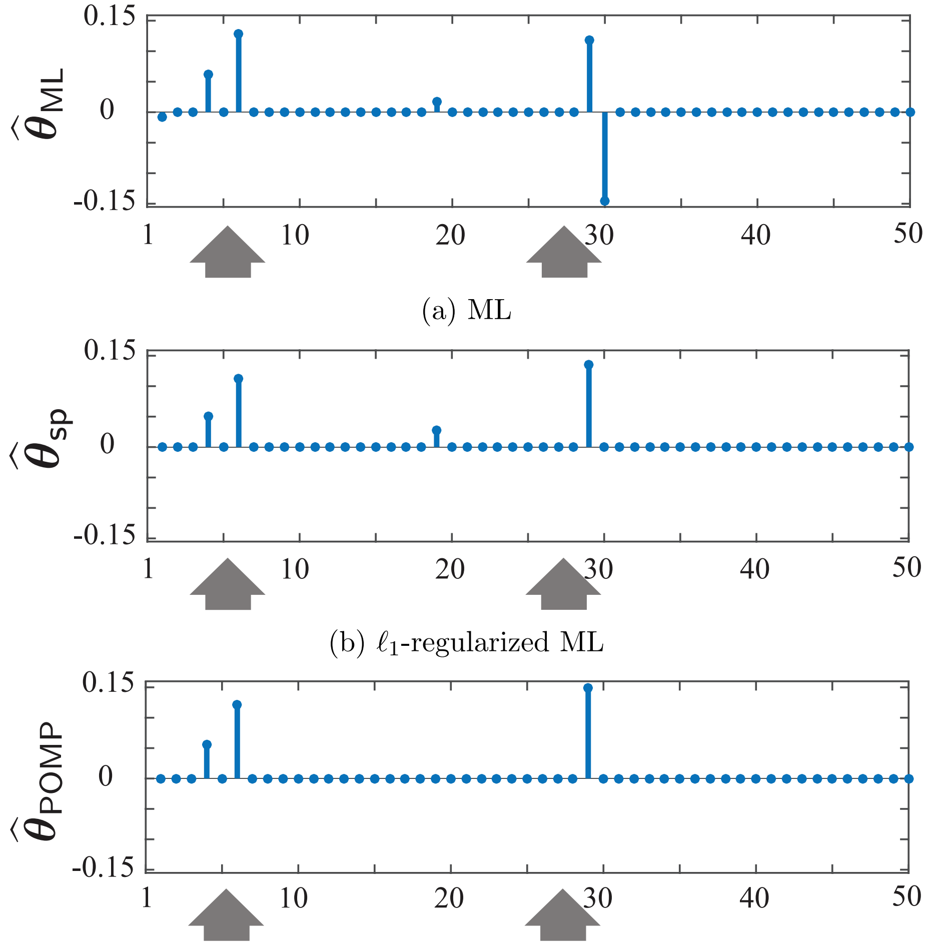


Figure 3.8: Segment of the RGC spiking data used in the analysis.



(a) ML

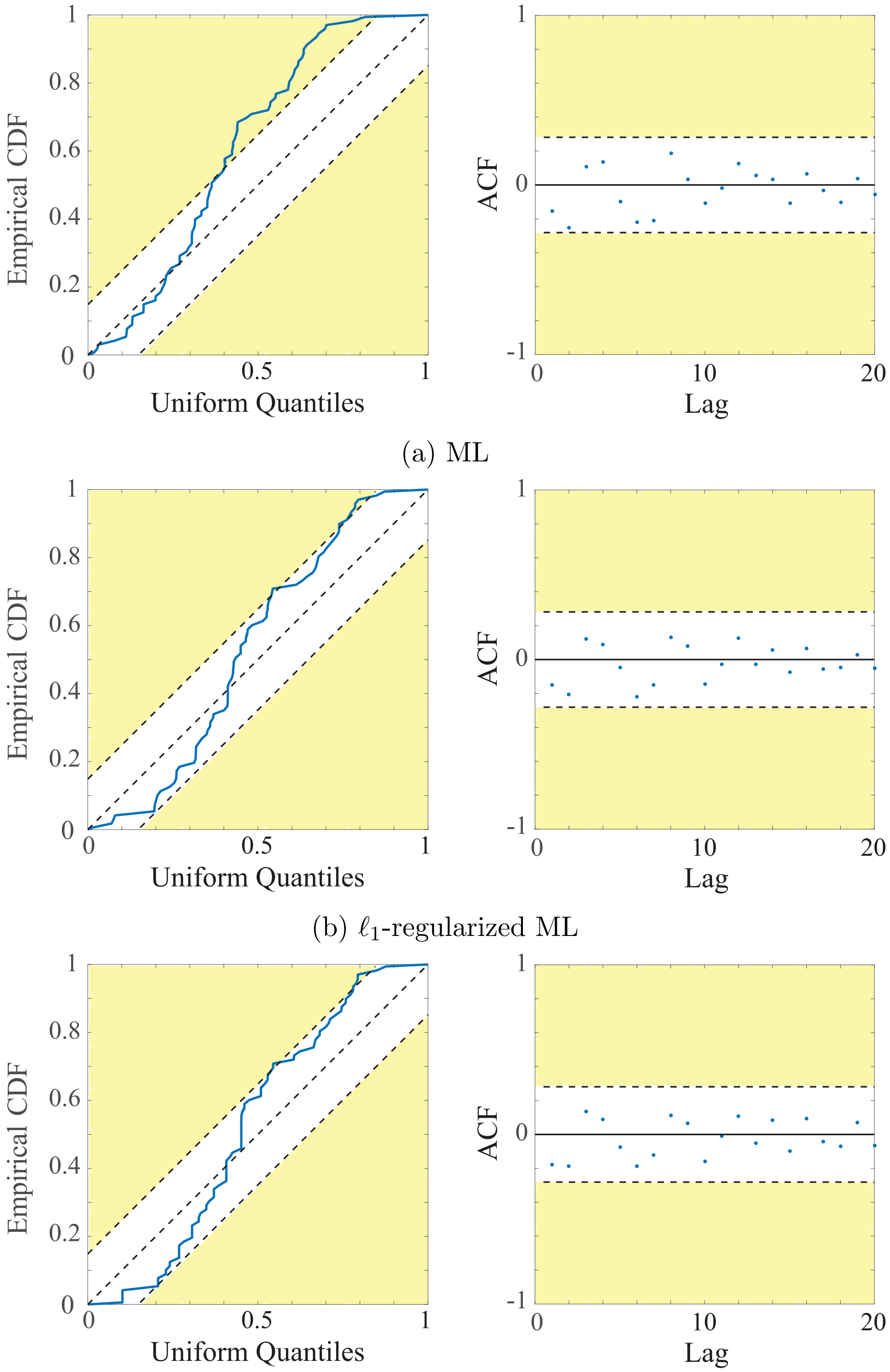
(b) ℓ_1-regularized ML

(c) POMP

Figure 3.9: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the RGC spiking parameters using the canonical self-exciting process model.

In order to capture the history dependence governing the spontaneous spiking activity of the RGC neuron, we model the spiking probability using two different link models to further corroborate the generalization of our results to models beyond the canonical self-exciting process studied in this chapter. First, we consider the canonical self-exciting process model. We have chosen $\pi_{\max} = 0.49$, $p = 50$ ($\Delta =$

25 ms) and $s_\star = 3$. The baseline parameter μ is estimated from the data and is set to be equal to half of empirical mean firing rate of the neuron.

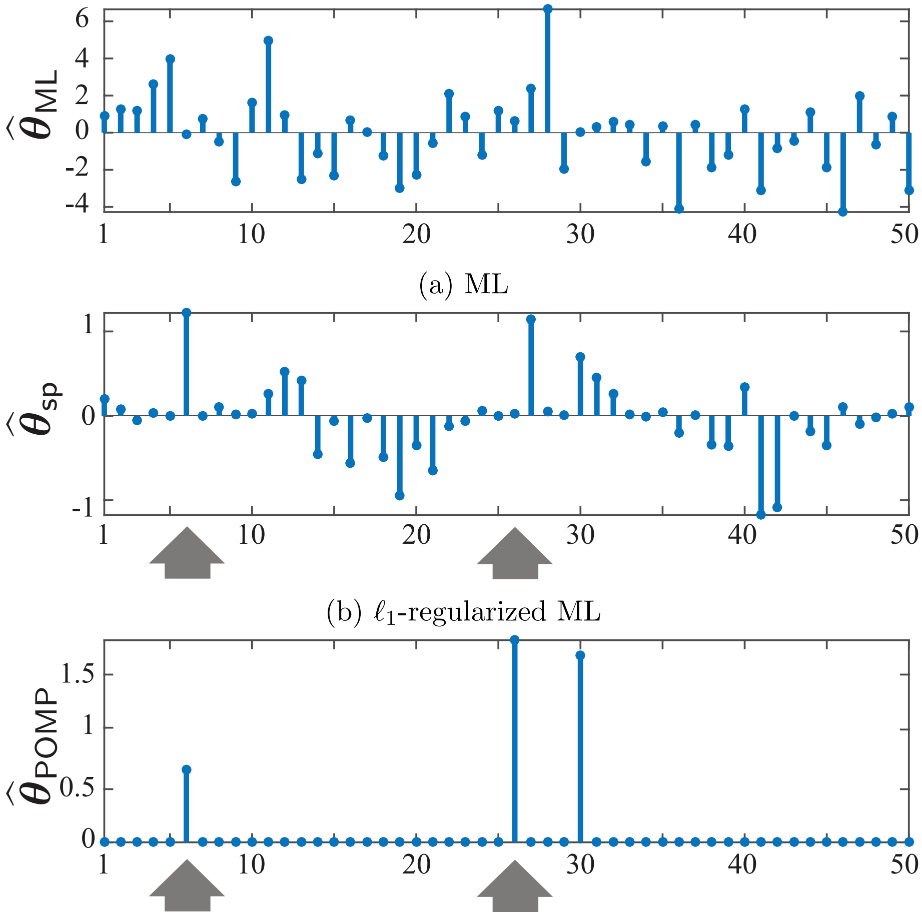


(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.10: KS and ACF tests at 95% confidence level, for the ML, ℓ_1-regularized ML and POMP estimates using the canonical self-exciting process model.

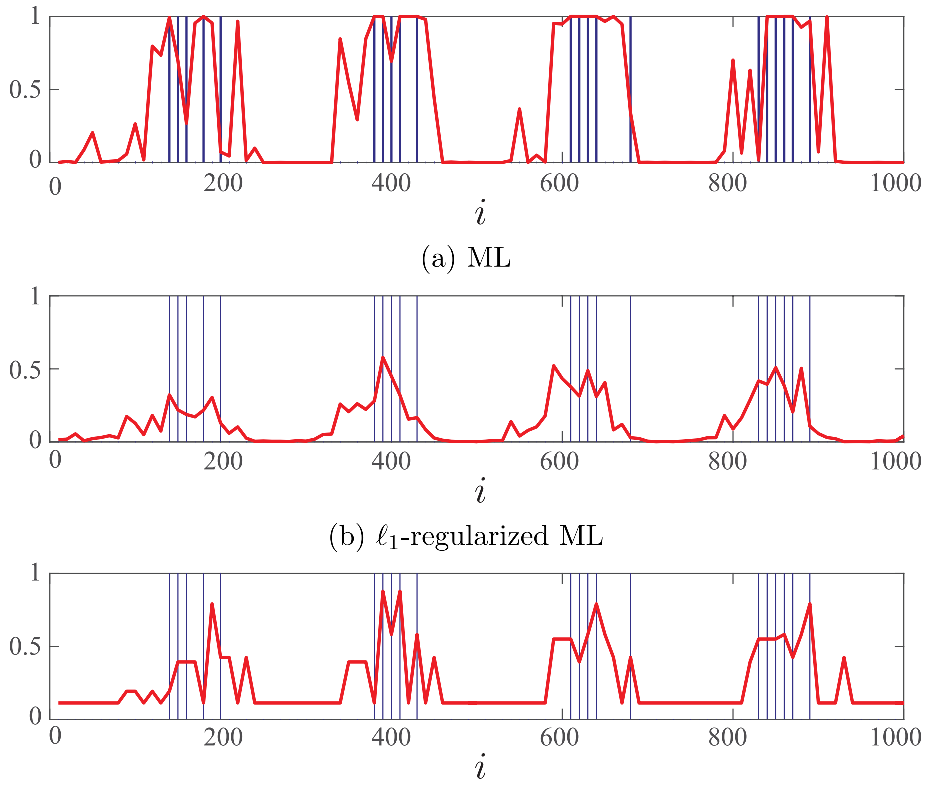


(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.11: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the RGC spiking parameters using the logistic link model.



(a) ML

(b) ℓ_1-regularized ML

(c) POMP

Figure 3.12: (a) ML, (b) ℓ_1-regularized ML, and (c) POMP estimates of the RGC spiking probability using the unconstrained logistic link model. Blue vertical lines show the locations of the spikes, and red traces show the estimated probabilities.

Figure 3.9 shows the estimated history components using the three estimators. All three estimates capture significant self-exciting history dependence components around the lags of 150 ms and 0.65–0.75 s (marked by the upward arrows). Invoking the foregoing argument for the LGN neuron regarding the power spectral density of the process (3.8), these estimated lag components are consistent with the empirical estimates of [225], as they indicate that the data can be characterized by a combination of $\frac{1}{150 \text{ ms}} = 6.66$ Hz bursts separated by gaps of length 0.65–0.75 s. The

ML estimator predicts an extra self-inhibitory (negative) component, which results in over-fitting the data. This phenomenon can be observed by noting that the ML estimate fails the KS test shown in Figure 3.10.

We will next consider a logistic link model of the form $\lambda_i = \frac{\exp(\mu + \boldsymbol{\theta}' x_{i-p}^{i-1})}{C + \exp(\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1})}$, with $C = 100$. This model is widely used in neuronal modeling literature (e.g., [206,207]), where the assumptions given by $(\star)$ are dropped and the optimization is performed in an unconstrained fashion. We adopt this approach and obtain all the estimates by dropping the assumptions of $(\star)$. Figure 3.11 shows the estimated history components using the unconstrained estimators. Compared to the canonical self-exciting process model with a linear link, both the regularized ML (Figure 3.11(b)) and POMP (Figure 3.11(c)) estimates capture similar significant self-exciting history dependence components, which are consistent across the two sets of estimates.

The KS and ACF test results for this case are very similar to Figure 3.10 are are thus omitted for brevity. In order to further inspect the goodness-of-fit of these methods, we plot the estimated spiking probabilities in Figure 3.12. The ML estimate shown in Figure 3.12(a) overfits the spiking events by rapidly saturating the rate to either 0 and 1, which results in undesired high rate estimates where there are no spikes. On the contrary, the regularized ML (Figure 3.12(b)) and POMP (Figure 3.12(c)) provide a more reliable estimate of the rates consistent with the spiking events. This analysis suggests that the sufficient assumptions of $(\star)$ are not necessary for the superior performance of the regularized and POMP estimators over that of ML.

3.5 Concluding Remarks

In this chapter, we studied the sampling properties of ℓ_1-regularized ML and greedy estimators for a canonical self-exciting process. The main theorems provide non-asymptotic sampling bounds on the number of measurements, which lead to stable recovery of the parameters of the process. To the best of our knowledge, our results are the first of this kind, and can be readily generalized to various other classes of self-exciting GLMs, such as processes with logarithmic or logistic links.

Compared to the existing literature, our results bring about two major contributions. First, we provide a theoretical underpinning for the advantage of ℓ_1-regularization in ML estimation as well as greedy estimation in problems involving binary observations. These methods have been used in neuroscience in an ad-hoc fashion. Our results establish the utility of these techniques by characterizing the underlying sampling trade-offs. Second, our analysis relaxes the widely-assumed hypotheses of i.i.d. covariates. This assumption is often violated when working with history-dependent data such as neural spiking data.

We also verified the validity of our theoretical results through simulation studies as well as application to real neuronal spiking data from mouse's LGN and ferret's RGC neurons. These results show that both the regularized ML and the greedy estimates significantly outperform the widely-used ML estimate. In particular, through making a connection with the spectrum of discrete point processes, we were able to quantify the estimation of the intrinsic firing frequency of LGN neurons. In the spirit of easing reproducibility, we have archived a MATLAB implementation of the estimators studied in this work using the CVX package [91] on the open source repository GitHub and made it publicly available [4].

One of the limitations of our analysis is the assumption that the spiking probabilities are bounded by 1/2, which results in loss of generality. This assumption is made for the sake of theoretical analysis in bounding the mixing rate of the canonical self-exciting process. Our numerical experiments suggest that it is not necessary for the operation of the ℓ_1-regularized and POMP estimators. We consider further inspection of the mixing properties of this process and thus relaxing this assumption as future work. Our future work also includes generalization of our analysis to multivariate GLMs, which will allow to infer network properties from multi-unit recordings of neuronal ensembles.

Chapter 4: Fast and Stable Signal Deconvolution via Compressible State-Space Models

Common biological measurements are in the form of noisy convolutions of signals of interest with possibly unknown and transient blurring kernels. Examples include EEG and calcium imaging data. Thus, signal deconvolution of these measurements is crucial in understanding the underlying biological processes. The objective of this chapter is to develop fast and stable solutions for signal deconvolution from noisy, blurred and undersampled data, where the signals are in the form of discrete events distributed in time and space. Due to their smoothness properties, in these application Gaussian state-space models exhibit poor performance in recovering the states and the sharp transitions. In this chapter we consider the problem of estimating compressible state space models, where the sparsity lies in the transitions of the states.

We introduce compressible state-space models as a framework to model and estimate such discrete events. These state-space models admit abrupt changes in the states and have a convergent transition matrix, and are coupled with compressive linear measurements. We consider a dynamic compressive sensing optimization problem and develop a fast solution, using two nested Expectation Maximization algorithms, to jointly estimate the states as well as their transition matrices. Under suitable sparsity assumptions on the dynamics, we prove optimal stability guarantees for the recovery of the states and present a method for the identification of the underlying discrete events with precise confidence bounds. We present simulation studies as well as application to calcium deconvolution and sleep spindle detection, which verify our theoretical results and show significant improvement over existing techniques. Our results show that by explicitly modeling the dynamics of the underlying signals, it is possible to construct signal deconvolution solutions that are scalable, statistically robust, and achieve high temporal resolution. Our proposed methodology provides a framework for modeling and deconvolution of noisy, blurred, and undersampled measurements in a fast and stable fashion, with potential application to a wide range of biological data.

4.1 Introduction

In many signal processing applications such as estimation of brain activity from magnetoencephalography (MEG) time-series [162], estimation of time-varying networks [120], electroencephalogram (EEG) analysis [151], calcium imaging [218], functional magnetic resonance imaging (fMRI) [53], and video compression [109], the signals often exhibit abrupt changes that are blurred through convolution with unknown kernels due to intrinsic measurement constraints. Extracting the underlying signals from blurred and noisy measurements is often referred to as signal deconvolution. Traditionally, state-space models have been used for such signal deconvolution problems, where the states correspond to the unobservable signals. Gaussian state-space models in particular are widely used to model smooth state transitions. Under normality assumptions, posterior mean filters and smoothers are optimal estimators, where the analytical solution is given respectively by the Kalman filter and the fixed interval smoother [14, 101].

When applied to observations from abruptly changing states, Gaussian state-space models exhibit poor performance in recovering sharp transitions of the states due to their underlying smoothing property. Although filtering and smoothing recursions can be obtained in principle for non-Gaussian state-space models, exact calculations are no longer possible. Apart from crude approximations like the extended Kalman filter, several methods have been proposed including numerical methods for low-dimensional states [117], Monte Carlo filters [103,117], posterior mode estimation [83, 84], and fully Bayesian smoothing using Markov chain Monte Carlo simulation [119,186]. In order to exploit sparsity, several dynamic compressed sensing (CS) techniques, such as the Kalman filtered CS algorithm, have been proposed that typically assume partial information about the sparse support or estimate it in a greedy and online fashion [51, 214, 215, 234, 239]. However, little is known about the theoretical performance guarantees of these algorithms.

In this chapter, we consider the problem of estimating state dynamics from noisy and undersampled observations, where the state transitions are governed by autoregressive models with compressible innovations. Motivated by the theory of CS, we employ an objective function formed by the ℓ_1-norm of the state innovations [17]. Unlike the traditional compressed sensing setting, the sparsity is associated with the dynamics and not the states themselves. In the absence of observation noise, the CS recovery guarantees are shown to extend to this problem [17]. However, in a realistic setting in the presence of observation noise, it is unclear how the CS recovery guarantees generalize to this estimation problem.

We will present stability guarantees for this estimator under a convergent state transition matrix, which confirm that the CS recovery guarantees can be extended to this problem. The corresponding optimization problem in its Lagrangian form is akin to the MAP estimator of the states in a linear state-space model where the innovations are Laplace distributed. This allows us to integrate methods from Expectation-Maximization (EM) theory and Gaussian state-space estimation to derive efficient algorithms for the estimation of states as well as the state transition matrix, which is usually unknown in practice. To this end, we construct two nested EM algorithms in order to jointly estimate the states and the transition matrix. The outer EM algorithm for state estimation is akin to the fixed interval smoother, and the inner EM algorithm uses the state estimates to update the state transition matrix [189]. The resulting EM algorithm is recursive in time, which makes the computational complexity of our method scale linearly with temporal dimension of the problem. This provides an advantage over existing methods based on convex optimization, which typically scale super-linearly with the temporal dimension.

Our results are related to parallel applications in spectral estimation, source localization, and channel equalization [79], where the measurements are of the form $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}$, with $\mathbf{Y}$ is the observation matrix, $\mathbf{X}$ denotes the unknown parameters, $\mathbf{A}$ is the measurement matrix, and $\mathbf{N}$ is the additive noise. These problems are referred to as Multiple Measurement Vectors (MMV) [63] and Multivariate Regression [152]. In these applications, solutions with row sparsity in $\mathbf{X}$ are desired. Recovery of sparse signals with Gaussian innovations is studied in [236]. Several recovery algorithms including the ℓ_1–ℓ_q minimization methods, subspace methods and greedy pursuit algorithms [67] have been proposed for support union recovery in this setup. Our contributions are distinct in that we directly model the state innovations as a compressible sequence, for recovery of which we present both sharp theoretical guarantees as well as fast algorithms from state-space estimation.

Finally, we provide simulation results as well as applications to two experimentally-acquired data sets: calcium imaging recordings of neuronal activity, and EEG data during sleep. In the former, the deconvolution problem concerns estimating the location of spikes given the temporally blurred calcium fluorescence, and in the latter, the objective is to detect the occurrence and onset of sleep spindles. Our simulation studies confirm our theoretical predictions on the performance gain obtained by compressible state-space estimation over those obtained by traditional estimators such as the basis pursuit denoising. Our real data analyses reveal that our compressible state-space modeling and estimation framework outperforms two of the commonly-used methods for calcium deconvolution and sleep spindle detection. In the spirit of easing reproducibility, we have made MATLAB implementations of our codes publicly available [5].

4.2 Methods

In this section we introduce the experimental procedures of recording the data for analysis and establish our problem formulation and notational conventions and present our main theoretical analysis and algorithm development.

4.2.1 Experimental Procedures

Surgery

2 hours before surgery, 0.1 cc dexamethasone (2 mg/ml, VetOne) was injected subcutaneously to reduce brain swelling during craniotomy. Anesthesia is induced with 4% isoflurane (Fluriso, VetOne) with a calibrated vaporizer (Matrx VIP 3000). During surgery, isoflurane level was reduced to and maintained at a level of 1.5%–2%. Body temperature of the animal is maintained at 36.0 degrees Celsius during surgery. Hair on top of head of the animal was removed using Hair Remover Face Cream (Nair), after which Betadine (Purdue Products) and 70% ethanol was applied sequentially 3 times to the surface of the skin before removing the skin. Soft tissues and muscles were removed to expose the skull. Then a custom designed 3D printed stainless headplate was mounted over left auditory cortex and secured with C&B-Metabond (Parkell). A craniotomy with a diameter of around 3.5 mm was then performed over left auditory cortex. A three layered cover slip was used as cranial window, which is made by gluing (NOA71, Norland Products) 2 pieces of 3 mm coverslips (64-0720 (CS-3R), Warner Instruments) with a 5 mm coverslip (64–0700 (CS-5R),

Warner Instruments). Cranial window was quickly dabbed in kwik-sil (World Precision Instruments) before mounted 3 mm coverslips facing down onto the brain. After kwik-sil cured, Metabond was applied to secure the position of the cranial window. Synthetic Black Iron Oxide (Alpha Chemicals) was then applied to the hardened Metabond surface. 0.05 cc Cefazolin (1 gram/vial, West Ward Pharmaceuticals) was injected subcutaneously when entire procedure was finished. After the surgery the animal was kept warm under heat light for 30 minutes for recovery before returning to home cage. Medicated water (Sulfamethoxazole and Trimethoprim Oral Suspension, USP 200 mg/40 mg per 5 ml, Aurobindo Pharms USA; 6 ml solution diluted in 100 ml water) substitute normal drinking water for 7 days before any imaging was performed.

Awake two-photon imaging

Spontaneous activity data of population of layer 2/3 auditory cortex (A1) neurons is collected from adult (3-month old) Thy1-GCaMP6s female mouse implanted with chronic window following the above procedure, using two-photon imaging. Acquisition is performed using a two-photon microscope (Thorlabs Bscope 2) equipped with a Vision 2 Ti:Sapphire laser (Coherent), equipped with a GaAsP photo detector module (Hamamatsu) and resonant scanners enabling faster high-resoluation scanning at 30–60 Hz per frame. The excitation wavelength was 920 nm. Regions ($\sim 300\ \mu\text{m}^2$) within A1 were scanned at 30 Hz through a 20x, 0.95 NA water-immersion objective (Olympus). During imaging the animal was head-fixed and awake. The microscope was rotated 45 degrees and placed over the left A1 where window was placed. An average image of field of view was generated by choosing a time window where minimum movement of the brain was observed and used as reference image for motion correction using TurboReg plugin in ImageJ. GCaMP6s positive cells are selected manually by placing a ring like ROI over each identified cell. Neuropil masks were generated by placing a 20 μm radius circular region over each cell yet excluding all cell soma regions. Traces of soma and neuropil were generated by averaging image intensity within respective masks at each time point. A ratio of 0.7 was used to correct for neuropil contamination.

Cell-attached patch clamp recordings and two-photon imaging

Recordings were performed in vitro in voltage clamp to simultaneously measure spiking activity and $\Delta F/F$. Thalamocortical slices containing A1 were prepared as previously described [140]. The extracellular recording solution consisted of artificial cerebral spinal fluid (ACSF) containing: 130 NaCl, 3 KCl, 1.25 KH2PO4, 20 NaHCO3, 10 glucose, 1.3 MgSO4, 2.5 CaCl2 (pH 7.35-7.4, in 95% O2 5% CO2). Action potentials were recorded extracellularly in loose-seal cell-attached configuration (seal resistance typically 20–30 MΩ) in voltage clamp mode. Borosilicate glass patch pipettes were filled with normal ACSF diluted 10%, and had a tip resistance of $\sim$ 3-5 MΩ in the bath. Data were acquired with a Multiclamp 700B patch clamp amplifier (Molecular Devices), low-pass filtered at 3-6 kHz, and digitized at 10 kHz using the MATLAB-based software. Action potentials were stimulated with a bipolar electrode placed in L1 or L23 to stimulate the apical dendrites of pyramidal cells (pulse duration 1-5 ms). Data were analyzed offline using MATLAB. Imaging was largely performed using a two-photon microscope (Ultima, Prairie Technologies) and a MaiTai DeepSee laser (SpectraPhysics), equipped with a GaAsP photo detector module (Hamamatsu) and resonant scanners enabling faster high-resoluation scanning at 30-60 Hz per frame. Excitation was set at 900 nm. Regions were scanned at 30 Hz through a 40x water-immersion objective (Olympus). Cells were manually selected as ring-like regions of interest (ROIs) that cover soma but exclude cell nuclei, and pixel intensity within each ROI was averaged to generate fluorescence over time and changes in fluorescence ($\Delta F/F$) were then calculated.

4.2.2 Problem Formulation and Theoretical Analysis

We consider the linear compressible state-space model given by

$$\mathbf{x}_t = \mathbf{\Theta}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \tag{4.1}$$

where $(\mathbf{x}_t)_{t=1}^T \in \mathbb{R}^p$ denote the sequence of unobservable states, $\mathbf{\Theta}$ is the state transition matrix satisfying $\|\mathbf{\Theta}\| < 1$, $\mathbf{w}_t \in \mathbb{R}^p$ is the state innovation sequence, $(\mathbf{y}_t)_{t=1}^T \in \mathbb{R}^{n_t}$ are the linear observations, $\mathbf{A}_t \in \mathbb{R}^{n_t \times p}$ denotes the measurement matrix, and $\mathbf{v}_t \in \mathbb{R}^{n_t}$ denotes the measurement noise. The main problem is to estimate the unobserved sequence $(\mathbf{x}_t)_{t=1}^T$ (and possibly $\mathbf{\Theta}$), given the sequence of observations $(\mathbf{y}_t)_{t=1}^T$. This problem is in general ill-posed, when $n_t < p$, for some t. We therefore need to make additional assumptions in order to seek a stable solution.

We assume that the state innovations are sparse (resp. compressible), i.e. $\mathbf{x}_t - \mathbf{\Theta}\mathbf{x}_{t-1}$ is s_t-sparse (resp. (s_t, ξ)-compressible) with $s_1 \gg s_t$ for $t \in [T]\backslash\{1\}$. Our theoretical analysis pertain to the compressed sensing regime where $1 \ll s_t < n_t \ll p$.

We assume that the rows of $\mathbf{A}_t$ are a subset of the rows of $\mathbf{A}_1$, i.e. $\mathbf{A}_t = (\mathbf{A}_1)_{n_t}$, and define $\widetilde{\mathbf{A}}_t = \sqrt{\frac{n_1}{n_t}}\mathbf{A}_t$. Other than its technical usefulness, the latter assumption helps avoid prohibitive storage of all the measurement matrices. In order to promote sparsity of the state innovations, we consider the dynamic ℓ_1-regularization (dynamic CS from now on) problem defined as

$$\underset{(\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}}{\text{minimize}} \quad \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1}\|_1}{\sqrt{s_t}} \quad \text{s.t.} \quad \|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_2 \leq \sqrt{\frac{n_t}{n_1}}\varepsilon. \tag{4.2}$$

where ε is an upper bound on the observation noise, i.e., $\|v_t\|_2 \leq \varepsilon$ for all t. Note that this problem is a variant of the dynamic CS problem introduced in [17]. We also consider the modified Lagrangian form of (4.2) given by

$$\underset{(\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}}{\text{minimize}} \quad \lambda \sum_{t=1}^{T} \frac{\|\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1}\|_1}{\sqrt{s_t}} + \frac{1}{n_t} \frac{\|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_2^2}{2\sigma^2}. \tag{4.3}$$

for some constants σ_2^2 and $\lambda \geq 0$. Note that if $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, n_t\sigma^2\mathbf{I})$, then Eq. (4.3) is akin to the maximum *a posteriori* (MAP) estimator of the states in (4.1), assuming that the *state* innovations were independent Laplace random variables with respective parameters $\lambda/\sqrt{s_t}$. We will later use this analogy to derive fast solutions to the optimization problem in (4.3).

Uniqueness and exact recovery of the sequence $(\mathbf{x}_t)_{t=1}^T$ in the absence of noise was proved in [17] for $\boldsymbol{\Theta} = \mathbf{I}$, by an inductive construction of dual certificates. The special case $\boldsymbol{\Theta} = \mathbf{I}$ can be considered as a generalization of the total variation (TV) minimization problem [168]. Our main result on stability of the solution of (4.2) is the following:

Theorem 7 (Stable Recovery of Activity in the Presence of Noise)**.** *Let* $(\mathbf{x}_t)_{t=1}^T \in \mathbb{R}^p$ *be a sequence of states with a known transition matrix* $\mathbf{\Theta} = \theta\mathbf{I}$, *where* $|\theta|< 1$ *and* $\widetilde{\mathbf{A}}_t$, $t \geq 1$ *satisfies RIP of order* $4s$ *with* $\delta_{4s} < 1/3$. *Suppose that* $n_1 > n_2 = n_3 = \cdots = n_T$. *Then, the solution* $(\widehat{\mathbf{x}}_t)_{t=1}^T$ *to the dynamic CS problem (4.2) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2 \leq \frac{1-\theta^T}{1-\theta}\left(12.6\left(1+\frac{1}{T}\sqrt{\frac{n_1}{n_2}}-\frac{1}{T}\right)\varepsilon + \frac{3}{T}\sum_{t=1}^{T}\frac{\sigma_{s_t}(\mathbf{x}_t - \mathbf{\Theta}\mathbf{x}_{t-1})}{\sqrt{s_t}}\right). \tag{4.4}$$

Proof Sketch. The proof of Theorem 7 is based on establishing a modified cone and tube constraint for the dynamic CS problem (4.2) and using the boundedness of the Frobenius norm of the inverse first-order differencing operator. Details of the proof are given in Appendix A.5.

Remark 1. The first term on the right hand side of Theorem 7 implies that the average reconstruction error of the sequence $(\mathbf{x}_t)_{t=1}^T$ is upper bounded proportional to the noise level ε, which implies the stability of the estimate. The second term is a measure of compressibility of the innovation sequence and vanishes when the sparsity condition is exactly met.

Remark 2. Similar to the general compressive sensing setup, the results of Theorem 7 are general and hold even when the sparsity assume is not exactly met, namely when the compressibility (sparsity) assumption is violated, the second term on the right hand side could be of order of a constant. However in order for $\widetilde{\mathbf{A}}_t$ to satisfy RIP of order $4s$ one requires $n \sim \mathcal{O}(s \log p)$ which in the absence of sparsity ($s \sim \mathcal{O}(p)$) reduces to the denoising regime.

Remark 3. The theory of LASSO suggests a choice of the regularization parameter $\lambda \sim \mathcal{O}\left(\sigma\sqrt{\frac{\log p}{n_t}}\right)$ [148]. In the absence of sparsity (compressibility) $s \sim \mathcal{O}(n)$, one can think of two possible scenarios: In the high-dimensional setup where $n \gg 1$, λ will be very small and (4.3) will be similar to solving the Maximum-Likelihood problem which is known to result in an unbiased estimator. In the low-dimensional case λ is chosen via cross-validation which adapts the problem to the sparsity level.

Remark 4. Finally, the choice of the ℓ_1-regularized maximum-likelihood estimation in (4.3) is motivated by its superior performance over ML estimation in the compressive sensing literature and the applications of interest in this chapter where exhibit sparse activity patterns in time. It is well-known that Laplace distribution is not a compressible distribution [94]. Therefore we have not made the assumption that the innovations $\mathbf{w}_t$ are Laplace distributed, but have made the observation that an ℓ_1-regularized ML estimator is akin to the MAP estimator for a Laplace state-space model. Theorem 7 suggests that a Laplace state-space model is asymptotically *as good as* any other compressible state-space model up to a constant factor in the error bounds.

4.2.3 Fast Iterative Solution via the EM Algorithm

Due to the high dimensional nature of the state estimation problem, algorithms with polynomial complexity exhibit poor scalability. Moreover, when the state transition matrix is not known, the dynamic CS optimization problem (4.3) is not convex in $\left((\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}\right)$. Therefore standard convex optimization solvers cannot be directly applied. This problem can be addressed by employing the Expectation-Maximization algorithm [189]. A related existing result considers weighted ℓ_1-regularization to adaptively capture the state dynamics [54]. Our approach is distinct in that we derive a fast solution to (4.3) via two nested EM algorithms, in order to jointly estimate the states and their transition matrix. The outer EM algorithm converts the estimation problem to a form suitable for the usage of the traditional Fixed Interval Smoothing (FIS) by invoking the EM interpretation of the Iterative Re-weighted Least Squares (IRLS) algorithms [18]. The inner EM algorithm performs state and parameter estimation efficiently using the FIS. We refer to our estimated as the Fast Compressible State-Space (FCSS) estimator.

The outer EM loop of FCSS

In [18], the authors established the equivalence of the IRLS algorithm as an instance of the EM algorithm for solving ℓ_1-minimization problems via the Normal/Independent (N/I) characterization of the Laplace distribution. Consider the ϵ-perturbed ℓ_1-norm as

$$\|\mathbf{x}\|_{1,\epsilon}= \sqrt{x_1^2+\epsilon^2}+\sqrt{x_2^2+\epsilon^2}+\cdots+\sqrt{x_p^2+\epsilon^2}. \tag{4.5}$$

Note that for $\epsilon = 0$, $\|\mathbf{x}\|_{1,\epsilon}$ coincides with the usual ℓ_1-norm. We define the ϵ-perturbed version of the dual problem (4.3) by

$$\underset{(\mathbf{x}_t)_{t=1}^T,\boldsymbol{\Theta}}{\text{minimize}} \quad \lambda \sum_{t=1}^{T} \frac{\|\mathbf{x}_t-\boldsymbol{\Theta}\mathbf{x}_{t-1}\|_{1,\epsilon}}{\sqrt{s_t}}+\frac{1}{n_t}\frac{\|\mathbf{y}_t-\mathbf{A}_t\mathbf{x}_t\|_2^2}{2\sigma^2}. \tag{4.6}$$

If instead of the $\ell_{1,\epsilon}$-norm, we had the square ℓ_2 norm, then the above problem could be efficiently solved using the FIS. The outer EM algorithm indeed transforms the problem of Eq. (4.6) into this form. Note that the ϵ-perturbation only adds a term of the order $\mathcal{O}(\epsilon p)$ to the estimation error bound of Theorem 7, which is negligible for small enough ϵ [18].

