Statistical Modeling of Long Memory and Uncontrolled Effects in Neural Recordings

Alexander Greaves-Tunnell

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Zaid Harchaoui, Chair

Don Percival

Jon Wakefield

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Statistical Modeling of Long Memory and Uncontrolled Effects
in Neural Recordings

Alexander Greaves-Tunnell

Chair of the Supervisory Committee:
Associate Professor Zaid Harchaoui
Department of Statistics

Scientific analyses of time series data are often formalized as statistical investigations targeting one or more aspects of a complex underlying dependence structure. In the multivariate time series setting, there are three main aspects of interest: dependence over time, between components of the multivariate observations, and across repeated trials of the experimental protocol. Classical methods for these data may not be equipped to account for issues such as long-range dependence or unexpected variation across experimental settings. On the other hand, it can be difficult to evaluate whether more recent methods, such as those that make use of deep neural networks, have made quantitatively verifiable progress towards alleviating these issues. There is thus an opportunity to develop tools that extend principled statistical perspectives to meet the demand of current practices and problems in applied time series analysis.

Motivated by these considerations, this dissertation develops methodology for the identification, estimation, and prediction of scientifically relevant features in the dependence structure of multivariate time series data. While these contributions apply to a wide range of data-analytical settings, corresponding to the broad prevalence of time series data across the sciences, they are motivated in particular by the challenges raised in the analysis of brain activity data. Neuroscientists measure the locally aggregated activity of cortical neurons as electromagnetic waveforms, recording from multiple locations on the brain surface and across

repeated recording trials for various subjects or conditions. The resulting data raise challenging, scientifically important questions that can be phrased in terms of the three aspects of time series dependence enumerated above.

We first address the topic of dependence over time through the lens of long-range dependent multivariate time series. We develop a statistical criterion for long memory in deep recurrent neural networks, thus offering a principled alternative to the heuristic evaluations based on synthetic data. We show negative results that suggest even deep recurrent neural network models explicitly designed to capture long-range dependencies fail to do so in a standard language modeling setting. This motivates the development of a model for multivariate long memory time series data, which we define in the frequency domain and apply to the analysis of brain activity in different states of consciousness.

Second, we develop a framework to model changes in the dependence structure, or functional connectivity, in recordings obtained from a repeated-stimulation experiment in the rhesus macaque cortex. The analysis flexibly captures nonlinear effects and incorporates information about the connectivity between brain regions prior to stimulation. It is further equipped to address questions of stability, informativeness, and similarity among the learned features. The method improves the prediction accuracy of stimulation-induced connectivity change and provides new insights on the factors mediating this response.

Finally, we continue our analysis of the brain-stimulation data to reveal previously unreported variation both between subjects and across experimental trials. We show that extensions to a simple regression model, either accounting for a more complex variance structure or estimating unmodeled confounders, can successfully mitigate the dramatic loss in predictive accuracy that results from applying the model to predict the results of previously unobserved experimental trials. Together, the contributions of this dissertation develop the capacity to detect, estimate, and predict scientifically meaningful aspects of the dependence structure in multivariate time series data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I gratefully acknowledge, first and foremost, my thesis advisor Zaid Harchaoui. It is humbling to reflect now, at the end of an academic journey that began in earnest in my second or third year of undergraduate study, on how much I had yet to learn at the start of my graduate career, and indeed how much I have yet to learn still. Zaid has been the single most influential figure in my intellectual life during this period. I am grateful for the challenge, for the respect of unflinchingly high standards, and for the curation of an intellectual passion and culture that I hope to maintain long after our work together has ended.

I have had the privilege of mentorship and support from many others along the way. Steven Miller at Williams College has been a warm and earnest mentor for over a decade. Christina Leslie at Memorial Sloan Kettering Cancer Center and Wing Hung Wong at Stanford showed me the possibilities of an academic path combining mathematical training with broad curiosity about the natural world. Máté Lengyel and Rich Turner at Cambridge CBL patiently supervised my first steps in my own research career. Emily Fox gave me the opportunity to continue and grow at the University of Washington. And most recently, it has been a privilege to collaborate with Julien Bloch, Azadeh Yazdan-Shahmorad, Ali Shojaie, and Eric Shea-Brown.

One of the great personal rewards of this experience has been the friendship of other students and labmates. My labmates, including Alex Tank, Corinne Jones, John Thickstun, Krishna Pillutla, and Lang Liu, are the implicit audience and standard for how I present my work. Meanwhile I have been lucky enough to meet many others, including Sean and Jenny Jewell, Vincent Roulet, Ema Perkovic, Samson Koelle, Sheridan Grant, and Max Schneider,

whose influence in my life has extended – and I hope will continue to extend – far beyond the limits of our time together in the department.

Finally, I owe a debt of gratitude beyond repayment to those who have supported me not as a student, but as a person on a long and uncertain journey far from home. To the Picardo family, whose friendship extends back to my first day at Williams; you have made Seattle feel like home. And most of all to my family, who through the constant hard work and occasional hard times have been the undisturbed foundation of culture, values, and love upon which my life, and therefore this work, is built. Thank you.

DEDICATION

For my parents,

Peter and Louise Greaves-Tunnell.

Chapter 1

INTRODUCTION

A great deal of methodological development in time series analysis is driven by the variety and complexity of motivating applications in the sciences. Such is the ubiquity of time series data gathered from natural and social phenomena that these applications span the history and breadth of science itself, from measurements of the Nile river dating from 622 AD (Toussoun, 1925) to micro-voltage fluctuations capturing neural activity thousands of times per second in a live brain (Yazdan-Shahmorad et al., 2018). Representation of sequentially ordered items such as written words or musical notes via dense vectors (Mikolov et al., 2013; Logan, 2000) renders such data amenable to time series analysis and thus brings further applications and their associated challenges into view. The steady advance of computational hardware capabilities for machine learning and ongoing refinement of measurement technologies in the sciences together give rise to datasets of unprecedented scale, coverage, and resolution.

Meanwhile, the methods developed to analyze these data are as diverse as the applications themselves. Deep neural networks dominate the state of the art for language and music modeling (Brown et al., 2020; Thickstun et al., 2018), yet it remains difficult to assess whether they have captured the statistical properties of interest in the underlying data (Takahashi and Tanaka-Ishii, 2017). On the other hand, analyses of neural recording data continue to use simple models, typically linear in a small number of theory-motivated features (Keller et al., 2018; Betzel et al., 2019), that may not make full use of the information available. Each in their own way, these unresolved questions point the way towards new opportunities for statistical investigation, and thus towards opportunities to more precisely define the broad scientific goals that motivate them.

From a statistical standpoint, the scientific objectives of a data analysis can often be cast in terms of the dependence structure of an underlying model. In the multivariate time series setting, such dependence may extend over time, between observed components, or across experimental replicates. This dissertation addresses instances of each of these cases as they arise in the analysis of modern sequence data, including natural language, music, and neural recordings. The latter application, focused in particular on multivariate waveforms obtained from a recording technique known as electrocorticography, features as a recurring source of scientific and statistical motivation for this research.

In this chapter, we introduce three concepts central to the work presented in this dissertation. First, in Section 1.1, we introduce the notion of long-range dependent time series from both the time and frequency domain perspectives, including their extension to the multivariate setting. Continuing the emphasis on multivariate time series, in Section 1.2 we discuss definitions and estimation of interaction, or connectivity, between the components. Finally, we give a brief introduction to electrocorticography in Section 1.3. We conclude with an overview of the main research contributions and an outline of the remaining chapters.

1.1 Long-range Dependence

For over a century, scientists have recorded data across a surprising range of disciplines that not only exhibit serial correlation, but whose correlations fail to decay at the rate implied by existing models for such phenomena. In astronomy, Newcomb (1895) observed that errors among consecutive observations dramatically inflate the standard error of the mean versus its expected $\sigma/\sqrt{n}$ behavior, a finding confirmed by subsequent statistical analyses (Pearson, 1902; Student, 1927) that rule out explanations via simple trends. Wheat yield data collected by Smith (1938) demonstrates an empirical law whose spatial autocorrelations violate assumptions of independence or even summability. This unusual behavior motivated subsequent development of space-time models with hyperbolically decaying autocorrelations (Whittle, 1956), and similar properties were explored early on in the development of the field of geostatistics (Matheron, 1973). Famously, the fluctuations of the Nile river were observed

by Hurst (1951) to exhibit nontrivial dependence over extremely long timescales and to make long excursions from the empirical mean, ultimately leading to his recommendation that the height of the Aswan High Dam far exceed the original plan based on conventional forecasts. In economics, Granger (1966) showed that many important variables have autocorrelations that admit a power-law representation in the frequency domain, subsequently motivating statistical models for such processes (Granger and Joyeux, 1980; Hosking, 1981) that remain prominent in econometric data analysis.

The statistical phenomenon underlying these discoveries is *long-range dependence*, a topic notable in time series analysis for its impact on theoretical foundations, statistical methodology, and in practical applications spanning a diverse range of scientific fields. Across what is now an expansive literature, long-range dependence (or *long memory*) is associated with several non-equivalent definitions. Our emphasis in this dissertation will be on the most common approach, which defines long memory in terms of the second-order dependence structure of a weakly stationary process. The definition can be specified in the time domain, in which case it concerns the autocovariance sequence, or in the frequency domain, in terms of the behavior of the spectral density function.

Long memory in the time and frequency domains. Weak stationarity of the scalar process $X_t \in \mathbb{R}, t \in \mathbb{Z}$ implies that the autocovariance $\mathrm{Cov}(X_t, X_{t+k})$ depends on the indices t and $t+k$ only through the lag k. We can thus denote the autocovariance sequence $\gamma_X(k) \triangleq \mathrm{Cov}(X_t, X_{t+k})$; throughout the text, we will drop the subscript when there is no confusion over the corresponding process. Long memory in the time domain is defined in terms of the slow decay of $\gamma(k)$.

Definition 1.1.1. Long memory in the time domain *(Pipiras and Taqqu, 2017). The second-order stationary process* $X_t \in \mathbb{R}, t \in \mathbb{Z}$ *has* long memory *if its autocovariance function* $\gamma(k)$ *satisfies*

$$\gamma(k) = L_\gamma(k) k^{2d-1}, \quad k = 0, 1, 2, \ldots,$$

where $d \in (0, \frac{1}{2})$ *and* $L_\gamma(k)$ *is a slowly varying function at infinity.*

The emphasis in this definition is on the k^{2d-1} term in the expression for the autocovariance, which guarantees that the sequence decays slowly (i.e. hyperbolically) as a function of the lag. The other term, $L_\gamma(k)$, functions in a merely technical way, namely to allow for a broader range of autocovariance functions while ensuring that $\gamma(k)$ still asymptotically behaves like k^{2d-1} as k tends to infinity. The slow decay of $\gamma(k)$ implies that relatively substantial correlations persist across long periods of time, justifying in a colloquial sense the use of the phrase "long memory."

In many scientific areas, it is preferable to work in the frequency domain. For example, in neuroscience prior knowledge of neuronal cell biology indicates that electrophysiological recordings will only contain scientifically relevant information within a certain frequency window (Cohen, 2014). The main quantity of interest in such analyses is the spectral density function $f(\lambda) \in \mathbb{C}$, obtained as the discrete Fourier transform of the autocovariance

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-ik\lambda} \gamma(k), \lambda \in (0, \pi].$$

The spectral density function represents the autocovariance sequence in a complex sinusoid basis; the magnitude of $f(\lambda)$ reflects the power of the autocovariance, viewed itself as a discrete-time signal, at the frequency λ. Long memory in the frequency domain is characterized by divergent behavior of the spectral density function near the origin.

Definition 1.1.2. Long memory in the frequency domain *(Pipiras and Taqqu, 2017). Let $X_t \in \mathbb{R}, t \in \mathbb{Z}$ be a second-order stationary process with spectral density function $f(\lambda)$. Then X_t has long memory if*

$$f(\lambda) = L_f(\lambda)\lambda^{-2d}, \quad \lambda \in (0, \pi],$$

where $d \in (0, \frac{1}{2})$ and $L_f(1/\lambda)$ is a slowly varying function at infinity.

As in the time domain definition, the second term in the expression above is most important, as it controls the asymptotic behavior of $f(\lambda)$. In this case, the relevant asymptotic regime is at low frequencies, as λ approaches zero. Near the origin, the spectral density be-

haves like a power law with exponent $-2d$; equivalently, the graph of the function in log-log coordinates is approximately linear with slope $-2d$ (see Figure 1.1, right panel).

It is important to note that Definitions 1.1.1 and 1.1.2 are not equivalent in general; Samorodnitsky (2016) provides counterexamples in both directions. Mutual implication requires the further condition that the slowly varying functions $L_\gamma(k)$ and $L_f(\lambda)$ are also quasi-monotone (Pipiras and Taqqu, 2017, Propositions 2.2.14, 2.2.16). However, some stochastic processes, such as those obtained by fractional differencing of white noise (Granger and Joyeux, 1980; Hosking, 1981), are known to satisfy both definitions.

Contrasting long vs. short memory time series. Both the time and frequency domain definitions of long memory imply that the absolute partial sums of the autocovariance function diverge; informally, we may write $\sum_{k=-\infty}^{\infty} |\gamma(k)| = \infty$. This yields a simple criterion by which we can define the contrasting case: a *short memory* process is one for which the autocovariance sequence is absolutely summable. Absolute summability of the autocovariance implies that the spectral density function is bounded and continuous at the origin, hence

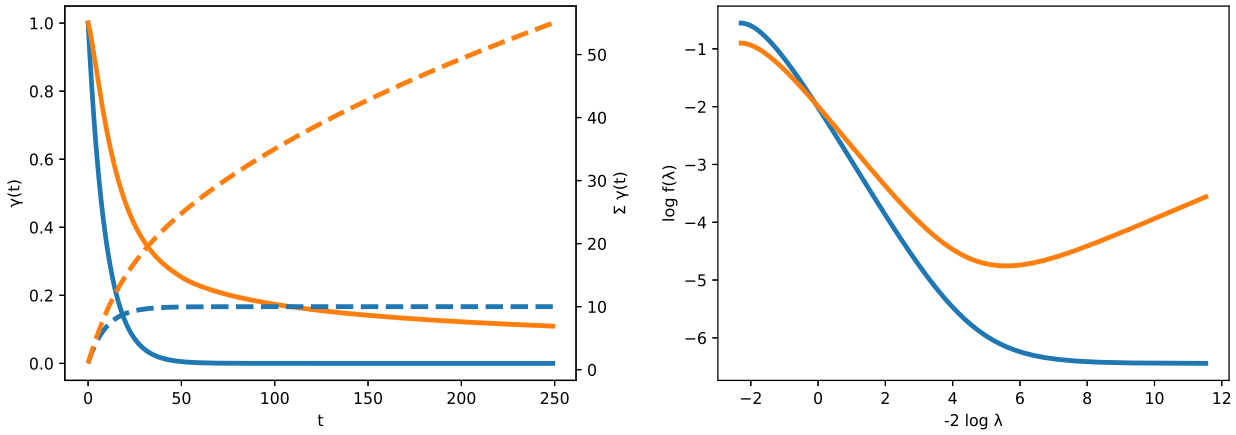

Figure 1.1: Time and frequency domain views of a short memory AR(1) process (blue, $d = 0$) and its long memory counterpart obtained by fractional differencing (orange, $d = 0.25$). *Left:* Autocorrelation sequences (solid lines) of the two processes, along with their partial sums (dotted lines). *Right:* Log-log plot of the spectral density function versus frequency.

short memory implies fundamentally different asymptotic behavior in both the time and frequency domains. Many well-known and widely used statistical time series models have short memory, including autoregressive moving average (ARMA) and Markov models (Brockwell and Davis, 1991). For an illustration of the contrast between long and short memory in both the time and frequency domains, see Figure 1.1.