The problem of Eq. (4.6) can be interpreted as a MAP problem: the first term corresponds to the state-space prior $-\log\ p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\Theta}) = -\log p_{s_t}(\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1})$, where $p_{s_t}(\mathbf{x}) \sim \exp\left(-\lambda\|\mathbf{x}\|_{1,\epsilon}/\sqrt{s_t}\right)$ denoting the ϵ-perturbed Laplace distribution; the second term is the negative log-likelihood of the data given the state, assuming a zero-mean Gaussian observation noise with covariance $\sigma^2\mathbf{I}$. Suppose that at the end of the l^{th} iteration, the estimates $(\widehat{\mathbf{x}}_t^{(l)})_{t=1}^T, \widehat{\boldsymbol{\Theta}}^{(l)}$ are obtained, given the observations $(\mathbf{y}_t)_{t=1}^T$. As it is shown in Appendix A.5.2, the outer EM algorithm transforms the optimization problem to:

$$\underset{(\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}}{\text{minimize}} \quad \frac{\lambda}{2}\sum_{j=1}^{p}\sum_{t=1}^{T} \frac{(\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1})_j^2 + \epsilon^2}{\sqrt{s_t}\sqrt{\left(\widehat{\mathbf{x}}_t^{(l)} - \widehat{\boldsymbol{\Theta}}^{(l)}\widehat{\mathbf{x}}_{t-1}^{(l)}\right)_j^2 + \epsilon^2}} + \sum_{t=1}^{T}\frac{1}{n_t}\frac{\|\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t\|_2^2}{2\sigma^2}. \tag{4.7}$$

in order to find $(\widehat{\mathbf{x}}_t^{(l+1)})_{t=1}^T, \widehat{\boldsymbol{\Theta}}^{(l+1)}$. Under mild conditions, convergence of the solution of (4.7) to that of (4.3) was established in [18]. The objective function of (4.7) is still not jointly convex in $\big((\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}\big)$. Therefore, to carry out the optimization, i.e. the outer M step, we will employ another instance of the EM algorithm, which we call the inner EM algorithm, to alternate between estimating of $(\mathbf{x}_t)_{t=1}^T$ and $\boldsymbol{\Theta}$.

The inner EM loop of FCSS

Let $\mathbf{W}_t^{(l)}$ be a diagonal matrix such that

$$(\mathbf{W}_t^{(l)})_{j,j} = {s_t}^{-1/2} \left\{ \left(\widehat{\mathbf{x}}_t^{(l)} - \widehat{\mathbf{\Theta}}^{(l)} \widehat{\mathbf{x}}_{t-1}^{(l)} \right)_j^2 + \epsilon^2 \right\}^{-1/2}.$$

Consider an estimate $\widehat{\mathbf{\Theta}}^{(l,m)}$, corresponding to the m^{th} iteration of the inner EM algorithm within the l^{th} M-step of the outer EM. In this case, Eq. (4.7) can be thought of the MAP estimate of the Gaussian state-space model given by:

$$\begin{array}{ll} \mathbf{x}_t = \widehat{\mathbf{\Theta}}^{(l,m)} \mathbf{x}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda} {\mathbf{W}_t^{(l)}}^{-1}\right) \\ \mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t, & \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, n_t \sigma^2 I) \end{array}, \qquad (4.8)$$

In order to obtain the inner E step, one needs to find the density of $(\mathbf{x}_t)_{t=1}^T$ given $(\mathbf{y}_t)_{t=1}^T$ and $\widehat{\mathbf{\Theta}}^{(l,m)}\}$. Given the Gaussian nature of the state-space in Eq. (4.8), this density is a multivariate Gaussian density, whose means and covariances can be efficiently computed using the FIS. For all $t \in [T]$, the FIS performs a forward Kalman filter and a backward smoother to generate [14, 174]:

$$\mathbf{x}_{t|T}^{(l,m+1)} := \mathbb{E}\left\{ \mathbf{x}_t | (\mathbf{y}_t)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l,m)} \right\},$$

$$\mathbf{\Sigma}_{t|T}^{(l,m+1)} := \mathbb{E}\left\{ \mathbf{x}_t \mathbf{x}_t' | (\mathbf{y}_t)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l,m)} \right\},$$

and

$$\mathbf{\Sigma}_{t-1,t|T}^{(l,m+1)} = \mathbf{\Sigma}_{t,t-1|T}^{(l,m+1)} = \mathbb{E}\left\{ \mathbf{x}_{t-1} \mathbf{x}_t' | (\mathbf{y}_t)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l,m)} \right\}.$$

Note that due to the quadratic nature of all the terms involving $(\mathbf{x}_t)_{t=1}^T$, the outputs of the FIS suffice to compute the expectation of the objective function in Eq. (4.7), i.e., the inner E step, which results in:

$$\begin{aligned}\underset{\boldsymbol{\Theta}}{\text{maximize}} - \frac{\lambda}{2}\left(\boldsymbol{\Theta}\left(\sum_{t=1}^{T}\mathbf{W}_t^{(l)}\left(\mathbf{x}_{t-1|T}^{(l,m+1)}\mathbf{x'}_{t-1|T}^{(l,m+1)} + \boldsymbol{\Sigma}_{t-1|T}^{(l,m+1)}\right)\right)\boldsymbol{\Theta}^T\right)\\ + \frac{\lambda}{2}\operatorname{Tr}\left(\boldsymbol{\Theta}\left(\sum_{t=1}^{T}\mathbf{W}_t^{(l)}\left(\mathbf{x}_{t-1|T}^{(l,m+1)}\mathbf{x'}_{t|T}^{(l,m+1)} + \mathbf{x}_{t|T}^{(l,m+1)}\mathbf{x'}_{t-1|T}^{(l,m+1)} + 2\boldsymbol{\Sigma}_{t-1,t|T}^{(l,m+1)}\right)\right)\right),\end{aligned} \tag{4.9}$$

to obtain $\widehat{\boldsymbol{\Theta}}^{(l,m)}$. The solution has a closed-form given by:

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}^{(l,m+1)} = &\left(\sum_{t=1}^{T}2\mathbf{W}_t^{(l)}\left(\mathbf{x}_{t-1|T}^{(l,m+1)}\mathbf{x'}_{t-1|T}^{(l,m+1)} + \boldsymbol{\Sigma}_{t-1|T}^{(l,m+1)}\right)\right)^{-1}\\ &\left(\sum_{t=1}^{T}\mathbf{W}_t^{(l)}\left(\mathbf{x}_{t-1|T}^{(l,m+1)}\mathbf{x'}_{t|T}^{(l,m+1)} + \mathbf{x}_{t|T}^{(l,m+1)}\mathbf{x'}_{t-1|T}^{(l,m+1)} + 2\boldsymbol{\Sigma}_{t-1,t|T}^{(l,m+1)}\right)\right).\end{aligned} \tag{4.10}$$

Algorithm 1 The Fast Compressible State-Space (FCSS) Estimator

```
procedure FCSS
    Initialize: $\widehat{\mathbf{\Theta}}^{(0)} = \mathbf{0}$, $(\widehat{\mathbf{x}}_t^{(0)} = \mathbf{0})_{t=1}^T$, $(\mathbf{W}_t^{(0)} = \mathbf{0})_{t=1}^T$.
    repeat
        $l = 0$ .
        Outer E-step:
        $\mathbf{W}_t^{(l)} = \mathsf{diag}\left\{ \frac{1}{\sqrt{s_t}\sqrt{\left(\widehat{\mathbf{x}}_t^{(l)} - \widehat{\mathbf{\Theta}}^{(l)}\widehat{\mathbf{x}}_{t-1}^{(l)}\right)_j^2 + \epsilon^2}} \right\}_{j=1}^p$ .
        Outer M-step:
        repeat
            $m = 0$ .
            Inner E-Step: Find the smoothed estimates $\mathbf{x}_{t|T}^{(l,m+1)}$, $\mathbf{\Sigma}_{t|T}^{(l,m+1)}$ and $\mathbf{\Sigma}_{t-1,t|T}^{(l,m+1)}$ using a Fixed Interval Smoother for
                $$\begin{cases} \mathbf{x}_t = \widehat{\mathbf{\Theta}}^{(l,m)}\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\lambda}{\mathbf{W}_t^{(l)}}^{-1}\right) \\ \mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, n_t\sigma^2 I) \end{cases}.$$
            Inner M-Step: Update $\widehat{\mathbf{\Theta}}^{(l,m+1)}$ via Eq. (4.10).
            $m \leftarrow m + 1$.
        until convergence criteria met
        Update the estimates:
            $$\widehat{\mathbf{\Theta}}^{(l+1)} \leftarrow \widehat{\mathbf{\Theta}}^{(l,m)}, \quad \left(\widehat{\mathbf{x}}_t^{(l+1)}\right)_{t=1}^T \leftarrow \left(\mathbf{x}_{t|T}^{(l,m)}\right)_{t=1}^T.$$
        $l \leftarrow l + 1$.
    until convergence criteria met
    $\widehat{\mathbf{\Theta}} \leftarrow \widehat{\mathbf{\Theta}}^{(l)}$.   $(\widehat{\mathbf{x}}_t)_{t=1}^T \leftarrow \left(\widehat{\mathbf{x}}_t^{(l)}\right)_{t=1}^T$
end procedure
```

Remark 1. The EM algorithm is known to converge linearly in the number of iterations [138]. In applications of interest in this chapter, the stability assumption results in finite time constants and the EM algorithm can be assumed to converge in finite number of steps. As a result the complexity of the FCSS algorithm is equal to complexity per EM iteration. By virtue of the FIS procedure, the compexity of the FCSS algorithm is *linear* in T, i.e., the observation duration. As we will show in Section 4.3, this makes the FCSS algorithm scale favorably when applied to long data sets. Each iteration of the FIS algorithm requires two inversions which is of complexity $\mathcal{O}(p^3)$. Similarly, updating $\boldsymbol{\Theta}$ requires an inversion which is of complexity $\mathcal{O}(p^3)$. Altogether, complexity of the FCSS algorithm amounts to $\mathcal{O}(p^3 T)$. For some applications of interest in this chapter, such as calcium deconvolution in the denoising regime and sleep spindle detection, the inversion is performed on a diagonal matrix and hence the complexity reduces to $\mathcal{O}(T)$.

Remark 2. In order to update $\boldsymbol{\Theta}$ in the inner M-step given by E. (4.10), we have not specifically enforced the condition $\|\boldsymbol{\Theta}\| < 1$ in the maximization step. This condition is required to obtain a convergent state transition matrix which results in the stability of the state dynamics. It is easy to verify that the set of matrices $\boldsymbol{\Theta}$ satisfying $\|\boldsymbol{\Theta}\| < 1 - \eta$, is a closed convex set for small positive η, and hence one can perform the maximization in (4.10) by projection onto this closed convex set. Alternatively, matrix optimization methods with operator norm constraints can be used [141]. We have avoided this technicality by first finding the global minimum and examining the largest eigenvalue. In the applications of interest in this chapter which follow next, the largest eigenvalue has always been found to be less than 1.

4.3 Results

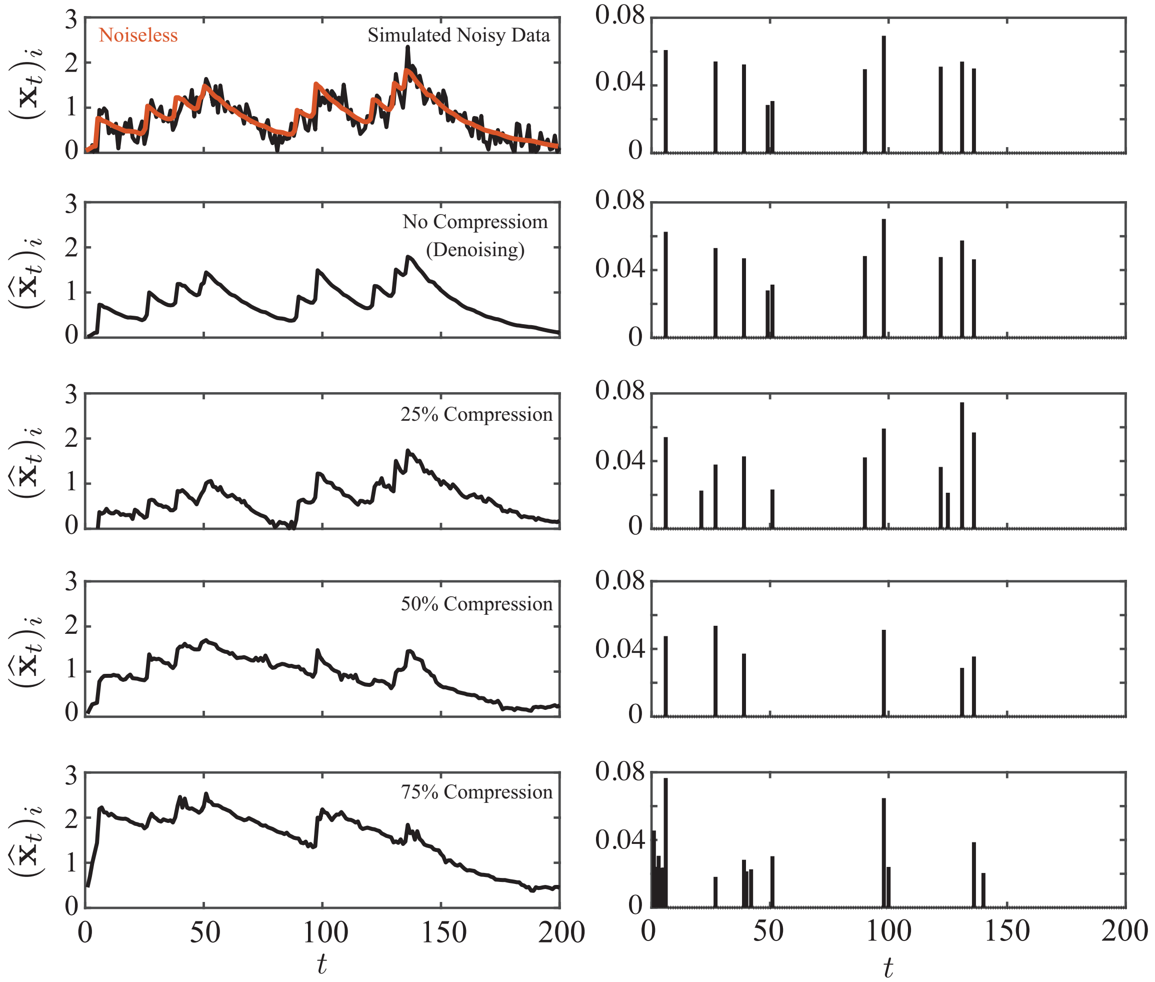
In this section, we study the performance of the FCSS estimator on simulated data as well real data from two-photon calcium imaging recordings of neuronal activity and sleep spindle detection from EEG.

4.3.1 Application to Simulated Data

We first apply the FCSS algorithm to simulated data and compare its performance with the Basis Pursuit Denoising (BPDN) algorithm. The parameters are chosen as $p = 200, T = 200, s_1 = 8, s_2 = 4, \epsilon = 10^{-10}$, and $\mathbf{\Theta} = 0.95\mathbf{I}$. We define the quantity $1 - n/p$ as the compression ratio. We refer to the case of $n_t = p$, i.e., no compression, as the denoising setting. The measurement matrix $\mathbf{A}$ is an $n_t \times p$ i.i.d. Gaussian random matrix, where n_t is chosen such that $\frac{s_t}{n_t}$ is a fixed ratio. An initial choice of $\lambda \geq 2\sqrt{2}\sigma\sqrt{\frac{\log p}{n_t}}$ is made inspired by the theory of LASSO [148], which is further tuned using two-fold cross-validation.

Figures 4.1–(a) and 4.1–(b) show the estimated states as well as the innovations for different compression ratios for one sample component. In the denoising regime, all the innovations (including the two closely-spaced components) are exactly recovered. As we take fewer measurements, the performance of the algorithm degrades as expected. However, the overall structure of the innovation sequence is captured even for highly compressed measurements.

Figure 4.2 shows the MSE comparison of the FCSS vs. BPDN, where the MSE is defined as $10 \log_{10} \frac{1}{T} \sum_{t=1}^{T} \|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2$. The FCSS algorithm significantly outperforms



(a) Reconstructed States (b) Reconstructed Spikes

Figure 4.1: Reconstruction results of FCSS on simulated data vs. compression levels. (a) reconstructed states, (b) reconstructed spikes. The FCSS estimates degrade gracefully as the compression level increases.

BPDN, especially at high SNR values. Figure 4.3 compares the performance of FCSS and BPDN on a sample component at a compression level of $n/p = 2/3$, in order to visualize the performance gain implied by Figure 4.3.

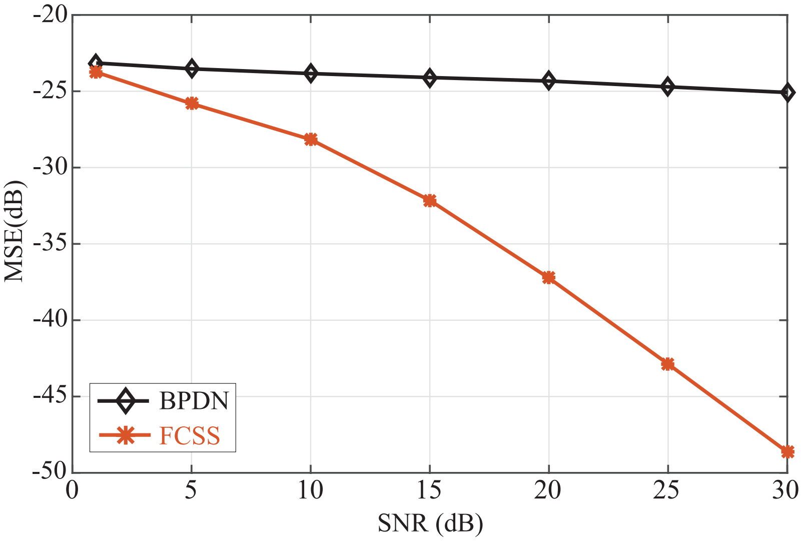


Figure 4.2: MSE vs. SNR comparison betwee FCSS and BPDN. The FCSS significantly outperforms the BPDN, even for moderate SNR levels.

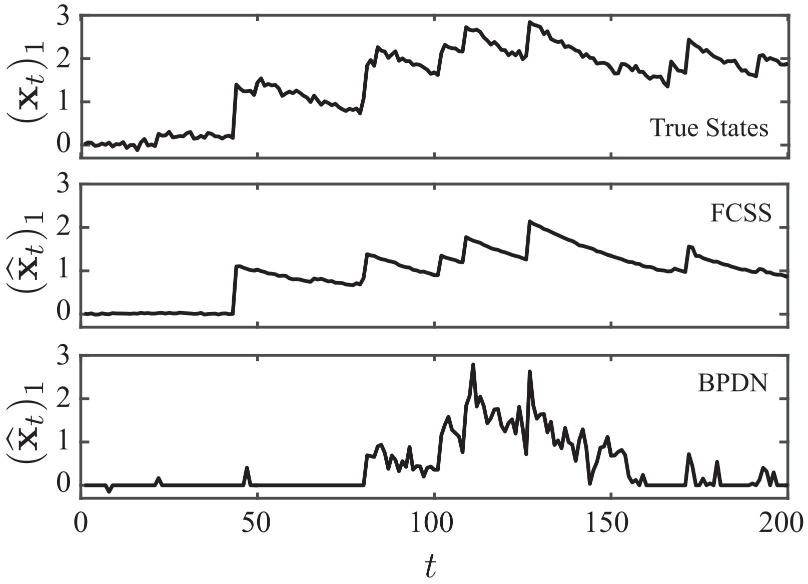


Figure 4.3: Example of state reconstruction results for FCSS and BPDN. Top: true states, Middle: FCSS state estimates, Bottom: BPDN estimates. The FCSS reconstruction closely follows the true state evolution, while the BPDN fails to capture the state dynamics.

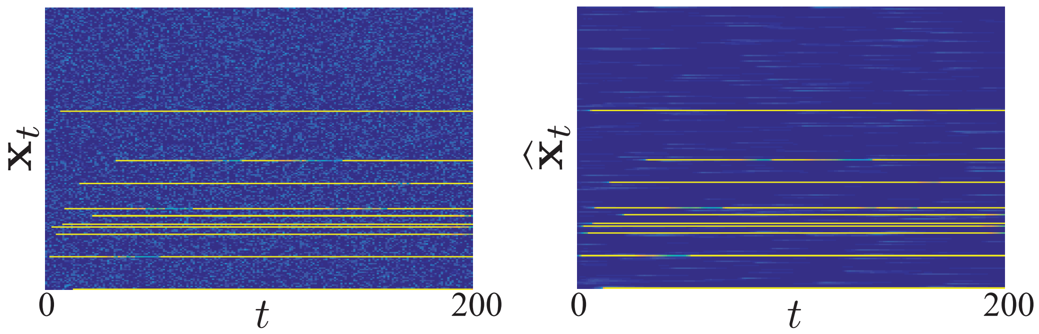


Figure 4.4: Raster plots of the observed and estimated states via FCSS. Left: noisy observations, Right: FCSS estimates. The FCSS significantly denoises the observed states.

Finally, Figure 4.4 shows the comparison of the estimated states for the entire simulated data in the denoising regime. As can be observed from the figure, the sparsity pattern of the states and innovations are captured while significantly denoising the observed states.

4.3.2 Application to Calcium Signal Deconvolution

Calcium imaging takes advantage of intracellular calcium flux to directly visualize calcium signaling in living neurons. This is done by using calcium indicators, which are fluorescent molecules that can respond to the binding of calcium ions by changing their fluorescence properties and using a fluorescence or two-photon microscope and a CCD camera to capture the visual patterns [190, 197]. Since spikes are believed to be the units of neuronal computation, inferring spiking activity from calcium recordings, referred to as calcium deconvolution, is an important problem in neural data analysis. Several approaches to calcium deconvolution have been proposed in the neuroscience literature, including model-free approaches such as sequential Monte Carlo methods [219] and model-based approaches such as non-negative deconvolution methods [166, 218]. These approaches require solving convex optimization problems, which do not scale well with the temporal dimension of the data. In addition, they lack theoretical performance guarantees and do not provide clear measures for assessing the statistical significance of the detected spikes.

In order to construct confidence bounds for our estimates, we employ recent results from high-dimensional statistics [211]. We first compute the confidence intervals around the outputs of the FCSS estimates using the node-wise regression procedure of [211], at a confidence level of $1 - \frac{\alpha}{2}$. We perform the node-wise regression separately for each time t. For an estimate $\widehat{\mathbf{x}}_t$, we obtain $\widehat{\mathbf{x}}_t^{\mathsf{u}}$ and $\widehat{\mathbf{x}}_t^{\mathsf{l}}$ as the upper and lower confidence bounds, respectively. Next, we partition the estimates into small segments, starting with a local minimum (trough) and ending in a local maximum (peak). For the i^{th} component of the estimate, let t_{min} and t_{max} denote the time index corresponding to two such consecutive troughs and peaks. If the difference $(\widehat{\mathbf{x}}_{t_{\mathsf{max}}}^{\mathsf{l}})_i - (\widehat{\mathbf{x}}_{t_{\mathsf{min}}}^{\mathsf{u}})_i$ is positive, the detected innovation component is declared significant (i.e., spike) at a confidence level of $1 - \alpha$, otherwise it is discarded (i.e., no spike). We refer to this procedure as Pruned-FCSS (PFCSS).

We first apply the FCSS algorithm for calcium deconvolution in a scenario where the ground-truth spiking is recorded *in vitro* through simultaneous electrophysiology (cell-attached patch clamp) and two-photon calcium imaging. The calcium trace as well as the ground-truth spikes are shown for a sample neuron in Figure 4.5–(a). The FCSS denoised estimate of the states (black) and the detected spikes (blue) using 95% confidence intervals (orange hulls) and the corresponding quantities for the

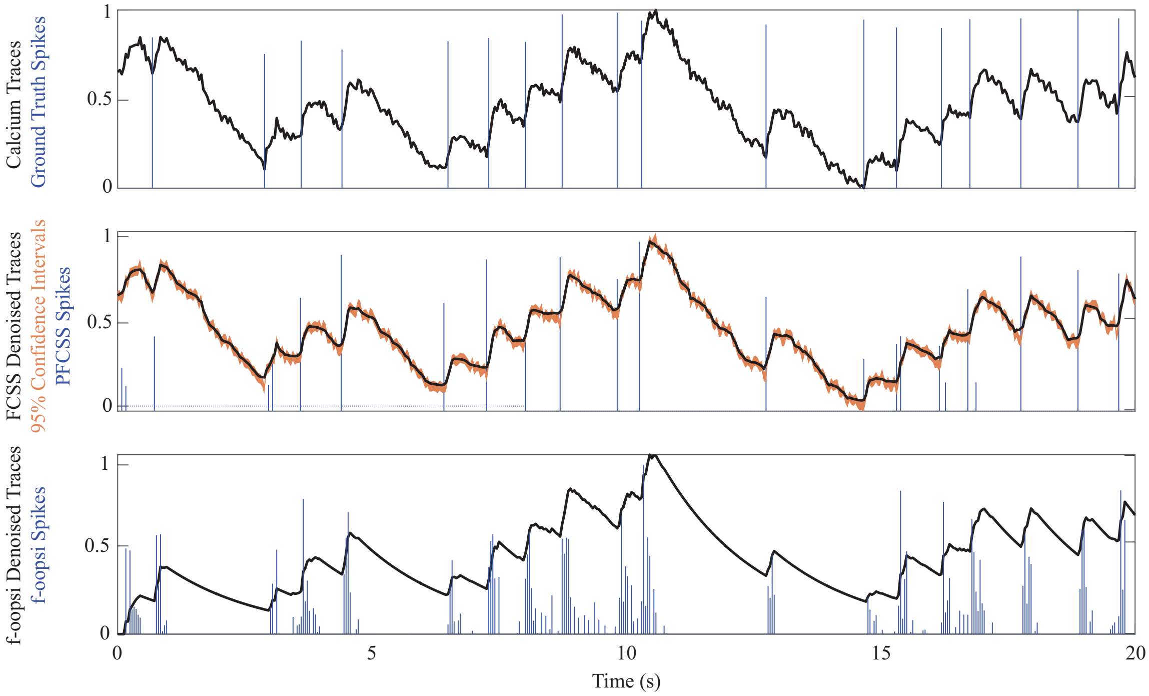
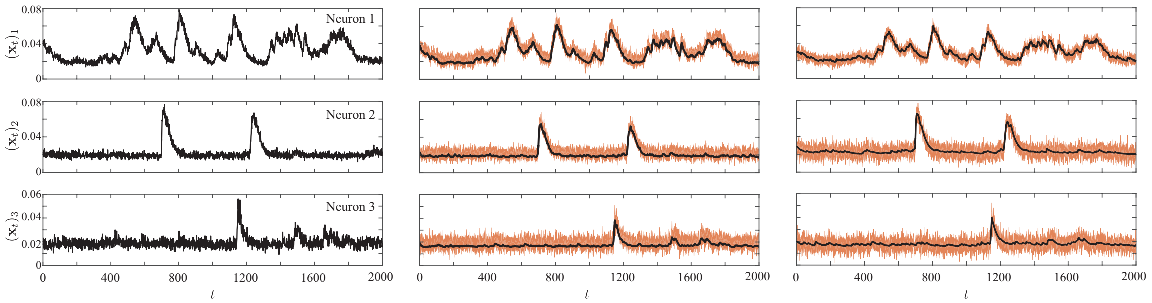


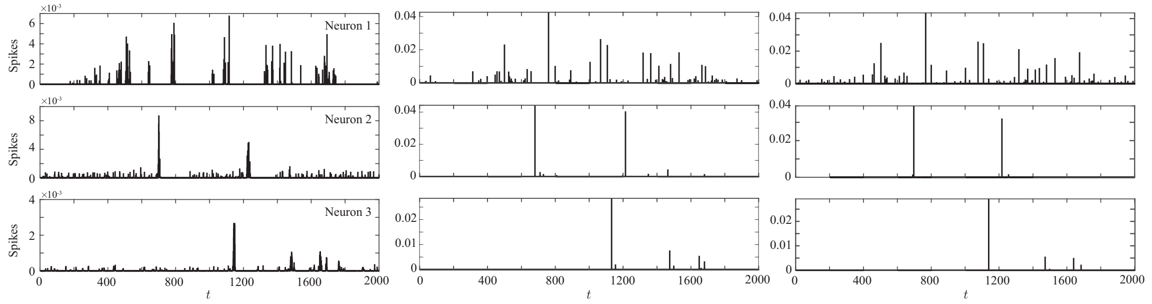
Figure 4.5: Ground-truth performance comparison between PFCSS and constrained f-oopsi. Top: the observed calcium traces (black) and ground-truth electrophysiology data (blue), Middle: PFCSS state estimates (black) with 95% confidence intervals (orange) and the detected spikes (blue), Bottom: the constrained f-oopsi state estimates (black) and the detected spikes (blue). The FCSS spike estimates closely match the ground-truth spikes with only a few false detections, while the constrained f-oopsi estimates contain significant clustered false detections.

constrained f-oopsi algorithm [166] are shown in Figures 4.5–(b) and –(c), respectively. Both algorithms detect the large dynamic changes in the data, corresponding to the spikes, which can also be visually captured in this case. However, in doing so, the f-oopsi algorithm incurs a high rate of false positive errors, manifested as clustered spikes around the ground truth events. Similar to f-oopsi, most state-of-the-art model-based methods suffer from high false positive rate, which makes the inferred spike estimates unreliable. Thanks to the aforementioned pruning process



(a) Raw Calcium Traces (b) Reconstructed States via FCSS $(n = p)$ (c) Reconstructed States via FCSS $(n/p = 2/3)$

Figure 4.6: Performance of FCSS on large-scale calcium imaging data. Top, middle and bottom rows correspond to three selected neurons labeled as Neuron 1, 2, and 3, respectively. (a) raw calcium traces, (b) FCSS reconstruction with no compression, (c) FCSS reconstruction with 2/3 compression ratio. Orange hulls show 95% confidence intervals. The FCSS significantly denoises the observed traces in both the uncompressed and compressed settings.



(a) Reconstructed Spikes via f-oopsi (b) Reconstructed Spikes via PFCSS $(n = p)$ (c) Reconstructed Spikes via PFCSS $(n/p = 2/3)$

Figure 4.7: Reconstructed spikes of PFCSS and constrained f-oopsi from large-scale calcium imaging data. Top, middle and bottom rows correspond to three selected neurons labeled as Neuron 1, 2, and 3, respectively. (a) constrained f-oopsi spike estimates, (b) PFCSS spike estimates with no compression, (c) PFCSS spike estimates with 2/3 compression ratio. The PFCSS estimates in both the uncompressed and compressed settings are sparse in time, whereas the constrained f-oopsi estimates are in the form of clustered spikes.

based on the confidence bounds, the PFCSS is capable of rejecting the insignificant innovations, and hence achieve a lower false positive rate. One factor responsible for this performance gap can be attributed to the underestimation of the calcium

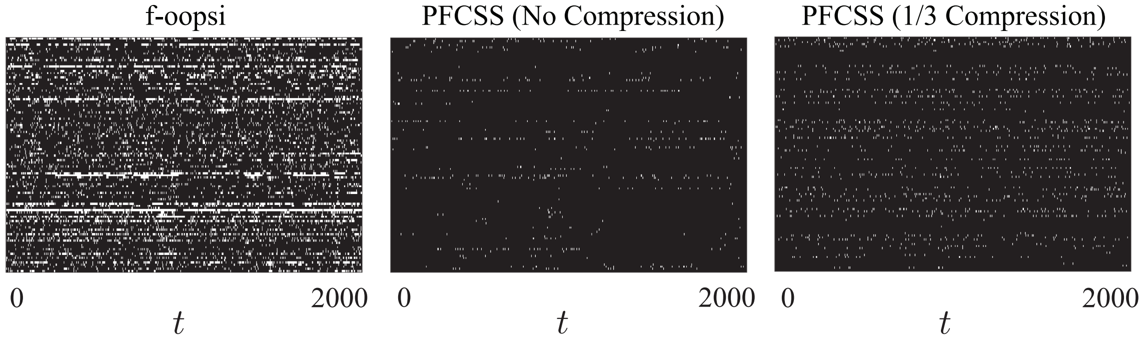


Figure 4.8: Raster plot of the estimated spikes from large-scale calcium imaging data. Left: the constrained f-oopsi estimates, Middle: PFCSS estimates with no compression, Right: PFCSS estimates with a $\frac{1}{3}$ compression ratio. The PFCSS estimates are spatiotemporally sparse, whereas the constrained f-oopsi outputs temporally clustered spike estimates.

decay rate in the transition matrix estimation step of f-oopsi. However, we believe the performance gain achieved by FCSS is mainly due to the explicit modeling of the sparse nature of the spiking activity by going beyond the Gaussian state-space modeling paradigm.

Next, we apply the FCSS algorithm to large-scale *in vivo* calcium imaging recordings, for which the ground-truth is not available due to measurement constraints. The data used in our analysis was recorded from 219 spontaneously active neurons in mouse auditory cortex. The two-photon microscope operates at a rate of 30 frames per second. We chose $T = 2000$ samples corresponding to 1 minute for the analysis. We chose $p = 108$ well-separated neurons visually. We estimate the measurement noise variance by appropriate re-scaling of the power spectral density in the high frequency bands where the signal is absent. We chose a value of $\epsilon = 10^{-10}$. It is important to note that estimation of the measurement noise variance is critical, since it affects the width of the confidence intervals and hence the detected spikes. Moreover, we estimate the baseline fluorescence by averaging the signal over values within a factor of 3 standard deviations of the noise. By inspecting Eq. (4.3), one can see a trade-off between the choice of λ and the estimate of the observation noise variance σ^2. We have done our analysis in both the compression regime, with a compression ratio of 1/3 ($n/p = 2/3$), and the denoising regime. The measurements in the compressive regime were obtained from applying i.i.d. Gaussian random matrices to the observed calcium traces. The latter is done to motivate the use of compressive imaging, as opposed to full sampling of the field of view.

Figure 4.6–(a) shows the observed traces for four selected neurons. The reconstructed states using FCSS in the compressive and denoising regimes are shown in Figures 4.6–(b) and –(c), respectively. The 90% confidence bounds are shown as orange hulls. The FCSS state estimates are significantly denoised while preserving the calcium dynamics. Figure 4.7 shows the detected spikes using constrained f-oopsi and PFCSS in both the compressive and denoising regimes. Finally, Figure 4.8 shows the corresponding raster plots of the reconstructed spikes for the entire ensemble of neurons. Similar to the preceding application on ground-truth date, the f-oopsi algorithm detects clusters of spikes, whereas the PFCSS procedure results in sparser spike detection. This results in the detection of seemingly more active neurons in the raster plot. However, motivated by the foregoing ground-truth analysis, we believe that a large fraction of these detected spikes may be due to false positive errors. Strikingly, even with a compression ratio of 1/3 the performance of the PFCSS is similar to the denoising case. The latter observation corroborates the feasibility of compressed two-photon imaging, in which only a random fraction of the field of view is imaged, which in turn can result in higher acquisition rates.

In addition to the foregoing discussion on the comparisons in Figures 4.5, 4.6, 4.7, and 4.8, two remarks are in order. First, the iterative solution at the core of FCSS is linear in the observation length and hence significantly faster than the batch-mode optimization procedure used for constrained f-oopsi. Our comparisons suggest that the FCSS reconstruction is at least 3 times faster than f-oopsi for moderate data sizes of the order of tens of minutes. Moreover, the vector formulation of FCSS allows for easy parallelization (without the need for GPU implementations), which allows simultaneous processing of ROI's without losing speed. As a numerical example the results of Figure 4.6–(b) took an average of 60-70 seconds to calculate for all $p = 108$ ROI's and $T = 2000$ frames on an Apple Macintosh desktop computer. Second, using only about two-thirds of the measurements achieves similar results by FCSS as using the full measurements.

4.3.3 Application to Sleep Spindle Detection

In this section we use compressible state-space models in order to model and detect sleep spindles. A sleep spindle is a burst of oscillatory brain activity manifested in the EEG that occurs during stage 2 non-rapid eye movement (NREM) sleep. It consists of stereotypical 12–14 Hz wave packets that last for at least 0.5 seconds [68]. The spindles occur with a rate of 2–5% in time, which makes their generation an appropriate candidate for compressible dynamics. Therefore, we hypothesize that the spindles can be modeled using a combination of few echoes of the response of a second order compressible state-space model. As a result, the spindles can be decomposed as sums of modulated sine waves.

In order to model the oscillatory nature of the spindles, we consider a second order autoregressive (AR) model where the pole locations are given by $ae^{-j2\pi\frac{f}{f_s}}$ and $ae^{+j2\pi\frac{f}{f_s}}$, where $0 < a < 1$ is a positive constant controlling the effective duration of the impulse response, f_s is the sampling frequency and f is a putative frequency accounting for the dominant spindle frequency. The equivalent state-space model for

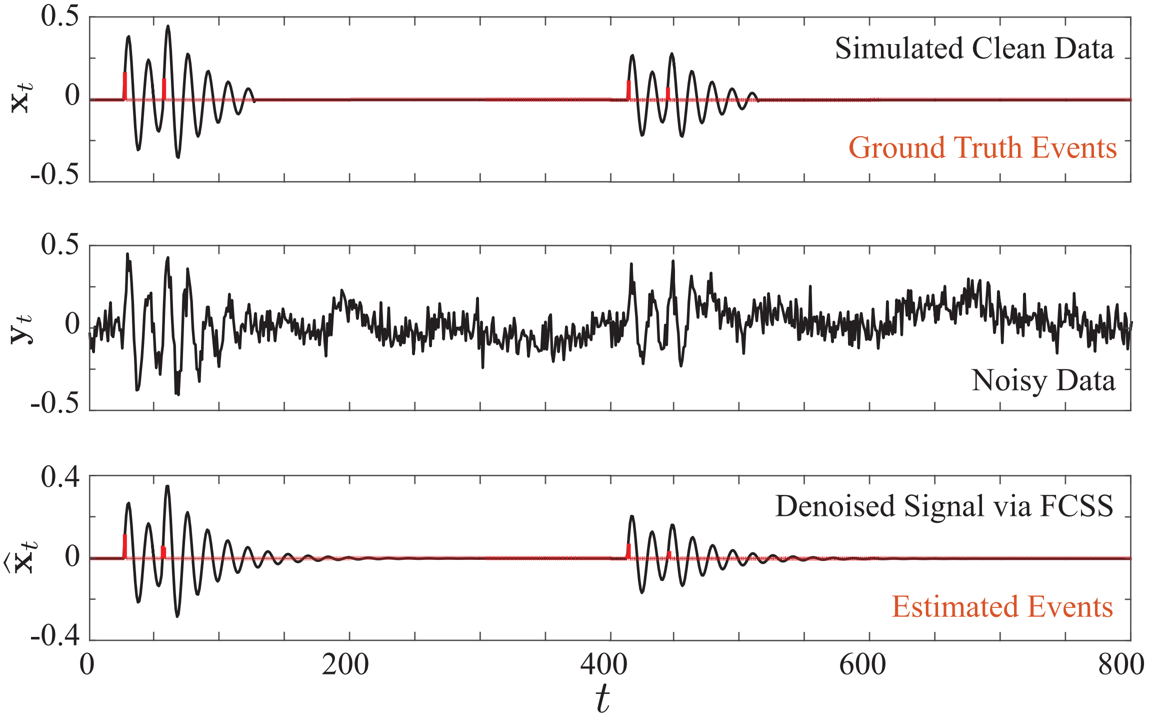


Figure 4.9: Performance of FCSS on simulated spindles. Top: simulated clean data (black) and ground-truth spindle events (red), Middle: simulated noisy data, Bottom: the denoised signal (black) and deconvolved spindle events (red). The FCSS estimates are significantly denoised and closely match the ground-truth data shown in the top panel.

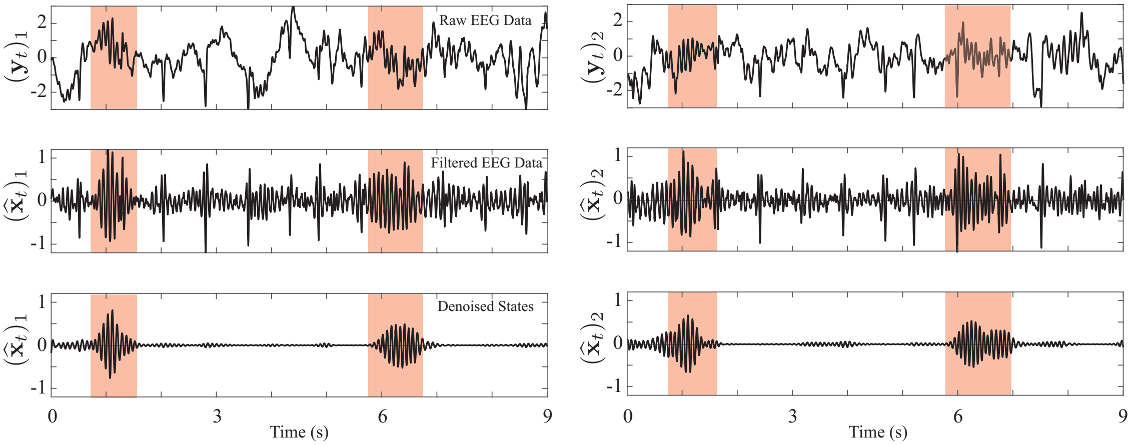


Figure 4.10: Performance comparison between FCSS and band-pass filtered EEG data. Left and right panels correspond to two selected electrodes labeled as 1 and 2, respectively. The orange blocks show the extent of the detected spindles by the expert. Top: raw EEG data, Middle: band-pass filtered EEG data in the 12–14 Hz band, Bottom: FCSS spindle estimates. The FCSS estimates closely match the expert annotations, while the band-pass filtered data contains significant signal components outside of the orange blocks.

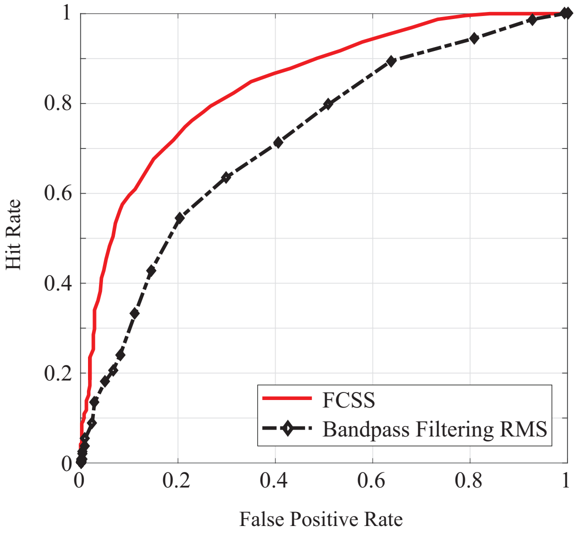


Figure 4.11: The ROC curves for FCSS (solid red) and bandpass-filtered RMS (dashed black). The FCSS outperforms the widely-used band-pass filtered RMS method as indicated by the ROC curves.

which the MAP estimation admits the FCSS solution is therefore:

$$\begin{aligned} \mathbf{x}_t &= 2a\cos\left(2\pi\tfrac{f}{f_s}\right)\mathbf{x}_{t-1} - a^2\mathbf{x}_{t-2} + \mathbf{w}_t, \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t. \end{aligned} \tag{4.11}$$

Note that impulse response of the state-space dynamics is given by $h_n = a^n \cos\left(2\pi\frac{f}{f_s}n\right) u_n$, which is a stereotypical decaying sinusoid. By defining the augmented state $\widetilde{\mathbf{x}}_t = [\mathbf{x}'_t, \mathbf{x}'_{t-1}]'$, Eq. (4.11) can be expressed in the canonical form:

$$\widetilde{\mathbf{x}}_t = \widetilde{\boldsymbol{\Theta}}\widetilde{\mathbf{x}}_{t-1} + \widetilde{\mathbf{w}}_t, \qquad \mathbf{y}_t = \widetilde{\mathbf{A}}_t\widetilde{\mathbf{x}}_t + \mathbf{v}_t, \tag{4.12}$$

where $\widetilde{\mathbf{w}}_t := [\mathbf{w}'_t, \mathbf{0}']'$, $\widetilde{\mathbf{A}}_t = \left[\mathbf{A}_t \,\middle|\, \mathbf{0}\right]$ and

$$\widetilde{\boldsymbol{\Theta}} = \left[\begin{array}{c:c} 2a\cos\left(2\pi\frac{f}{f_s}\right)\mathbf{I} & -a^2\mathbf{I} \\ \hdashline \mathbf{I} & \mathbf{0} \end{array}\right].$$

Eq. (4.9) can be used to update $\widetilde{\boldsymbol{\Theta}}$ in the M step of the FCSS algorithm. However, $\widetilde{\boldsymbol{\Theta}}$ has a specific structure in this case, determined by a and f, which needs to be taken into account in the optimization step. Let $\phi =: 2a\cos\left(2\pi\frac{f}{f_s}\right)$ and $\psi = a^2$, and let

$$\sum_{t=1}^{T} \widetilde{\mathbf{W}}_t^{(l)} \left(\widetilde{\mathbf{x}}_{t-1|T}^{(l,m+1)} \widetilde{\mathbf{x}}'^{(l,m+1)}_{t-1|T} + \widetilde{\boldsymbol{\Sigma}}_{t-1|T}^{(l,m+1)}\right) =: \left[\begin{array}{c:c} \mathbf{A} & \mathbf{B} \\ \hdashline \mathbf{C} & \mathbf{D} \end{array}\right],$$

$$\sum_{t=1}^{T} \widetilde{\mathbf{W}}_t^{(l)} \left(\widetilde{\mathbf{x}}_{t-1|T}^{(l,m+1)} \widetilde{\mathbf{x}}'^{(l,m+1)}_{t|T} + \widetilde{\mathbf{x}}_{t|T}^{(l,m+1)} \widetilde{\mathbf{x}}'^{(l,m+1)}_{t-1|T} + 2\widetilde{\boldsymbol{\Sigma}}_{t-1,t|T}^{(l,m+1)} \right) =: \left[\begin{array}{c:c} \mathbf{E} & \mathbf{F} \\ \hdashline \mathbf{G} & \mathbf{H} \end{array} \right].$$

Then, Eq. (4.9) is equivalent to maximizing

$$\underset{\phi,\psi}{\text{maximize}} \quad \frac{\lambda}{2}(\phi^2 \text{Tr}(\mathbf{A}) - \phi\psi \text{Tr}(\mathbf{B}+\mathbf{C}) + \psi^2 \text{Tr}(\mathbf{D}) + \text{Tr}(\mathbf{A})) - \tfrac{\lambda}{2}(\phi \text{Tr}(\mathbf{E}) - \psi \text{Tr}(\mathbf{G}) + \text{Tr}(\mathbf{F})). \tag{4.13}$$

subject to $0 \leq \psi \leq 1$ and $\phi^2 \leq 4\psi$, which can be solved using interior point methods. In our implementation, we have imposed additional priors of $f \sim \mathsf{Uniform}(12, 14)$ Hz and $a \sim \mathsf{Uniform}(0.95, 0.99)$, which simplifies the constraints on ϕ and ψ to

$$0.95^2 \leq \psi \leq 0.99^2, \quad \text{and} \quad 4\psi \cos^2\left(2\pi \tfrac{12}{f_s}\right) \leq \phi^2 \leq 4\psi \cos^2\left(2\pi \tfrac{14}{f_s}\right).$$

Given the convexity of the cost function in (4.13), one can conclude from the KKT conditions that if the global minimum is not achieved inside the region of interest it must be achieved on the boundaries.

Figure 4.9, top panel, shows two instances of simulated spindles (black traces) with parameters $f_s = 200$ Hz, $f = 13$ Hz and $a = 0.95$, with the ground truth events generating the wave packets shown in red. The middle panel shows the noisy version of the data with an SNR of -7.5 dB. The noise was chosen as white Gaussian noise plus slowly varying (2 Hz) oscillations to resemble the slow oscillations in real EEG data. As can be observed the simulated signal exhibits visual resemblance to real spindles, which verifies the hypothesis that spindles can be decomposed into few combinations of wave packets generated by a second order AR model. The third panel, shows the denoised data using FCSS, which not only is successful in detecting the the ground-truth dynamics (red bars), but also significantly denoises the data.

We next apply FCSS to real EEG recordings from stage-2 NREM sleep. Manual scoring of sleep spindles can be very time-consuming, and achieving accurate manual scoring on a long-term recording is a highly demanding task with the associated risk of decreased diagnosis. Although automatic spindle detection would be attractive, most available algorithms sensitive to variations in spindle amplitude and frequency that occur between both subjects and derivations, reducing their effectiveness [150, 183]. Moreover most of these algorithms require significant pre- and post-processing and manual tuning. Examples include algorithms based on Empirical Mode Decomposition (EMD) [52, 82, 232], data-driven Bayesian methods [20], and machine learning approaches [73, 216]. Unlike our approach, none of the existing methods consider modeling the generative dynamics of spindles, as transient sparse events in time, in the detection procedure.

The data used in our analysis is part of the recordings in the DREAMS project [69], recorded using a 32-channel polysomnograpgh. We have used the EEG channels in our analysis. The data was recorded at a rate of $f_s = 200$ Hz for 30 minutes. The data was scored for sleep spindles independently by two experts. We have used expert annotations to separate regions which include spindles for visualization purposes. For comparison purposes, we use a bandpass filtered version of the data within 12–14 Hz, which is the basis of several spindle detection algorithms [61, 104, 183], hallmarked by the widely-used bandpass filtered root-mean-square (RMS) method [183].

Figure 4.10 shows the detection results along with the bandpass filtered version of the data for two of the EEG channels. The red bars show the expert markings of the onset and offset of the spindle events. The FCSS simultaneously captures the spindle events and suppressed the activity elsewhere, whereas the bandpass filtered data produces significant activity in the 12–14 Hz throughout the observation window, resulting in high false positives. To quantify this observation, we have computed the ROC curves of the FCSS and bandpass filtering followed by root mean square (RMS) computation in Figure 4.11, which confirms the superior performance of the FCSS algorithm over the data set. The annotations of one of the experts has been used for as the ground truth benchmark.

4.4 Discussion

In this section, we discuss the implication of our techniques in regard to the application domains as well as existing methods.

4.4.1 Connection to existing literature in sparse estimation

Contrary to the traditional compressive sensing, our linear measurement operator does not satisfy the RIP [17], despite the fact that $\mathbf{A}_t$'s satisfy the RIP. Nevertheless, we have extended the near-optimal recovery guarantees of CS to our compressible state-space estimation problem via Theorem 7. Closely related problems to our setup are the super-resolution and sparse spike deconvolution problems [48, 74], in which abrupt changes with minimum separation in time are resolved in fine scales using coarse (lowpass filtered) frequency information, which is akin to working in the compressive regime.

Theoretical guarantees of CS require the number of measurements to be roughly proportional to the sparsity level for stable recovery [148]. These results do not readily generalize to the cases where the sparsity lies in the dynamics, not the states per se. Most of the dynamic compressive sensing techniques such as Kalman filtered compressed sensing, assume partial information about the support or estimate them in a greedy and often ad-hoc fashion [51, 214, 215, 234, 239]. As another example, the problem of recovering discrete signals which are approximately sparse in their gradients using compressive measurements, has been studied in the literature using Total Variation (TV) minimization techniques [146,168]. For one-dimensional signals, since the gradient operator is not orthonormal, the Frobenius operator norm of its inverse grows linearly with the discretization level [146]. Therefore, stability results of TV minimization scale poorly with respect to discretization level. In higher dimensions, however, the fast decay of Haar coefficients allow for near-optimal theoretical guarantees [62]. A major difference of our setup with those of CS for TV-minimization is the *structured* and *causal* measurements, which unlike the non-causal measurements in [146], do not result in an overall measurement matrix satisfying RIP. We have considered dynamics with convergent transition matrices, in order to generalize the TV minimization approach. To this end, we showed that using the state-space dynamics one can infer temporally global information from local and causal measurements. Another closely related problem is the fused lasso [204] in which sparsity is promoted both on the covariates and their differences.

4.4.2 Application to calcium deconvolution

In addition to scalability and the ability to detect abrupt transitions in the states governed by discrete events in time (i.e., spikes), our method provides several other benefits compared to other spike deconvolution methods based on state-space models, such as the constrained f-oopsi algorithm. First, our sampling-complexity trade-offs are known to be optimal from the theory of compressive sensing, whereas no performance guarantee exists for constrained f-oopsi. Second, we are able to construct precise confidence intervals on the estimated states, whereas constrained f-oopsi does not produce confidence intervals over the detected spikes. A direct consequence of these confidence intervals is estimation of spikes with high fidelity and low false alarm. Third, our comparisons suggest that the FCSS reconstruction is at least 3 times faster than f-oopsi for moderate data sizes of the order of tens of minutes. Finally, our results corroborate the possibility of using compressive measurement for reconstruction and denoising of calcium traces. From a practical point of view, a compressive calcium imaging setup can lead to higher scanning rate as well as better reconstructions, which allows monitoring of larger neuronal populations [165]. Due to the structured nature of our sampling and reconstruction schemes, we can avoid prohibitive storage problems and benefit from parallel implementations.

4.4.3 Application to sleep spindle detection

Another novel application of our modeling and estimating framework is to case sleep spindle generation as a second-order dynamical system governed by compressive innovations, for which FCSS can be efficiently used to denoise and detect the spindle events. Our modeling framework suggest that spectrotemporal spindle dynamics cannot be fully captured by just pure sinusoids via bandpass filtering, as the data consistently contains significant 12–14 Hz oscillations almost everywhere (See Figure 4.10). Therefore, using the bandpass filtered data for further analysis purposes clearly degrades the performance of the resulting spindle detection and scoring algorithms. The FCSS provides a robust alternative to bandpass filtering in the form of model-based denoising.

In contrast to state-of-the-art methods for spindle detection, our spindle detection procedure requires minimal pre- and post-processing steps. We expect similar properties for higher order AR dynamics, which form a useful generalization of our methods for deconvolution of other transient neural signals. In particular, K-complexes during the stage 2 NREM sleep form another class of transient signals with high variability. A potential generalization of our method using higher order models can be developed for simultaneous detection of K-complexes and spindles.

4.5 Concluding Remarks

In this chapter, we considered estimation of compressible state-space models, where the state innovations consist of compressible discrete events. For dynamics with convergent state transition dynamics, using theory of compressed sensing we provided an optimal error bound and stability guarantees for the dynamic ℓ_1-regularization algorithm which is akin to the MAP estimator for a Laplace state-space model. We also developed a fast and low-complexity iterative algorithm, namely FCSS, for estimation of the states as well as their transition matrix. We further verified the validity of our theoretical results through simulation studies as well as application to spike deconvolution from calcium traces and detection of sleep spindles from EEG data. Our methodology has two unique major advantages: first, we have proven theoretically why our algorithm performs well, and characterized its error performance. Second, we have developed a fast algorithm, with guaranteed convergence to a solution of the deconvolution problem, which for instance, is ~ 3 times faster than the widely-used f-oopsi algorithm in calcium deconvolution applications.

While we focused on two specific application domains, our modeling and estimation techniques can be generalized to apply to broader classes of signal deconvolution problems: we have provided a framework to model transient phenomena which are driven by sparse generators in time domain, and whose event onsets are of importance. Examples include heart beat dynamics and rapid changes in the covariance structure of neural data (e.g., epileptic seizures). In the spirit of easing reproducibility, we have made a MATLAB implementation of our algorithm publicly available [5].

Chapter 5: Multiplicative Updates for Optimization Problems with Dynamics

In this chapter we consider the problem of optimizing general convex objective functions with nonnegativity constraints. Using the Karush-Kuhn-Tucker (KKT) conditions for the nonnegativity constraints we will derive fast multiplicative update rules for several problems of interest in signal processing, including nonnegative deconvolution, point-process smoothing, ML estimation for Poisson observations, nonnegative least squares and nonnegative matrix factorization (NMF). Our algorithm can also account for temporal and spatial structure and regularization . We will analyze the performance of our algorithm on simultaneously recorded neuronal calcium imaging and electrophysiology data.

5.1 Introduction

The advent of big data has given rise to new challenges in signal processing. Fast and scalable solvers for solving large optimization problems remains a big challenge of optimization theory. In this chapter we consider the problem of solving general optimization problems under nonnegativity constraints. Such optimization problems arise in many applications of interest. Examples include nonnegative matrix factorization for images of objects [125], Poisson image reconstruction [223], point process smoothing for stimulus-response experiments in neurophysiology [191], nonnegative least squares [124] and nonnegative calcium deconvolution [111]. In this chapter we will use the KKT conditions [36] to provide a unified framework for solving such optimization problems with nonnegativity constraints. As we will see these conditions naturally lead to multiplicative updates with suitable convergence in many applications.

Multiplicative updates have been used for solving ML and MAP estimation as well as KL-divergence minimization. Many of these algorithms are special cases of the so-called proximal backward-forward scheme [154]. These algorithms try to find fixed points of a set of equations resulting from setting gradients of the objective function to zero. A With the help of parallel computing and graphics processing units (GPUs), these iterative methods can be solved very fast. Therefore, they become increasingly important. An important application of these multiplicative updates is the Richardson-Lucy (RL) algorithm for image deconvolution [133], which is widely used in astronomy and microscopy [187]. The RL algorithm recovers the ML estimate of a sample under Poisson statistics [177].

Multiuplicative updates are commonly contrasted with gradient descent methods. Their update steps do not necessarily follow the direction of the steepest descent. Multiplicative updates are argued to be insensitive to noise and more flexible [229]. Despite fast early convergenece multiplicative updates are claimed to converge slowly in later stages [222]. However, this argument has been refuted for Poisson image reconstruction [229], the Weiszfeld problem [154] and NMF [126] by showing their equivalence to a Majorization Minimization (MM) algorithm which has linear convergence in iterations [226]. In contrast, both multiplicative updates and gradient descent based algorithms such as the proximal-gradient method have sublinear rate of convergence [154] in general. Moreover, with specific choices of the stepsize, in many cases such as the Weiszfeld problem these algorithms have proven to be equivalent [154] . These findings suggest that slow convergence of multiplicative updates in some cases is due to absence of strong convexity in the objective function.

An advantage of multiplicative updates over gradient descent based algorithms is their flexibility in terms of adapting to the objective functions without the need for calculation dual functions or tuning extra parameters such as the step-size. Despite the recent breakthroughs in choosing these parameters [116], each step in calculation of the step size is usually as costly as an iteration of the algorithm which is not as effective for big data problems. In addition many problems such as image reconstruction and calcium deconvolution [166] are spatially separable and are easily parallelized.

Finally, temporal dynamics and penalization play an important role in signal recovery from noisy data. Examples include state-space estimations, video reconstruction and total variation denoising problems. Apart from special cases, the solutions to these problems are generally batch mode and computationally demanding. In this chapter we provide a unified framework for generalizations of multiplicative updates to the problems with nonnegativity constraints and dynamics by adapting the update rules to different forms of penalties. We have empirically found that multiplicative updates show superior convergence properties and speed to gradient descent methods for models that include dynamics and penalization.

5.2 Problem Formulation

We consider a convex optimization problem of the form

$$\underset{\mathbf{X}\succeq 0}{\text{minimize}}\, \mathcal{F}(\mathbf{X}) := \mathcal{L}(\mathbf{X}) + \lambda \mathcal{P}(\mathbf{X}), \tag{5.1}$$

where $\mathcal{L}(.)$ denotes a convex objective function and $\mathcal{P}(.)$ denotes a suitable penalty function. Typically $\mathcal{L}(.)$ is a negative log-likelihood and $\mathcal{P}(.)$ is a smooth norm. Additionally we make the assumption that both $\mathcal{L}$ and $\mathcal{P}$ are differentiable with respect to $\mathbf{X}$ on the positive orthant,

Among the algorithms used for solving (5.1) one can name the primal-dual algorithm and proximal gradient method. For specific choices of the penalty functions ℓ_1 and ℓ_2 (Tikhonov) regularization several fast algorithms exist. However these algorithms cannot be easily generalized to arbitrary penalties or temporal dynamics. In some cases such as the gradient based methods they require knowledge of the proximal map or have extra parameters such as the step size to be tuned and chosen. Calculation of the step size is usually as costly as a few iterations of the algorithm and could slow them down. However, our approach to solving (5.1) does not require tuning of extra parameters and is very simple to implement. We will next discuss our solution.

5.3 Solution to the Main Optimization Problem

In this section we will introduce our solution to (5.1) via multiplicative updates. The Lagrangian form of (5.1) is given by

$$\underset{\mathbf{X},\mathbf{S}\succeq 0}{\text{minimize}}\, \mathcal{F}(\mathbf{X}) + \mathbf{S} \odot \mathbf{X}. \tag{5.2}$$

Assuming convexity and zero duality gap, the KKT conditions for (5.2) can be expressed as

$$\mathbf{X}^{\star} \succeq 0, \;\; \mathbf{S}^{\star} \succeq 0, \tag{5.3}$$

$$\mathbf{S}^{\star} \odot \mathbf{X}^{\star} = \mathbf{0}, \tag{5.4}$$

$$\nabla_{\mathbf{X}}\mathcal{F}(\mathbf{X}) + \mathbf{S} = \mathbf{0}. \tag{5.5}$$

In the rest of the chapter, we drop the subscripts and arguments whenever they can be understood from the context. Multiplying (5.5) by $\mathbf{X}$ and using (5.4) we obtain:

$$\nabla\mathcal{F}(\mathbf{X}) \odot \mathbf{X} = \mathbf{0}. \tag{5.6}$$

Our solution to (5.1) looks for a positive fixed point of (5.6). Therefore giving us the multiplicative update rule

$$\mathbf{X}^{(k+1)} \leftarrow \left(\nabla\mathcal{F}(\mathbf{X}^{(k)})\right)^{-} \oslash \left(\nabla\mathcal{F}(\mathbf{X}^{(k)})\right)^{+} \odot \mathbf{X}^{(k)}. \tag{5.7}$$

In all application introduced in this chapter we initialize the algorithm with a positive solution, the choice of which depends on the application. The update rule will then ensure the solution remains positive. In order to provide more insight into our algorithm we will next provide several examples and applications.

In applications of interest in this chapter we consider temporal dynamics in $\mathbf{X}$, hence referring to our algorithm by FAst DEconvolution (FADE) algorithm. In the spirit of easing reproducibility, we have made MATLAB implementations of our codes publicly available [6].

5.4 Examples and Application to Real Data

In this Section we will provide examples of the multiplicative updates in different applications of interest.

5.4.1 Nonnegative Deconvolution

In its simplest form the nonnegative deconvolution problem can be formalized by considering the state-space model given by

$$\mathbf{x}_t = \mathbf{\Theta}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \tag{5.8}$$

where $\mathbf{w}_t \succeq 0$ models the innovations at time $t \in [T]$. Usually, the observation noise is assumed to be i.i.d normal, i.e. $\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{\Sigma}_t)$ and the measurement matrices $\mathbf{A}_t$

are assumed to conserve positivity. For this problem we can identify $\mathbf{W} = \mathbf{W}_{[T]}$ and

$$\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T} \|\mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t\|_{\mathbf{\Sigma}_t}^2 = \sum_{t=1}^{T} \left\| \mathbf{y}_t - \mathbf{A}_t \sum_{\tau=0}^{t-1} \mathbf{\Theta}^{\tau} \mathbf{w}_{t-\tau} \right\|_{\mathbf{\Sigma}_t}^2 ,$$

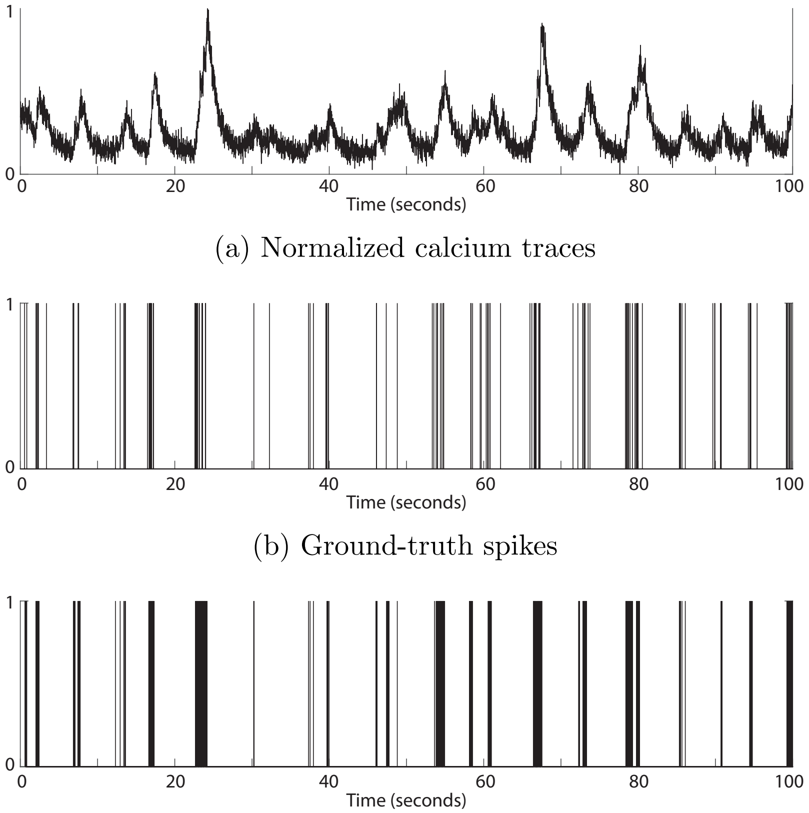
from which we can calculate

$$\begin{aligned} (\nabla_{\mathbf{w}_t} \mathcal{L}(\mathbf{W}))^+ &= \sum_{\tau \geq t} \left(\mathbf{\Theta}^{\tau - t}\right)^T \mathbf{A}_\tau^T \mathbf{\Sigma}_\tau^{-1} \mathbf{y}_\tau, \\ (\nabla_{\mathbf{w}_t} \mathcal{L}(\mathbf{W}))^- &= \sum_{\tau \geq t} \left(\mathbf{\Theta}^{\tau - t}\right)^T \mathbf{A}_\tau^T \mathbf{\Sigma}_\tau^{-1} \mathbf{A}_\tau \mathbf{x}_\tau. \end{aligned}$$

Typically one can use a smooth norm in order to enforce prior assumptions on the spikes, for example one can use a sparsity inducing prior $\mathcal{P} = \|\mathbf{W}\|_{1,1}$, for which $(\nabla \mathcal{P})^+ = \mathbf{1}$ and $(\nabla \mathcal{P})^- = \mathbf{0}$. The choice of the penalty function on the spikes is arbitrary and could differ from application to application. In applications where such information is not readily available, one would like to enforce minimal assumptions on the spikes and hence would want to enforce non-informative priors. The most famous example of such priors is known as Jeffrey's prior [106]. However this problem is an active area of research as there is no unanimously agreed upon choice of non-informative priors.

5.4.2 Application to Calcium Deconvolution

Calcium imaging is used to visualize currents associated with action potentials in living neurons. This is done using fluorescent molecules that change their fluorescence



(a) Normalized calcium traces

(b) Ground-truth spikes

(c) Deconvolved spikes

Figure 5.1: Application of the FADE algorithm to calcium deconvolution problem.

properties upon binding calcium, and using a one- or two-photon fluorescence microscope to record these changes [190, 197]. Inferring action potentials (spikes) from calcium recordings, referred to as calcium deconvolution, is an important problem in neural data analysis. For the special case of calcium imaging we have $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, $\mathbf{A}_t = \mathbf{I}$ and $\mathbf{\Theta} = \theta\mathbf{I}$. Here the baseline is assumed to have been estimated and subtracted separately, but can be estimated similarly. We refer to [111] for details on estimation of the unknown parameters σ^2 and θ and a list of methods used for calcium deconvolution. These approaches require solving convex optimization problems, which do not scale well with the temporal dimension of the data.

Figure 5.1 shows application of the FADE algorithm to simultaneously recorded imaging and electrophysiology data. The algorithm has covnverged (less than 0.5% change in spikes) in 28 iterations. The data is a 100 second interval from the spikefinder challenge [7] (dataset 3, neuron 1). We have used an AR(2) model and an $\ell_{0.5,1}$ penalty on the spikes in order to enforce temporal sparsity. The spikes have been obtained by simply thresholding the deconvolved spikes at 3σ, where σ is the estimated standard deviation of the observation noise. A comparison of the performance of our algorithm with many other methods is provided on the spikefinder challenge website [7].

One can use spatial regularization on elements of $\mathbf{w}_t$ in this setup as well as compressive sensing regimes for when $\mathbf{A}$ satisfies the restricted isometry property RIP [111]. We refer to [111] for a more detailed discussion.

5.4.3 Poisson Image Reconstruction and Point Process Smoothing

State-space models with Poisson observations have also been studied in many applications of interest. In neuroscience, temporal dynamics of stimulus-response experiments in neurophysiology have been modeled using a Poisson state-space model. In emission tomography, dynamics of the photons hitting the detectors can be modeled with Poisson noise models. Without loss of generality we consider the state-space model given by

$$\mathbf{x}_t = \mathbf{\Theta}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t \sim \text{Poisson}\left(\phi\left(\mathbf{A}\mathbf{x}_t + \mathbf{b}_t\right)\right), \tag{5.9}$$

where $\mathbf{w}_t \succeq 0$ and $\mathbf{b}_t \succeq 0$ model the spikes and baseline rates at time $t \in [T]$ respectively and $\phi(.)$ is a bijective convex function. Common examples include $\phi(x) = \exp(x)$, $\phi(x) = \frac{\exp(x)}{1+\exp(x)}$ and $\phi(x) = x$. We assume the latter in our derivations due to space considerations.

Several approaches have been proposed in the literature for finding the MAP solution to (5.9). We refer to [98] for a detailed list of these methods. In [191] the authors used the maximum a posteriori derivation of the Kalman filter and proposed an approximate expectation maximization (EM) approach to this problem by Gaussian approximations of the posterior likelihood. This EM approach has several shortcomings. First, it requires solving a nonlinear system of equations which could potentially be computationally costly. Second, it only accounts for Gaussian spikes. Third its performance heavily depends on the Poisson rate model, especially when the rates are small, which is the usual case for spiking activities. In these cases usually $\phi(x) = \exp(x)$ is considered for stability of approximations. Moreover due to nonlinear recursive filtering nature of the problem, the performance of the Gaussian approximation quickly degrades as the dimension of the latent space goes beyond 2 or 3. Similarly, in [98] the authors proposed SPIRAL which uses a Gaussian approximation to $\mathcal{L}$ and is a gradient-based solution to (5.9). Except for the special cases of ℓ_1 and TV penalties, calculation of the Gaussian model is tedious leading to slow convergence. In [200] the authors introduce a variational auto-encoder (gradient descent based) model to retrieve the low-dimensional temporal factors.

In applications such as fluorescence microscopy, it is also common to use to use variance stabilizing transforms [98] such as square root filtering [179] in order to make Gaussian approximations to the Poisson distribution. In the high photon regime such transformations are not necessary as one can use infinite divisibility property of the Poisson distribution for Gaussian approximations. However one would then need to deal with complications arising from equality of the mean and the covariance matrices for such approximations. In contrast, our algorithm gives an exact solution, is fast, can account for any rate model and suitably scales with the problem dimensions.

The Gaussian approximations could then be used as an input to a Kalman smoother if the innovations (spikes) follow a half-normal or Gaussian distribution. Despite the fact that our solutions are faster, exact and do not involve approximations, for Gaussian state-spaces the Kalman smoother provides a smoothed estimate of the covariances which could be used for building confidence intervals, whereas the covariances are not a direct output of the multiplicative updates.

Considering the MAP estimator for $\mathbf{W} = \mathbf{W}_{[T]}$ we can identify

$$\mathcal{L}(\mathbf{W}) = \sum_{t=1}^{T} \mathbf{1}^T \left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right) - \mathbf{y}_t^T \log\left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right), \tag{5.10}$$

for which we have

$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^+ = \sum_{\tau \geq t} \left(\mathbf{\Theta}^{\tau - t}\right)^T \mathbf{A}_\tau^T \mathbf{1},$$

and

$$\left(\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{W})\right)^- = \sum_{\tau\geq t}\left(\boldsymbol{\Theta}^{\tau-t}\right)^T \mathbf{A}_\tau^T\left(\mathbf{y}_t \oslash \left(\mathbf{A}\mathbf{x}_t + \mathbf{b_t}\right)\right).$$

The penalty function and the corresponding terms can be calculated similar to the nonnegative deconvolution problem. A similar update rule can be derived for the baseline. The special case of $\boldsymbol{\Theta} = \mathbf{0}$ (no dynamics with the convention $\mathbf{0}^0 = \mathbf{I}$) and $\lambda = 0$ (no penalization) is known as the Richardson-Lucy (RL) iterations. The RL algorithm has also been used with TV seminorm regularization in [70]. Similar to the RL algorithm we can use FADE for blind deconvolution, when the measurement matrix $\mathbf{A}$ is unknown. In this setup one can alternatively update $\mathbf{A}$ and $\mathbf{X}$. We can also used FADE, for estimation of GLM models for self-exciting point process models [113].

5.4.4 Combination with Other Constraints

In many applications of interest the optimization problem could also include several inequality constraints. For example in fluorescence microscopy the maximum changes of the fluorescence level with respect to baseline (also referred to as $\frac{\boldsymbol{\Delta}\mathbf{F}}{\mathbf{F}}$) is controlled by the properties of the indicator in use. In these situations we need to satisfy the KKT conditions for the extra constraints. Here we will introduce an adaptive method in order to achieve this goal. Consider the modified problem setup of Section 5.4.3

given by

$$\begin{aligned}&\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t = \mathbf{\Theta}\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_{t-1} + \mathbf{w}_t \\ &\mathbf{y}_t \sim \text{Poisson}\Big(\mathbf{A}b_t(1 + \left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t)\Big)\end{aligned} \quad , \tag{5.11}$$

where $b_t \geq 0$ denotes the known baseline fluorescence at time t, on top of which $\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t$ lies. In addition to nonnegativity constraints we need to account for the following constraints

$$\left(\tfrac{\mathbf{\Delta F}}{\mathbf{F}}\right)_t \preceq c_f \quad \text{for all } t \ \cdot \tag{$\star$}$$

The constant c_f is a characteristic of the indicator used and is assumed to be known. In order to enforce ($\star$) we proceed as in Algorithm 2.