Long memory in multivariate time series. It is common to obtain data as a multivariate (or multiple) time series, wherein the measurement at each time index is vectorial rather than scalar. In this setting, the weakly stationary process $X_t \in \mathbb{R}^p$ is characterized by a matrix-valued autocovariance sequence $\Gamma(k) \in \mathbb{R}^{p\times p}$ and matrix-valued spectral density function $f(\lambda) \in \mathbb{C}^{p\times p}$. While multivariate notions of long memory were first considered some time ago (Robinson, 1994; Lobato, 1997), the general case of bivariate long memory and its subsequent multivariate extension are somewhat more recent (Robinson, 2008; Kechagias and Pipiras, 2015). We will use the frequency-domain definition from the latter.

Definition 1.1.3. Multivariate long memory *(Kechagias and Pipiras, 2015). A weakly stationary process $X_t \in \mathbb{R}^p$ has long memory if for every $j, k \in \{1, ..., p\}$ the $(j, k)^{th}$ component $f_{jk}(\lambda)$ of its spectral density can be written as*

$$f_{jk}(\lambda) = G_{jk}(\lambda)|\lambda|^{-(d_j+d_k)}$$

such that

$$G_{jk}(\lambda) \to g_{jk}e^{i\phi_{jk}} \quad as\ |\lambda| \to 0 \tag{1.1}$$

for some $d \in (0, 1/2)^p$.

The multivariate LRD spectrum consists of two components: a long memory vector d controlling the behavior of the spectral components near zero frequency, and a complex matrix-valued function $G(\lambda)$ that specifies the behavior at higher frequencies. Basic conditions on $G(\lambda)$ are enforced to ensure the validity of $f(\lambda)$ as a spectral density function.

1.2 Interactions in Multivariate Time Series

The memory of a multivariate time series characterizes its dependence structure across time, but it remains to address the question of dependence between the component series. Characterization of this dependence structure provides information as to how and to what degree these components interact, in the sense of influencing the behavior of other series. In neuroscience applications involving multiple time series of neural activity recorded from different areas of the brain, statistical interaction between the component series is known as *functional connectivity*. Functional connectivity is distinct from the literal, anatomical connection between brain regions, yet it reproduces similar network structures and is known to be informative for network-level neural behaviors such as the response to stimulation (Wang et al., 2013; Younce et al., 2021). Statistical interactions between multiple time series thus play a central role in modern analyses of neuroscience data.

Directed vs. undirected measures of interaction. Interactions between time series can refer either to a directed relationship, in which component i influences component j but not necessarily vice versa, or an undirected relationship, measuring the overall association between the series. The simplest undirected measure of interaction in a stationary time series is given by the autocovariance $\Gamma(k)$. The (i, j) entry of $\Gamma(k)$ represents the linear dependence between the two series at lag k; it is zero at all lags for uncorrelated components. In the frequency domain, the Fourier transform of the cross-covariance $\Gamma_{ij}(k)$ yields the cross-spectrum $S_{ij}(\lambda)$, which decomposes the cross-covariance across frequencies. The cross-spectrum is complex in general as the off-diagonal terms of the autocovariance are not necessarily an even function of the lag. Therefore, it is typical to work with the normalized magnitude of the cross spectrum,

$$C_{ij}(\lambda) = \frac{|S_{ij}(\lambda)|^2}{S_{ii}(\lambda)S_{jj}(\lambda)},$$

known as the coherence. We have $|C_{ij}(\lambda)| \leq 1$ by the Cauchy-Schwarz inequality.

The most popular directed measure of interaction in multiple time series is Granger causality, which despite its name does not denote a causal relationship (Imbens and Rubin, 2015) but instead is defined in terms of prediction: if component i is useful for forecasting component j, then a directed association from i to j is established. Typically, the prediction question at the heart of this definition is assessed by estimation of a linear vector autoregressive (VAR) model, where a zero coefficient at index (i, j) for all lags yields a conclusion of no Granger causality from i to j. Granger causality has also been studied in the frequency domain (Geweke, 1982) and remains a popular choice for inferring directed functional connectivity in neuroscience (Seth et al., 2015).

Network view of time series connectivity. A key perspective in multiple time series is that the dependency structure among components, represented in either the time or frequency domains, can be viewed as an edge-weighted graph in which nodes represent the components and edges their corresponding pairwise dependency measure. This view extends the framework of graphical modeling in multivariate statistics (Lauritzen, 1996) to multiple time series analysis. Dahlhaus (2000) established that a graphical model in which edges encode *partial correlation* relations among components satisfies the global Markov property, and thus can be used to reason about the conditional independence structure of the joint distribution of observed variables.

Let $\{X_{it}\}_{t=-\infty}^{\infty}$ and $\{X_{jt}\}_{t=\infty}^{\infty}$ be two components of the time series $X_t \in \mathbb{R}^d$, and let $\tilde{X}_t \in \mathbb{R}^{d-2}$ represent the time series with components i and j removed. The partial correlation between these components at time t, denoted as $\rho_{ij}(k)$, is given by $\rho_{ij}(k) = \mathrm{Cov}(\varepsilon_{it}, \varepsilon_{j,t+k})$, where $\varepsilon_{it} = X_{it} - P_{\mathrm{sp}(\{\tilde{X}_u\}_{u=-\infty}^{\infty})}(X_{it})$ and $P_{\mathrm{sp}(Z)}(Y)$ denotes orthogonal projection of the random variable Y onto the linear span of the random variable Z. Colloquially, then, the partial correlation thus quantifies the second-order dependence that obtains between components i and j after controlling for the (linear) influence of all other observed components. In practice, the partial correlation structure is estimated through its frequency-domain analogue, the *partial coherence*, defined as the rescaled cross-spectral component $\tilde{S}_{ij}(\lambda)$ of $\tilde{S}(\lambda)$, the

spectral density function of the bivariate process $(\varepsilon_{it}, \varepsilon_{jt})^T \in \mathbb{R}^2$.

Estimation of graphical structure from time series data is key in many scientific data analyses, as it reveals how information may propagate around the elements of the complex system under investigation. In neuroscience, functional connectivity networks reveal distinct patterns of coordination across disparate brain regions during the performance of different tasks (Hermundstad et al., 2013), and atypical patterns of functional connectivity are an emerging indicator of psychiatric disorders (Garrity et al., 2007).

Despite the prominence of functional connectivity as an object of statistical and scientific investigation, there is little consensus on how functional connectivity is defined as a statistical relation between the recorded activity in different brain regions, and in practice a wide variety of working definitions are observed (Bastos and Schoffelen, 2016; Noble et al., 2019). A recent comparative analysis of functional connectivity metrics (Mohanty et al., 2020) found that while the correlation failed to capture known patterns of coordination across brain regions, several other measures, including standard and wavelet coherences, were more successful and largely consistent in their results. From the scientific standpoint, this suggests a relative deprioritization of the debate over which single metric best characterizes functional connectivity in favor of defining in each case an analytical framework well suited to the problem at hand. Multiple metrics may be used in the same analysis, as for example in the simultaneous coherence and partial coherence analysis of Sun et al. (2004), with replication across metrics a positive indicator for the scientific robustness of the result.

Estimation. Given an observed sequence $(x_1, x_2, \ldots, x_T)$, the autocovariance $\Gamma(k)$ may be estimated for $0 \leq k < T$ by

$$\widehat{\Gamma}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \hat{\mu})(x_{t+k} - \hat{\mu})^T,$$

where $\hat{\mu}$ is the empirical mean (Brockwell and Davis, 1991, §11.2). Since this sum runs over only $T-k$ observations, less data is available for longer lags k and the estimator consequently

suffers from higher variance. Frequency domain estimators are based on the periodogram $I(\lambda_j)$, defined at Fourier frequencies $\lambda_j = 2\pi j/T$ as

$$I(\lambda_j) = d_j d_j^*,$$
$$d_j = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} \exp(-it\lambda_j) x_t,$$

where the d_j are obtained from the discrete Fourier transform of the observed sequence and d_j^* denotes the conjugate transpose. Under assumptions of short memory, the periodogram is asymptotically unbiased but inconsistent (Brockwell and Davis, 1991). The typical solution is to smooth the periodogram ordinates by convolving across frequencies with a local window function. Despite the complexity apparent in its definition, Dahlhaus (2000) showed that the partial coherence can be straightforwardly computed as the rescaled inverse of the spectral density.