Algorithm 2 Multiplicative Updates with Adaptive Regularization

```
procedure MULTIPLICATIVE UPDATES
    Initialize: $\mathcal{P}\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_t = \|\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_{[T]}\|_{\infty,\infty}$, $\lambda = 0$, $\lambda_0 = 0.01$, $i = 0$.
    repeat
        if $\max\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_t \geq c_f$ and $i = 0$ then
            $\lambda \leftarrow \lambda_0$, $i \leftarrow 1$
        end if
        if $\lambda > 0$ then
            Set $\lambda \leftarrow \lambda \frac{\left\|\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_{[T]}\right\|_{\infty,\infty}}{c_f}$
        end if
        Update $\mathbf{W}$.
    until convergence criteria met
end procedure
```

The main idea behind Algorithm 2 is that when the constraints are violated the complimentary slackness condition should be met for the optimal dual variable λ in Lagrangian form of the problem, meaning that the optimal solution should satisfy $\|\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_{[T]}\|_{\infty,\infty} = c_f$, which is equivalent to finding a fixed point of updates for the dual (regularization) variable λ.

5.5 Other Examples

5.5.1 Dynamic Nonnegative Least Square (NLS)

The NLS problem can in general be formulated

$$
\begin{aligned}
&\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{V}, \qquad \mathbf{V} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \\
&\mathcal{L}(\mathbf{X}) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_2^2,\ (\nabla \mathcal{L})^+ = \mathbf{A}^T \mathbf{Y},\ \nabla (\mathcal{L})^- = \mathbf{A}^T \mathbf{A} \mathbf{Y}.
\end{aligned}
$$

The most famous algorithm for solving the NLS problem is the active set method [124] which does not account for temporal dynamics in $\mathbf{x}_t$ or other forms of penalty. In these settings our update rules are very similar to the nonnegative deconvolution problem. A very useful example from the compressed sensing literature is the Multiple Measurement Vector (MMV) problem (without the positivity constraint) [57]. A commonly used penalty in this setup is the $\|\mathbf{X}\|_{2,1}$ which enforces row sparsity.

5.5.2 Dynamic Nonnegative Matrix Factorization (NMF)

The NMF problem is very similar to the NLS problem except that the matrix **A** is not known. In this case we can alternatively update our estimates of **A** and **X** [35].

$$
\begin{aligned}
&\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{V}, \qquad \mathbf{V} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \\
&\mathcal{L}(\mathbf{X}) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_2^2, \\
&(\nabla_{\mathbf{X}}\mathcal{L})^+ = \mathbf{A}^T\mathbf{Y}, \quad (\nabla_{\mathbf{X}}\mathcal{L})^- = \mathbf{A}^T\mathbf{A}\mathbf{Y} \\
&(\nabla_{\mathbf{A}}\mathcal{L})^+ = \mathbf{Y}\mathbf{X}^T, \quad (\nabla_{\mathbf{A}}\mathcal{L})^- = \mathbf{Y}\mathbf{X}^T\mathbf{X}
\end{aligned}
$$

In the absence of penalization or dynamics we recover the multiplicative updates of [125]. Our update rules can also account for the dynamic case where

$$
\begin{aligned}
&\mathbf{X}_t = \alpha\mathbf{X}_{t-1} + \mathbf{W}_t, \qquad \mathbf{W}_t \succeq \mathbf{0} \\
&\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t + \mathbf{V}_t, \qquad \mathbf{V}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I})
\end{aligned}
$$

For example one can account for sparsely changing temporal factors by considering a Laplacian distribution on $\mathbf{W}_t$.

5.6 Concluding Remarks

In this chapter we considered convex optimization problems with nonnegativity constraints and provided unified multiplicative updates for them using the KKT conditions. These updates are easy to implement and parallelizable on a CPU. They do not require tuning of extra parameters such as the step size, exhibit fast convergence in practice and can account for temporal dynamics and smooth penalties without slowing down.

Although in the absence of convexity the KKT conditions no longer hold, we have empirically observed that our updates exhibit good performance when the problem has simple nonconvexities. As an example one can model calcium saturation in the calcium deconvolution problem by adopting the calcium hill model given by $\mathbf{y}_t = \alpha \frac{\mathbf{x}_t}{\mathbf{x}_t + \mathbf{c}} + \mathbf{v}_t$ [233]. These observations suggest that suitable initializations result in convergence to a suitable local minimum. As another example one can combine the multiplicative updates with the IRLS algorithm [55] for ℓ_q, $q < 1$ minimization problems. The convergence of the IRLS algorithm was shown in the literature by showing an equivalence to a special case of the EM algorithm [18]. We applied this generalization to calcium imaging data using a nonconvex penalty.

Finally, the positivity constraint can easily be relaxed in the general form of the problems in two ways: First, any variable $\mathbf{X}$ can be decomposed into $\mathbf{X} = \mathbf{X}^{+} - \mathbf{X}^{-}$, where both $\mathbf{X}^{+}$ and $\mathbf{X}^{-}$ are positive. Second, generalized positivity and negativity could be defined with respect to the boundary of the convex set of feasible solutions, i.e. any point point inside/outside the feasibility set could be considered as positive/negative. Generalized positive and negative terms in the decompositions could be redefined similarly. Therefore by looking for a generalized positive fixed point of the gradient of the log-likelihood, the multiplicative updates can be generalized to a larger class of problems with not necessarily positivity constraints. We leave full details of these extensions and examples and their convergence properties to future work.

Chapter 6: Megapixel Two-photon imaging at kHz Framerates

Two-photon laser scanning microscopy enables high-resolution imaging within scattering specimens such as the brain, but the sequential acquisition of image voxels fundamentally limits its speed. We developed a two-photon imaging technique that scans lines of excitation across a focal plane at multiple angles, recovering diffraction-limited images from relatively few incoherently multiplexed measurements. We use a static image as a prior for image reconstruction to track rapid brightness changes in neural activity sensors, and sparsity as a prior to track diffusing particles at over 1.4 billion voxels per second. We imaged glutamate sensor transients across hundreds of individually resolved dendritic spines in mouse cortex at framerates over 1kHz. This method surpasses a physical limit on the speed of sequential two-photon imaging imposed by fluorescence lifetime, enabling recordings that are not possible by raster scanning.

6.1 Introduction

The study of brain function relies on measurement tools that achieve high spatial resolution over large volumes at high rates. Activity travels through neural circuits on a timescale of milliseconds. Interacting elements such as neurons or synapses are scattered over distances many times their size, requiring imaging volumes containing millions of voxels. The living brain is opaque, meaning that optical tools for monitoring brain activity must be insensitive to light absorption and scattering. Two-photon imaging achieves this insensitivity by using nonlinear absorption to confine fluorescence excitation to the high-intensity focus of a laser and prevent excitation by scattered light. All emitted fluorescence can be assigned to that focus regardless of scattering, without forming an optical image. Instead, an image is produced by scanning the focus in space. However, this serial approach to image acquisition creates a tradeoff between achievable framerates and pixel counts per frame. Common fluorophores have fluorescence lifetimes of approximately 3 ns, and brighter fluorophores tend to have longer lifetimes [198, 199]. Consequently, approximately 10 ns must pass between consecutive measurements to obtain distinct samples. The maximum achievable framerate for a 1-megapixel field of view (FOV) under raster-scanning fluorescence imaging is therefore approximately 100 Hz. In practice, pixel rates have been further limited by factors such as photodamage, fluorophore saturation, and scanner technology [231].

Calcium indicator transients have been widely adopted as a proxy for neuronal and synaptic activity [58, 201, 231], in part because their large amplitude and slow speed can be recorded at low frame rates [161]. However, calcium transients are imperfect reporters of the biophysical quantities underlying neuronal communication, such as action potentials and synaptic events [97, 161, 192]. Faster reporters and imaging methods are needed to more directly monitor such signals at speeds commensurate with computations in the brain.

The pixel rate bottleneck for raster imaging can be avoided by more efficient sampling [171, 231]. In most activity imaging paradigms, a dense pixel-based representation of the sample is recorded then reduced to a lower-dimensional space, e.g. by selecting regions of interest [166, 218]. By sampling this low-dimensional representation more directly, the equivalent result can be obtained with many fewer measurements. Several methods have been developed that can record from sample volumes more efficiently than raster scanning:

Random Access imaging using Acousto-Optic Deflectors/Lenses4 [42,143] enables sampling of any subset of points in the sampling region, with a fixed access time required to move the excitation focus between points. If the desired points are sparse enough in space, the time saved by not sampling the intervening area significantly outweighs the access time costs. Multifocal multiphoton (MMP) methods scan a fixed pattern of focal points through the sample, allowing multiple subvolumes to be acquired simultaneously. If a single-pixel detector is used, the resulting image volumes sampled by each focus are superposed [230]. A spatially resolved detector such as a camera, eye [30] or multianode photomultiplier [114] can also be used, but the resolution of resulting images is significantly degraded by scattering. Extended Depth of Field (EDoF) methods collapse the axial dimension of the sample, allowing projections of volumes to be acquired at the rate of two-dimensional images. EDoF excitation has been achieved at high resolutions using scanned Bessel beams [34, 72, 131, 202]. Several other imaging methods involving patterned two-photon illumination have been described [80, 171].

Except random access imaging, the above techniques share a common feature: Signals from different locations in the sample are deliberately mixed, enabling a given volume to be measured with fewer samples. For most analyses, the underlying signals are then unmixed computationally. We refer to this approach as "Projection Microscopy" because it deliberately projects multiple resolution elements into each measurement. Projection Microscopy has advantages over random access imaging when the specimen can move unpredictably, such as in awake animals or diffusing particles, when regions of interest are not well approximated by single points, such as cell membranes, or when it is difficult to rapidly select regions of interest, such as dendritic spines.

Recovery of signals from mixed measurements is common in imaging. Several methods combine images with distinct optical transfer functions to improve resolution [95, 102, 149, 170]. In all imaging methods, finite resolution can cause pixels to contain signals from multiple sources, such as a neuron and its surrounding neuropil. It is common for analyses to explicitly model these mixed sources (e.g. [166]). Source recovery is often posed as an optimization problem, using implicit or explicit regularization to impose a desired statistical structure on the recovered signals, such as independence [38] or sparsity [166]. Compressive Sensing [77] is a framework for the acquisition and unmixing of signals that admit a sparse representation in some basis. By acquiring mixed signals and regularizing for sparsity during recovery, structured systems with many fewer measurements than unknowns can be accurately recovered if the measurements are conducted appropriately [50]. Highly coherent measurements (i.e. ones that tend to mix a given source with others in the same way) make recovery of the underlying sources ambiguous, while incoherent measurements can guarantee accurate recovery [50, 81]. The prior information used for recovery can take many forms, such as the distribution or dynamics of the sources [19] [111]. In medical imaging and microscopy, compressive sensing using sparsity has greatly increased imaging speed and reduced radiation dose [56, 134, 159].

In this Chapter, we recover neural activity from highly incoherent mixed measurements using strong priors obtained from a detailed structural image of the sample and a model of indicator dynamics. These priors constrain the space of solutions enough that regularization for sparsity is unnecessary. We also image and track diffusing particles within a more canonical compressive sensing framework. The goal of our work has been to advance optical neurophysiology by developing a practical high-speed two-photon microscope. This microscope retains the high spatial resolution and insensitivity to scattering of conventional two-photon imaging, uses known spatiotemporal structure of samples to increase framerate, is insensitive to sample motion, enables highly accurate source recovery by performing efficient incoherent measurements, and adapts easily to a variety of experiments.

6.2 Results

6.2.1 Scanned Line Angular Projection Microscopy

We built a microscope that scans line foci across the focal plane at four different angles, obtaining linear projections of the sample. We chose line foci because they:

1. Are simple to produce optically,

2. efficiently sample a compact area by scanning

3. achieve diffraction-limited spatial resolution

4. produce two-photon excitation more efficiently than non-contiguous foci of the same area

5. are a low-coherence basis for the sample space.

We named this approach Scanned Line Angular Projection Microscopy (SLAPMi). SLAPMi samples the entire FOV with four line scans. The frame time is proportional to the resolution, compared to resolution squared for a raster scan, resulting in greatly increased frame rates. For example, SLAPMi images a 250 × 250 μm FOV with diffraction-limited optical resolution at a framerate of 1016 Hz, corresponding to over 1.49 billion pixels per second, recovered from 5 million multiplexed measurements.

To reduce optical power at the sample and enable control over degree of parallelization, we included a spatial light modulator (SLM) in an amplitude modulation geometry. This configuration selects an arbitrary pattern in the focal plane for imaging, and discards the remaining excitation light, making SLAPMi a random access microscope. This reduces excitation power in sparse samples and allows users to artificially introduce sparsity into densely labeled samples. Unlike AOD-based imaging, the SLAPMi framerate is independent of the number of pixels imaged, up to the entire field of view. When imaging sufficiently sparse samples, the SLM and other optimizations for excitation efficiency allow SLAPMi to use average powers lower than conventional raster scanning.

6.2.2 Particle Localization and Tracking

Recovering images from mixed measurements requires identifying a representation of the sample with rank lower than the number of measurements. For sparsely labeled samples, this representation can be the nonzero pixels in the image [16, 134]. We demonstrated this regime by localizing and tracking fluorescent particles. Each particle produces a bump of signal on each of the four scan axes corresponding to its position, and each frame consists of the superposition of signals for all particles in view. If particles are sufficiently sparse, three measurement axes are sufficient to localize all particles in the FOV.

Using the microscopes known projection matrix (see 6.3), we can recover the maximum likelihood image by incorporating dynamics into Richardson-Lucy (RL) deconvolution as discussed in Chapter 5 [133, 177]. For large numbers of labeled pixels, maximum likelihood reconstruction results in spurious peaks in the recovered image, which are reduced by solvers that enforce sparsity.

As a demonstration, we recorded the motion of thousands of 500 nm fluorescent beads within a 250 μm (diameter) × 250 μm (depth) imaging volume (87 equally-spaced planes; 1.016kHz frame rate, 5080 measurements per plane, 10 Hz volume rate), (Figure 6.1, supplemental movie 1). Particles were readily tracked within the recovered volume movies using established tracking methods for volumetric data (Figure 6.2).

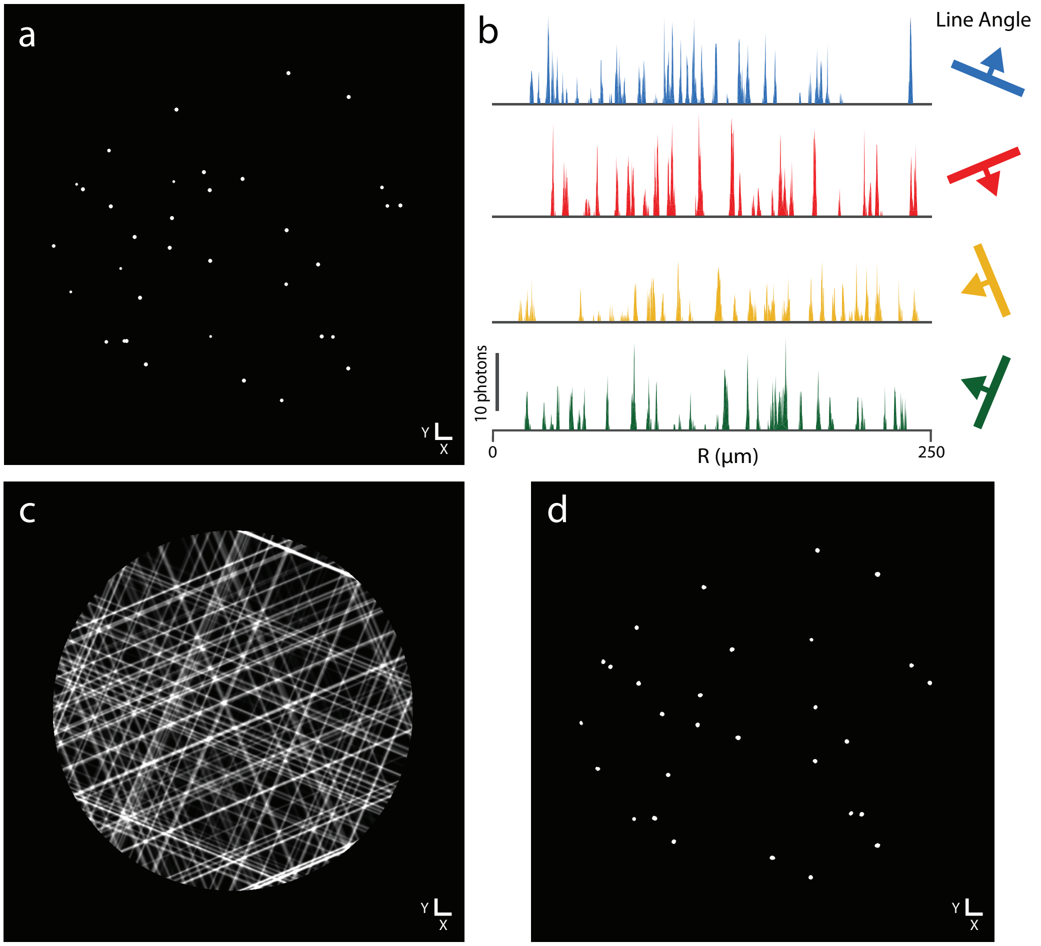


Figure 6.1: Particle localization using SLAPMi: a) 2D raster image of 500nm fluorescent beads on a glass surface. b) 2D SLAPMi measurement of the same sample, consisting of projections along the four scan axes. R denotes position on the scan axis. c) Backprojection of the measurements in b, corresponding to one iteration of RL deconvolution. d) Recovered image following 20 iterations of pruned RL deconvolution

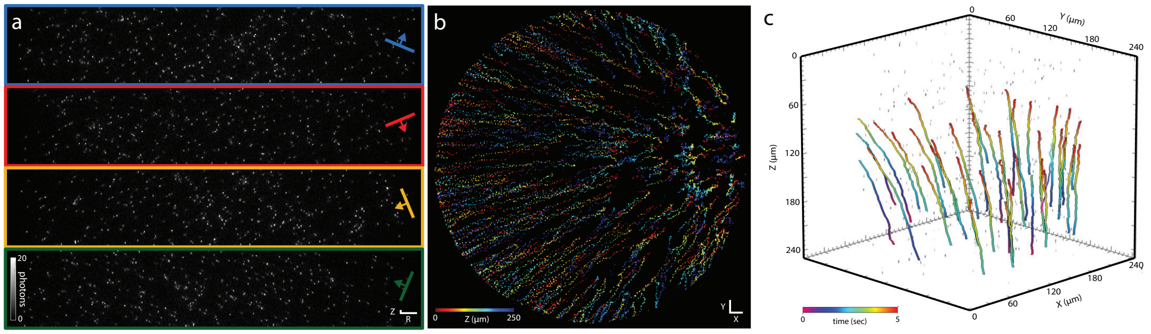


Figure 6.2: Particle tracking using SLAPMi: a) 3D SLAPMi measurement of a 250 × 250 × 250 250 μm volume of 500 nm fluorescent beads in water. b) Superposition of 40 volume images acquired with SLAPMi, showing convection and diffusion of beads. Color denotes depth. c) Example particle tracks obtained with a common software package

6.2.3 Imaging Neural Activity

To recover neural activity with SLAPMi, we adopted a sample representation in which fixed spatial components vary in brightness over time [166, 218]. The spatial components are obtained from a separate raster-scanned volume image. We identify compartments of labelled neurons with a manually trained pixel classifier (Ilastik [195]), and a skeletonization-based algorithm that divides neurites into short segments (Figure 6.3-b), resulting in up to 1000 segments per plane (see Methods). Source recovery consists of assigning an intensity to each segment at each frame according to the following model:

$$\begin{aligned}
&\left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_t = \theta \left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_{t-1} + \mathbf{w}_t \\
&\mathbf{y}_t \sim \mathrm{Poisson}\Big(\mathbf{A} b_t (1 + \left(\frac{\Delta \mathbf{F}}{\mathbf{F}}\right)_t)\Big) \\
&\text{subject to}\, \mathbf{W} \succeq \mathbf{0},
\end{aligned}$$

where $\mathbf{Y}$ are the measurements (number of lixels × frames), $\mathbf{P}$ is the projection matrix and $\mathbf{S}$ is the segmented image obtained by a 2P raster scan. $\mathbf{P}$ and $\mathbf{S}$ are measured separately. We estimate $\mathbf{W}$, the innovations, by minimizing the negative log-likelihood as discussed in chapter 5. Importantly, in this paradigm, regularization is unnecessary because the segmentation is low rank.

6.2.4 In Vitro Validation

We validated SLAPMi in experiments with rat hippocampal cultures under conditions where ground truth activity was known. In the first of these experiments, we co-cultured cells expressing a cytosolic fluorophore (tdTomato) with cells expressing the glutamate sensor Venus-iGluSnFR. The mixed culture was imaged using a single detector, while stimulating with a field electrode. Stimulation triggers transients only in Venus-iGluSnFR-expressing cells, and no change in tdTomato brightness. We collected separate two-channel raster images to verify the identity of the imaged cells, and quantified recovered signals in tdTomato-expressing cells to assess spatial crosstalk.

SLAPMi reliably reported stimulation-induced transients only in Venus-iGluSnFR labeled sample voxels (Figure 6.3, supplemental movie 2).

In the second experiment, we imaged yGluSnFR-expressing neurons while uncaging glutamate at two locations at different times, to assess the timing precision of recovered signals (Figure 6.4, supplemental movie 3). We quantified the onset time of transients at each pixel of the reconstructed image.

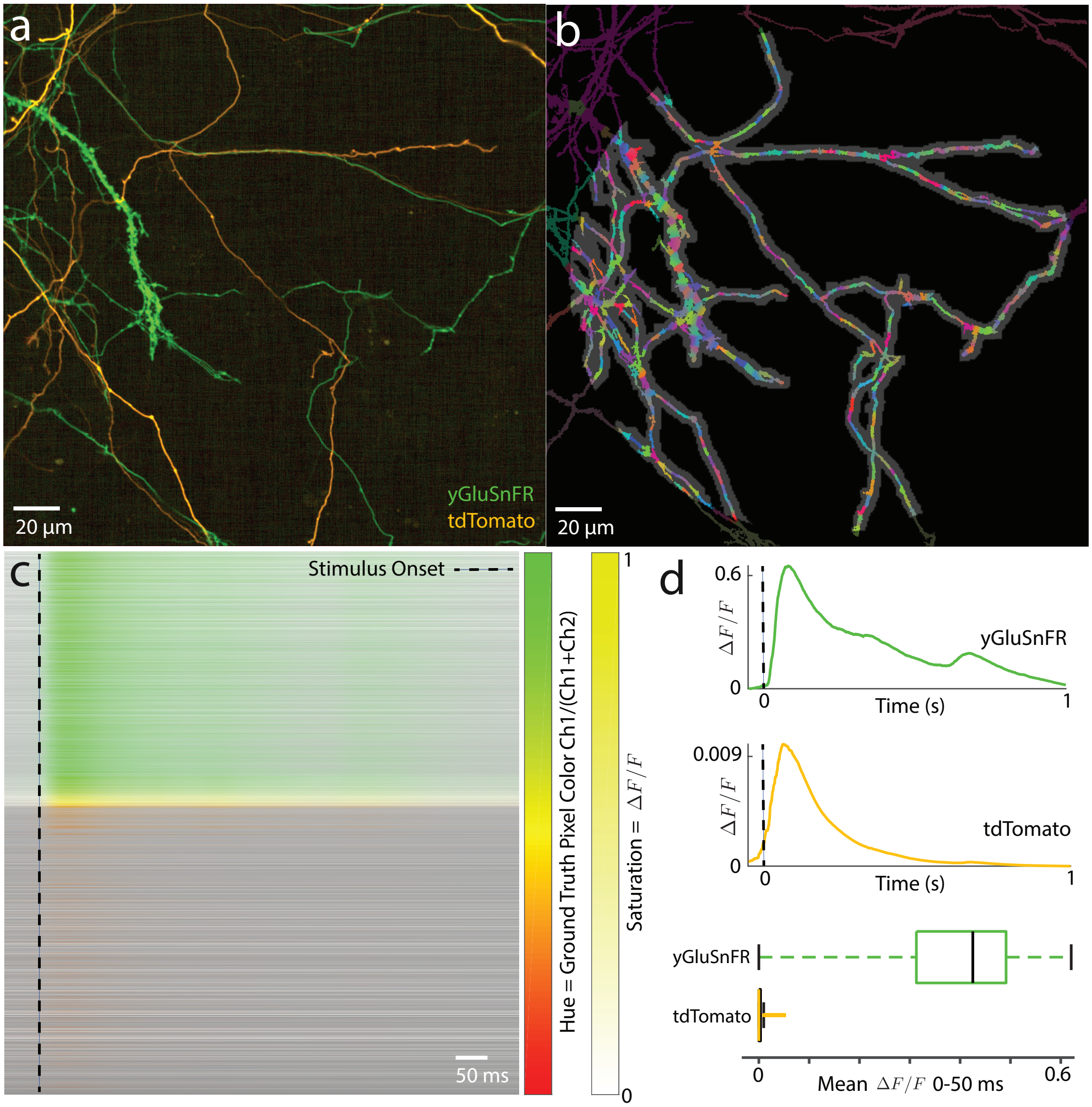


Figure 6.3: In Vitro validation, red/green experiment: a) Superposed 2D raster image of cultures expressing yGluSnFR (Green, channel 1) and tdTomato (Red, both channels). The activity was imaged only on channel 1, b) Segmentation for a). The SLM boundaries are tinted in gray, c) Recovered pixelwise $\frac{\Delta F}{F}$ for a). Pixels are sorted according to their color (redness). The value (darkness) represents square root of F_0, d) Mean $\frac{\Delta F}{F}$ for most green and most red pixels (1000 pixels each). Top and Middle: Spatial mean $\frac{\Delta F}{F}$ for most green/red pixels. Bottom: Temporal mean $\frac{\Delta F}{F}$ (0-50 ms after stimulus) quantiles for most green/red pixels.

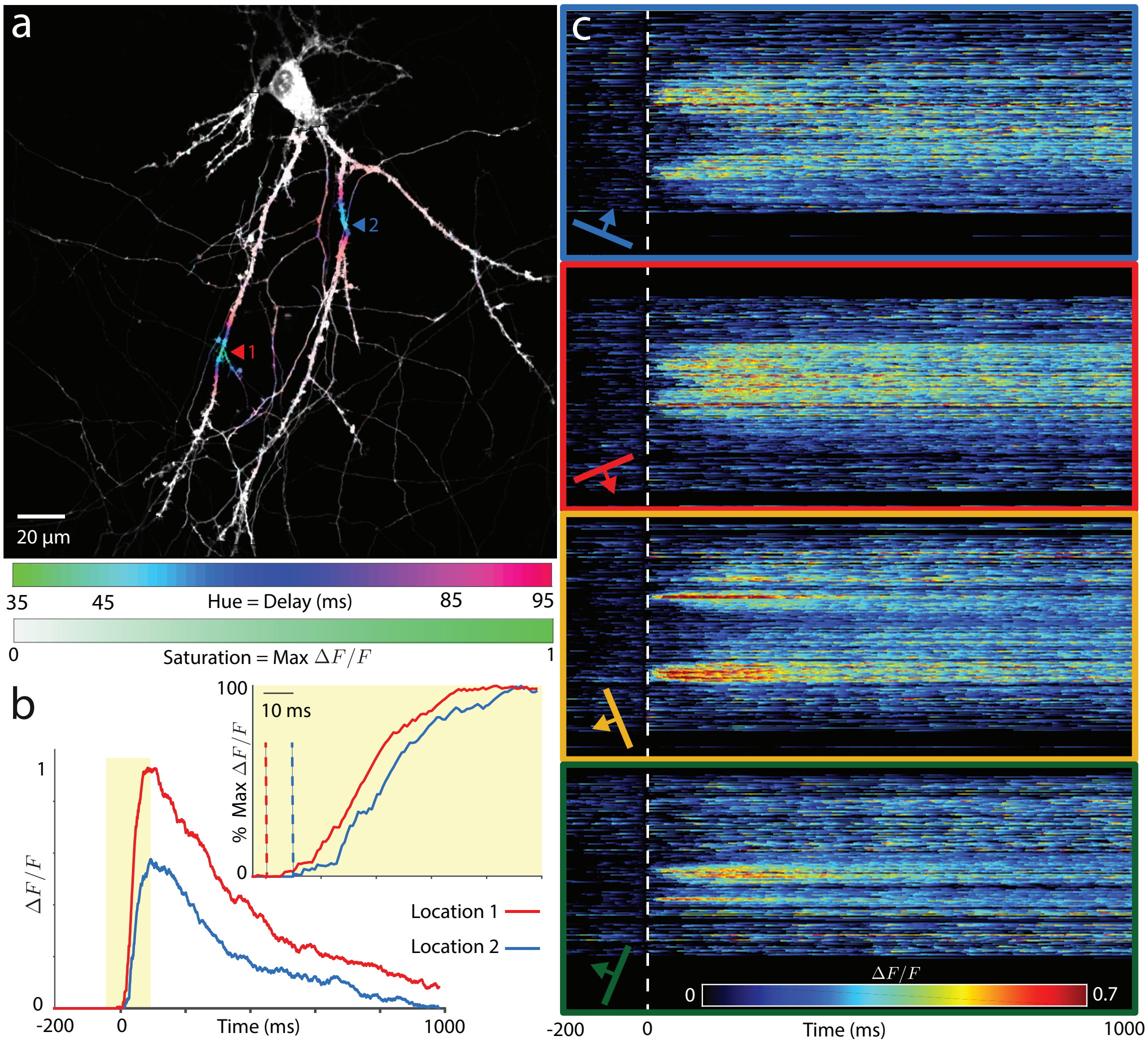


Figure 6.4: In Vitro validation, two-spot uncaging experiment: a) Raster image for 2spot uncaging, with uncaging locations (1 and 2) denoted by red and blue arrows, colored according to delay to half-max $\frac{\Delta F}{F}$, b) Dff traces for the two uncaging locations. Inset: zoomed in $\frac{\Delta F}{F}$ traces normalized to peak $\frac{\Delta F}{F}$. Dashed lines denote the stimulus onset, c) 2D SLAPMi lixel-space measurement of the sample, consisting of projections along the four scan axes, showing diffusion. The dashed line marks the stimulus onset in uncaging location 1.

6.2.5 In Vivo Activity Imaging

SLAPMi inherits the resolution enhancement and insensitivity to scattering that make two-photon imaging effective in vivo. However, we were concerned that tissue heating from light absorption could limit the practicality of two-photon projection microscopy. Economy of illumination power is critical for biological imaging [122]. Conventional two-photon imaging is limited by brain heating under common configurations [167]. Two-photon projection microscopy is even more limited by heating, because higher degrees of parallelization require a matched increase in power to maintain nonlinear excitation efficiency. Lower degrees of parallelization make two-photon excitation more efficient and source recovery more effective. In general, multiphoton projection methods benefit by using the lowest degree of parallelization compatible with an experiment's required framerate [167, 193].

To reduce optical power at the sample and control degree of parallelization, we included a spatial light modulator (SLM) at an intermediate focal plane in an amplitude modulation geometry. This configuration selects an arbitrary pattern in the focal plane for imaging, and discards the remaining excitation light, thereby combining benefits of random access imaging and projection microscopy. Unlike AOD-based methods, random access SLAPMi has no access time. By retaining a buffer area surrounding each region of interest, SLAPMi remains insensitive to sample motion with no reduction in imaging rate. When imaging sparse samples such as dendrites, the SLM and other optimizations for excitation efficiency allow us to use average powers lower than conventional raster scanning.

6.3 Methods

6.3.1 Solver

The solver is a slight modification of the one introduced in Chapter 5, where the baseline was assumed to be known. The baseline is meant to capture model mismatches in the data due to imperfect alignment and motion correction. Optionally, we allow a regularization term to impose prior knowledge of the indicator (maximum $\Delta F/F$) though this was not necessary for the in vitro and in vivo datasets presented here. The RL iterations are known to amplify noise for large number of iterations, especially in very sparse samples such as the particle localization and tracking data. In such scenarios a small ℓ_q-regularization ($q \leq 1$) proved to be helpful in pruning such artifacts. (Figure 6.1). We also implemented an optional damping, which is known to be useful in suppressing the artifacts due to low photon counts [222].

The only user-supplied parameters involved in recovery are:

- The decay time constant, a property of the indicator,
- the number of segments (not needed),
- the convergence threshold of the solver, or, number of multiplicative iterations,
- the damping parameter which determines the level of suppression of the objective function for the low-photon counts.

Mixed measurements can introduce spurious correlations into recovered signals, even when the mixing matrix is invertible, because measurement noise is shared among recovered sources. In simulations of SLAPMi imaging and source recovery, we accurately recovered signal amplitudes and correlations between sources without significant bias (See Section 6.3.2). Recovery was robust to errors in segmentation and kinetics. In addition to the existing solver, we tried several solvers and models including:

- A Kalman filtering and smoothing approach by fitting a Gaussian approximation to the state space model [191]. The point process smoothing is very slow and exhibits weak approximation abilities for ambient dimensions higher than a few.
- Nonnegative least squares with temporal dynamics, which did not suit the Poisson noise model.
- Gradient descent methods such as spiral [98], which happened to be sensitive to the step size.
- We also considered a multiplicative baseline model and a rank 1 baseline model for the data, which showed inferior performance.

6.3.2 Simulations

We evaluated performance of the solver under different conditions using simulated data. We used the delta PSF for the simulations. Unless otherwise stated, we use the following default parameters: p = 500 segments, 40 random generations of the dynamics, uniformly distributed between $0.5 - 5$. The solver was initialized by a constant positive solution for the spikes and a time constant of 100 frames, We assumed a photon count of 100 Photons per frame per segment which is close to the measured datasets,T = 500 frames with one spike per segments at frame 100, segments placed randomly within a 500 × 500 pixel circle at the center of the field of view and the maximum allowed $\Delta F/F$ for the solver set to 10. Segments are assumed to be 10 ×10 pixel squares. The null distributions were obtained by random shuffling of time points for each seed.

Evaluation metric: we report the Pearson correlation between the estimated signal and its ground-truth counterpart over time for each segment.

In the first simulation (Figure 6.5) we evaluated the effect of sample brightness on performance of the solver. This was done by fixing the expected Poisson rates per segment per frame to vary in the range 0.1 to 10^5. As a comparison, typical values of this parameter in the recorded datasets in this work are around 10-100. Increasing the photon budget improves the reconstructions.

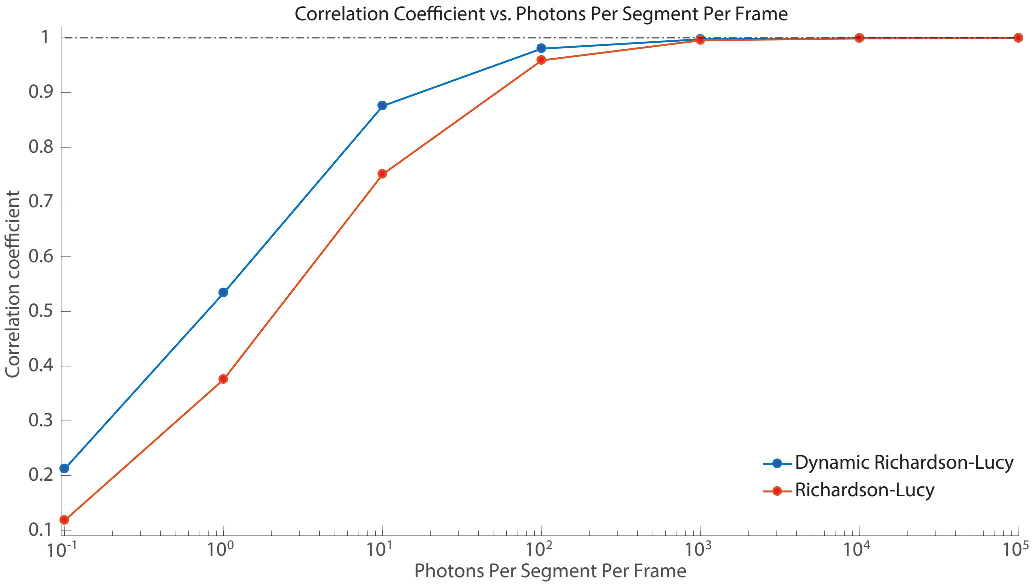


Figure 6.5: Effect of sample brightness on the solver.

In the second simulation (Figure 6.6), we changed the number of sources in the range 10-1000, in both the generation of the ground truth and solver. Increasing the number of sources decreases solver performance.

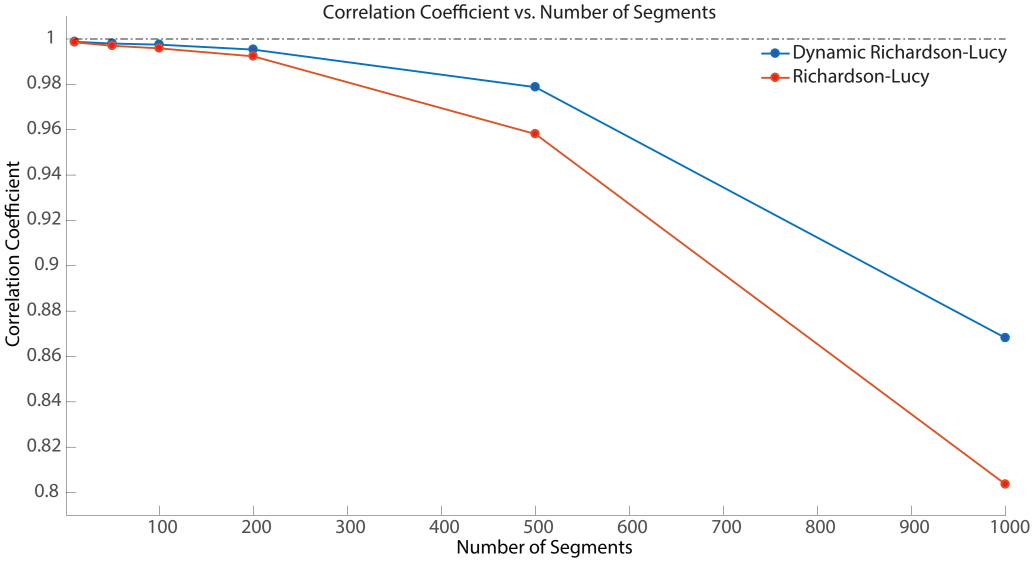


Figure 6.6: Effect of number of sources on the solver.