In the long memory case, more serious challenges arise in estimation of the covariance and spectral density. Indeed, the slow decay of serial correlations leads to significant challenges for even the simplest estimators for long memory time series: the sample mean converges at the rate $T^{d-0.5}$ as opposed to the usual $T^{-0.5}$ rate, implying severe issues when the process has long memory towards the higher end of the stationary regime (Hosking, 1996). This in turn affects the asymptotics of the sample autocovariance, both in terms of the rate and limiting distribution. Interestingly, in the case of the multivariate sample autocovariance these aspects separate over two distinct regimes: in the strong long memory regime $0.25 \leq d_j \leq 0.5$, the convergence rate accelerates from $T^{d-0.5}$ to $T^{-0.5}$ as d_j decreases; in the weak long memory regime $0 \leq d_j < 0.25$, the convergence rate settles and the limiting distribution changes from nonnormal to normal as d_j decreases (Chung, 2002, Corollary 1). The situation is similarly challenging in the frequency domain: in contrast to the short memory case, the long memory periodogram is asymptotically biased, and even at Fourier frequencies a finite collection of periodogram ordinates have a nonstandard, non-i.i.d. limiting distribution

(Beran et al., 2013, §4.6.3). Fortunately, these issues tend to be restricted to a region near the origin, leaving the higher-frequency ordinates to behave approximately as in the short memory case.

Alternatives beyond the plug-in estimators for the autocovariance and spectral density include recursive and likelihood-based methods. For linear autoregressive models, the Yule-Walker equations define both an estimation strategy for the model parameters and a recursive formula for computation of the autocovariance sequence given a set of parameter estimates. Optimal convergence rates were established by Cai et al. (2010) for a tapered maximum likelihood estimator of the covariance under a model of independent Gaussian observations, then extended to the case of serial short memory dependence. In the frequency domain, likelihood-based methods are typically based on the Whittle approximation to the Gaussian likelihood (Whittle, 1953), and they can be penalized to provide component-wise data-adaptive smoothness of the spectrum, a feature not available to periodogram-based estimators (Dai and Guo, 2004; Krafty and Collinge, 2013).

1.3 Electrocorticography: Neural Recordings from the Cortical Surface

Throughout this dissertation, we emphasize in particular the analysis of a specific type of multivariate waveform data generated by a method in experimental neuroscience known as *electrocorticography*, or ECoG. The data consist of voltage measurements from a set of electrodes, typically arranged in a grid pattern and embedded in a biocompatible material, that are placed directly on the surface of the brain. At a high level, these measurements represent a summary of the local "activity" in the brain with high temporal and spatial resolution. Our recurring emphasis on this data, however, motivates a slightly more detailed discussion of ECoG and its associated concepts.

Neuronal activity and local field potentials. The outermost layer of neural tissue in the mammalian brain, the cerebral cortex, contains the regions responsible for many complex information processing activities, including sensory processing, motor planning, and abstract

conceptual association (Lerner and Schenk, 2014). The elementary processing units at the cellular level are neurons, which are highly interconnected via specialized anatomical structures. Neuronal networks are extremely dense, containing on average 10^4 cells and several kilometers of connective material per cubic millimeter (Gerstner and Kistler, 2002). Neurons communicate by electrochemical signaling, introducing or removing specific compounds from the area immediately outside the cell, and this "activity" gives rise to measurable currents due to the polarization of these compounds. At any location suitably close to a population of active neurons, their currents superimpose and generate a potential, which is measured in terms of microvolts (μV) and referred to as the *local field potential*, or LFP (Buzsáki et al., 2012).

Scientific applications of ECoG data. The method of recording LFP directly from the cortical surface via grid electrodes is called electrocorticography. The immediate proximity of the recording electrode to the neural populations of interest yields a signal of significantly higher quality versus recording modalities based on sensors outside the skull, and as a result ECoG data are prized in signal processing and statistical applications. For example, ECoG has proven to be a highly successful signal modality in the brain-computer interface (BCI) literature, which emphasizes real-time decoding of recorded brain signals to control computers or external devices (Leuthardt et al., 2006; Miller et al., 2007). Separately, statistical analyses of ECoG data from rhesus macaque monkeys show that this information can be used to classify different states of consciousness (Yanagawa et al., 2013; Wen and Liu, 2016).

More recently, ECoG has been combined with stimulation techniques to understand how the brain responds to perturbations from its baseline activity (Yazdan-Shahmorad et al., 2016; Caldwell et al., 2019). The network perspective on multivariate time series is particularly relevant for these studies, as a long-term scientific objective is to predict changes in functional connectivity resulting from various interventions, and ultimately to design interventions to achieve targeted network reorganization. This could have significant impact in clinical practice, as aberrant functional connectivity underlies a variety of severe neural

disorders (Garrity et al., 2007; Nakai et al., 2021).

Advantages and limitations of ECoG. The main advantage of ECoG is the quality of the LFP signal and its relative lack of vulnerability to measurement artifacts (Nicolas-Alonso and Gomez-Gil, 2012). The source of this advantage is also its limitation; since the electrodes must be placed directly on the cortical surface, ECoG is necessarily an invasive method requiring access inside the skull of a live subject. This has limited ECoG studies to two main sources of subjects: humans undergoing brain surgery, typically for intractable epilepsy, and monkey subjects raised and trained in highly specialized labs. Thus ECoG datasets tend to contain few replicates across different subjects, and it is often difficult or impossible to collect more data should the need arise.

1.4 Contributions and Thesis Outline

The work presented in this dissertation is divided into three chapters. At a high level, each in turn emphasizes one of the three dimensions of dependence in multivariate time series data: first, along the time axis via long memory; second, across components in terms of the coherence; and third, between entire series obtained by replicated experimental trials.

Long memory analysis of language, music, and ECoG data. In Chapter 2 we present two complementary pieces of work concerned with long memory in modern sequence data. First, we develop a framework for addressing the question of long memory in language and music data based on semiparametric estimation and hypothesis testing of the long memory parameter d. This framework is extended to study whether recurrent deep neural networks trained on such data capture this property, where we find evidence for a negative conclusion. Second, we complement this negative result with a positive modeling contribution: a flexible model for the multivariate long memory spectral density. We demonstrate the model on an ECoG dataset from monkey subjects, showing that it can reproduce and extend the results of a previous analysis.

This chapter is joint work with Zaid Harchaoui. The first part was presented at the Statistics for Learning and Data Science conference and subsequently published at the International Conference on Machine Learning (Greaves-Tunnell and Harchaoui, 2019), both in Summer 2019. The second part was presented at the Joint Statistical Meetings in Summer 2021.

Prediction of stimulation-induced connectivity changes. In Chapter 3 we turn our attention to connectivity in multivariate time series, and specifically to the changes in band-limited coherence between ECoG signals induced by brain stimulation. We propose to frame this problem in terms of prediction of the change over two different, experimentally relevant timescales that may depend on both aspects of the stimulation procedure and the original state of the underlying coherence network. We equip this approach with a resampling-based assessment of the stability of the penalized estimator, importance rankings for features or feature groups based on their predictive contribution, and similarity scores for features based on their estimated components in an additive model. Our results show substantial improvements in prediction accuracy over a baseline method and provide novel perspective on the importance of the baseline coherence network in predicting stimulation-induced changes.

This chapter is part of a project with UW Bioengineering and includes collaboration from Julien Bloch, Zaid Harchaoui, Eric Shea-Brown, Ali Shojaie, and Azadeh Yazdan-Shahmorad. A preliminary version was presented at the Society for Neuroscience conference in Winter 2021, and a journal-length version is now under review (Bloch et al., 2021).

Characterizing experimental heterogeneity in neural recordings. In Chapter 4 we address the question of variation at the level of experimental replicates, that is, how observations collected from repeated trials of an ECoG experiment over time may deviate from a naive assumption of group-wise independence. We demonstrate that a prediction rule estimated under the assumption of group independence performs very poorly when the data are stratified at the group level, illustrating the practical consequences of this

statistical consideration. We survey and implement a broad variety of regression techniques that attempt to account for group-level heterogeneity under various assumptions on the data generating process. We find promising results for models accounting for heteroskedastic noise, and for methods that attempt to adjust for unmodeled confounders across trials.

This chapter is joint work with Zaid Harchaoui.

Chapter 2

TESTING AND ESTIMATION OF LONG-RANGE DEPENDENT REPRESENTATIONS FOR MULTIVARIATE TIME SERIES

2.1 Introduction

Modern analyses of sequence data aim to characterize complex phenomena by the statistical properties of an idealized underlying data-generating process. In applications ranging from natural language and music to neuroscience, the data have consistently been found to exhibit long-range dependence. It is consequently expected that models adequate to explain and predict these sequences will capture this property. Whereas this criterion may have been simple to evaluate for classical tools in time series analysis, a dramatic expansion in data collection and computing infrastructure has driven the field towards models that are significantly more challenging to analyze. Standard training corpuses for natural language or music data contain millions of observations (Marcus et al., 1993; Thickstun et al., 2017), while methodological advances in neuroscience enable high-frequency recording from electrode arrays spanning large regions of the cortical surface (Yazdan-Shahmorad et al., 2018). The sequence models most often used to model this data, particularly recurrent neural networks (RNNs), are routinely trained with millions of parameters (Melis et al., 2017).

Beyond their popularity and empirical success, RNN models are a notable case to study as a great deal of their historical development has focused explicitly on the challenge of capturing long-range dependencies (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997a). Despite this focus, the current standard for evaluation is highly empirical; if a model's capacity to represent long-range dependence is measured at all, it is typically evaluated heuristically against some set of tasks in which success is taken an indicator of "memory" in a colloquial

sense. Though undoubtedly informative, such heuristics are rarely defined with respect to an underlying mathematical or statistical property of interest, nor do they necessarily have any correspondence to the data on which the models are subsequently trained.

The result of these developments is a state of current practice in which solid analytical tools lag significantly behind the current empirical methodology. There is thus an opportunity to address the following questions from a statistical perspective: Do the models fit to these data learn to represent long-range dependencies? Can it be established quantitatively that this property is indeed present in the data itself? Can we specify a model that explicitly represents long-range dependencies while also flexibly capturing higher-frequency behavior?

The broad aim of this chapter is to develop statistically principled tools connecting the current state of practice in applied time series analysis to the mathematical foundations of long-range dependence developed in the stochastic process literature. We first develop and illustrate a method for the estimation, visualization, and hypothesis testing of long memory in RNNs, based on an approach that mathematically defines and directly estimates long-range dependence as a property of a multivariate time series. We thus contextualize a core objective of sequence modeling with deep networks against a well-developed but as-yet-unexploited literature on long memory processes. We subsequently propose a model for the spectral density function of a multivariate long-range dependent time series. We provide an analysis demonstrating the relation of this model to recent theoretical development of multivariate LRD processes and establishing an estimation procedure through an alternating minimization framework.

We complement these developments with a variety of empirical analyses. For the long memory analysis of RNNs, we report experimental results obtained on a wide-ranging collection of real-world music and language data, confirming the (often strong) long-range dependencies that are observed by practitioners. However, we find that this property is not adequately captured by a variety of RNNs trained to benchmark performance on a language dataset. Meanwhile, the frequency-domain model is applied to a public repository of ECoG recordings and assessed in the context of both previous statistical analysis and

domain-specific expectations based on a well-established literature.

The remainder of this chapter has the following structure. In Section 2 we give an account of the historical and current literature related to our work. In Section 3 we demonstrate a statistical testing procedure for long memory in language and music data. In Section 4 we define and evaluate a statistical criterion for long memory in RNN models trained on these data. Code corresponding to these two sections is publicly available at https://github.com/alecgt/RNN_long_memory. In Section 5 we define and analyze a frequency-domain model for multivariate long-range dependent time series. In Section 6 we apply this model to ECoG recordings from a rhesus macaque in different states of consciousness.

2.2 Related Work

The motivation and context for the work presented in this chapter can be traced to three separate themes in the literature: the historical emphasis on long-range dependence in the development of recurrent neural network models, statistical approaches to the analysis of natural language and music data, and mathematical constructions of long-range dependent processes that gave rise to formal estimation and testing procedures. This section discusses each of these themes, along with the statistical modeling background and connections to neuroscientific data that motivate the second half of the chapter.

2.2.1 Definitions and evaluation of "long-range dependence" in the RNN literature

The RNN literature is notable for its early and explicit focus on learning to represent long-range dependencies in sequence data. Somewhat curiously, despite this focus, the definition of "long-range dependence" in this literature is itself lacking in mathematical detail, and consequently a wide variety of potentially non-equivalent and typically indirect methods have been used to evaluate its presence in a given model. The seminal reference in this literature is that of Bengio et al. (1994), who state "[a] task displays long-range dependence if prediction of the desired output at time t depends on input presented at an earlier time $\tau \ll t$." Even ignoring its imprecision, the suggested definition is extremely broad. In the simple setting of

linear one-step prediction of a stationary process $X_t \in \mathbb{R}$, the optimal coefficients $\boldsymbol{\phi}_t \in \mathbb{R}^t$ are given by projection of X_{t+1} onto the linear span of $X_1, X_2, ..., X_t$, equivalently represented by the Yule-Walker equations $\Gamma_t \boldsymbol{\phi}_t = \gamma_t$, (Brockwell and Davis, 1991, §5.2) where the real Toeplitz matrix Γ_t has entries $(\Gamma_t)_{ij} = \mathrm{Cov}(X_{i-j}, X_0)$ for $i, j = 1, \ldots, t$ and $\gamma_t \in \mathbb{R}^t$ has entries $\gamma_{tj} = \mathrm{Cov}(X_j, X_0)$ for $j = 1, \ldots, t$. The optimal prediction is $\widehat{X}_{t+1} = \sum_j \phi_{tj} X_j$ and is thus seen to depend on every X_j, $j = 1, \ldots, t$ for which $\phi_{tj} \neq 0$. It is thus the lack of long-range dependence, rather than its presence, that is a special case under this definition.

This perspective was the foundation for analysis of the "vanishing gradient problem," which describes a dilemma for gradient-based estimation of the parameters in a recurrent neural network (Bengio et al., 1994; Pascanu et al., 2013). The RNN recursions

$$\begin{aligned} y_t &= \sigma(h_t) \\ h_t &= W_h \sigma(h_{t-1}) + W_x x_t + b, \end{aligned}$$

are considered as a nonlinear dynamical system in the hidden state h_t with parameters $W_h \in \mathbb{R}^{d \times d}$, $W_x \in \mathbb{R}^{d \times p}$, $b \in \mathbb{R}^d$, and element-wise nonlinear mapping σ. The proposed definition of long-range dependence motivates the concept of a "robustly latched" system, in which the hidden state h_t remains in the subset of the basin of some attractor on which the eigenvalues of the Jacobian are strictly less than one. The relevance of this property derives from the fact that a robustly latched system remains in the same basin of attraction so long as x_t is bounded, so that it colloquially can be said to remember past information despite long intervening sequences of irrelevant input (Bengio et al., 1994, Theorem 3). The dilemma arises upon considering the gradient of a loss function $\mathcal{E}_t = \ell(h_t)$ with respect to one of the parameters, here denoted by θ:

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \frac{\partial \mathcal{E}_t}{\partial h_t} \frac{\partial h_t}{\partial \theta} = \frac{\partial \mathcal{E}_t}{\partial h_t} \sum_{k=0}^{t-1} \frac{\partial h_t}{\partial h_{t-k}} \frac{\partial h_{t-k}}{\partial \theta},$$

where the second equality results from evaluation of the total derivative of h_t considered as a

function of the parameter θ. The term $\frac{\partial h_t}{\partial h_{t-k}}$ is itself a product that can be upper-bounded by a term decaying exponentially to zero with k under the assumption that the model is robustly latched. The contribution of information from time $t-k$ to the gradient thus vanishes, so that it is difficult to estimate a model in which the output y_t depends significantly on x_τ for $\tau \ll t$.