In the third simulation (Figure 6.7), we evaluated performance using different estimated decay time constants. The ground truth activity was generated using a time constant of 100 frames. We varied the time-constant in the range 0-1000. The solver performs best when the correct time constant is used. The solver is insensitive to underestimation of the time-constant, as this can be compensated with additional spikes but does reduce the denoising benefit of the dynamics model. A time constant of 0 corresponds to normal Richardson-Lucy iterations. Substantial overestimation of the time constant degrades the performance of the solver. The null distributions for different time constants is shown in Figure 6.8.

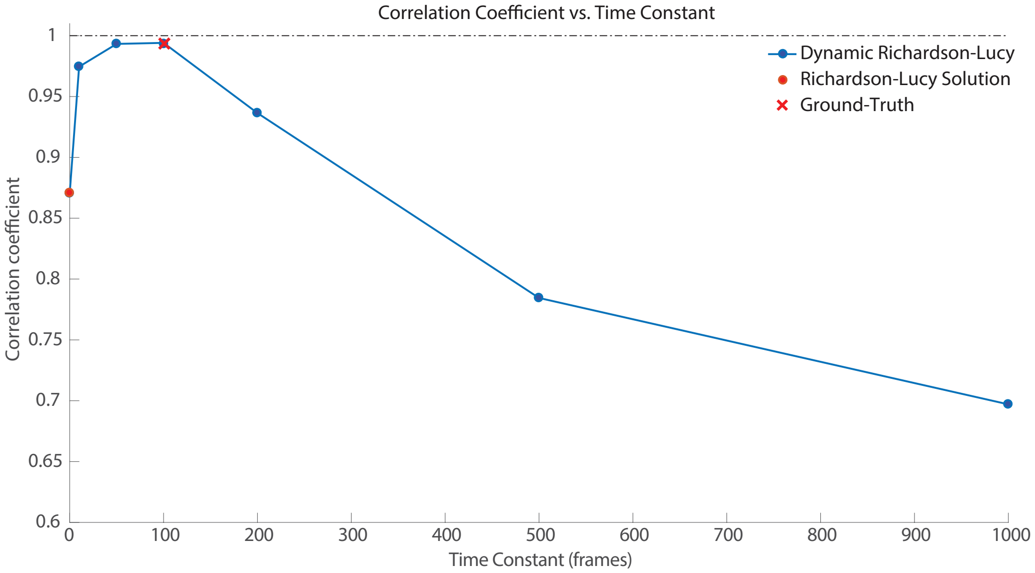


Figure 6.7: The effect of erroneous decay time-constant on the solver.

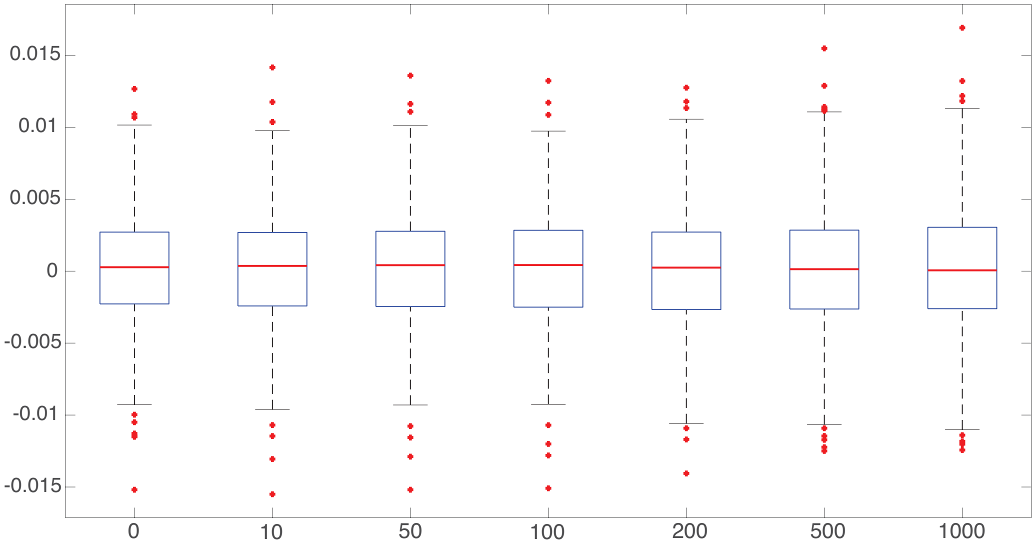


Figure 6.8: Null distributions for different time-constants.

In the fourth simulation (Figure 6.9), we evaluated the performance of the solver in the case where not all sources in the field of view were included in the segmentation. We added 0-50% additional unsegmented sources. This is meant to simulate fluorescence activity within the 'ON' region of the SLM but not detected in the reference image, a possibility in sensors with very low baseline fluorescence. Increasing unsegmented activity degrades the performance of the solver.

The null distributions for different levels of unsegmented activities are shown in Figures 6.10 and 6.11.

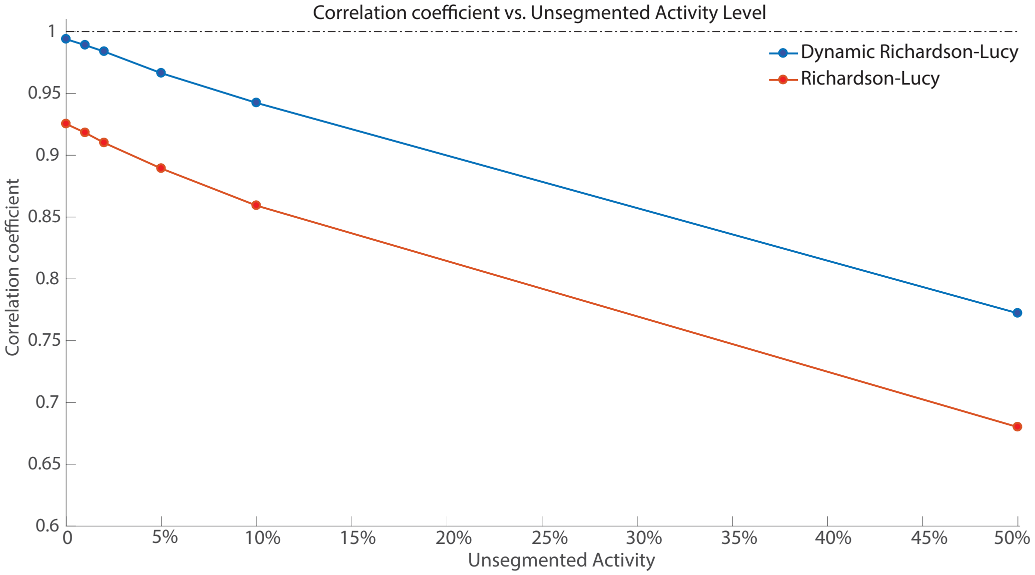


Figure 6.9: The effect of unsegmented activity on the reconstructions.

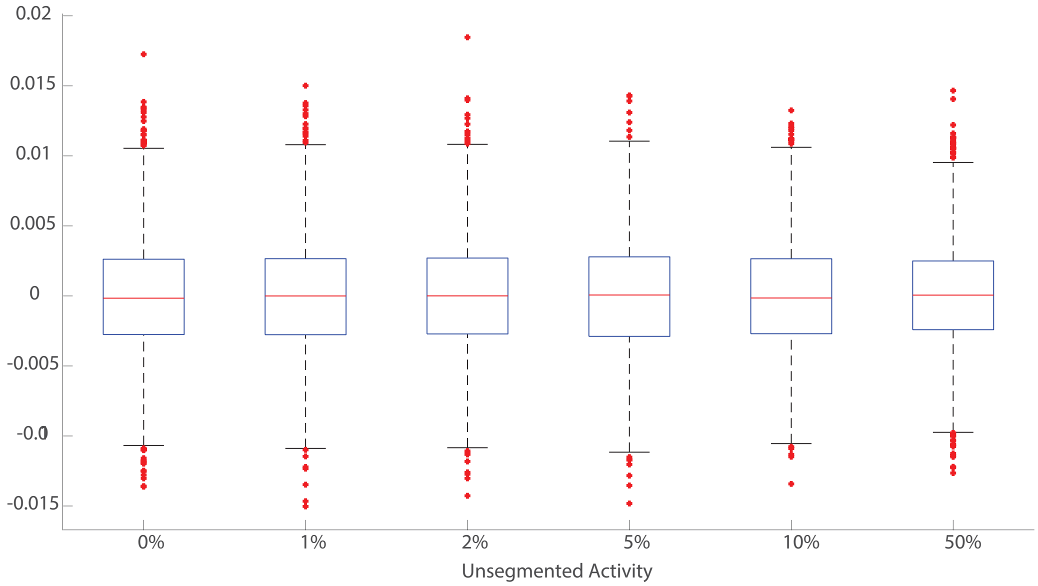


Figure 6.10: Null distributions for different levels of unsegmented activity.

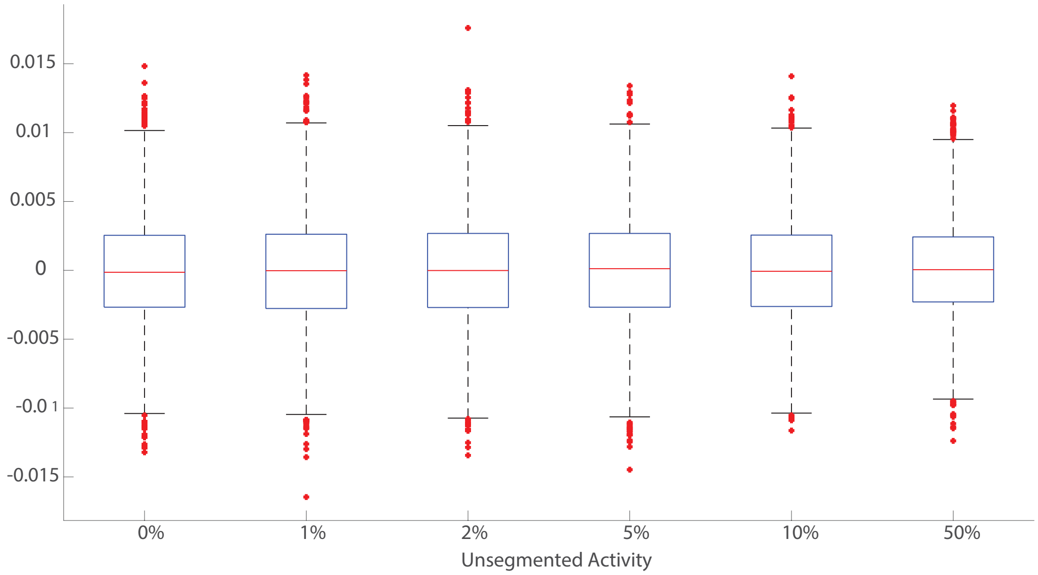


Figure 6.11: Null distributions for the unsegmented activity for the frame by frame solver.

In the fifth simulation (Figure 6.12) we evaluated sensitivity of the solver to alignment errors. In order to do so, we shifted the reference image by 0-10 pixels after registration. Better alignment improves the reconstructions.

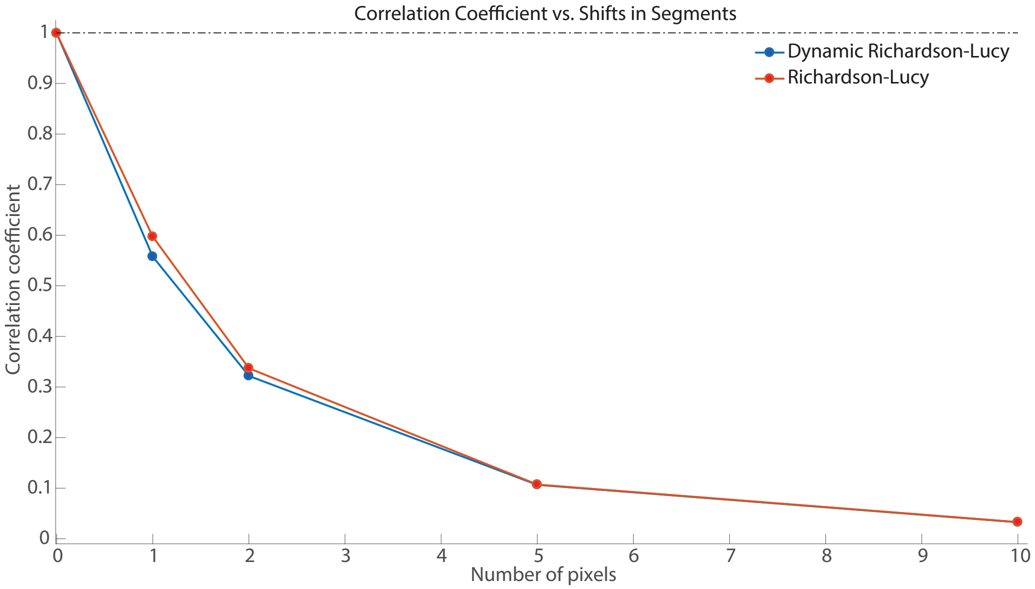


Figure 6.12: The effect of alignment (shifts in pixels) on the reconstructions.

In the sixth simulation (Figure 6.13) we evaluated performance of the solver when the reconstruction segmentations do not match the ground-truth activity segmentations. Alternate segmentations lead to a degraded performance which can be partially compensated for by using a finer segmentation than the ground truth.

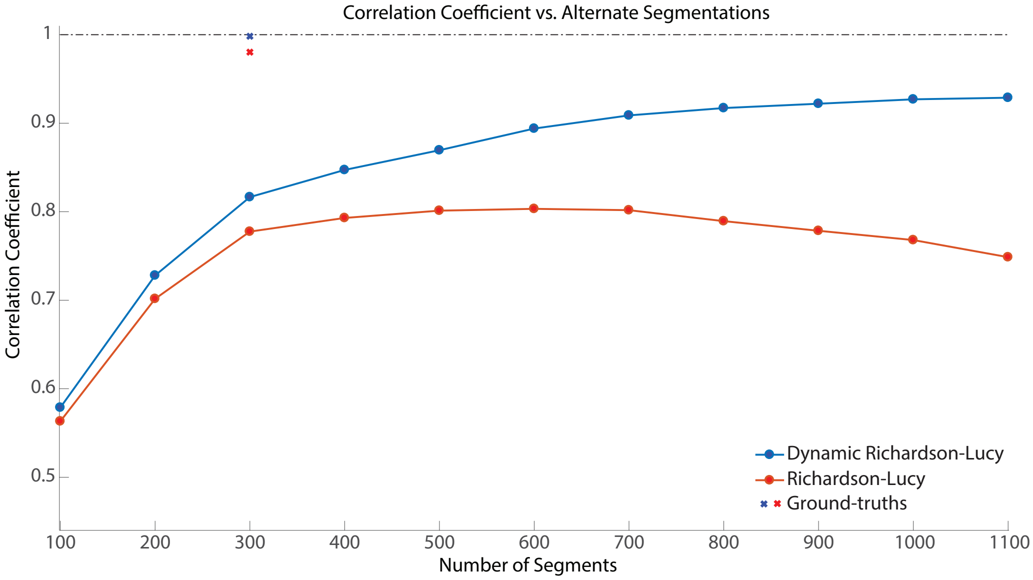


Figure 6.13: The effect of alternate segmentations on the reconstructions.

We next quantified the undesired correlations introduced to the activities (Figures 6.14 and 6.15) as a result of the solver for the dynamic RL solver compared to the RL iterations. We generated 3000 frames of activity with random spiking patterns using vines and the extended onion method [128], resulting in correlated activity with correlations in the range -0.6 -0.6. As can be noted, recovered activities using the dynamic RL iterations significantly improves.

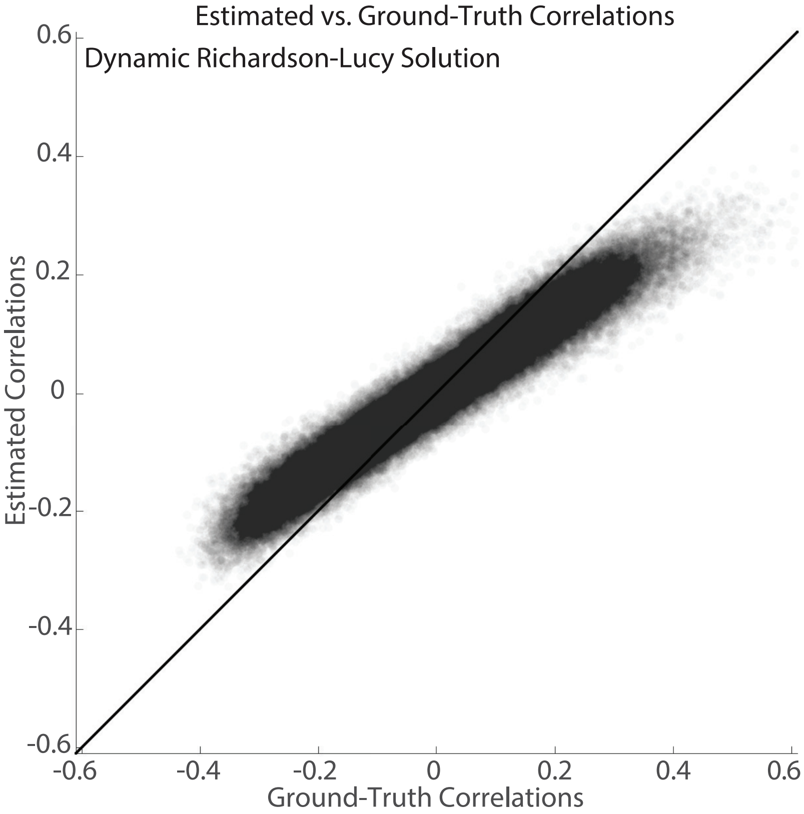


Figure 6.14: Recovered activity correlations using the dynamic solver .

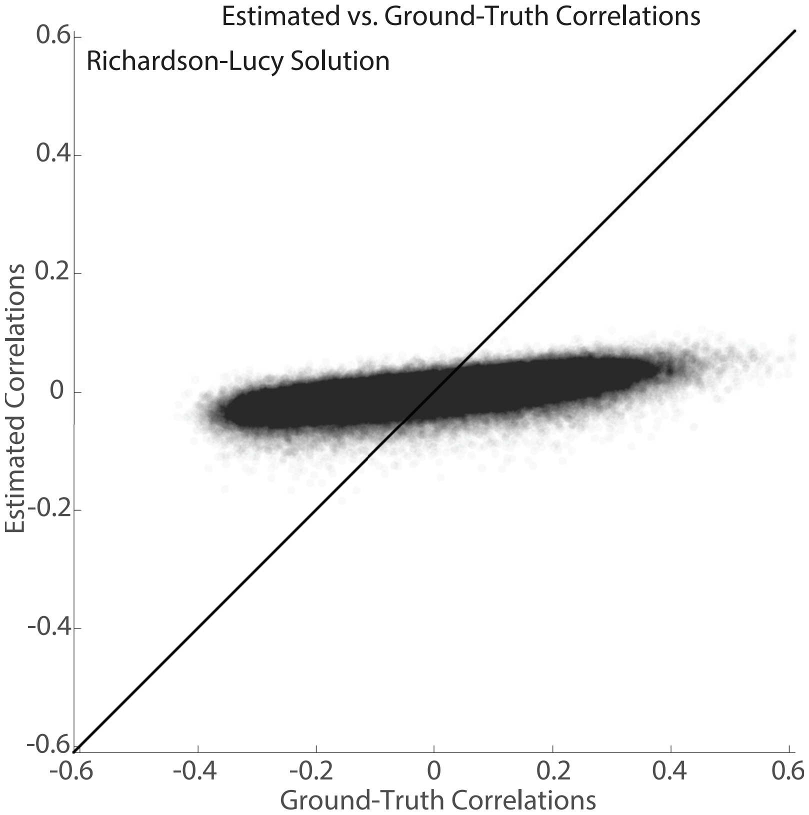


Figure 6.15: Recovered activity correlations using the frame by frame solver.

6.3.3 Software

Software was written in Matlab and LabView FPGA. SLAPMi interfaces with ScanImage (Vidrio Technologies) to perform raster scanning.

6.3.4 Measurement Matrix

The projection matrix (**P**) is measured in an automated calibration step using a thin ($\ll$1μm) fluorescent film. Images of the excitation focus in the film, collected by a camera, allow a correspondence to be made between the positions of galvanometer scanners and the location of the resulting line focus. The raster scanning focus is also mapped, allowing us to create a model of the line foci transformed into the space of the sample image obtained by the raster scan.

6.3.5 Motion Registration and Alignment

Recorded SLAPMi data are spatially registered to compensate for sample motion. As with raster imaging, translations of a single resolution element can be sufficient to impact recovery of activity in fine structures, necessitating precise registration. Accurate registration and source recovery rely on accurate, minimally warped reference raster images. SLAPMi was designed to efficiently interface with freely available ScanImage software (Vidrio Technologies), which is used to collect raster stacks. These images can be warped by many factors, including nonlinearity in the scan pattern, sample motion, and the rolling shutter artifact of the raster scan. We estimate and compensate for warping by collecting two sets of reference images interleaved, one with each of the two galvos acting as the fast axis. To compensate for motion and activity variations during reference image acquisition we obtain a large number of stacks and rely on consensus between aligned stacks to reject these artifacts. Registration of SLAPMi recordings is performed by identifying the 3D translation that maximizes the sum of 1-dimensional correlations (or optionally, Dynamic Time Warping distances) between the recorded signal and the expected projection of the reference image on each of the four scan axes. If the SLM is not used, this objective is maximized using cross-correlations, where the SLM is used, we perform an iterative multiscale grid search.

6.4 Concluding Remarks

In vivo imaging techniques are becoming increasingly specialized, with different methods best suited for different organisms and experimental parameters [231]. Until now, random access imaging has been the most effective method for imaging hundreds of target sites, spanning hundreds of microns, hundreds of times per second, in scattering tissue. SLAPMi performs such measurements in highly dynamic samples at rates exceeding 1kHz. SLAPMi improves upon existing projection microscopy techniques by having lower coherence, higher frame rates, and random access excitation. Methods that scan a static multifocal pattern or Bessel beam have tended to mix sample voxels coherently, making unmixing difficult where objects overlap in their measurements. SLAPMi's angular projections, in contrast, ensure that no two voxels are always mixed together in measurements. Scanned Bessel beam imaging records volume projections as fast as 2D scans, but the 2D scan rate is not increased. SLAPMi records from planes with just four 1D scans. The approach of scanning excitation patterns across an SLM dramatically reduces power usage, and could be used similarly with other projection schemes provided the excitation pattern lies in the focal plane.

SLAPMi has several limitations. Recovered traces become less precise as the number of distinct objects e.g. neurons, imaged increases above 1000. As with all projection methods, certain sample structures are adversarial to source recovery due to the compressed nature of the measurements. In particular, regions where many distinct objects are packed closely together may not be uniquely determined by

SLAPMi measurements, and might not be accurately recovered even when the total number of objects imaged is less than 1000. Dim objects surrounded by extremely bright objects are difficult to recover, because Poisson noise originating from bright sources may exceed signal from the dim source on all projection axes. The SLM allows dense or bright regions to be selectively dimmed, and can ameliorate these issues. SLAPMi has a maximum frame rate limited by the 2D galvanometer scanners. In samples compatible with random access imaging having a very small number of target sites, random access imaging may allow higher frame rates.

SLAPMi achieves the same lateral resolution as raster scanning, but axial sectioning is reduced, as two photon intensity of lines drops off as $\frac{1}{z}$, compared to $\frac{1}{z^2}$ for points. In general we find SLAPMi produces extremely robust results without parameter tuning in a wide variety of sample preparations, making it a practical alternative to other imaging methods.

Chapter 7: Conclusions and Future Work

In this thesis we revisited and made improvements over several theoretical aspects of compressive sensing for nonlinear and dynamic models.

From a theoretical perspective, in Chapters 2 and 3 we derived minimax optimal sampling-complexity tradeoffs for autoregressive processes, point processes and generalized linear models where the covariates do not satisfy the conventional i.i.d. assumptions. We consider extension of these theoretical results to multivariate processes as future work. The results on point processes were motivated by their applications in characterizing the self-exciting and history dependence nature of neural spiking activities. Our results on autoregressive processes started from a class project for the Compressive Sensing class taught by Professor Piya Pal at UMD. The main idea behind the theoretical guarantees came from the theoretical results on convergence of eigenvalues of covariance matrices that I learned during the course I took on adaptive filter theory taught by Professor Ali Olfat at University of Tehran.

From an algorithm design point of view, in Chapter 4 we introduced the idea of compressible state-space models with sparse innovations and provided a fast, optimal recovery algorithm for such models. These models have huge applications in modeling biological signals. Many beautiful intuitions about such state-space models came from suggestions of Professor Prakash Narayan during my research proposal exam at UMD. We consider generalization of these models to heavy-tailed innovations as future work.

Finally, in Chapter 6 we developed a two-photon imaging technique that scans lines of excitation across the sample at multiple angles, recovering high-resolution images from relatively few incoherently multiplexed measurements. By combining traditional Poisson image reconstruction techniques with temporal dynamics we managed to reconstruct neural activity in behaving animals at framerates higher than 1 kHz for Megapixel fields of view. This research was in collaboration with Janelia research campus. Many intuitions about the reconstruction algorithms and the experiment came from the suggestions of several group leaders at Janelia. Combined with the evolution of fast sensors we consider studying neuronal dynamical systems using SLAPMi as future work.

Appendix A: Proof of Theoretical Results

A.1 Proofs of Main Theorems for Autoregressive Processes

A.1.1 The Restricted Strong Convexity of the matrix of covariates

The first element of the proofs of both Theorems 3 and 6 is to establish the Restricted Strong Convexity (RSC) for the matrix $\mathbf{X}$ of covariates formed from the observed data. First, we investigate the closely related Restricted Eigenvalue (RE) condition. Let $[\lambda_{\mathsf{min}}(s), \lambda_{\mathsf{max}}(s)]$ be the smallest interval containing the singular values of $\frac{1}{n}(\mathbf{X}_S'\mathbf{X}_S)$, where $\mathbf{X}_S$ is a sub-matrix $\mathbf{X}$ over an index set S of size s.

Definition 2 (Restricted Eigenvalue Condition)**.** *A matrix $\mathbf{X}$ is said to satisfy the RE condition of order s if $\lambda_{\mathsf{min}}(s) > 0$.*

Although the RE condition only restricts $\lambda_{\mathsf{min}}(s)$, in the following analysis we also keep track of $\lambda_{\mathsf{max}}(s)$, which appears in some of the bounds. Establishing the RSC for $\mathbf{X}$ proceeds in a sequence of lemmas (Lemmas 7–5 culminating in Lemma 6). We first show that the RE condition holds for the true covariance of an AR process:

Lemma 1 (from [92])**.** *Let $\mathbf{R} \in \mathbb{R}^{k\times k}$ be the $k \times k$ covariance matrix of a stationary process with power spectral density $S(\omega)$, and denote its maximum and minimum eigenvalues by $\phi_{\max}(k)$ and $\phi_{\mathsf{min}}(k)$, respectively. Then, $\phi_{\max}(k)$ is increasing in k,*

$\phi_{\mathsf{min}}(k)$ *is decreasing in* k*, and we have*

$$\phi_{\mathsf{min}}(k) \downarrow \inf_{\omega} S(\omega), \quad \textit{and} \quad \phi_{\mathsf{max}}(k) \uparrow \sup_{\omega} S(\omega). \tag{A.1}$$

This result gives us the following corollary:

Corollary 3 (Singular Value Spread of $\mathbf{R}$)**.** *Under the sufficient stability assumption, the singular values of the covariance* $\mathbf{R}$ *of an AR process lie in the interval* $\left[\frac{\sigma_{\mathsf{w}}^2}{8\pi}, \frac{\sigma_{\mathsf{w}}^2}{2\pi\eta^2}\right]$.

Proof. For an AR(p) process

$$S(\omega) = \frac{1}{2\pi} \frac{\sigma_{\mathsf{w}}^2}{|1 - \sum_{\ell=1}^{p} \theta_\ell e^{-j\ell\omega}|^2}.$$

Combining $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$ with Lemma 7 proves the claim. □

Note that by Lemma 7, the result of Corollary 7 not only holds for AR processes, but also for *any* stationary process satisfying $\inf_\omega S(\omega) > 0$ and $\sup_\omega S(\omega) < \infty$, i.e., a process with finite spectral spread.

We next establish conditions for the RE condition to hold for the empirical covariance $\widehat{\mathbf{R}}$:

Lemma 2. *If the singular values of* $\mathbf{R}$ *lie in the interval* $[\lambda_{\mathsf{min}}, \lambda_{\mathsf{max}}]$*, then* $\mathbf{X}$ *satisfies the RE condition of order* $s_\star$ *with parameters* $\lambda_{\mathsf{min}}(s_\star) = \lambda_{\mathsf{min}} - ts_\star$ *and* $\lambda_{\mathsf{max}}(s_\star) = \lambda_{\mathsf{max}} + ts_\star$*, where* $t = \max_{i,j} |\widehat{R}_{ij} - R_{ij}|$.

Proof. Let $\widehat{\mathbf{R}} = \frac{1}{n}(\mathbf{X}^T\mathbf{X})$. For every $s_\star$-sparse $\boldsymbol{\theta}$ we have

$$\boldsymbol{\theta}'\widehat{\mathbf{R}}\boldsymbol{\theta} \geq \boldsymbol{\theta}'\mathbf{R}\boldsymbol{\theta} - t\|\boldsymbol{\theta}\|_1^2 \geq (\lambda_{\mathsf{min}} - ts_\star)\|\boldsymbol{\theta}\|_2^2,$$

$$\boldsymbol{\theta}'\widehat{\mathbf{R}}\boldsymbol{\theta} \leq \boldsymbol{\theta}'\mathbf{R}\boldsymbol{\theta} + t\|\boldsymbol{\theta}\|_1^2 \leq (\lambda_{\mathsf{max}} + ts_\star)\|\boldsymbol{\theta}\|_2^2,$$

which proves the claim. □

We will next show that t can be suitably controlled with high probability. Before doing so, we state a key result of Rudzkis [181] regarding the concentration of second-order empirical sums from stationary processes:

Lemma 3. *Let $\mathbf{x}_{-p+1}^n$ be samples of a stationary process which satisfies*

$$x_k = \sum_{j=-\infty}^{\infty} b_{j-k} w_j, \tag{A.2}$$

where w_k's are i.i.d random variables with

$$|\mathbb{E}(|w_j|^k)| \leq (\tilde{c}\sigma_{\mathsf{w}})^k k!\,,\ \ k = 2, 3, \cdots, \tag{A.3}$$

for some constant $\tilde{c}$ and

$$\sum_{j=-\infty}^{\infty} |b_j| < \infty. \tag{A.4}$$

Then, the biased sample autocorrelation given by

$$\widehat{r}_k^b = \frac{1}{n+k} \sum_{i,j=1, j-i=k}^{n+k} x_i x_j$$

satisfies

$$\mathbb{P}(|\widehat{r}_k^b - r_k^b| > t) \leq c_1(n+k)\exp\left(-\frac{c_2}{\sigma_{\mathsf{w}}} \frac{t^2(n+k)}{c_3\sigma_{\mathsf{w}}^3 + t^{3/2}\sqrt{n+k}}\right), \tag{A.5}$$

for positive absolute constants c_1, c_2 and c_3 which are independent of the dimensions of the problem. In particular, if $x_k = w_k$, i.e., a sub-Gaussian white noise process, c_3 vanishes.

Proof. The lemma is a special case of Theorem 4 under Condition 2 of Remark 3 in [181]. For the special case of $x_k = w_k$, the constant H in Lemma 7 of [181] and hence c_3 vanish. □

Using the result of Lemma 3, we can control t and establish the RE condition for $\widehat{\mathbf{R}}$ as follows:

Lemma 4. *Let m be a positive integer. Then, $\mathbf{X}$ satisfies the RE condition of order $(m+1)s$ with a constant $\lambda_{\mathsf{min}}/2$ with probability at least*

$$1 - c_1 p^2 (n+p) \exp\left(- \frac{c_4 \sqrt{\frac{n}{s}}}{1 + c_5 \frac{n+p}{\left(\frac{n}{s}\right)^{3/2}}} \right), \tag{A.6}$$

where c_1 is the same as in Lemma 3, $c_4 = \frac{c_2}{\sigma_{\mathsf{w}}}\sqrt{\frac{\lambda_{\mathsf{min}}}{2(m+1)}}$ and $c_5 = \frac{c_3 \sigma_{\mathsf{w}}^3}{\left(\frac{\lambda_{\mathsf{min}}}{2(m+1)}\right)^{3/2}}$.

Proof. First, note that for the given AR process, condition (A.2) is verified by the Wold decomposition of the process, condition (A.3) results from the sub-Gaussian assumption on the innovations, and condition (A.4) results from the stability of the process. Noting that

$$\widehat{R}_{i,i+k} = \frac{1}{n}\sum_{i=1}^{n} x_i x_{i+k} = \frac{1}{n}\sum_{i,j=1, j-i=k}^{n+k} x_i x_j = \frac{n+k}{n}\widehat{r}_k^b, \tag{A.7}$$

for $i = 1, \cdots, n$ and $k = 0, \cdots, p-1$, Eq. (A.5) implies:

$$\mathbb{P}\left(|\widehat{R}_{i,i+k} - R_{i,i+k}|> \tau\right) \leq c_1(n+k)\exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+k)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}}\right). \tag{A.8}$$

By the union bound and $k \leq p$, we get:

$$\mathbb{P}\left(\max_{i,j}|\widehat{R}_{ij} - R_{ij}|> \tau\right) \leq c_1 p^2(n+p)\exp\left(-\frac{c_2\sqrt{\tau n}}{\frac{c_3\sigma_{\mathsf{w}}^4(n+p)}{\tau^{3/2}n^{3/2}} + \sigma_{\mathsf{w}}}\right). \tag{A.9}$$

Choosing $\tau = \frac{\lambda_{\mathsf{min}}}{2(m+1)s}$ and invoking the result of Lemma 2 establishes the result of the lemma. □

We next define the closely related notion of the Restricted Strong Convexity (RSC):

Definition 3 (Restricted Strong Convexity [148])**.** *Let*

$$\mathbb{V} := \{\mathbf{h} \in \mathbb{R}^p | \|\mathbf{h}_{S^c}\|_1 \leq 3\|\mathbf{h}_S\|_1 + 4\|\boldsymbol{\theta}_{S^c}\|_1\}. \tag{A.10}$$

Then, $\mathbf{X}$ *is said to satisfy the RSC condition of order* s *if there exists a positive* $\kappa > 0$ *such that*

$$\frac{1}{n}\mathbf{h}'\mathbf{X}'\mathbf{X}\mathbf{h} = \frac{1}{n}\|\mathbf{X}\mathbf{h}\|_2^2 \geq \kappa\|\mathbf{h}\|_2^2, \quad \forall \mathbf{h} \in \mathbb{V}. \tag{A.11}$$

The RSC condition can be deduced from the RE condition according to the following result:

Lemma 5 (Lemma 4.1 of [31])**.** *If* $\mathbf{X}$ *satisfies the RE condition of order* $s_\star = (m+1)s$ *with a constant* $\lambda_{\mathsf{min}}((m+1)s)$*, then the RSC condition of order* s *holds with*

$$\kappa = \lambda_{\mathsf{min}}((m+1)s)\left(1 - 3\sqrt{\frac{\lambda_{\mathsf{max}}(ms)}{m\lambda_{\mathsf{min}}\left((m+1)s\right)}}\right)^2. \tag{A.12}$$

We can now establish the RSC condition of order s for $\mathbf{X}$:

Lemma 6. *The matrix of covariates $\mathbf{X}$ satisfies the RSC condition of order s with a constant $\kappa = \frac{\sigma_{\mathsf{w}}^2}{16\pi}$ with probability at least*

$$1 - c_1 p^2 (n+p) \exp\left(-\frac{c_\eta \sqrt{\frac{n}{s}}}{1 + c'_\eta \frac{n+p}{\left(\frac{n}{s}\right)^{3/2}}}\right), \tag{A.13}$$

where $c_\eta = \frac{c_2 \eta}{\sqrt{16\pi(72+\eta^2)}}$ and $c'_\eta = \frac{c_3(16\pi(72+\eta^2))^{3/2}}{\eta^3}$.

Proof. Choosing $m = \lceil \frac{72}{\eta^2} \rceil$, and using Lemmas 2, 4, and 5 establishes the result. Note that if $x_k = w_k$, i.e., a sub-Gaussian white noise process, then c_3 and hence c'_η vanish. □

We are now ready prove Theorems 3 and 6.

A.1.2 Proof of Theorem 3

We first establish the so-called vase (cone) condition for the error vector $\mathbf{h} = \widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}$:

Lemma 7. *For a choice of the regularization parameter $\gamma_n \geq \|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty = \frac{2}{n}\|\mathbf{X}'\left(\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta}\right)\|_\infty$, the optimal error $\mathbf{h} = \widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}$ belongs to the vase*

$$\mathbb{V} := \{\mathbf{h} \in \mathbb{R}^p | \|\mathbf{h}_{S^c}\|_1 \leq 3\|\mathbf{h}_S\|_1 + 4\|\boldsymbol{\theta}_{S^c}\|_1\}. \tag{A.14}$$

Proof. Using several instances of the triangle inequality we have:

$$\begin{aligned}
0 &\geq \frac{1}{n}\left(\|\mathbf{x}_1^n - \mathbf{X}(\boldsymbol{\theta}+\mathbf{h})\|_2^2 - \|\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta}\|_2^2\right) + \gamma_n\left(\|\boldsymbol{\theta}+\mathbf{h}\|_1 - \|\boldsymbol{\theta}\|_1\right)\\
&\geq -\frac{1}{n}\|\mathbf{X}^T\left(\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta}\right)\|_\infty \|\mathbf{h}\|_1 + \gamma_n\left(\|\boldsymbol{\theta}_S + \mathbf{h}_{S^c} + \mathbf{h}_S + \boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}\|_1\right)\\
&\geq -\frac{\gamma_n}{2}(\|\mathbf{h}_{S^c}\|_1 + \|\mathbf{h}_S\|_1) + \gamma_n\left(\|\boldsymbol{\theta}_S + \mathbf{h}_{S^c}\|_1 - \|\mathbf{h}_S + \boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}\|_1\right)\\
&= -\frac{\gamma_n}{2}(\|\mathbf{h}_{S^c}\|_1 + \|\mathbf{h}_S\|_1) + \gamma_n(\|\boldsymbol{\theta}_S\|_1 + \|\mathbf{h}_{S^c}\|_1 - \|\mathbf{h}_S\|_1 - \|\boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}_{S^c}\|_1 - \|\boldsymbol{\theta}_S\|_1)\\
&= \frac{\gamma_n}{2}(\|\mathbf{h}_{S^c}\|_1 - 3\|\mathbf{h}_S\|_1 - 4\|\boldsymbol{\theta}_{S^c}\|_1).
\end{aligned}$$

□

The following result of Negahban et al. [148] allows us to characterize the desired error bound:

Lemma 8 (Theorem 1 of [148]). *If $\mathbf{X}$ satisfies the RSC condition of order s with a constant $\kappa > 0$ and $\gamma_n \geq \|\nabla\mathfrak{L}(\boldsymbol{\theta})\|_\infty$, then any optimal solution $\widehat{\boldsymbol{\theta}}_{\ell_1}$ satisfies*

$$\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2 \leq \frac{2\sqrt{s}\gamma_n}{\kappa} + \sqrt{\frac{2\gamma_n\sigma_s(\boldsymbol{\theta})}{\kappa}}. \tag{$\star$}$$

In order to use Lemma 8, we need to control $\gamma_n = \|\nabla\mathfrak{L}(\boldsymbol{\theta})\|_\infty$. We have:

$$\nabla\mathfrak{L}(\boldsymbol{\theta}) = \frac{2}{n}\mathbf{X}'(\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta}), \tag{A.15}$$

It is easy to check that by the uncorrelatedness of the innovations w_k's, we have

$$\mathbb{E}\left[\nabla\mathfrak{L}(\boldsymbol{\theta})\right] = \frac{2}{n}\mathbb{E}\left[\mathbf{X}'(\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta})\right] = \frac{2}{n}\mathbb{E}\left[\mathbf{X}'\mathbf{w}_1^n\right] = \mathbf{0}. \tag{A.16}$$

Eq. (A.16) is known as the orthogonality principle. We next show that $\nabla\mathfrak{L}(\boldsymbol{\theta})$ is concentrated around its mean. We can write

$$(\nabla\mathfrak{L}(\boldsymbol{\theta}))_i = \frac{2}{n}\mathbf{x}_{-i+1}^{n-i'}\mathbf{w}_1^n,$$

and observe that the jth element in this expansion is of the form $y_j = x_{n-i-j+1}w_{n-j+1}$. It is easy to check that the sequence y_1^n is a martingale with respect to the filtration given by

$$\mathcal{F}_j = \sigma\left(\mathbf{x}_{-p+1}^{n-j+1}\right),$$

where $\sigma(\cdot)$ denote the sigma-field generated by the random variables $x_{-p+1}, x_{-p+2}, \cdots, x_{n-j+1}$. We use the following concentration result for sums of dependent random variables [213]:

Lemma 9. *Fix $n \geq 1$. Let Z_j's be sub-Gaussian $\mathcal{F}_j$-measurable random variables, satisfying for each $j = 1, 2, \cdots, n$,*

$$\mathbb{E}\left[Z_j|\mathcal{F}_{j-1}\right] = 0, \quad \textit{almost surely},$$

then there exists a constant c such that for all $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} Z_j - \mathbb{E}[Z_j]\right| \geq t\right) \leq \exp\left(-\frac{nt^2}{c^2}\right).$$

Proof. This is a special case of Theorem 3.2 of [213] or Lemma 3.2 of [212], for sub-Gaussian-weighted sums of random variables. The constant c depends on the sub-Gaussian constant of Z_i's. □

Since y_j's are a product of two independent sub-Gaussian random variables, they are sub-Gaussian as well. Lemma 9 implies that

$$\mathbb{P}\left(|\nabla \mathfrak{L}(\boldsymbol{\theta})_i| \geq t\right) \leq \exp\left(-\frac{nt^2}{c_0^2 \sigma_{\mathsf{w}}^4}\right). \tag{A.17}$$

where $c_0^2 := \frac{c^2}{\sigma_{\mathsf{w}}^4}$ is an absolute constant. By the union bound, we get:

$$\mathbb{P}\Big(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq t\Big) \leq \exp\left(-\frac{t^2 n}{c_0^2 \sigma_{\mathsf{w}}^4} + \log p\right). \tag{A.18}$$

Let d_4 be any positive integer. Choosing $t = c_0 \sigma_{\mathsf{w}}^2 \sqrt{1+d_4}\sqrt{\frac{\log p}{n}}$, we get:

$$\mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq c_0 \sigma_{\mathsf{w}}^2 \sqrt{1+d_4}\sqrt{\frac{\log p}{n}}\right) \leq \frac{2}{n^{d_4}}.$$

Hence, a choice of $\gamma_n = d_2\sqrt{\frac{\log p}{n}}$ with $d_2 := c_0 \sigma_{\mathsf{w}}^2 \sqrt{1+d_4}$, satisfies $\gamma_n \geq \|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty$ with probability at least $1 - \frac{2}{n^{d_4}}$. Let $d_0 := \frac{(3+d_4)^2}{c_\eta^2}$ and $d_1 = \frac{4c'_\eta(3+d_4)}{c_\eta}$. Using Lemma 6, the fact that $n > s \max\{d_0(\log p)^2, d_1(p \log p)^{1/2}\}$ by hypothesis, and $p > n$ we have that the RSC of order s hold for $\kappa = \frac{\sigma_{\mathsf{w}}^2}{16\pi}$ with a probability at least $1 - \frac{2c_1}{p^{d_4}} - \frac{1}{p^{d_4}}$. Combining these two assertions, the claim of Theorem 1 follows for $d_3 = 32\pi c_0\sqrt{1+d_4}$. ■

A.1.3 Proof of Theorem 6

The proof is mainly based on the following lemma, adopted from Theorem 2.1 of [235], stating that the greedy procedure is successful in obtaining a reasonable $s^\star$-sparse approximation, if the cost function satisfies the RSC:

Lemma 10. *Let $s^\star$ be a constant such that*

$$s^\star \geq 4\rho s \log 20\rho s, \tag{A.19}$$

and suppose that $\mathfrak{L}(\boldsymbol{\theta})$ satisfies RSC of order $s^\star$ with a constant $\kappa > 0$. Then, we have

$$\left\| \widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{OMP}} - \boldsymbol{\theta}_S \right\|_2 \leq \frac{\sqrt{6}\varepsilon_{s^\star}}{\kappa},$$

where $\eta_{s^\star}$ satisfies

$$\varepsilon_{s^\star} \leq \sqrt{s^\star + s} \|\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty. \tag{A.20}$$

Proof. The proof is a specialization of the proof of Theorem 2.1 in [235] to our setting with the spectral spread $\rho = 1/4\eta^2$. □

In order to use Lemma 6, we need to bound $\|\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty$. We have:

$$\mathbb{E}\left[\nabla \mathfrak{L}(\boldsymbol{\theta}_S)\right] = \frac{1}{n}\mathbb{E}\left[\mathbf{X}'(\mathbf{x}_1^n - \mathbf{X}\boldsymbol{\theta}_S)\right] = \frac{1}{n}\mathbb{E}\left[\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_S)\right] = \mathbf{R}(\boldsymbol{\theta} - \boldsymbol{\theta}_S) \leq \frac{\sigma_{\mathsf{w}}^2}{2\pi\eta^2}\varsigma_s(\boldsymbol{\theta})\mathbf{1},$$

where in the second inequality we have used (A.16), and the last inequality results from Corollary 3. Let d_4' be any positive integer. Using the result of Lemma 9 together with the union bound yields:

$$\mathbb{P}\left(\|\nabla\mathfrak{L}(\boldsymbol{\theta}_S)\|_\infty \geq c_0\sigma_{\mathsf{w}}^2\sqrt{1+d_4'}\sqrt{\frac{\log p}{n}}+\frac{\sigma_{\mathsf{w}}^2\varsigma_s(\boldsymbol{\theta})}{2\pi\eta^2}\right)\leq\frac{2}{n^{d_4'}}.$$

Hence, we get the following concentration result for $\varepsilon_{s^\star}$:

$$\mathbb{P}\left(\varepsilon_{s^\star}\geq\sqrt{s^\star+s}\left(c_0\sigma_{\mathsf{w}}^2\sqrt{1+d_4'}\sqrt{\frac{\log p}{n}}+\frac{\sigma_{\mathsf{w}}^2\varsigma_s(\boldsymbol{\theta})}{2\pi\eta^2}\right)\right)\leq\frac{2}{n^{d_4'}}. \tag{A.21}$$

Noting that by (A.35) we have $s^\star+s\leq\frac{4s\log s}{\eta^2}$. Let $d_0'=\frac{4(3+d_4')^2}{\eta^2c_\eta^2}$ and $d_1'=\frac{16c_\eta'(3+d_4)}{c_\eta}$. By the hypothesis of $\varsigma_s(\boldsymbol{\theta})\leq As^{1-\frac{1}{\xi}}$ for some constant A, and invoking the results of Lemmas 6 and 6, we get:

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)}-\boldsymbol{\theta}_S\right\|_2\leq d_2'\sqrt{\frac{s\log s\log p}{n}}+d_2''\sqrt{s\log s}\varsigma_s(\boldsymbol{\theta})\leq d_2'\sqrt{\frac{s\log s\log p}{n}}+d_2''\frac{\sqrt{\log s}}{s^{\frac{1}{\xi}-\frac{3}{2}}},$$

where $d_2'=\frac{16\pi c_0\sqrt{24(1+d_4')}}{\eta}$ and $d_2''=\frac{A}{\pi\eta^3}$, with probability at least $1-\frac{2c_1}{p^{d_4'}}-\frac{1}{p^{d_4'}}-\frac{2}{n^{d_4'}}$. Finally, we have:

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)}-\boldsymbol{\theta}\right\|_2=\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)}-\boldsymbol{\theta}_S+\boldsymbol{\theta}_S-\boldsymbol{\theta}\right\|_2\leq\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{OMP}}^{(s^\star)}-\boldsymbol{\theta}_S\right\|_2+\|\boldsymbol{\theta}_S-\boldsymbol{\theta}\|_2.$$

Choosing $d_3'=2d_2''$ completes the proof. ■

A.1.4 Proof of Proposition 3

Consider the event defined by

$$\mathcal{A} := \left\{ \max_{i,j} |\widehat{R}_{ij} - R_{ij}| \leq \tau \right\}.$$

Eq. (A.9) in the proof of Lemma 4 implies that:

$$\mathbb{P}(\mathcal{A}^c) \leq c_1 p^2 (n+p) \exp\left(-\frac{c_2 \sqrt{\tau n}}{\frac{c_3 \sigma_{\mathsf{w}}^4 (n+p)}{\tau^{3/2} n^{3/2}} + \sigma_{\mathsf{w}}} \right).$$

By choosing τ as in the proof of Theorem 3, we have

$$\begin{aligned} \mathcal{R}^2_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}) \leq \mathcal{R}^2_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\ell_1}) &= \sup_{\mathcal{H}} \left(\mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2 \right] \right) \\ &\leq \mathbb{P}(\mathcal{A}) d_3^2 \frac{s \log p}{n} + \sup_{\mathcal{H}} \mathbb{E}_{A^c} \left[\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2 \right] \\ &\leq d_3^2 \frac{s \log p}{n} + 8(1-\eta)^2 c_1 \exp\left(-\frac{c_2 \sqrt{\tau n}}{\frac{c_3 \sigma_{\mathsf{w}}^4 (n+p)}{\tau^{3/2} n^{3/2}} + \sigma_{\mathsf{w}}} + 3 \log p \right), \end{aligned}$$

where the second inequality follows from Theorem 3, and the third inequality follows from the fact that $\|\widehat{\boldsymbol{\theta}}_{\ell_1} - \boldsymbol{\theta}\|_2^2 \leq 4(1-\eta)^2$ by the sufficient stability assumption. For $n > s \max\{d_0 (\log p)^2, d_1 (p \log p)^{1/2}\}$, the first term will be the dominant, and thus we get $\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}}) \leq 2 d_3 \sqrt{\frac{s \log p}{n}}$, for large enough n.

As for a lower bound on $\mathcal{R}_{\mathsf{est}}(\widehat{\boldsymbol{\theta}}_{\mathsf{minimax}})$, we take the approach of [89] by constructing a family of AR processes with sparse parameters $\boldsymbol{\theta}$ for which the minimax risk is optimal modulo constants. In our construction, we assume that the innovations are Gaussian. The key element of the proof is the Fano's inequality:

Lemma 11 (Fano's Inequality)**.** *Let $\mathcal{Z}$ be a class of densities with a subclass $\mathcal{Z}^\star$ of densities $f_{\boldsymbol{\theta}_i}$, parameterized by $\boldsymbol{\theta}_i$, for $i \in \{0, \cdots, 2^M\}$. Suppose that for any two distinct $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}^\star$, $\mathcal{D}_{\mathsf{KL}}(f_{\boldsymbol{\theta}_1} \| f_{\boldsymbol{\theta}_2}) \leq \beta$ for some constant β. Let $\widehat{\boldsymbol{\theta}}$ be an estimate of the parameters. Then*

$$\sup_j \mathbb{P}(\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j | H_j) \geq 1 - \frac{\beta + \log 2}{M}, \tag{A.22}$$

where H_j denotes the hypothesis that $\boldsymbol{\theta}_j$ is the true parameter, and induces the probability measure $\mathbb{P}(.|H_j)$.

Consider a class $\mathcal{Z}$ of AR processes with s-sparse parameters over any subset $S \subset \{1, 2, \cdots, p\}$ satisfying $|S| = s$, with parameters given by

$$\theta_\ell = \pm e^{-m} \mathbb{1}_S(\ell), \tag{A.23}$$

where m remains to be chosen. We also add the all zero vector $\boldsymbol{\theta}$ to $\mathcal{Z}$. For a fixed S, we have $2^s + 1$ such parameters forming a subfamily $\mathcal{Z}_S$. Consider the maximal collection of $\binom{p}{s}$ subsets S for which any two subsets differ in at least $s/4$ indices. The size of this collection can be identified by $A(p, \frac{s}{4}, s)$ in coding theory, where $A(n, d, w)$ represents the maximum size of a binary code of length n with minimum distance d and constant weight w [135]. We have

$$A(p, \tfrac{s}{4}, s) \geq \frac{p^{\frac{7}{8}s - 1}}{s!},$$

for large enough p (See Theorem 6 in [90]). Also, by the Gilbert-Varshamov bound [135], there exists a subfamily $\mathcal{Z}_S^\star \subset \mathcal{Z}_S$, of cardinality $|\mathcal{Z}_S^\star| \geq 2^{\lfloor s/8 \rfloor} + 1$, such that any two distinct $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}_S^\star$ differ at least in $s/16$ components. Thus for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Z}^\star := \bigcup\limits_S \mathcal{Z}_S^\star$, we have

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \geq \frac{1}{4}\sqrt{s}e^{-m} =: \alpha, \tag{A.24}$$

and $|\mathcal{Z}^\star| \geq \frac{p^{\frac{7}{8}s-1}}{s!} 2^{\lfloor s/8 \rfloor}$. For an arbitrary estimate $\widehat{\boldsymbol{\theta}}$, consider the testing problem between the $\frac{p^{\frac{7}{8}s-1}}{s!} 2^{\lfloor s/8 \rfloor}$ hypotheses $H_j : \boldsymbol{\theta} = \boldsymbol{\theta}_j \in \mathcal{Z}^\star$, using the minimum distance decoding strategy. Using Markov's inequality we have

$$\sup_{\mathcal{Z}} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right] \geq \sup_{\mathcal{Z}^\star} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right] \geq \frac{\alpha}{2} \sup_{\mathcal{Z}^\star} \mathbb{P}\left(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 \geq \frac{\alpha}{2}\right) = \frac{\alpha}{2} \sup_j \mathbb{P}\left(\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_j | H_j\right). \tag{A.25}$$

Let $f_{\boldsymbol{\theta}_j}$ denote joint probability distribution of $\{x_k\}_{k=1}^n$ conditioned on $\{x_k\}_{k=-p+1}^0$ under the hypothesis H_j. Using the Gaussian assumption on the innovations, for $i \neq j$, we have

$$\begin{aligned}
\mathcal{D}_{\mathsf{KL}}(f_{\boldsymbol{\theta}_i} \| f_{\boldsymbol{\theta}_j}) &\leq \sup_{i \neq j} \mathbb{E}\left[\log \frac{f_{\boldsymbol{\theta}_i}}{f_{\boldsymbol{\theta}_j}} | H_i\right] \\
&\leq \sup_{i \neq j} \mathbb{E}\left[-\frac{1}{2\sigma_{\mathsf{w}}^2} \sum_{k=1}^n \left(\left(x_k - \boldsymbol{\theta}_i' \mathbf{x}_{k-p}^{k-1}\right)^2 - \left(x_k - \boldsymbol{\theta}_j' \mathbf{x}_{k-p}^{k-1}\right)^2\right) \Big| H_i\right] \\
&\leq \sup_{i \neq j} \frac{n}{2\sigma_{\mathsf{w}}^2} \mathbb{E}\left[\left((\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)' \mathbf{x}_{k-p}^{k-1}\right)^2 \Big| H_i\right] \\
&= \frac{n}{2\sigma_{\mathsf{w}}^2} \sup_{i \neq j} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)' \mathbf{R} (\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \\
&\leq \frac{n\lambda_{\mathsf{max}}}{2\sigma_{\mathsf{w}}^2} \sup_{i \neq j} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|_2^2 \leq \frac{nse^{-2m}}{64\pi\eta^2} =: \beta.
\end{aligned} \tag{A.26}$$

Using Lemma 11, (A.24), (A.25) and (A.26) yield:

$$\sup_{\mathcal{Z}} \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2\right] \geq \frac{\sqrt{s}e^{-m}}{8}\left(1 - \frac{2\left(\frac{nse^{-2m}}{64\pi\eta^2} + \log 2\right)}{s\log p}\right).$$

for p large enough so that $\log p \geq \frac{\log s - \frac{9}{8}}{\frac{3}{8} - \frac{1}{s}}$. Choosing $m = \frac{1}{2}\log\left(\frac{n}{8\pi\eta^2\log p}\right)$ gives us the claim of Proposition 3 with $L = \frac{d_3}{\eta\sqrt{2\pi}}$ for large enough s and p such that $s\log p \geq \log(256)$. The hypothesis of $s \leq \frac{1-\eta}{\sqrt{8\pi}\eta}\sqrt{\frac{n}{\log p}}$ guarantees that for all $\boldsymbol{\theta} \in \mathcal{Z}^\star$, we have $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta$. ■

A.1.5 Generalization to stable AR processes

We consider relaxing the sufficient stability assumption of $\|\boldsymbol{\theta}\|_1 \leq 1 - \eta < 1$ to $\boldsymbol{\theta}$ being in the set of stable AR processes. Given that the set of all stable AR processes is not necessarily convex, the LASSO and OMP estimates cannot be obtained by convex optimization techniques. Nevertheless, the results of Theorems 3 and 6 can be generalized to the case of stable AR models:

Corollary 4. *The claims of Theorems 3 and 6 hold when* $\boldsymbol{\Theta}$ *is replaced by the set of stable AR processes, except for possibly slightly different constants.*

Proof. Note that the stability of the process guarantees boundedness of the power spectral density. The result follows by simply replacing the bounds $\left[\frac{\sigma_{\mathrm{w}}^2}{8\pi}, \frac{\sigma_{\mathrm{w}}^2}{2\pi\eta^2}\right]$ on the singular values of the covariance matrix $\mathbf{R}$ in Corollary 3 by $[\inf_\omega S(\omega), \sup_\omega S(\omega)]$. □

A.2 Proofs of Main Theorems for GLM's

A.2.1 Roadmap of the Proofs

This appendix contains the proofs of Theorems 5 and 6, as well as Corollaries 1 and 2. Before presenting the proofs, we establish some of the basic properties of the canonical self-exciting process (Proposition 5) as well as our notational conventions as preliminaries. We then state a key result, namely Lemma 12, which is at the core of the proofs of Theorems 5 and 6. The proofs are presented via a sequence of three propositions (Propositions 1–6) based on existing results in the literature, in conjunction with Proposition 5 and Lemma 12. Therefore, Appendix A.2 is stand-alone modulo the proofs of Proposition 5 and Lemma 12.

The proofs of Proposition 5 and Lemma 12 are presented in Appendix A.3. In particular, the proof of Lemma 12 follows from two propositions (Propositions 7 and 8). Therefore, Appendix A.3 is stand-alone modulo the proofs of Propositions 7 and 8.

While Proposition 7 is a well-known result, Proposition 8 requires a careful proof, which is presented in Appendix A.4 and relies on an existing result on the concentration of dependent random variables (Proposition 9).

A.2.2 Preliminaries

We state some useful properties of the canonical self-exciting process in the form of the following proposition:

Proposition 5. *[Properties of the Canonical Self-Exciting Process] The canonical self-exciting process is stationary and we have*

$$
\begin{aligned}
\pi_\star &= \frac{\mu}{1-\mathbf{1}'\boldsymbol{\theta}} > 0, \quad \mu > 0 \Rightarrow \mathbf{1}'\boldsymbol{\theta} < 1, \quad \mu + \mathbf{1}'\boldsymbol{\theta}^+ < 1,\\
S(\omega) &= \frac{1}{2\pi}\left(\pi_\star^2\delta(\omega) + \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2\left|1-\Theta(\omega)\right|^2}\right),\\
S(\omega) &\geq \frac{\pi_\star(1-\pi_\star)}{2\pi(1+2\pi_{\max})^4} =: \kappa_l,
\end{aligned}
$$

where $\pi_\star$ denotes the stationary probability of spiking, $S(\omega)$ denotes the power spectral density of the process, and $\boldsymbol{\theta}^\pm = \max\{\pm\boldsymbol{\theta}, \mathbf{0}\}$.

Proof. The proof is given in Appendix A.3. □

The stationarity gap of $1-\mathbf{1}'\boldsymbol{\theta}$ plays an important role in controlling the convergence rate of the process to its stationary distribution. Throughout the proof, we will also use the notation

$$
\mathbb{S}_p(t) := \{\boldsymbol{\nu} \mid \|\boldsymbol{\nu}\|_p = t\}.
$$

to denote the p-norm ball of radius t. For simplicity of notation, we also define the n-sample empirical expectation as follows:

$$
\hat{\mathbb{E}}_n\{f(x_\cdot)\} := \frac{1}{n}\sum_{i=1}^{n} f(x_i)
$$

for any measurable function $f(x_\cdot)$. Note that the subscript $x_\cdot$ refers to an index in the set $\{1, 2, \cdots, n\}$.

A.2.3 Establishing the Restricted Strong Convexity

The proof of Theorems 5 and 6 relies on establishing the Restricted Strong Convexity (RSC) for the negative log-likelihood given by (3.1). Recall that if the log-likelihood is twice differentiable with respect to $\boldsymbol{\theta}$, the RSC property implies the existence of a lower quadratic bound on the negative log-likelihood:

$$\mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\psi}, \boldsymbol{\theta}) := \mathfrak{L}(\boldsymbol{\theta} + \boldsymbol{\psi}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\psi}' \nabla \mathfrak{L}(\boldsymbol{\theta}) \geq \kappa \|\boldsymbol{\psi}\|_2^2, \tag{A.27}$$

for a positive constant $\kappa > 0$ and all $\boldsymbol{\psi} \in \mathbb{R}^p$ satisfying:

$$\|\boldsymbol{\psi}_{S^c}\|_1 \leq 3\|\boldsymbol{\psi}_S\|_1 + 4\|\boldsymbol{\theta}_{S^c}\|_1. \tag{A.28}$$

for any index set $S \subset \{1, 2, \cdots, p\}$ of cardinality s. The latter condition is known as the cone constraint.

The following key lemma establishes the Restricted Strong Convexity condition for the canonical self-exciting process:

Lemma 12 (Restricted strong convexity of the canonical self-exciting process)**.** *Let $\mathbf{x}_{-p+1}^n$ denote a sequence of samples from the canonical self-exciting process with parameters $\{\mu, \boldsymbol{\theta}\}$ satisfying the conditions given by ($\star$). Then, for $n \geq d_1 s^{2/3} p^{2/3} \log p$, the negative log-likelihood function $\mathfrak{L}(\boldsymbol{\theta})$ satisfies the RSC property with a positive constant $\kappa > 0$ with probability at least $1 - 2\exp\left(-\frac{c\kappa^2 n^3}{s^2 p^2}\right)$, for some constant c, and both κ and c are only functions of d_1, c_1, and $\pi_{\max}$.*

Proof. The proof is given in Appendix A.3. □

Lemma 12 can be viewed as the key result in the proofs of Theorems 5 and 6 which follow next.

A.2.4 Proof of Theorem 5

Given the results of Lemma 12 and Proposition 1, it only remains to establish an upper bound on γ_n. To this end, we establish a suitable upper bound on $\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty$ which holds with high probability and provides the appropriate scaling of γ_n. From Eq. (3.1), we have

$$\nabla \mathfrak{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left[x_i - (\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1}) \right] \frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_i (1 - \lambda_i)}. \tag{A.29}$$

We proceed in two steps:

Step 1. We first show that

$$\mathbb{E}\left[\nabla \mathfrak{L}(\boldsymbol{\theta})\right] = \mathbf{0}. \tag{A.30}$$

To see this, we use the law of iterated expectations on the ith term as follows:

$$\begin{aligned} \mathbb{E}\left[[x_i - (\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1})] \frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_i (1 - \lambda_i)} \right] &= \mathbb{E}\left[\mathbb{E}\left[x_i - (\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1}) \frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_i (1 - \lambda_i)} \middle| \mathbf{x}_{i-p}^{i-1} \right] \right] \\ &= \mathbb{E}\left[\underbrace{\mathbb{E}\left[x_i - (\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1}) \middle| \mathbf{x}_{i-p}^{i-1} \right]}_{0} \frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_i (1 - \lambda_i)} \right] = 0. \end{aligned} \tag{A.31}$$

Summing over i, establishes (A.30).

Step 2. We next show that the summation given by (A.29) is concentrated around its mean. The iterated expectation argument used in establishing (A.31) implies that the sequence

$$\left\{ \left[x_i - (\mu + \boldsymbol{\theta}'\mathbf{x}_{i-p}^{i-1})\right] \frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_i(1-\lambda_i)} \right\}_{i=1}^{n}$$

is a martingale with respect to the filtration given by

$$\mathcal{F}_i = \sigma\left(\mathbf{x}_{-p+1}^{i}\right),$$

where $\sigma(\cdot)$ denote the sigma-field generated by the random variables in its argument.

By Lemma 9 we have

$$\mathbb{P}\left(|(\nabla \mathfrak{L}(\boldsymbol{\theta}))_i| \geq t\right) \leq \exp(-cnt^2). \tag{A.32}$$

By the union bound, we get:

$$\mathbb{P}\Big(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq t \Big) \leq \exp(-ct^2 n + \log p). \tag{A.33}$$

Choosing $t = \sqrt{\frac{1+\alpha_1}{c}}\sqrt{\frac{\log p}{n}}$ for some $\alpha_1 > 0$ yields

$$\begin{aligned} \mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta})\|_\infty \geq \sqrt{\frac{1+\alpha_1}{c}}\sqrt{\frac{\log p}{n}} \right) &\leq 2\exp(-\alpha_1 \log p) \\ &\leq \frac{2}{n^{\alpha_1}}. \end{aligned} \tag{A.34}$$

Hence, a choice of $\gamma_n = d_2\sqrt{\frac{\log p}{n}}$ with $d_2 := \sqrt{\frac{1+\alpha_1}{c}}$ satisfies (1.16) with probability at least $1-\frac{2}{n^{\alpha_1}}$. Combined with the result of Lemma 12 for $n > d_1 s^{2/3} p^{2/3} \log p$, we have that the RSC is satisfied with a constant κ with a probability at least $1-\frac{1}{p^{\alpha_2}} \geq 1-\frac{1}{n^{\alpha_2}}$ for some constant α_2. The latter results in conjunction with Proposition 1 establishes the claim of Theorem 1. ■

Remark. The choice of $\pi_{\min}$ does not affect the proof of Theorem 5, and can be chosen as 0 in defining the set $\boldsymbol{\Theta}$, thereby relaxing the first inequality in $(\star)$. However, as we will show below, the assumption of $\pi_{\min} > 0$ is required for the proof of Theorem 6.

A.2.5 Proof of Theorem 6

The proof is mainly based on the following proposition, adopted from Theorem 2.1 of [235], stating that the greedy procedure is successful in obtaining a reasonable $s^\star$-sparse approximation, if the cost function satisfies the RSC:

Proposition 6. *Suppose that $\mathfrak{L}(\boldsymbol{\theta})$ satisfies RSC with a constant $\kappa > 0$. Let $s^\star$ be a constant such that*

$$s^\star \geq \frac{4s}{\pi_{\min}^2 \kappa} \log \frac{20s}{\pi_{\min}^2 \kappa} = \mathcal{O}(s \log s), \tag{A.35}$$

Then, we have

$$\left\| \widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}} - \boldsymbol{\theta}_s \right\|_2 \leq \frac{\sqrt{6}\epsilon_{s^\star}}{\kappa},$$

where $\epsilon_{s^\star}$ satisfies

$$\epsilon_{s^\star} \leq \sqrt{s^\star + s} \|\nabla \mathfrak{L}(\boldsymbol{\theta}_s)\|_\infty. \tag{A.36}$$

Proof. The proof is a specialization of the proof of Theorem 2.1 in [235] to our setting. □

Recall that Lemma 12 establishes the RSC for the negative log-likelihood function. In order to complete the proof of Theorem 6, it only remains to upper bound $\|\nabla \mathfrak{L}(\boldsymbol{\theta}_s)\|_\infty$. Let $\lambda_{i,s} := \mu + \boldsymbol{\theta}_s' \mathbf{x}_{i-p}^{i-1}$. We have

$$\begin{aligned}
\mathbb{E}\left[\nabla \mathfrak{L}(\boldsymbol{\theta}_s)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[x_i - (\mu + \boldsymbol{\theta}_s' \mathbf{x}_{i-p}^{i-1})\right]\frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_{i,s}(1-\lambda_{i,s})}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[(\boldsymbol{\theta}-\boldsymbol{\theta}_s)'\mathbf{x}_{i-p}^{i-1}\middle|\mathbf{x}_{i-p}^{i-1}\right]\frac{\mathbf{x}_{i-p}^{i-1}}{\lambda_{i,s}(1-\lambda_{i,s})}\right] \\
&\leq c_2 \sigma_s(\boldsymbol{\theta})\mathbf{1}.
\end{aligned}$$

where we have used the fact that $0 \leq x_i \leq 1$ for all i, and $c_2 := \frac{1}{\pi_{\min}(1-\pi_{\max})}$. Invoking the result of Proposition 9 together with the union bound yields:

$$\mathbb{P}\left(\|\nabla \mathfrak{L}(\boldsymbol{\theta}_s)\|_\infty \geq c_1\sqrt{\frac{\log p}{n}} + c_2\sigma_s(\boldsymbol{\theta})\right) \leq \frac{2}{n^{\beta_1}}.$$

for some constants c_1 and β_1. Hence, we get the following concentration result for $\epsilon_{s^\star}$:

$$\mathbb{P}\left(\epsilon_{s^\star} \geq \sqrt{s^\star + s}\left(c_1\sqrt{\frac{\log p}{n}} + c_2\sigma_s(\boldsymbol{\theta})\right)\right) \leq \frac{2}{n^{\beta_1}}. \tag{A.37}$$

Noting that by (A.35) we have $s^\star + s = \mathcal{O}(s \log s) \leq c_0 s \log s$, for some constant c_0, and invoking the result of Lemma 12, we get:

$$\left\|\widehat{\boldsymbol{\theta}}_{\mathsf{POMP}}^{(s^\star)} - \boldsymbol{\theta}_S\right\|_2 \leq d_2'\sqrt{\frac{s\log s \log p}{n}} + d_3' s \log s \sigma_s(\boldsymbol{\theta}) \leq d_2'\sqrt{\frac{s\log s\log p}{n}} + d_3'\frac{\log s}{s^{\frac{1}{\xi}-2}},$$

where $d_2' = \sqrt{c_0}c_1$ and $d_3' = \sqrt{c_0}c_2$. with probability $\left(1 - \exp\left(-\frac{c\kappa^2 n^3}{s^2(\log s)^2 p^2}\right)\right)\left(1 - \frac{2}{n^{\beta_1}}\right)$. Choosing $n > d_1' s^{2/3}(\log s)^{2/3} p^{2/3} \log p$ establishes the claimed success probability of Theorem 6. Finally, we have:

$$\left\|\widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}} - \boldsymbol{\theta}\right\|_2 = \left\|\widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}} - \boldsymbol{\theta}_s + \boldsymbol{\theta}_s - \boldsymbol{\theta}\right\|_2 \leq \left\|\widehat{\boldsymbol{\theta}}^{(s^\star)}_{\mathsf{POMP}} - \boldsymbol{\theta}_s\right\|_2 + \|\boldsymbol{\theta}_s - \boldsymbol{\theta}\|_2.$$

Using $\|\boldsymbol{\theta}_s - \boldsymbol{\theta}\|_2 \leq \sigma_s(\boldsymbol{\theta}) = \mathcal{O}\left(s^{1-\frac{1}{\xi}}\right)$ completes the proof.

■

A.2.6 Proofs of Corollaries 1 and 2

Proof of Corollary 1. The claim is a direct consequence of the boundedness of covariates and can be treated by replacing $\boldsymbol{\theta}$ with the augmented parameter vector $[\mu, \boldsymbol{\theta}']'$ and augmenting the covariate vectors with an initial component of 1. The reader can easily verify that all the proof steps can be repeated in the same fashion. □

Proof of Corollary 2. The claim is a direct consequence of the boundedness of covariates which results in $\phi(\cdot)$ being Lipschitz and hence the stationarity of the underlying process. Moreover, for twice-differentiable $\phi(\cdot)$, the proof of Lemma 12 in Appendix A.2 can be generalized in a straightforward fashion. The reader can easily verify that all the remaining portions of the proofs of the main theorems can be repeated for such $\phi(\cdot)$ in a similar fashion to that of the canonical self-exciting process. □

A.3 Proofs of Proposition 5 and Lemma 12

A.3.1 Proof of Proposition 5

The canonical self-exciting process can be viewed as a Markov chain with states $X_i = \mathbf{x}_{i-p}^{i-1}$. Since each x_i has two possible values, there are 2^p possible states. This Markov chain is irreducible since transition from any state to any other state is possible in at most p steps. Also, transition from an all-zero state to itself is possible. Hence the chain is aperiodic as well. This implies that there exists a stationary distribution for the Markov chain. We also know that if $\{X_i\}_{i=1}^{\infty}$ is a stationary Markov Chain, then for any functional $f(.)$, $\{f(X_i)\}_{i=1}^{\infty}$ is a strictly stationary stochastic process (SSS). Therefore the canonical self-exciting process and the spiking probability sequence λ_1^n are both SSS. In particular, we have

$$\pi_\star := \mathbb{E}[x_i] = \mathbb{E}\left[\mathbb{E}\left[x_i|\lambda_i\right]\right] = \mathbb{E}[\lambda_i] = \mu + \pi_\star \mathbf{1}'\boldsymbol{\theta}.$$

Hence, the stationary probability $\pi_\star$ satisfies:

$$\pi_\star = \frac{\mu}{1 - \mathbf{1}'\boldsymbol{\theta}}.$$

In order to prove the first two inequalities, we make the necessary assumption that the baseline rate μ is positive, due to the non-degeneracy assumption. In order to highlight the necessity of this condition, consider a sample path which contains p successive zeros starting from index $i+1$ to $i+p$, corresponding to an all-zero covariate vector $\mathbf{x}_{i+1}^{i+p}$ (note that this sample path will almost surely occur). We then have $\lambda_{i+p+1} = \mu + \boldsymbol{\theta}'\mathbf{x}_{i+1}^{i+p} = \mu$. Therefore, if μ is not positive, the process becomes degenerate.