The framework and conclusions of this analysis, including the underlying definition of long-range dependence, have been of central importance in subsequent methodological developments. Many seek to modify the recurrence relations of the hidden state so that the vanishing gradient problem is (potentially) ameliorated. A prominent example is the Long Short-Term Memory (LSTM) of Hochreiter and Schmidhuber (1997a), which augments the simple RNN model by introduction of a cell state c_t,

$$
\begin{aligned}
c_t &= \alpha_t \circ f_\theta(h_{t-1}, x_t) + \beta_t \circ c_{t-1} \\
\alpha_t &= \sigma(A_\alpha h_{t-1} + B_\alpha x_t + b_\alpha) \\
\beta_t &= \sigma(A_\beta h_{t-1} + B_\beta x_t + b_\beta)
\end{aligned}
$$

where $\circ$ denotes the Hadamard product and α_t, β_t are known as the "input" and "forget" gates, respectively. The cell state partial derivatives $\frac{\partial c_t}{\partial c_{t-k}}$, analogous to the hidden state partial derivatives in the simple RNN, are in this case at least not guaranteed to decay exponentially in k.

A complementary probabilistic perspective on the RNN hidden state is available through the lens of *iterated random functions* (Diaconis and Freedman, 1999). Iterated random functions define a stochastic process X_t on $(X, \mathcal{X})$ by the recurrence relation $X_t = f_{Z_t}(X_{t-1})$, where the potentially nonlinear mapping f depends on Z_t, which is itself a stochastic process on $(Z, \mathcal{Z})$. This definition generalizes to the nonlinear case the study of random coefficient autoregressions, for which $Z_t = (A_t, b_t)$ and $f_{Z_t}(x) = A_t x + b_t$. Assuming Z_t is strictly stationary and ergodic, a sufficient condition for the existence of a unique stationary solution to the above recurrence is "contraction on average," i.e. the existence of a measurable function

$z \mapsto K_z$ satisfying $d(f_z(x), f_z(x')) \leq K_z d(x, x')$, $\mathbb{E}[\max\{\log K_{Z_0}, 0\}] < \infty$, and $\mathbb{E}[\log K_{Z_0}] < 0$ for all $(x, x', z) \in X \times X \times Z$, where d is a metric on X. Meanwhile, contraction in the L_p norm, $\|d(f_{Z_0}(x), f_{Z_0}(x'))\|_p \leq \alpha d(x, x')$ for some $\alpha \in (0, 1)$, is sufficient for X_t to admit moments of order $p \geq 1$ (Douc et al., 2014, §4.3). Despite their suitability as a model for the hidden state, iterated random functions are not widely known in the RNN literature, which has tended to prefer deterministic analyses.

Finally, evaluation of methods or architectures proposed to improve RNN capture of long-range dependencies has tended to prioritize the demonstration of high predictive performance on a synthetic task. This approach originates in a preceding literature on learnable "grammars" for finite-state automata (Tomita, 1982). In the RNN context, it stems from the identification by Bengio et al. (1994) of minimal tasks that satisfy heuristic criteria for long-range dependence. Prominent examples include sequence classification when the subsequence relevant to the class is followed by a long realization of white noise (the "latch" problem) or computation of the parity of a temporally distant binary substring (the "parity problem). A simple demonstration by Hochreiter and Schmidhuber (1997b) showed that these tasks can often be solved quickly by random parameter search, casting doubt on their informativeness. However, these and related synthetic tasks have remained popular in the literature (Mikolov et al., 2015; Pascanu et al., 2013). In other cases, despite explicit modeling emphasis on long-range dependence, evaluation is focused entirely on the application of interest, such as language or music modeling, where it is assumed that the data are long-range dependent and that improved performance indicates more successful capture of this property (Mikolov et al., 2015; Boulanger-Lewandowski et al., 2012).

2.2.2 Statistical analyses of natural language and music

In this chapter we focus in particular on RNNs as the main components of stochastic models for language or music. The statistical analysis of these data, however, predates the entire history of RNNs and offers a complementary viewpoint by building from a descriptive analysis focused on the specific data-generating process of interest.

In the case of natural language, the unifying perspective in this literature is of written text as realizations of a discrete process $X_t \in \mathcal{V}$, with $\mathcal{V}$ the finite set of words (i.e. the vocabulary) in the language. Early analyses revealed empirical "laws" corresponding to simple characterizations of this process, most notably Zipf's law (Zipf, 1949), which identifies a power-law relation $f(r) \propto r^{-1}$ between the rank r and corresponding frequency $f(r)$ of words in English text. Shannon (1951) initiated an information-theoretic investigation of natural language in terms of the entropy rate

$$\bar{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

where $H(X)$ is the entropy $H(X) = -\int p(x) \log p(x) dx$ of the random variable X with density p. Such statistical characterizations of natural language provide a reference against which to evaluate text generated by RNN models. While RNN-generated text is generally found to reproduce Zipf's law, often after minimal training (Takahashi and Tanaka-Ishii, 2017), even state-of-the-art language models fail to achieve the expected entropy rate (Braverman et al., 2020).

Assuming stationarity of the generating process, a notion of long-range dependence can be formulated in terms of the decay of the mutual information at a fixed lag

$$I(X_t, X_{t-k}) = H(X_t) - H(X_t|X_{t-k}),$$

where $H(X|Y)$ denotes the conditional entropy $H(X|Y) = -\int p(x,y) \log p(x|y) dx dy$. Lin and Tegmark (2017) construct a simple recursive grammar model that attains slow decay of the mutual information between words, analogous to long memory. This formulation is extended by Braverman et al. (2020) to evaluate long-range dependence in an RNN language model, via the conditional mutual information

$$I(\widehat{X}_t, X_{1:t-k}|X_{t-k+1:t-1}) = H(\widehat{X}_t|X_{t-k+1:t-1}) - H(\widehat{X}_t|X_{1:t-1}). \tag{2.1}$$

Here, $\widehat{X}_t$ denotes the language model prediction at time t, which is generated conditionally given $X_{1:t-1}$. It follows from basic rules of the conditional entropy that this quantity is non-negative and equal to zero if and only if $\widehat{X}_t$ is conditionally independent of $X_{1:t-k}$ given $X_{t-k+1:t-1}$.

The main disadvantage of analyses based on mutual information is that it is difficult to estimate. The first conditional mutual information term in Eq. (2.1) involves a sum over the conditional density $p(\widehat{X}_t|X_{t-k+1:t-1})$, which in turn requires marginalization of the language model density $p(\widehat{X}_t|X_{1:t-1})$ over the distribution of word k-grams. Despite the massive size of modern language corpora, this is infeasible even for modest k, to say nothing of the lengths required for a reasonable assessment of long-range dependence. Even in less data-intensive settings, the mutual information is not straightforward to estimate. The standard plug-in estimator is biased, and the variance can be large when the data have poor coverage of the true support of the generating distribution, as is inevitable for word sequences drawn from a large vocabulary (Paninski, 2003).

Unlike language, music (and audio data in general) is naturally modeled as a continuous-time process, discretized for digital representation at a frequency high enough to cover the standard range of human perception. The effects of this discretization, along with key musical properties, are often easier to study in the frequency domain. Discretization raises the familiar problem of aliasing, whereby the spectral density f_X of the discretized process X_t is related to that of the original, continuous-time process Y_τ via $f_X(\lambda) = \sum_{k=-\infty}^{\infty} f_Y(\lambda + 2\pi k/\Delta\tau)$, with $\Delta\tau$ the discretization interval. Meanwhile, musical harmonies and the timber of instruments yield unique profiles of peaks in the (log-)periodogram computed from the duration of a note or chord (Beran, 2003).

The question of long-range dependence in music data was raised by Voss and Clarke (1975), who identified "$1/f$" behavior of the spectrum across a broad range of genres and styles. The process $X_t \in \mathbb{R}$ is considered a "$1/f$-type process" if its spectral density function $f_X(\lambda)$ satisfies $f_X(\lambda) \propto \lambda^\alpha$ for some $\alpha < 0$ (Percival and Walden, 2006). The process is stationary for $-1 < \alpha < 0$ and nonstationary for $\alpha \leq -1$, in which case the notion of the

spectral density is extended by relation to a valid spectral density function corresponding to a stationary process after suitable differencing. Recent work on the transcription of musical scores from raw audio has established a perspective that has significant overlap with language modeling. The introduction of large, accurately annotated datasets such as MusicNet (Thickstun et al., 2017) has enabled the training of large-scale recurrent or convolutional neural network models that achieve state-of-the-art prediction performance (Sigtia et al., 2015; Thickstun et al., 2018).