The third inequality follows from the fact that for a covariate vector $\mathbf{x}_{i+1}^{i+p}$ with a support matching that of $\boldsymbol{\theta}^{+}$ we have $\lambda_{i+p+1} = \mu + \boldsymbol{\theta}'\mathbf{x}_{i+1}^{i+p} = \mu + \mathbf{1}'\boldsymbol{\theta}^{+}$, which should be a valid probability. Moreover, the inequality is strict since the stationary probability $\pi_\star = \frac{\mu}{1-\mathbf{1}'\boldsymbol{\theta}}$ must be well-defined.

We will next calculate the power spectral density of the process. Let $\boldsymbol{r}_{-\infty}^{\infty}$ and $\boldsymbol{c}_{-\infty}^{\infty}$ denote the autocorrelation and autocovariance values of the process, respectively. By the stationarity of the process we have:

$$r_k = \mathbb{E}\left[x_{\cdot+k}x_{\cdot}\right] = \mathbb{E}\left[x_k x_0\right] = \mathbb{E}\left[\mathbb{E}\left[x_k x_0 | \mathbf{x}_{-\infty}^{k-1}\right]\right] = \mathbb{E}\left[\mu x_0 + \boldsymbol{\theta}'\mathbf{x}_{k-p}^{k-1}x_0\right] = \mu\pi_\star + \boldsymbol{\theta}'\boldsymbol{r}_{k-p}^{k-1}.$$

for $k > 0$. Similarly, by subtracting the means we have the following identity for the autocovariance:

$$c_k = \boldsymbol{\theta}'\boldsymbol{c}_{k-p}^{k-1}. \tag{A.38}$$

A straightforward calculation gives $c_0 = \pi_\star - \pi_\star^2$. Eq. (A.38) resembles the Yule-Walker equations for an AR process of order p with parameter $\boldsymbol{\theta}$ and the innovations variance given by $\sigma^2 = \frac{\pi_\star - \pi_\star^2}{(1-\mathbf{1}'\boldsymbol{\theta})^2}$. Thus, the power spectral density of the canonical self-exciting process can be expressed as:

$$S(\omega) = \frac{1}{2\pi}\left(\pi_\star^2\delta(\omega) + \frac{\pi_\star - \pi_\star^2}{\left(1-\mathbf{1}'\boldsymbol{\theta}\right)^2\left|1-\Theta(\omega)\right|^2}\right). \tag{A.39}$$

We have $1 - \mathbf{1}'\boldsymbol{\theta} \leq 1 + \|\boldsymbol{\theta}\|_1$. Moreover,

$$\begin{aligned}
|1 - \Theta(\omega)| &= \left| 1 - \sum_k \theta_k e^{-j\omega k} \right| \\
&\leq 1 + \|\boldsymbol{\theta}\|_1 = 1 + \|\boldsymbol{\theta}^+\|_1 + \|\boldsymbol{\theta}^-\|_1 \\
&\leq 1 + 2(\pi_{\max} - \mu) \leq 1 + 2\pi_{\max},
\end{aligned}$$

which implies the lower bound on $S(\omega)$. ■

A.3.2 Proof of Lemma 12

The proof is inspired by the elegant treatment of Negahban et al. [148]. The major difficulty in the proof lies in the high inter-dependence of the covariates and observations.

Noticing that the negative log-likelihood (3.1) is twice differentiable, a second order Taylor expansion of the negative log-likelihood (3.1) around $\boldsymbol{\theta}$ yields:

$$\begin{aligned}
\mathfrak{D}_{\mathfrak{L}}(\boldsymbol{\psi}, \boldsymbol{\theta}) &= \mathfrak{L}(\boldsymbol{\theta} + \boldsymbol{\psi}) - \mathfrak{L}(\boldsymbol{\theta}) - \boldsymbol{\psi}' \nabla \mathfrak{L}(\boldsymbol{\theta}) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i \frac{\left(\boldsymbol{\psi}' \mathbf{x}_{i-p}^{i-1}\right)^2}{\left(\mu + \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1} + \nu(\boldsymbol{\psi}' \mathbf{x}_{i-p}^{i-1})\right)^2} + \frac{1}{n} \sum_{i=1}^{n} (1 - x_i) \frac{\left(\boldsymbol{\psi}' \mathbf{x}_{i-p}^{i-1}\right)^2}{\left(1 - \mu - \boldsymbol{\theta}' \mathbf{x}_{i-p}^{i-1} - \nu(\boldsymbol{\psi}' \mathbf{x}_{i-p}^{i-1})\right)^2} \\
&\geq \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{\psi}' \mathbf{x}_{i-p}^{i-1}\right)^2,
\end{aligned}$$

for some $\nu \in [0, 1]$. The inequality follows from the fact that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \nu\boldsymbol{\psi}$ satisfy $(\star)$, and hence:

$$\mu + \boldsymbol{\theta}'\mathbf{x}_{i-p}^{i-1} + \nu\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1} \leq \pi_{\max} < 1,$$
$$1 - \mu - \boldsymbol{\theta}'\mathbf{x}_{i-p}^{i-1} - \nu\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1} \leq 1 - \pi_{\min} < 1.$$

The result of the Lemma 12 is equivalent to proving that

$$\hat{\mathbb{E}}_n\left[\left(\boldsymbol{\psi}'x_{\cdot-p}^{\cdot-1}\right)^2\right] \geq \kappa\|\boldsymbol{\psi}\|_2^2 \tag{A.40}$$

holds with probability greater than $1-2\exp\left(-\frac{c\kappa^2n^3}{s^2p^2}\right)$. Since both sides of (A.40) are quadratic in $\boldsymbol{\psi}$, the statement is equivalent to proving

$$\hat{\mathbb{E}}_n\left[(\boldsymbol{\psi}'\mathbf{x}_{\cdot-p}^{\cdot-1})^2\right] \geq \kappa,$$

for all $\|\boldsymbol{\psi}\|_2 \in \mathbb{S}_2(1)$. We establish this in two steps:

Step 1. First, we show that the statement holds for the true expectation:

$$\mathbb{E}\Big[\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot-p}^{\cdot-1}\right)^2\Big] \geq \kappa_l > 0 \tag{A.41}$$

for some κ_l which will be specified below, for all $\|\boldsymbol{\psi}\|_2 \in \mathbb{S}_2(1)$. To establish the inequality (A.41), we use the following result:

Proposition 7. *Let $\mathbf{R} \in \mathbb{R}^{p\times p}$ be the $p \times p$ covariance matrix of a stationary process with power spectral density $S(\omega)$, and denote its maximum and minimum eigenvalues by $\lambda_{\max}(p)$ and $\lambda_{\min}(p)$ respectively then $\lambda_{\max}(p)$ is increasing in p, $\lambda_{\min}(p)$ is* *decreasing in p and we have*

$$\lambda_{\mathsf{min}}(p) \downarrow \inf_{\omega} S(\omega), \quad \textit{and} \;\; \lambda_{\mathsf{max}}(p) \uparrow \sup_{\omega} S(\omega). \tag{A.42}$$

Proof. This is a well-known result in stochastic processes. See [93] for a proof and detailed discussions. □

Using Proposition 7, we can lower-bound $\mathbb{E}\left[\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot:-p}^{\cdot:-1}\right)^2\right]$ by:

$$\mathbb{E}\left[\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot:-p}^{\cdot:-1}\right)^2\right] = \boldsymbol{\psi}'\mathbf{R}\boldsymbol{\psi} \geq \lambda_{\min}(p) \geq \inf_{\omega} S(\omega).$$

Next, using Proposition 5 the bound of Eq. (A.41) follows for

$$\kappa_l := \frac{\pi_\star(1-\pi_\star)}{2\pi(1+2\pi_{\max})^4}.$$

Step 2. We now show that the empirical and the true expectations of $\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot:-p}^{\cdot:-1}\right)^2$ are close enough to each other. Let

$$\mathfrak{D}_{\boldsymbol{\psi},n} := \hat{\mathbb{E}}_n\left[\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot:-p}^{\cdot:-1}\right)^2\right] - \mathbb{E}\left[\left(\boldsymbol{\psi}'\mathbf{x}_{\cdot:-p}^{\cdot:-1}\right)^2\right].$$

and

$$\mathfrak{D}_n := \sup_{\boldsymbol{\psi}\in\mathbb{S}_2(1)} \left|\mathfrak{D}_{\boldsymbol{\psi},n}\right|.$$

The final step in proving Lemma 12 is given by the following proposition:

Proposition 8. *We have*

$$\mathbb{P}\left[\mathfrak{D}_n \geq \frac{\kappa_l}{4}\right] \leq 2\exp\left(-\frac{c\kappa_l^2 n^3}{s^2 p^2}\right), \tag{A.43}$$

for some constant c.

Proof. The proof is given in Appendix A.4. □

Finally, the statement of Lemma 12 follows from Proposition 8 by taking $\kappa = \kappa_l/4$.

■

A.4 Proof of Proposition 8

In order to establish the concentration inequality of Eq. (A.43), we need to invoke a result from concentration of dependent random variables. We proceed in two steps:

Step 1. We first establish a geometric property of $\mathfrak{D}_n$, namely its $\mathcal{O}(\frac{sp}{n})$-Lipschitz property with respect to the normalized Hamming metric. Recall that the normalized Hamming metric between two sequences $\mathbf{x}_1^n$ and $\mathbf{y}_1^n$ is defined as $d(\mathbf{x}_1^n, \mathbf{y}_1^n) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i \neq y_i)$.

First, by evaluating the first order optimality conditions of the solution $\widehat{\boldsymbol{\theta}}_{\mathsf{sp}}$, it can be shown that the error vector $\boldsymbol{\psi} = \widehat{\boldsymbol{\theta}}_{\mathsf{sp}} - \boldsymbol{\theta}$ satisfies the inequality:

$$\|\boldsymbol{\psi}_{S^c}\|_1 \leq 3\|\boldsymbol{\psi}_S\|_1 + 4\|\boldsymbol{\theta}_{S^c}\|_1,$$

with S denoting the support of the best s-term approximation to $\boldsymbol{\theta}$ (see for example [148]). By the assumption of $\sigma_S(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{s})$, we can choose a constant c_0 such that

$\sigma_S(\boldsymbol{\theta}) \leq c_0\sqrt{s}$. Hence,

$$\|\boldsymbol{\psi}\|_1 \leq 4\|\boldsymbol{\psi}_S\|_1+\sigma_s(\boldsymbol{\theta}) \leq (4+c_0)\sqrt{s}\|\boldsymbol{\psi}_S\|_2\leq (4+c_0)\sqrt{s} \tag{A.44}$$

where we have used the fact that $\|\boldsymbol{\psi}_S\|_1\leq \sqrt{s}\|\boldsymbol{\psi}_S\|_2\leq \sqrt{s}$ for all $\boldsymbol{\psi} \in \mathbb{S}_2(1)$. Therefore for all $i \in \{1, 2, \cdots, n\}$, we have:

$$0 \leq \left(\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1}\right)^2 \leq \|\boldsymbol{\psi}\|_1^2 \leq (4+c_0)^2 s. \tag{A.45}$$

We first prove the claim for $\mathfrak{D}_{\boldsymbol{\psi},n}$. To establish the latter, we need to prove

$$\frac{1}{n}\left|\sum_{i=1}^{n}\left(\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1}\right)^2 - \left(\boldsymbol{\psi}'\mathbf{y}_{i-p}^{i-1}\right)^2\right| \leq Cd(\mathbf{x}_{-p+1}^{n}, \mathbf{y}_{-p+1}^{n}),$$

for some $C = \mathcal{O}(\frac{sp}{n})$, or equivalently

$$\left|\sum_{i=1}^{n}\left(\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1}\right)^2 - \left(\boldsymbol{\psi}'\mathbf{y}_{i-p}^{i-1}\right)^2\right| \leq C' \sum_{i=-p+1}^{n} \mathbf{1}(x_i \neq y_i), \tag{A.46}$$

for some $C' = \mathcal{O}(s)$. Let us start by setting the values of $\mathbf{x}_{-p+1}^{n}$ equal to those of $\mathbf{y}_{-p+1}^{n}$ and iteratively change x_j to $1-x_j$ for all indices j where $x_j \neq y_j$ to obtain the configuration given by $\mathbf{x}_{-p+1}^{n}$. For each such change (say x_j to $1-x_j$), the left hand side changes by at most

$$\left|\sum_{i=1}^{n}\left(\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1}\right)^2_{|x_j=1} - \left(\boldsymbol{\psi}'\mathbf{x}_{i-p}^{i-1}\right)^2_{|x_j=0}\right| \leq \left(\boldsymbol{\psi}'\mathbf{x}_{j-p}^{j-1}\right)^2 + 2\sum_{i\neq j}|\psi_{i-j}|\|\boldsymbol{\psi}\|_1\leq 3\|\boldsymbol{\psi}\|_1^2\leq 3(4+c_0)^2 s,$$

where we have used the inequality given by Eq. (A.45). Hence, the C can be taken as $3(4+c_0)^2 sp/n$ and the claim of the proposition for $\mathfrak{D}_{\boldsymbol{\psi},n}$ follows. A very similar argument can be used to extend the claim to $\mathfrak{D}_n$. Let $\boldsymbol{\psi}^\star := \boldsymbol{\psi}^\star(\mathbf{x}_{-p+1}^n)$ be the $\boldsymbol{\psi}$ for which the supremum in the definition of $\mathfrak{D}_n$ is achieved (such a choice of $\boldsymbol{\psi}$ exists by the Weierstrass extreme value theorem). Since $\boldsymbol{\psi}^\star$ also satisfies (A.44), a similar argument shows that $\mathfrak{D}_n$ is $\mathcal{O}(\frac{sp}{n})$-Lipschitz (with possibly different constants).

Step 2. Next, we establish the concentration of $\mathfrak{D}_n$ around zero. Let $H = [\mathbf{x}_{i-p}^{i-2}, 1]$ and $\widehat{H} = [\mathbf{x}_{i-p}^{i-2}, 0]$ be two vectors (history components) of length p which only differ in their last component, and let the mixing coefficient $\bar{\eta}_{ij}$ for $j \geq i$ be defined as:

$$\bar{\eta}_{ij} = \|p(\mathbf{x}_j^n|H) - p(\mathbf{x}_j^n|\widehat{H})\|_{TV}, \tag{A.47}$$

with $\|\cdot\|_{TV}$ denoting the total variation difference of the probability measures induced on $\{0,1\}^{n-j+1}$. Also, let

$$\eta_{ij} = \sup_{H,\widehat{H}} \bar{\eta}_{ij},$$

and

$$Q_{n,i} := 1 + \eta_{i,i+1} + \cdots + \eta_{i,n}.$$

We now invoke Theorem 1.1 of [121] in the form of the following proposition:

Proposition 9. *If $\mathfrak{D}_n$ is C-Lipschitz and $q := \max_{1\leq i\leq n} Q_{n,i}$, then*

$$\mathbb{P}\left[|\mathfrak{D}_n - \mathbb{E}[\mathfrak{D}_n]| \geq t\right] \leq 2\exp\left(\frac{-2nt^2}{qC^2}\right).$$

Proof. The proof is identical to the beautiful treatment of [121] when specializing the underlying function of the variables $\mathbf{x}_{-p+1}^{i}$ to be $\mathfrak{D}_n$. □

As we showed in Step 1, $C = C'sp/n$, for some constant C'. Now, we have

$$\eta_{ij} \leq 2^{n-j+1} |\pi_{\max}^{n-j+1} - \pi_{\min}^{n-j+1}| \leq (2\pi_{\max})^{n-j+1} \,,$$

where we have used the fact that each element of the measures $p(\mathbf{x}_j^n|H)$ and $p(\mathbf{x}_j^n|\widehat{H})$ satisfies the assumption $(\star)$ and that the size of the state space $\{0,1\}^{n-j+1}$ is given by 2^{n-j+1}. By the assumption $(\star)$, we have $\eta_{ij} \leq \rho^{n-j+1}$ for $\rho := 2\pi_{\max} < 1$. Hence, $Q_{n,i} \leq \frac{1}{1-\rho}$ for all i, and $q \leq \frac{1}{1-\rho}$ by definition. Using the result of Proposition 9, we get:

$$\mathbb{P}\left[\mathfrak{D}_n \geq \mathbb{E}[\mathfrak{D}_n] + \frac{\kappa_l}{2}\right] \leq 2\exp\left(\frac{-n^3\kappa_l^2(1-\rho)}{2C's^2p^2}\right). \tag{A.48}$$

It only remains to show that the expectation in (A.48) can be suitably bounded. Note that by a similar concentration argument for $\mathfrak{D}_{\psi^\star,n}$, we have:

$$\mathbb{E}[\mathfrak{D}_n] = \mathbb{E}[|\mathfrak{D}_{\psi^\star,n}|] = \int_0^\infty \left(1 - F_{|\mathfrak{D}_{\psi^\star,n}|}(t)\right) dt \leq \int_0^\infty 2\exp\left(-\frac{2(1-\rho)n^3t^2}{C's^2p^2}\right) dt = 2\sqrt{\frac{C'\pi}{(1-\rho)}}\frac{ps}{n^{3/2}}.$$

Thus choosing $n \geq d_1 s^{2/3} p^{2/3} \log p$, for some positive constant d_1, $\mathbb{E}[\mathfrak{D}_n]$ drops as $1/\log^{3/2} p$, and will be smaller than $\kappa_l/4$ for large enough p. Hence, combined with (A.48) and by defining $c := \frac{1-\rho}{2C'}$ we have:

$$\mathbb{P}\left[\mathfrak{D}_n \geq \frac{\kappa_l}{4}\right] \leq 2\exp\left(\frac{-cn^3\kappa_l^2}{s^2p^2}\right),$$

which establishes the claim of Proposition 8. ■

A.5 Proof of Main Theorem on Compressible State-Spaces

A.5.1 Proof of Theorem 7

Proof. The main idea behind the proof is establishing appropriate cone and tube constraints [50]. In order to avoid unnecessary complications we assume $s_1 \gg s_2 = \cdots = s_T$ and $n_1 \gg n_2 = \cdots = n_T$. Let $\widehat{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{g}_t$, $t \in [T]$ be an arbitrary solution to the primal form (4.2). We define $\mathbf{z}_t = \mathbf{x}_t - \theta \mathbf{x}_{t-1}$ and $\mathbf{h}_t = \mathbf{z}_t - \widehat{\mathbf{z}}_t$ for $t \in [T]$. For a positive integer p, let $[p] := \{1, 2, \cdots, p\}$. For an arbitrary set $V \subset [p]$, $\mathbf{x}_V$ denotes the vector $\mathbf{x}$ restricted to the indices in V, i.e. all the components outside of V set to zero. We can decompose $(\mathbf{h}_t)_{S_t^c} = (\mathbf{h}_t)_{I_1} + (\mathbf{h}_t)_{I_2} + \cdots + (\mathbf{h}_t)_{I_{r_t}}$, where $r_t = \lfloor p/4s_t \rfloor$, and $(\mathbf{h}_t)_{I_1}$ is the $4s_t$-sparse vector corresponding to $4s_t$ largest-magnitude entries remaining in $(\mathbf{h}_t)_{S_t^c}$, $(\mathbf{h}_t)_{I_2}$ is the $4s_t$-sparse vector corresponding to $4s$ largest-magnitude entries remaining in $(\mathbf{h}_t)_{S_t^c} - (\mathbf{h}_t)_{I_1}$ and so on. By the optimality of $(\widehat{\mathbf{z}}_t)_{t=1}^T$ we have

$$\sum_{t=1}^{T} \frac{\|\mathbf{z}_t + \mathbf{h}_t\|_1}{\sqrt{s_t}} \leq \sum_{t=1}^{T} \frac{\|\mathbf{z}_t\|_1}{\sqrt{s_t}}.$$

Using several instances of triangle inequality we have

$$\sum_{t=1}^{T} \frac{-\sigma_{s_t}(\mathbf{z}_t) + \|(\mathbf{h}_t)_{S_t^c}\|_1 - \|(\mathbf{h}_t)_{S_t}\|_1 + \|(\mathbf{z}_t)_{S_t}\|_1}{\sqrt{s_t}} \leq \sum_{t=1}^{T} \frac{\|\mathbf{z}_t + \mathbf{h}_t\|_1}{\sqrt{s_t}} \leq \sum_{t=1}^{T} \frac{\|\mathbf{z}_t\|_1}{\sqrt{s_t}} = \sum_{t=1}^{T} \frac{\|(\mathbf{z}_t)_{S_t}\|_1 + \sigma_{s_t}(\mathbf{z}_t}{\sqrt{s_t}}$$

which after re-arrangement yields the cone condition given by

$$\sum_{t=1}^{T} \frac{\|(\mathbf{h}_t)_{S_t^c}\|_1}{\sqrt{s_t}} \leq \sum_{t=1}^{T} \frac{\|(\mathbf{h}_t)_{S_t}\|_1 + \|(\mathbf{z}_t)_{S_t^c}\|_1}{\sqrt{s_t}}. \tag{A.49}$$

Also, by the definition of partitions $(I_j)_{j=1}^{r_t}$ we have

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{j=2}^{r_t} \|(\mathbf{h}_t)_{I_j}\|_2 &\leq \sum_{t=1}^{T}\sum_{j=2}^{r_t} 2\sqrt{s_t}\|(\mathbf{h}_t)_{I_j}\|_\infty \\
&\leq \sum_{t=1}^{T}\sum_{j=2}^{r_t} \frac{\|(\mathbf{h}_t)_{I_{j-1}}\|_1}{2\sqrt{s_t}} = \sum_{t=1}^{T} \frac{\|(\mathbf{h}_t)_{S_t^c}\|_1}{2\sqrt{s_t}} \\
&\leq \sum_{t=1}^{T} \frac{\|(\mathbf{h}_t)_{S_t}\|_1 + \sigma_{s_t}(\mathbf{z}_t)}{2\sqrt{s_t}} \leq \sum_{t=1}^{T} \frac{\|(\mathbf{h}_t)_{S_t}\|_2}{2} + \frac{\sigma_{s_t}(\mathbf{z}_t)}{2\sqrt{s_t}}.
\end{aligned} \tag{A.50}$$

Moreover, using the feasibility of both $\mathbf{x}_t$ and $\widehat{\mathbf{x}}_t$ we have the tube constraints

$$\begin{aligned}
&\|\mathbf{y}_1 - \mathbf{A}_1\mathbf{x}_1\|_2 \leq \varepsilon \Rightarrow \|\theta\left(\mathbf{y}_1\right)_{[n_2]} - \theta\mathbf{A}_2\mathbf{x}_1\|_2 \leq \theta\varepsilon, \\
&\|\mathbf{y}_2 - \mathbf{A}_2\mathbf{x}_2\|_2 \leq \sqrt{\tfrac{n_2}{n_1}}\varepsilon,
\end{aligned}$$

from which we conclude $\|\mathbf{y}_2 - \theta\left(\mathbf{y}_1\right)_{[n_2]} - \mathbf{A}_2\mathbf{z}_2\|_2 \leq (1+\theta)\varepsilon$. Similarly $\|\mathbf{y}_2 - \theta\left(\mathbf{y}_1\right)_{[n_2]} - \mathbf{A}_2\widehat{\mathbf{z}}_2\|_2 \leq (1+\theta)\varepsilon$. Therefore the triangle inequality yields $\|\mathbf{A}_2\mathbf{h}_2\|_2 \leq 2(1+\theta)\varepsilon$. Similarly for all $t \in [T]\backslash\{2\}$, we have the tighter bound

$$\|\mathbf{A}_t\mathbf{h}_t\|_2 \leq 2(1+\theta)\sqrt{\frac{n_t}{n_1}}\varepsilon, \tag{A.51}$$

which is a consequence of having fewer measurements for $t \in [T]\backslash\{2\}$. In conjunction, (A.49), (A.50), and (A.51) yield

$$\begin{aligned}
2(1+\theta)\left(T+\sqrt{\frac{n_1}{n_2}}-1\right)\varepsilon &\geq \|\mathbf{A}_1\mathbf{h}_1\|_2+\sum_{t=2}^{T}\sqrt{\frac{n_1}{n_t}}\|\mathbf{A}_t\mathbf{h}_t\|_2 \\
&\geq \sum_{t=1}^{T}\|\widetilde{\mathbf{A}}_t\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2-\sum_{t=1}^{T}\sum_{j=2}^{r_t}\|\widetilde{\mathbf{A}}_t\left(\mathbf{h}_t\right)_{I_j}\|_2 \\
&\geq \sqrt{1-\delta_{4s}}\sum_{t=1}^{T}\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2-\frac{\sqrt{1+\delta_{4s}}}{2}\sum_{t=1}^{T}\sum_{j=2}^{r_t}\|\left(\mathbf{h}_{I_j,t}\right)\|_2 \\
&\geq \sqrt{1-\delta_{4s}}\sum_{t=1}^{T}\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2-\frac{\sqrt{1+\delta_{4s}}}{2}\sum_{t=1}^{T}\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2+\frac{\sigma_{s_t}(\mathbf{z}_t)}{\sqrt{s_t}} \\
&\geq \left(\sqrt{1-\delta_{4s}}-\frac{\sqrt{1+\delta_{4s}}}{2}\right)\sum_{t=1}^{T}\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2-\frac{\sqrt{1+\delta_{4s}}}{2}\sum_{t=1}^{T}\frac{\sigma_{s_t}(\mathbf{z}_t)}{\sqrt{s_t}}.
\end{aligned}$$

Therefore after rearrangement for $\delta_{4s} < 1/3$

$$\sum_{t=1}^{T}\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2\leq 8.37(1+\theta)\left(T+\sqrt{\frac{n_1}{n_2}}-1\right)\varepsilon+\frac{5}{2}\sum_{t=1}^{T}\frac{\sigma_{s_t}(\mathbf{z}_t)}{\sqrt{s_t}}.$$

Next, using (A.50) yields

$$\begin{aligned}
\sum_{t=1}^{T}\|\mathbf{h}_t\|_2 &\leq \sum_{t=1}^{T}\sum_{j=2}^{r_t}\|\left(\mathbf{h}_t\right)_{I_j}\|_2+\|\left(\mathbf{h}_t\right)_{S_t\cup I_1}\|_2 \\
&\leq 12.55\left(T+\sqrt{\frac{n_1}{n_2}}-1\right)\varepsilon+3\sum_{t=1}^{T}\frac{\sigma_{s_t}(\mathbf{z}_t)}{\sqrt{s_t}}. \qquad \text{(A.52)}
\end{aligned}$$

By definition we have $\mathbf{h}_t = \mathbf{g}_t - \theta \mathbf{g}_{t-1}$ for $t \in [T]$ with $\mathbf{g}_0 = \mathbf{0}$. Therefore by induction we have $\mathbf{g}_t = \sum_{j=1}^{t} \theta^{t-j} \mathbf{h}_j$ or in matrix form

$$\mathcal{G} := \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_t \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{I} & 0 & \cdots & 0 \\ \theta \mathbf{I} & \mathbf{I} & \cdots & 0 \\ \theta^2 \mathbf{I} & \theta \mathbf{I} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{T-1}\mathbf{I} & \theta^{T-2}\mathbf{I} & \cdots & \mathbf{I} \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_t \end{bmatrix} =: \mathcal{A}\mathcal{H}.$$

Using several instances of the triangle inequality we get:

$$\begin{aligned} \sum_{t=1}^{T} \|\mathbf{g}_t\|_2 &= \sum_{t=1}^{T} \left\| \sum_{j=1}^{t} \theta^{t-j} \mathbf{h}_j \right\|_2 \leq \sum_{t=1}^{T} \sum_{j=1}^{t} \theta^{t-j} \left\| \mathbf{h}_j \right\|_2 \\ &\leq \sum_{j=1}^{T-1} \theta^j \sum_{t=1}^{T} \|\mathbf{h}_t\|_2 = \frac{1-\theta^T}{1-\theta} \sum_{t=1}^{T} \|\mathbf{h}_t\|_2, \end{aligned}$$

which in conjunction with (A.52) completes the proof. □

A.5.2 The Expectation Maximization Algorithm

In this section we give a short overview of the EM algorithm and its connection to iteratively re-weighted least squares (IRLS) algorithms. More details can be found in [18] and the references therein. Given the observations $\mathbf{y}$, the goal of the EM algorithm is to find the ML estimates of a set of parameters $\mathbf{\Theta}$ by maximizing the likelihood $\mathfrak{L}(\mathbf{\Theta}) := p(\mathbf{y}|\mathbf{\Theta})$. Such maximization problems are typically intractable, but often become significantly simpler by introducing a latent variable $\mathbf{u}$. The EM algorithm connects solving the ML problem to maximizing $\widetilde{\mathfrak{L}}(\boldsymbol{\Theta}) := p(\mathbf{y}, \mathbf{u}|\boldsymbol{\Theta})$, if one knew $\mathbf{u}$.

Consider the state-space model:

$$\begin{aligned} &\mathbf{x}_t = \boldsymbol{\Theta}\mathbf{x}_{t-1} + \tfrac{\boldsymbol{\omega}_t}{\sqrt{\mathbf{u}_t}}, \\ &\mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \end{aligned} \tag{A.53}$$

where $\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{I})$, $\mathbf{u}_t$ is a positive i.i.d. random vector, and the square root operator and division of the two vectors are understood as element-wise operations. Let $\delta_{t,j}^2 := (\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1})_j^2$ for $j = 1, 2, \cdots, p$. For an appropriate choice of the density of $(\mathbf{u}_t)_j$ denoted by $p_U(\cdot)$, we have [18]:

$$p(\tfrac{\boldsymbol{\omega}_t}{\sqrt{\mathbf{u}_t}}|\boldsymbol{\Theta}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\Theta}) \propto \exp\left(-\lambda \textstyle\sum_{j=1}^p \kappa(\delta_{t,j}^2)\right),$$

where

$$\kappa(z) := -2\ln\left(\int_0^\infty u^{n/2} e^{-uz/2} p_U(u) du\right), \forall z \geq 0, \tag{A.54}$$

and $\kappa'(z)$ is a completely monotone function [123]. Random vectors of the form $\mathbf{w}_t = \frac{\boldsymbol{\omega}_t}{\sqrt{\mathbf{u}_t}}$ are known as Normal/Independent [123]. Note that a choice of $\kappa(z) = \sqrt{z^2 + \epsilon^2}$ results in the ϵ-perturbed Laplace distributions used in our model [18]. Given T observations $(\mathbf{y}_t)_{t=1}^T \in \mathbb{R}^{n_t}$ and conditionally independent samples $(\mathbf{x}_t)_{t=1}^T \in \mathbb{R}^p$, we denote the objective function of the MAP estimator by $\mathfrak{L}((\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta})$, that is $\log \mathfrak{L}((\mathbf{x}_t)_{t=1}^T, \boldsymbol{\Theta}) = \sum_{t=1}^T \log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\Theta}) + \log p\left(\mathbf{x}_t|\mathbf{x}_{t-1}, \boldsymbol{\Theta}\right)$. Consider the current estimates $\left\{(\widehat{\mathbf{x}}_t^{(l)})_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)}\right\}$ at iteration l. Then:

$$\begin{aligned}
&\log \mathfrak{L}((\mathbf{x}_t)_{t=1}^T, \mathbf{\Theta}) - \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{\Theta}) \\
&= \sum_{t,j=1}^{T,p} \log \left(\int_{(\mathbf{u}_t)_j} p\left((\boldsymbol{\omega}_t)_j, (\mathbf{u}_t)_j | \mathbf{\Theta} \right) d\,(\mathbf{u}_t)_j \right) \\
&= \sum_{t,j=1}^{T,p} \log \left(\int_{(\mathbf{u}_t)_j} \frac{p\left((\mathbf{u}_t)_j \middle| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right)}{p\left((\mathbf{u}_t)_j \middle| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right)} p\left((\boldsymbol{\omega}_t)_j, (\mathbf{u}_t)_j | \mathbf{\Theta} \right) d\,(\mathbf{u}_t)_j \right) \\
&\geq \sum_{t,j=1}^{T,p} \int_{(\mathbf{u}_t)_j} p\left((\mathbf{u}_t)_j \middle| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right) \log \left(\frac{p\left((\boldsymbol{\omega}_t)_j, (\mathbf{u}_t)_j | \mathbf{\Theta} \right)}{p\left((\mathbf{u}_t)_j \middle| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right)} \right) d\,(\mathbf{u}_t)_j \\
&= \sum_{t,j=1}^{T,p} \mathbb{E}_{(\mathbf{u}_t)_j \left| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right.} \Big\{ \log p((\boldsymbol{\omega}_t)_j, (\mathbf{u}_t)_j | \mathbf{\Theta}) \Big\} + C, \qquad \text{(A.55)}
\end{aligned}$$

where the inequality follows from Jensen's inequality and the constant C accounts for terms which do not depend on $\mathbf{\Theta}$. The so called Q-function is defined as:

$$\begin{aligned}
&Q\Big((\mathbf{x}_t)_{t=1}^T, \mathbf{\Theta} \Big| \left(\widehat{\mathbf{x}}_t^{(l)} \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \Big) := \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{\Theta}) \\
&+ \sum_{t,j=1}^{T,p} \mathbb{E}_{(\mathbf{u}_t)_j \left| \left((\widehat{\mathbf{x}}_t^{(l)})_j \right)_{t=1}^T, \widehat{\mathbf{\Theta}}^{(l)} \right.} \Big\{ \log p((\boldsymbol{\omega}_t)_j, (\mathbf{u}_t)_j | \mathbf{\Theta}) \Big\}. \qquad \text{(A.56)}
\end{aligned}$$

The EM algorithm maximizes the lower bound given by the Q-function of (A.56) instead of the log-likelihood itself. Moreover for all $t \in [T], j \in [p]$ and $\kappa(z) =$

$\sqrt{z^2+\epsilon^2}$ we have [123]:

$$\mathbb{E}_{(\mathbf{u}_t)_j \Big| \left(\left(\widehat{\mathbf{x}}_t^{(l)}\right)_j\right)_{t=1}^{T}, \widehat{\boldsymbol{\Theta}}^{(l)}} \Big\{ \log p((\boldsymbol{\omega}_t)_j\,, (\mathbf{u}_t)_j\,|\boldsymbol{\Theta}) \Big\} = -\frac{\lambda}{2} \frac{(\mathbf{x}_t - \boldsymbol{\Theta}\mathbf{x}_{t-1})_j^2 + \epsilon^2}{\sqrt{(\widehat{\mathbf{x}}_t^{(l)} - \widehat{\boldsymbol{\Theta}}^{(l)}\widehat{\mathbf{x}}_{t-1}^{(l)})_j^2 + \epsilon^2}},$$

which after replacement results in the state-space model given by (4.8). This expectation gets updated in the outer EM loop using the *final* outputs of the inner loop. The outer EM algorithm can thus be summarized as forming the Q-function (E-step) and maximizing over $\boldsymbol{\Theta}$ (M-step), which is known to converge to a stationary point due to its ascent property [18]. As discussed in Section 4.2.3, the outer M-step is implemented by another instance of the EM algorithm by alternating between Fixed Interval Smoothing (E-step) and updating $\boldsymbol{\Theta}$ (M-step).

Appendix B: Statistical Tests of Goodness of Fit

In this appendix, we will give an overview of the statistical goodness-of-fit tests for assessing the accuracy of the AR model estimates. A detailed treatment can be found in [127].

B.1 Goodness-Of-Fit Tests forAutoregressive Models

B.1.1 Residue-based tests

Let $\widehat{\boldsymbol{\theta}}$ be an estimate of the parameters of the process. The residues (estimated innovations) of the process based on $\widehat{\boldsymbol{\theta}}$ are given by

$$e_k = x_k - \widehat{\boldsymbol{\theta}}^T \mathbf{x}_{k-p}^{k-1}, \qquad i = 1, 2, \cdots, n.$$

The main idea behind most of the available statistical tests is to quantify how close the sequence $\{e_i\}_{i=1}^n$ is to an i.i.d. realization of a known distribution F_0 which is most likely absolutely continuous . Let us denote the empirical distribution of the n-samples by $\widehat{F}_n$. If the samples are generated from F_0 the Glivenko-Cantelli theorem suggests that:

$$\sup_t |\widehat{F}_n(t) - F_0(t)| \xrightarrow{\text{a.s.}} 0.$$

That is, for large n the empirical distribution $\widehat{F}_n$ is uniformly close to F_0. The Kolmogorov-Smirnov (KS) test, Cramér-von Mises (CvM) criterion and the Anderson-Darling (AD) test are three measures of discrepancy between $\widehat{F}_n$ and F_0 which are easy to compute and are sufficiently discriminant against alternative distributions. More specifically, the limiting distribution of the following three random variables are known: The KS test statistic

$$K_n := \sup_t |\widehat{F}_n(t) - F_0(t)|,$$

the CvM statistic

$$C_n := \int (\widehat{F}_n(t) - F_0(t))^2 dF_0(t),$$

and the AD statistic

$$A_n := \int \frac{(\widehat{F}_n(t) - F_0(t))^2}{F_0(t)\,(1 - F_0(t))} dF_0(t).$$

For large values of n, the Glivenko-Cantelli theorem also suggests that these statistics should be small. A simple calculation leads to the following equivalent for the statistics:

$$K_n = \max_{1\le i\le n} \max\left\{ \left|\frac{i}{n} - F_0(e_i)\right|, \left|\frac{i-1}{n} - F_0(e_i)\right| \right\},$$

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0(e_i) - \frac{2i-1}{2n} \right)^2,$$

and

$$nA_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \Big(\log F_0(e_i) + \log \Big(1 - F_0(e_i) \Big) \Big).$$

B.1.2 Spectral domain tests for Gaussian AR processes

The aforementioned KS, CvM and AD tests all depend on the distribution of the innovations. For Gaussian AR processes, the spectral versions of these tests are introduced in [15]. These tests are based on the similarities of the periodogram of the data and the estimated power-spectral density of the process. The key idea is summarized in the following lemma:

Lemma 13. *Let $S(\omega)$ be the (normalized) power-spectral density of stationary process with bounded spectral spread, and $\widehat{S}_n(\omega)$ be the periodogram of the n samples of a realization of such a process, then for all ω we have:*

$$\sqrt{n}\left(2\int_0^{\omega}\left(\widehat{S}_n(\lambda)-S(\lambda)\right)d\lambda\right)\xrightarrow{\mathsf{d.}}\mathcal{Z}(\omega), \tag{B.1}$$

where $\mathcal{Z}(\omega)$ is a zero-mean Gaussian process.