2.2.3 Hypothesis testing for long memory processes

Mathematical formalization of long memory stochastic processes paved the way for statistical hypothesis testing procedures. The classical parametric approach involves joint estimation of the parameters $\Theta = \{\Phi, \theta, d\}$ in the ARFIMA model $\Phi(B)(1-B)^{-d}X_t = \theta(B)Z_t$ by maximum likelihood. Consistency and asymptotic normality of the MLE under the assumption of Gaussian Z_t and $d > 0$ were established by Li and McLeod (1986). However, the computational challenge of calculating this estimator and the relative rigidity of the parametric assumption motivate investigation of semiparametric, frequency-domain estimators.

The main challenge with estimators defined in the frequency domain is the poor behavior of the long memory periodogram. Whereas the periodogram ordinates of a stationary short-range dependent process are asymptotically unbiased, asymptotically uncorrelated at Fourier frequencies $\lambda_j = 2\pi j/T$, $j = 1, \ldots, T$, and satisfy

$$\left(\frac{I_X(\lambda_{j1})}{f_X(\lambda_{j1})}, \ldots, \frac{I_X(\lambda_{jk})}{f_X(\lambda_{jk})}\right) \to_d (Z_1, \ldots, Z_k),$$

where $\lambda_{j1}, \ldots, \lambda_{jk}$ are distinct Fourier frequencies and $Z_1, \ldots, Z_k$ are i.i.d. standard exponential random variables, their long memory counterparts enjoy none of these properties (Brockwell and Davis, 1991; Beran et al., 2013). The issue originates in a classical formula for the product of discrete Fourier transform ordinates $d_j = (2\pi T)^{-1/2} \sum_{t=1}^{T} X_t \exp(it\lambda_j)$

(Priestley, 1981):

$$\mathbb{E}[d(\lambda_j)\overline{d(\lambda_j)}] = \int_{-\pi}^{\pi} K_T(\lambda - \lambda_j) f_X(\lambda) d\lambda,$$

where $K_T(\lambda) = (2\pi T)^{-1}(\sin(T\lambda/2)/\sin(\lambda/2))^2$ is the Fejér kernel. The convolution converges uniformly to f_X when f_X is continuous on $[-\pi, \pi]$, as in the short-range dependent case (Stein and Shakarchi, 2011), but it can be shown that a bias term depending on λ results under long memory conditions (Beran et al., 2013).

Estimation of the long memory parameter is often semiparametric, in that the model for the data consists of the finite-dimensional parameter of interest d and an infinite-dimensional nuisance parameter f_U specifying the behavior of the spectrum away from the origin. By restricting focus to an interval near the origin (so-called "narrowband" estimation), we can leverage the local behavior of the spectrum $f_X(\lambda) \approx c_f|1 - \exp(-i\lambda)|^{-2d}$ to reduce the nuisance parameter to the scalar $c_f \neq 0$. One approach to estimating d then comes from consideration of the logarithm of this local behavior, which yields an equation linear in d. Replacing the unknown f_X with the periodogram and assuming $\log I_X(\lambda_j) \approx \log f_X(\lambda_j) + \log Z_j$ yields the regression equation

$$\log I_X(\lambda_j) = \log c_f - 2d \log|1 - \exp(i\lambda_j)| + \log Z_j$$

from which d can be estimated by least squares (Geweke and Porter-Hudak, 1983). A second approach makes use of the Whittle approximation to the time-domain Gaussian likelihood $\mathcal{L}_W(\theta) \approx 2n^{-1}\sum_{j=1}^{T}(\log f_X(\lambda_j) + I_X(\lambda_j)/f_X(\lambda_j))$ (Whittle, 1953). Restricting to a region near the origin enables the approximation $f_X(\lambda) \approx c_f\lambda^{-2d}$, where we have also used $|1 - \exp(i\lambda)|/\lambda \to 1$ as $\lambda \to_+ 0$. The parameter c_f can be profiled out to yield an approximate profile likelihood $K(d)$, from which the "local Whittle" or "Gaussian semiparametric" estimator is computed as $\hat{d} = \arg\min_{d\in(-1/2,1/2)} K(d)$ (Kunsch, 1987). Asymptotic normality of the Gaussian semiparametric estimator was established by Robinson (1995).

The Gaussian semiparametric estimator (GSE) admits a straightforward extension to the multivariate case $X_t \in \mathbb{R}^p$, which is of relevance to the particular applications in this chapter.

It is again based on restriction of the Whittle likelihood to the m lowest Fourier frequencies,

$$\mathcal{L}^{\text{LW}}(G, d) \approx \frac{1}{m} \sum_{j=1}^{m} \left[\log\det \Lambda_j(d) G \Lambda_j^*(d) + \mathbf{Tr}\left[\left(\Lambda_j(d) G \Lambda_j^*(d) \right)^{-1} I(\lambda_j) \right] \right],$$

where we define $\Lambda(d) = \text{diag}(\lambda^{-d} e^{i(\pi-\lambda)/2})$ and $G \in \mathbb{R}^{p \times p}$ is real, symmetric, and positive-definite. The profile likelihood is obtained by first computing $\widehat{G}(d)$ as the solution to $\partial \mathcal{L}^{\text{LW}} / \partial G = 0$ and substituting into the above equation. The long memory vector $d \in \mathbb{R}^p$ is estimated as

$$\hat{d}^{\text{GSE}} = \arg\min_{d \in \Theta} \log\det \widehat{G}(d) - 2 \sum_{i=1}^{p} d_i \frac{1}{m} \sum_{j=1}^{m} \log \lambda_j, \tag{2.2}$$

over the feasible set $\Theta = (-\frac{1}{2}, \frac{1}{2})^p$. A key result due to Shimotsu (2007) establishes that the estimator $\hat{d}^{\text{GSE}}$ is consistent and asymptotically normal under mild conditions, with

$$\sqrt{m}(\hat{d}^{\text{GSE}} - d_0) \to_d \mathcal{N}(0, \Omega^{-1}), \tag{2.3}$$

where

$$\Omega = 2 \left[I_p + G \odot G^{-1} + \frac{\pi^2}{4} (G \odot G^{-1} - I_p) \right],$$

d_0 is the true long memory, and $\odot$ denotes the Hadamard product.

Some recent work applying the long memory perspective to RNN models complements or builds from the contributions of this chapter. Belletti et al. (2018) and Zhao et al. (2020) both analyze conditions on the input sequence x_t and parameterization of the recurrent architecture such that the hidden state sequence h_t is guaranteed to have short memory. However, in proposing modifications to RNN architecture to improve the memory, both papers abandon the long memory framework and instead focus on a different criterion: Belletti et al. (2018) analyze the mutual information between hidden states under a Gaussian assumption, while Zhao et al. (2020) enforce the slow decay of coefficient magnitudes over time in a linearized version of the network.

2.2.4 Frequency-domain modeling and applications in neuroscience

Most models for long memory processes are defined in the time domain, often by straightforward extension of a class of short memory models. Common examples include the (vector) ARFIMA model (Tsay, 2010; Sowell, 1992) and the linear autoregressive conditionally heteroskedastic (LARCH) model (Robinson, 1991), which extend ARMA and ARCH processes such that the observations or conditional variances, respectively, provably exhibit long memory. While popular in certain fields, these models may be difficult to estimate and can be an awkward choice in applications preferring a frequency-domain representation of the signal. In these settings, one simple and appealing approach is given by the fractional exponential (FEXP) model (Beran, 1993), which proposes an additive structure for the univariate long memory spectral density:

$$f_X(\lambda) = g(\lambda)^{1-2d} \exp\left(\sum_{j=0}^{p} \eta_j h_j(\lambda)\right).$$

The model parameters are the weights η_j and long memory parameter d; the functions h_j and g are selected in advance such that the h_j are continuous and symmetric around $\lambda = 0$ and $g : [-\pi, \pi] \to \mathbb{R}_+$ satisfies $\lim_{x\to 0} g(x)/x = 1$. An approximate likelihood is constructed via the relation $I_X(\lambda)/f(\lambda) \approx Z$, which converts the problem to that of maximum likelihood estimation in a generalized linear model with logarithmic link function and multiplicative exponential errors. Unfortunately, this approach does not extend readily to the multivariate setting.