The explicit formula for the covariance function of $\mathcal{Z}(.)$ is calculated in [15]. Lemma 13 suggests that for a good estimate $\widehat{\boldsymbol{\theta}}$ which admits a power spectral density $S(\omega;\widehat{\boldsymbol{\theta}})$, one should get a (*close* to) Gaussian process replacing $S(\omega)$ with $S(\omega;\widehat{\boldsymbol{\theta}})$ in (B.1). The spectral form of the CvM, KS and AD statistics can thus be characterized given an estimate $\widehat{\boldsymbol{\theta}}$.

B.2 Goodness-Of-Fit Tests for Point Process Models

In this appendix, we will give an overview of the statistical tools used to assess the goodness-of-fit of point process models. A detailed treatment can be found in [206].

B.2.1 The Time-Rescaling Theorem

Let $0 < t_1 < t_2 < \cdots$ be a realization of a continuous point process with conditional intensity $\lambda(t) > 0$, i.e. t_k is the first instance at which $N(t_k) = k$. Define the transformation

$$z_k := Z(t_k) = \int_{t_{k-1}}^{t_k} \lambda(t)dt. \tag{B.2}$$

Then, the transformed point process with events occurring at $t'_k = \sum_{i=1}^{k} z_k$ corresponds to a homogeneous Poisson process with rate 1. Equivalently, $z_1, z_2, \cdots$ are *i.i.d exponential* random variables. The latter can be used to construct statistical tests for the goodness-of-fit.

B.2.2 The Komlogorov-Smirnov Test for Homogeneity

Suppose that we have obtained the rescaled process through (B.2) with the *estimated* conditional intensity. When applying the time-rescaling theorem to the discretized process, if the estimated conditional intensity is close to its true value, the rescaled process is expected to behave as a homogeneous Poisson process with rate 1. The Kolmogorov-Smirnov (KS) test can be used to check for the homogeneity of the process. Let z_k's be the rescaled times and define the transformed rescaled times by the inverse exponential CDF:

$$u_k := 1 - e^{-z_k}.$$

If the true conditional intensity was used to rescale the process, the random variables u_k must be i.i.d. $\mathsf{Uniform}(0,1]$ distributed. The KS test plots the empirical qualities of u_k's versus the true quantiles of the uniform density given by $b_k = \frac{k-1/2}{J}$, where J is the total number of observed spikes. If the conditional intensity is well estimated, the resulting curve must lie near the 45° line. The asymptotic statistics of the KS distribution can be used to construct confidence intervals for the test. For instance, the 95% and 99% confidence intervals are approximately given by $\pm\frac{1.36}{\sqrt{J}}$ and $\pm\frac{1.63}{\sqrt{J}}$ hulls around the 45° line, respectively.

B.2.3 The Autocorrelation Function Test for Independence

In order to check for the independence of the resulting rescaled intervals z_k, the following transformation is used:

$$v_k = \Phi^{-1}(u_k)$$

where Φ is the standard Normal CDF. If the true conditional intensity was used to rescale the process, then v_k's would be i.i.d. Gaussian and their uncorrelatedness would imply independence. The Autocorrelation Function (ACF) of the variables v_k must then be close to the discrete delta function. The 95% and 99% confidence intervals can be considered using the asymptotic statistics of the sample ACF, approximately given by $\pm\frac{1.96}{\sqrt{J}}$ and $\pm\frac{2.575}{\sqrt{J}}$, respectively.

Remark. The binning size used for discretizing the data can potentially affect the ISI distribution of the time-rescaled process. In order to avoid these issues, we have used the empirical ISI distribution estimated from a large realization of the process (estimated from the training data) as the null hypothesis for both tests (performed on the test data).

Bibliography

[1] Cushing, ok wti spot price fob dataset. (Date last accessed 14-December-2015).

[2] Regional integrated transportation information system (ritis). (Date last accessed 27-December-2015).

[3] Regional integrated transportation information system (ritis). (Date last accessed 27-December-2015).

[4] *MATLAB implementation of algortihms for estimation of self-exciting point process moels.* Available on GitHub Repository: `https://github.com/kaazemi/PPSelf`, 2016.

[5] *MATLAB implementation of the FCSS algorithm.* Available on GitHub Repository: `https://github.com/kaazemi/FCSS`, 2016.

[6] *MATLAB implementation of the FCSS algorithm.* Available on GitHub Repository: `https://github.com/kaazemi/FADE`, 2016.

[7] *SpikeFinder Challenge.* Available at: `http://spikefinder.codeneuro.org/`, 2016.

[8] Mohammed S Ahmed and Allen R Cook. *Analysis of freeway traffic time-series data by using Box-Jenkins techniques.* Number 722. 1979.

[9] Samir A Ahmed and Allen R Cook. *Application of time-series analysis techniques to freeway incident detection.* Number 841. 1982.

[10] Carlos D Aizenman and David J Linden. Regulation of the rebound depolarization and spontaneous firing patterns of deep nuclear neurons in slices of rat cerebellum. *Journal of Neurophysiology*, 82(4):1697–1709, 1999.

[11] Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

[12] Hirotugu Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(1):203–217, 1970.

[13] Htrotugu Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[14] Brian Anderson and John B Moore. Optimal filtering. *Prentice-Hall Information and System Sciences Series, Englewood Cliffs: Prentice-Hall, 1979*, 1979.

[15] Theodore Wilbur Anderson. Goodness-of-fit tests for autoregressive processes. *Journal of time series analysis*, 18(4):321–339, 1997.

[16] Nicholas Antipa, Sylvia Necula, Ren Ng, and Laura Waller. Single-shot diffuser-encoded light field imaging. In *Computational Photography (ICCP), 2016 IEEE International Conference on*, pages 1–11. IEEE, 2016.

[17] Demba Ba, Behtash Babadi, Patrick Purdon, and Emery Brown. Exact and stable recovery of sequences of signals with sparse increments via differential ℓ_1-minimization. 2012.

[18] Demba Ba, Behtash Babadi, Patrick L Purdon, and Emery N Brown. Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Transactions on Signal Processing*, 62(1):183–195, 2014.

[19] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Bayesian compressive sensing using laplace priors. *IEEE Transactions on Image Processing*, 19(1):53–63, 2010.

[20] Behtash Babadi, Scott M McKinney, Vahid Tarokh, and Jeffrey M Ellenbogen. Diba: a data-driven bayesian algorithm for sleep spindle detection. *IEEE Transactions on Biomedical Engineering*, 59(2):483–493, 2012.

[21] Kareem E Baddour and Norman C Beaulieu. Autoregressive modeling for fading channel simulation. *IEEE Transactions on Wireless Communications*, 4(4):1650–1662, 2005.

[22] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[23] Riccardo Barbieri, Eric C Matten, AbdulRasheed A Alabi, and Emery N Brown. A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability. *American Journal of Physiology-Heart and Circulatory Physiology*, 288(1):H424–H435, 2005.

[24] Riccardo Barbieri, Michael C Quirk, Loren M Frank, Matthew A Wilson, and Emery N Brown. Construction and analysis of non-poisson stimulus-response models of neural spiking activity. *Journal of neuroscience methods*, 105(1):25–37, 2001.

[25] Jaume Barceló, Lidin Montero, Laura Marqués, and Carlos Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, (2175):19–27, 2010.

[26] Maurice S Bartlett. Statistical estimation of density functions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 245–254, 1963.

[27] MS Bartlett. The spectral analysis of point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 264–296, 1963.

[28] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[29] Mark F Bear, Barry W Connors, and Michael A Paradiso. *Neuroscience*, volume 2. Lippincott Williams & Wilkins, 2007.

[30] Jörg Bewersdorf, Rainer Pick, and Stefan W Hell. Multifocal multiphoton microscopy. *Optics letters*, 23(9):655–657, 1998.

[31] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

[32] Aaron G Blankenship and Marla B Feller. Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nature Reviews Neuroscience*, 11(1):18–29, 2010.

[33] Joanna Borowska, Stuart Trenholm, and Gautam B Awatramani. An intrinsic neural oscillator in the degenerating mouse retina. *The Journal of Neuroscience*, 31(13):5000–5012, 2011.

[34] EJ Botcherby, R Juškaitis, and T Wilson. Scanning two photon fluorescence microscopy with extended depth of field. *Optics communications*, 268(2):253–260, 2006.

[35] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[36] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[37] Emery N Brown, Robert E Kass, and Partha P Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature neuroscience*, 7(5):456–461, 2004.

[38] Emery N Brown, David P Nguyen, Loren M Frank, Matthew A Wilson, and Victor Solo. An analysis of neural receptive field plasticity by point process adaptive filtering. *Proceedings of the National Academy of Sciences*, 98(21):12261–12266, 2001.

[39] ERVRL Brown, Riccardo Barbieri, Valérie Ventura, R Kass, and L Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

[40] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[41] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[42] A Bullen, SS Patel, and P Saggau. High-speed, random-access fluorescence microscopy: I. high-resolution optical recording with voltage-sensitive dyes and ion indicators. *Biophysical journal*, 73(1):477–491, 1997.

[43] John Parker Burg. Maximum entropy spectral analysis. In *37th Annual International Meeting*. Society of Exploration Geophysics, 1967.

[44] Ana Calabrese, Joseph W Schumacher, David M Schneider, Liam Paninski, and Sarah MN Woolley. A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds. *PloS one*, 6(1):e16104, 2011.

[45] Emmanuel J Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians Madrid, August 22–30*, pages 1433–1452, 2006.

[46] Emmanuel J Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15:257–325, 2006.

[47] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[48] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[49] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[50] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[51] Avishy Carmi, Pini Gurfil, and Dimitri Kanevsky. Methods for sparse signal recovery using kalman filtering with embedded pseudo-measurement norms and quasi-norms. *IEEE Transactions on Signal Processing*, 58(4):2405–2409, 2010.

[52] Leonardo Causa, Claudio M Held, Javier Causa, Pablo A Estévez, Claudio A Perez, Rodrigo Chamorro, Marcelo Garrido, Cecilia Algarín, and Patricio Peirano. Automated sleep-spindle detection in healthy children polysomnograms. *IEEE Transactions on Biomedical Engineering*, 57(9):2135–2146, 2010.

[53] Catie Chang and Gary H Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

[54] Adam S Charles and Christopher J Rozell. Dynamic filtering of sparse signals using reweighted ℓ_1. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6451–6455. IEEE, 2013.

[55] Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on*, pages 3869–3872. IEEE, 2008.

[56] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. *Medical physics*, 35(2):660–663, 2008.

[57] Jie Chen and Xiaoming Huo. Sparse representations for multiple measurement vectors (mmv) in an over-complete dictionary. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 4, pages iv–257. IEEE, 2005.

[58] Xiaowei Chen, Ulrich Leischner, Nathalie L Rochefort, Israel Nelken, and Arthur Konnerth. Functional mapping of single spines in cortical neurons in vivo. *Nature*, 475(7357):501–505, 2011.

[59] Zhe Chen, David F Putrino, Soumya Ghosh, Riccardo Barbieri, and Emery N Brown. Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(2):121–135, 2011.

[60] Stephen Clark. Traffic prediction using multivariate nonparametric regression. *Journal of transportation engineering*, 129(2):161–168, 2003.

[61] Z Clemens, D Fabo, and P Halasz. Overnight verbal memory retention correlates with the number of sleep spindles. *Neuroscience*, 132(2):529–535, 2005.

[62] Albert Cohen, Ronald DeVore, Pencho Petrushev, and Hong Xu. Nonlinear approximation and the space bv($\mathbb{R}^2$). *American Journal of Mathematics*, pages 587–628, 1999.

[63] Shane F Cotter, Bhaskar D Rao, Kjersti Engan, and Kenneth Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *Signal Processing, IEEE Transactions on*, 53(7):2477–2488, 2005.

[64] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[65] Ralph B D'Agostino. *Goodness-of-fit-techniques*, volume 68. CRC press, 1986.

[66] DJ Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Springer Science & Business Media, 2007.

[67] Mike E Davies and Yonina C Eldar. Rank awareness in joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(2):1135–1146, 2012.

[68] Luigi De Gennaro and Michele Ferrara. Sleep spindles: an overview. *Sleep medicine reviews*, 7(5):423–440, 2003.

[69] Stéphanie Devuyst, Thierry Dutoit, Patricia Stenuit, and Myriam Kerkhofs. Automatic sleep spindles detectionoverview and development of a standard proposal assessment method. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1713–1716. IEEE, 2011.

[70] Nicolas Dey, Laure Blanc-Feraud, Christophe Zimmer, Pascal Roux, Zvi Kam, Jean-Christophe Olivo-Marin, and Josiane Zerubia. Richardson–lucy algorithm with total variation regularization for 3d confocal microscope deconvolution. *Microscopy research and technique*, 69(4):260–266, 2006.

[71] David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[72] Pascal Dufour, Michel Piché, Yves De Koninck, and Nathalie McCarthy. Two-photon excitation fluorescence microscopy with a high depth of field using an axicon. *Applied optics*, 45(36):9246–9252, 2006.

[73] Fazil Duman, Aykut Erdamar, Osman Erogul, Ziya Telatar, and Sinan Yetkin. Efficient sleep spindle detection algorithm with decision tree. *Expert Systems with Applications*, 36(6):9980–9985, 2009.

[74] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[75] Mike Egesdal, Chris Fathauer, Kym Louie, Jeremy Neuman, George Mohler, and Erik Lewis. Statistical and stochastic modeling of gang rivalries in Los Angeles. *SIAM Undergraduate Research Online*, 3:72–94, 2010.

[76] Stephen John Eglen, Michael Weeks, Mark Jessop, Jennifer Simonotto, Tom Jackson, and Evelyne Sernagor. A data repository and analysis framework for spontaneous neural activity recordings in developing retina. *GigaScience*, 3(1):3, 2014.

[77] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications.* Cambridge University Press, 2012.

[78] Kaveh Farokhi Sadabadi. *Vehicular traffic modelling, data assimilation, estimation and short term travel time prediction.* PhD thesis, University of Maryland, College Park, 2014.

[79] Ian J Fevrier, Saul B Gelfand, and Michael P Fitz. Reduced complexity decision feedback equalization for multipath channels with large delay spreads. *Communications, IEEE Transactions on*, 47(6):927–937, 1999.

[80] Jeffrey J Field, Keith A Wernsing, Scott R Domingue, Alyssa M Allende Motz, Keith F DeLuca, Dean H Levi, Jennifer G DeLuca, Michael D Young, Jeff A Squier, and Randy A Bartels. Superresolved multiphoton microscopy with spatial frequency-modulated imaging. *Proceedings of the National Academy of Sciences*, 113(24):6605–6610, 2016.

[81] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[82] Luay Fraiwan, Khaldon Lweesy, Natheer Khasawneh, Heinrich Wenz, and Hartmut Dickhaus. Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier. *Computer methods and programs in biomedicine*, 108(1):10–19, 2012.

[83] Sylvia Frühwirth-Schnatter. Applied state space modelling of non-gaussian time series using integration-based kalman filtering. *Statistics and Computing*, 4(4):259–269, 1994.

[84] Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of time series analysis*, 15(2):183–202, 1994.

[85] John Galbraith, Victoria Zinde-Walsh, et al. Autoregression-based estimators for arfima models. Technical report, CIRANO, 2001.

[86] John W Galbraith and Victoria Zinde-Walsh. On some simple, autoregression-based estimation and identification techniques for arma models. *Biometrika*, 84(3):685–696, 1997.

[87] George L Gerstein and Nelson Y-S Kiang. An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophysical Journal*, 1(1):15, 1960.

[88] George L Gerstein and Benoit Mandelbrot. Random walk models for the spike activity of a single neuron. *Biophysical journal*, 4(1 Pt 1):41, 1964.

[89] Alexander Goldenshluger and Assaf Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *Annals of statistics*, pages 417–444, 2001.

[90] Ronald L Graham and Neil Sloane. Lower bounds for constant weight codes. *Information Theory, IEEE Transactions on*, 26(1):37–43, 1980.

[91] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[92] Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*, volume 321. Univ of California Press, 1958.

[93] Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*, volume 321. Univ of California Press, 2001.

[94] Rémi Gribonval, Volkan Cevher, and Mike E Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012.

[95] Mats GL Gustafsson. Extended resolution fluorescence microscopy. *Current opinion in structural biology*, 9(5):627–628, 1999.

[96] Fang Han and Han Liu. Transition matrix estimation in high dimensional time series. In *ICML (2)*, pages 172–180, 2013.

[97] Jiang Hao and Thomas G Oertner. Depolarization gates spine calcium transients and spike-timing-dependent potentiation. *Current opinion in neurobiology*, 22(3):509–515, 2012.

[98] Zachary T Harmany, Roummel F Marcia, and Rebecca M Willett. Spiral out of convexity: Sparsity-regularized algorithms for photon-limited imaging. In *IS&T/SPIE Electronic Imaging*, pages 75330R–75330R. International Society for Optics and Photonics, 2010.

[99] Jarvis Haupt, Waheed U Bajwa, Gil Raz, and Robert Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.

[100] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[101] Simon S Haykin. *Adaptive filter theory.* Pearson Education India, 2008.

[102] Rainer Heintzmann and Christoph Cremer. Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating. In *Proc. SPIE*, volume 3568, page 15, 1999.

[103] Markus Hürzeler and Hans R Künsch. Monte carlo approximations for general state-space models. *Journal of Computational and graphical Statistics*, 7(2):175–193, 1998.

[104] Eero Huupponen, Germán Gómez-Herrero, Antti Saastamoinen, Alpo Värri, Joel Hasan, and Sari-Leena Himanen. Development and comparison of four sleep spindle detection methods. *Artificial intelligence in medicine*, 40(3):157–170, 2007.

[105] Ching-Kang Ing, Ching-Zong Wei, et al. Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, 33(5):2423–2474, 2005.

[106] Harold Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*, volume 186, pages 453–461. The Royal Society, 1946.

[107] Soren Johansen. Likelihood-based inference in cointegrated vector autoregressive models. *OUP Catalogue*, 1995.

[108] Edward G Jones, Mircea Steriade, and David McCormick. *The thalamus.* Plenum Press New York, 1985.

[109] Hong Jung and Jong Chul Ye. Motion estimated and compensated compressed sensing dynamic magnetic resonance imaging: What we can learn from video compression techniques. *International Journal of Imaging Systems and Technology*, 20(2):81–98, 2010.

[110] Abbas Kazemipour, Behtash Babadi, and Min Wu. Sparse estimation of self-exciting point processes with application to LGN neural modeling. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 478–482. IEEE, 2014.

[111] Abbas Kazemipour, Ji Liu, Krystyna Solarana, Daniel Nagode, Patrick Kanold, Min Wu, and Behtash Babadi. Fast and stable signal deconvolution via compressible state-space models. *IEEE Transactions on Biomedical Engineering*, 2017.

[112] Abbas Kazemipour, Sina Miran, Piya Pal, Behtash Babadi, and Min Wu. Sampling requirements of stable autoregressive estimation. *Preprint*.

[113] Abbas Kazemipour, Min Wu, and Behtash Babadi. Robust estimation of self-exciting generalized linear models with application to neuronal modeling. *IEEE Transactions on Signal Processing*, 2017.

[114] Ki Hean Kim, Christof Buehler, Karsten Bahlmann, Timothy Ragan, Wei-Chung A Lee, Elly Nedivi, Erica L Heffer, Sergio Fantini, and Peter TC So. Multifocal multiphoton microscopy based on multianode photomultiplier tubes. *Optics express*, 15(18):11658–11678, 2007.

[115] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput Biol*, 7(3):e1001110, 2011.

[116] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[117] Genshiro Kitagawa. A self-organizing state-space model. *Journal of the American Statistical Association*, pages 1203–1215, 1998.

[118] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

[119] Leonhard Knorr-Held. Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26(1):129–144, 1999.

[120] Mladen Kolar, Le Song, Amr Ahmed, and Eric P Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, pages 94–123, 2010.

[121] Leonid Aryeh Kontorovich, Kavita Ramanan, et al. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

[122] P Philippe Laissue, Rana A Alghamdi, Pavel Tomancak, Emmanuel G Reynaud, and Hari Shroff. Assessing phototoxicity in live fluorescence imaging. *Nature Methods*, 14(7):657–661, 2017.

[123] Kenneth Lange and Janet S Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198, 1993.

[124] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

[125] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[126] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[127] Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 150. Wiley New York et al, 1986.

[128] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[129] Lauren C Liets, Bruno A Olshausen, Guo-Yong Wang, and Leo M Chalupa. Spontaneous activity of morphologically identified ganglion cells in the developing ferret retina. *The Journal of neuroscience*, 23(19):7343–7350, 2003.

[130] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[131] Rongwen Lu, Wenzhi Sun, Yajie Liang, Aaron Kerlin, Jens Bierfeld, Johannes Seelig, Daniel E Wilson, Benjamin Scholl, Boaz Mohar, Masashi Tanimoto, et al. Video-rate volumetric functional imaging of the brain at synaptic resolution. *BioRxiv*, page 058495, 2016.

[132] Steven J Luck, Leonardo Chelazzi, Steven A Hillyard, and Robert Desimone. Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of neurophysiology*, 77(1):24–42, 1997.

[133] Leon B Lucy. An iterative technique for the rectification of observed distributions. *The astronomical journal*, 79:745, 1974.

[134] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic resonance in medicine*, 58(6):1182–1195, 2007.

[135] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error correcting codes*, volume 16. Elsevier, 1977.

[136] Michael E Mann and Jeffrey Park. Oscillatory spatiotemporal signal detection in climate studies: A multiple-taper spectral domain approach. *Advances in geophysics*, 41:1–132, 1999.

[137] S Lawrence Marple Jr. Digital spectral analysis with applications. *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1987, 512 p.*, 1, 1987.

[138] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[139] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[140] Xiangying Meng, Joseph PY Kao, Hey-Kyoung Lee, and Patrick O Kanold. Visual deprivation causes refinement of intracortical circuits in the auditory cortex. *Cell reports*, 12(6):955–964, 2015.

[141] Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

[142] Delaram Motamedvaziri, Mohammad H Rohban, and Venkatesh Saligrama. Sparse signal recovery under Poisson statistics. *arXiv preprint arXiv:1307.4666*, 2013.

[143] KM Naga Srinivas Nadella, Hana Roš, Chiara Baragli, Victoria A Griffiths, George Konstantinou, Theo Koimtzis, Geoffrey J Evans, Paul A Kirkby, and R Angus Silver. Random-access scanning microscopy for 3d imaging in awake behaving animals. *Nature methods*, 2016.

[144] Yuval Nardi and Alessandro Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

[145] Deanna Needell and Joel A Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[146] Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013.

[147] Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[148] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[149] Mark AA Neil, Rimas Juškaitis, and Tony Wilson. Method of obtaining optical sectioning by using structured light in a conventional microscope. *Optics letters*, 22(24):1905–1907, 1997.

[150] Antoine Nonclercq, Charline Urbain, Denis Verheulpen, Christine Decaestecker, Patrick Van Bogaert, and Philippe Peigneux. Sleep spindle detection through

amplitude–frequency normal modelling. *Journal of neuroscience methods*, 214(2):192–203, 2013.

[151] Paul L Nunez and Brian A Cutillo. *Neocortical dynamics and human EEG rhythms.* Oxford University Press, USA, 1995.

[152] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, pages 1–47, 2011.

[153] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[154] Daniel P Palomar and Yonina C Eldar. *Convex optimization in signal processing and communications.* Cambridge university press, 2010.

[155] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Comp. in Neural Systems*, 15(4):243–262, 2004.

[156] Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, 165:493–507, 2007.

[157] Thomas W Parks and C Sidney Burrus. *Digital filter design.* Wiley-Interscience, 1987.

[158] Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.

[159] Nicolas C Pégard, Hsiou-Yuan Liu, Nick Antipa, Maximillian Gerlock, Hillel Adesnik, and Laura Waller. Compressive light-field microscopy for 3d neural activity recording. *Optica*, 3(5):517–524, 2016.

[160] Donald B Percival and Andrew T Walden. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[161] Darcy S Peterka, Hiroto Takahashi, and Rafael Yuste. Imaging voltage in neurons. *Neuron*, 69(1):9–21, 2011.

[162] James W Phillips, Richard M Leahy, and John C Mosher. Meg-based imaging of focal neuronal current sources. *Medical Imaging, IEEE Transactions on*, 16(3):338–348, 1997.

[163] Jonathan W Pillow, Yashar Ahmadian, and Liam Paninski. Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. *Neural Comp.*, 23(1):1–45, 2011.

[164] Didier Pinault, Nathalie Leresche, Stephane Charpier, Jean-Michel Deniau, Christian Marescaux, Marguerite Vergnes, and Vincenzo Crunelli. Intracellular recordings in thalamic neurones during spontaneous spike and wave discharges in rats with absence epilepsy. *The Journal of Physiology*, 509(2):449–456, 1998.

[165] Eftychios A Pnevmatikakis and Liam Paninski. Sparse nonnegative deconvolution for compressive calcium imaging: algorithms and phase transitions. In *Advances in Neural Information Processing Systems*, pages 1250–1258, 2013.

[166] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016.

[167] Kaspar Podgorski and Gayathri Ranganathan. Brain heating induced by near-infrared lasers during multiphoton microscopy. *Journal of neurophysiology*, 116(3):1012–1023, 2016.

[168] Clarice Poon. On the role of total variation in compressed sensing. *SIAM Journal on Imaging Sciences*, 8(1):682–720, 2015.

[169] Donald Stephen Poskitt. Autoregressive approximation in nonstandard situations: the fractionally integrated and non-invertible cases. *Annals of the Institute of Statistical Mathematics*, 59(4):697–725, 2007.

[170] Stephan Preibisch, Fernando Amat, Evangelia Stamataki, Mihail Sarov, Robert H Singer, Eugene Myers, and Pavel Tomancak. Efficient bayesian-based multiview deconvolution. *nature methods*, 11(6):645–648, 2014.

[171] Robert Prevedel, Aart J Verhoef, Alejandro J Pernía-Andrade, Siegfried Weisenburger, Ben S Huang, Tobias Nöbauer, Alma Fernández, Jeroen E Delcour, Peyman Golshani, Andrius Baltuska, et al. Fast volumetric calcium imaging across multiple cortical layers using sculpted light. *Nature methods*, 13(12):1021, 2016.

[172] Maxim Raginsky, Rebecca M Willett, Zachary T Harmany, and Roummel F Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Transactions on Signal Processing*, 58(8):3990–4002, 2010.

[173] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[174] Herbert E Rauch, CT Striebel, and F Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.

[175] Holger Rauhut, Justin Romberg, and Joel A Tropp. Restricted isometries for partial random circulant matrices. *Applied and Computational Harmonic Analysis*, 32(2):242–254, 2012.

[176] Patricia Reynaud-Bouret, Emmanuel Roy, et al. Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin*, 13(5):883–896, 2007.

[177] William Hadley Richardson. Bayesian-based iterative method of image restoration. *JOSA*, 62(1):55–59, 1972.

[178] Peter M Robinson. *Time series with long memory.* Oxford University Press, 2003.

[179] Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999.

[180] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[181] R Rudzkis. Large deviations for estimates of spectrum of stationary series. *Lithuanian Mathematical Journal*, 18(2):214–226, 1978.

[182] Hailin Sang and Yan Sun. Simultaneous sparse model selection and coefficient estimation for heavy-tailed autoregressive processes. *Statistics*, 49(1):187–208, 2015.

[183] P Schimicek, J Zeitlhofer, P Anderer, and B Saletu. Automatic sleep-spindle detection procedure: aspects of reliability and validity. *Clinical EEG and neuroscience*, 25(1):26–29, 1994.

[184] Benjamin Scholl, Andrew YY Tan, Joseph Corey, and Nicholas J Priebe. Emergence of orientation selectivity in the mammalian visual pathway. *The Journal of Neuroscience*, 33(26):10616–10624, 2013.

[185] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[186] Neil Shephard and Michael K Pitt. Likelihood analysis of non-gaussian measurement time series. *Biometrika*, 84(3):653–667, 1997.

[187] Lawrence A Shepp and Yehuda Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE transactions on medical imaging*, 1(2):113–122, 1982.

[188] Ritei Shibata. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, pages 147–164, 1980.

[189] Robert H Shumway and David S Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

[190] Diana Smetters, Ania Majewska, and Rafael Yuste. Detecting action potentials in neuronal populations with calcium imaging. *Methods*, 18(2):215–221, 1999.

[191] Andrew Smith and Emery N Brown. Estimating a state-space model from point process observations. *Neural Comp.*, 15(5):965–991, 2003.

[192] Aleksander Sobczyk, Volker Scheuss, and Karel Svoboda. Nmda receptor subunit-dependent [ca2+] signaling in individual hippocampal dendritic spines. *Journal of Neuroscience*, 25(26):6037–6046, 2005.

[193] Nicholas James Sofroniew, Daniel Flickinger, Jonathon King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife*, 5:e14472, 2016.

[194] SOMPONG Sombati and ROBERT J Delorenzo. Recurrent spontaneous seizure activity in hippocampal neuronal networks in culture. *Journal of neurophysiology*, 73(4):1706–1711, 1995.

[195] Christoph Sommer, Christoph Straehle, Ullrich Koethe, and Fred A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 230–233. IEEE, 2011.

[196] Petre Stoica and Randolph L Moses. *Introduction to spectral analysis*, volume 1. Prentice hall Upper Saddle River, 1997.

[197] Christoph Stosiek, Olga Garaschuk, Knut Holthoff, and Arthur Konnerth. In vivo two-photon calcium imaging of neuronal networks. *Proceedings of the National Academy of Sciences*, 100(12):7319–7324, 2003.

[198] SJ Strickler and Robert A Berg. Relationship between absorption intensity and fluorescence lifetime of molecules. *The Journal of chemical physics*, 37(4):814–822, 1962.

[199] George Striker, Vinod Subramaniam, Claus AM Seidel, and Andreas Volkmer. Photochromicity and fluorescence lifetimes of green fluorescent protein. *The Journal of Physical Chemistry B*, 103(40):8612–8617, 1999.

[200] David Sussillo, Rafal Jozefowicz, LF Abbott, and Chethan Pandarinath. Lfads-latent factor analysis via dynamical systems. *arXiv preprint arXiv:1608.06315*, 2016.

[201] Gergely Szalay, Linda Judák, Gergely Katona, Katalin Ócsai, Gábor Juhász, Máté Veress, Zoltán Szadai, András Fehér, Tamás Tompa, Balázs Chiovini, et al. Fast 3d imaging of spine, dendritic, and neuronal assemblies in behaving animals. *Neuron*, 92(4):723–738, 2016.

[202] Gabrielle Thériault, Martin Cottet, Annie Castonguay, Nathalie McCarthy, and Yves De Koninck. Extended two-photon microscopy in live samples with bessel beams: steadier focus, faster volume scans, and simpler stereoscopic imaging. *Frontiers in cellular neuroscience*, 8, 2014.

[203] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[204] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[205] Nicolas X Tritsch, Eunyoung Yi, Jonathan E Gale, Elisabeth Glowatzki, and Dwight E Bergles. The origin of spontaneous activity in the developing auditory system. *Nature*, 450(7166):50–55, 2007.

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

[207] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[208] M Tsodyks, Tal Kenet, Amiram Grinvald, and A Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science*, 286(5446):1943–1946, 1999.

[209] Henry C Tuckwell. *Introduction to theoretical neurobiology: volume 2, nonlinear and stochastic theories*, volume 8. Cambridge University Press, 2005.

[210] Gaetano Valenza, Luca Citi, Enzo Pasquale Scilingo, and Riccardo Barbieri. Point-process nonlinear models with Laguerre and Volterra expansions: Instantaneous assessment of heartbeat dynamics. *IEEE Transactions on Signal Processing*, 61(11):2914–2926, 2013.

[211] Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[212] Sara A van de Geer. *Empirical Processes in M-estimation.* Cambridge university press, 2000.

[213] Sara A van de Geer. On Hoeffding's inequality for dependent random variables. In Herold Dehling and Walter Philipp, editors, *Empirical Process Techniques for Dependent Data*. Springer, 2001.

[214] Namrata Vaswani. Kalman filtered compressed sensing. In *2008 15th IEEE International Conference on Image Processing*, pages 893–896. IEEE, 2008.

[215] Namrata Vaswani. Ls-cs-residual (ls-cs): compressive sensing on least squares residual. *IEEE Transactions on Signal Processing*, 58(8):4108–4120, 2010.

[216] Errikos M Ventouras, Efstratia A Monoyiou, Periklis Y Ktonas, Thomas Paparrigopoulos, Dimitris G Dikeos, Nikos K Uzunoglu, and Constantin R Soldatos. Sleep spindle detection using artificial neural networks trained with filtered time-domain eeg: a feasibility study. *Computer methods and programs in biomedicine*, 78(3):191–207, 2005.

[217] David Vere-Jones. Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–62, 1970.

[218] Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of neurophysiology*, 104(6):3691–3704, 2010.

[219] Joshua T Vogelstein, Brendon O Watson, Adam M Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential monte carlo methods. *Biophysical journal*, 97(2):636–655, 2009.

[220] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[221] Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007.

[222] Richard L White. Image restoration using the damped richardson-lucy method. In *1994 Symposium on Astronomical Telescopes & Instrumentation for the 21st Century*, pages 1342–1348. International Society for Optics and Photonics, 1994.

[223] Rebecca M Willett, Zachary T Harmany, and Roummel F Marcia. Poisson image reconstruction with total variation regularization. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4177–4180. IEEE, 2010.

[224] Kam Chung Wong, Ambuj Tewari, and Zifan Li. Regularized estimation in high dimensional time series under mixing conditions. *arXiv preprint arXiv:1602.04265*, 2016.

[225] Rachel OL Wong, Markus Meister, and Carla J Shatz. Transient period of correlated bursting activity during development of the mammalian retina. *Neuron*, 11(5):923–938, 1993.

[226] Chong Wu, Can Yang, Hongyu Zhao, and Ji Zhu. On the convergence of the em algorithm: From the statistical perspective. *arXiv preprint arXiv:1611.00519*, 2016.

[227] Wei-Biao Wu and Ying Nian Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.

[228] Hong-ping Xu, Moran Furman, Yann S Mineur, Hui Chen, Sarah L King, David Zenisek, Z Jimmy Zhou, Daniel A Butts, Ning Tian, Marina R Picciotto, et al. An instructive role for patterned spontaneous retinal activity in mouse visual map development. *Neuron*, 70(6):1115–1127, 2011.

[229] Ming Yan, A Bui, Jason Cong, and Luminita A Vese. General convergent expectation maximization (em)-type algorithms for image reconstruction. *Inverse Problems and Imaging*, 7(3):1007–1029, 2013.

[230] Weijian Yang, Jae-eun Kang Miller, Luis Carrillo-Reid, Eftychios Pnevmatikakis, Liam Paninski, Rafael Yuste, and Darcy S Peterka. Simultaneous multi-plane imaging of neural circuits. *Neuron*, 89(2):269–284, 2016.

[231] Weijian Yang and Rafael Yuste. In vivo imaging of neural activity. *Nature Methods*, 14(4):349–359, 2017.

[232] Zhihua Yang, Lihua Yang, and Dongxu Qi. Detection of spindles in sleep eegs using a novel algorithm based on the hilbert-huang transform. In *Wavelet Analysis and Applications*, pages 543–559. Springer, 2006.

[233] Ryohei Yasuda, Esther A Nimchinsky, Volker Scheuss, Thomas A Pologruto, Thomas G Oertner, Bernardo L Sabatini, and Karel Svoboda. Imaging calcium concentration dynamics in small neuronal compartments. *Sci STKE*, 2004(219):pl5, 2004.

[234] Jinchun Zhan and Namrata Vaswani. Time invariant error bounds for modified-cs-based sparse signal sequence recovery. *IEEE Transactions on Information Theory*, 61(3):1389–1409, 2015.

[235] Tong Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.

[236] Zhilin Zhang and Bhaskar D Rao. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *Selected Topics in Signal Processing, IEEE Journal of*, 5(5):912–926, 2011.

[237] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006.

[238] Lingjiong Zhu. Nonlinear Hawkes processes. *arXiv preprint arXiv:1304.7531*, 2013.

[239] Justin Ziniel and Philip Schniter. Dynamic compressive sensing of time-varying signals via approximate message passing. *IEEE transactions on signal processing*, 61(21):5270–5284, 2013.