Even in the short-range dependent case, the periodogram has poor asymptotic properties and is thus of limited value as an estimator for the spectral density. The standard solution is to construct an estimator by smoothing of the periodogram ordinates over a local window of frequencies (Brillinger, 2001). However, this approach is somewhat restrictive in the multivariate case as the same degree of smoothing is applied to all components of the cross-spectrum. An alternative framework modeling the Cholesky decomposition of the

multivariate spectral density function (Dai and Guo, 2004) allows for different degrees of smoothing component-wise and has been adopted in nonparametric approaches based on spline regression (Rosen and Stoffer, 2007; Krafty and Collinge, 2013). These models explicitly assume short-range dependence of the generating process.

The motivation to model multivariate, long-range dependent data in the spectral domain derives from a growing body of literature in computational neuroscience, which has identified power-law behavior consistent with long memory in the observed spectra across a wide range of electrophysiological data collection modalities and experimental settings (Timmermann et al., 2019; Wen and Liu, 2016; Maxim et al., 2005). While the biological origins of these long memory dynamics remain under investigation, they are believed to be generated through mechanisms distinct from the more classically analyzed oscillatory patterns (He et al., 2010). Furthermore, measures of long-range dependence have been shown to be predictive of specific patterns in neural activity, including states of consciousness or pathological conditions such as Alzheimer's disease (Yanagawa et al., 2013; Maxim et al., 2005).

The statistical tools used to study this data have been limited to date. Maxim et al. (2005) model individual fMRI voxels as a fractional Gaussian noise. An ad-hoc estimator for the long memory is developed by Wen and Liu (2016) in terms of the resampled time series $\tilde{X}_t^f$ with resampling rate f and corresponding periodogram $\tilde{I}_X^f(\lambda)$. The authors define $S^f(\lambda) = \sqrt{\tilde{I}_X^f(\lambda)\tilde{I}_X^{1/f}(\lambda)}$ and estimate the long memory component $\widehat{S}(\lambda)$ of the spectrum as the pointwise median of the functions $S^{f_1}(\lambda), \ldots, S^{f_n}(\lambda)$, where $f_1, \ldots, f_n$ are a collection of non-integer resampling factors. We emphasize that in each case, the methods assume that the data is a scalar time series, despite their subsequent application to multivariate fMRI, MEG, or ECoG data. The multivariate case is handled via repeated, independent univariate analysis of the components, corresponding to the implicit assumption that all cross-spectral terms are identically zero and precluding an analysis of interactions between signals.

2.2.5 Contributions

This chapter develops a perspective drawing on the themes of both Sections 2.2.2 and 2.2.3 to analyze long-range dependence in RNN models of natural language and music data. We focus in particular on a set of highly popular RNN models whose motivation originates with the vanishing gradient analysis of Bengio et al. (1994). These include the LSTM of Hochreiter and Schmidhuber (1997a), the structurally constrained recurrent network (SCRN) of Mikolov et al. (2015), and the recurrent additive network of Levy et al. (2018).

In contrast to the heuristic performance-based approaches developed in the RNN literature, we follow Takahashi and Tanaka-Ishii (2017) in evaluating the models in terms of the relevant statistical property of the data on which they are trained. Here, the main point of contrast is that we focus on long memory, as opposed to definitions based on the mutual information (Lin and Tegmark, 2017; Braverman et al., 2020), and thereby avoid the attendant limitations that have confined those approaches to approximate or inconclusive analyses. The key idea is to use word or music embeddings - real, vector-valued representations of the sequence to be modeled - which can then be analyzed as a multivariate time series. This is not merely a convenient workaround; such embeddings are ubiquitous across modern, state-of-the-art applications of deep neural network models to these data.

Our analysis of the trained RNN models focuses on the hidden state as a stochastic process. Rather than analyze this process in terms of its recurrence equations, which is attempted under various assumptions by Belletti et al. (2018) and Zhao et al. (2020), we estimate the long memory from sample paths generated by the trained model. We study a simple transformation of the estimator proposed by Shimotsu (2007), which retains the asymptotic normality used to calibrate a testing procedure, and which as a semiparametric estimator allows for direct investigation of the long memory without needing to specify the higher-frequency behavior of the model.

We obtain negative results with regard to long memory in trained RNN models that motivate a positive modeling contribution in the second half of the chapter. Unlike Belletti

et al. (2018) and Zhao et al. (2020), who propose new developments for the RNN architecture at the expense of changing focus away from a long memory analysis, we develop a frequency-domain approach. We illustrate the model in a detailed re-analysis of the ECoG data from Wen and Liu (2016), both reproducing their results on the marginal short and long-range dependence properties of the data and extending the analysis to reveal interactions between the components via the partial coherence.

2.3 Long Memory in Language and Music

In the first two sections of this chapter, we develop a framework for visualization and hypothesis testing of long memory in RNNs, based on an approach that mathematically defines and directly estimates long-range dependence as a property of a multivariate time series. Much of the development of deep recurrent neural networks has been motivated by the goal of finding good representations and models for text and audio data. It is therefore scientifically appropriate to first ask whether long memory is even a relevant consideration in the context of these applications.

Representation of language and music via vector embeddings

A key methodological perspective throughout this section is the identification of raw language and music data with real, vector-valued time series observed at a regular interval. This perspective is only novel from the standpoint of statistical testing for language data; it is not introduced out of convenience, but rather as a recognition of what has become a nearly universal approach to modeling sequence data with RNNs.

In language modeling, the basic unit of observation is a word "token," which may represent a word in the English vocabulary or a unit of punctuation. Extremely rare words or those not typically assumed to belong to the finite vocabulary, such as proper names, are assigned to a common `UNK` token, for "unknown." While classical approaches to language modeling attempted to estimate conditional distributions directly on this state space, for example through Markov models, much greater empirical success has been obtained by iden-

tifying each token with a unique vector in a high-dimensional Euclidean space. Such "word embeddings" were first justified heuristically (Mikolov et al., 2013) but subsequently shown to correspond to weight vectors in a matrix factorization of the pointwise mutual information between word-context pairs (Levy and Goldberg, 2014).

Audio data, including music, is real-valued but univariate in its raw form, corresponding to air pressure measurements sampled at a discrete but very high rate, typically around 44kHz. This corresponds to the Nyquist rate for the upper frequency range of human perception of speech, which extends to around 20 kHz (Monson et al., 2012). Classically, multivariate representations of this data derive from a time-frequency decomposition such as the short-time Fourier or wavelet transforms. These representations are often engineered such that the parts of the spectrum most relevant for human hearing are increased in saliency, as is the case for mel-frequency cepstral coefficients (Logan, 2000). As with language modeling, more recent empirical success has been driven in part by learning vector representations for a windowed sequence of raw audio data as part of an end-to-end model for prediction (Thickstun et al., 2018).

Optimization. Relatively little discussion of optimization procedures for the problem in Eq. (2.2) is available in the time series literature. For instance, we are not aware of any proof that the objective is convex in the multivariate setting. To compute the estimator $\hat{d}^{\mathrm{GSE}}$, we apply L-BFGS-B, a quasi-Newton algorithm that handles box constraints (Byrd et al., 1995). L-BFGS-B is an iterative algorithm requiring the gradient of the objective; this is derived in Appendix A.

Bandwidth selection. The choice of the bandwidth parameter m determines the tradeoff between bias and variance in the estimator: at small m the variance may be high due to few data points, while setting m too large can introduce bias by accounting for the behavior of the spectral density function away from the origin.

When it is possible to simulate from the target process, as will be the case when we

evaluate criteria for long memory in recurrent neural networks, we can naturally control the variance simply by simulating long sequences and computing a dense estimate of the periodogram. Without knowledge of the shape of the spectral density function, however, it is difficult to know how to set the bandwidth to avoid bias, and thus we prefer the relatively conservative setting of $m = \sqrt{T}$. This choice is justified by a bias study for the bandwidth parameter, which is given in Appendix A.

Total memory

It is common for sequence embeddings and RNN hidden layers to have hundreds of dimensions, and thus long memory estimation for these sequences naturally occurs in a high-dimensional setting. This topic is virtually unexplored in the time series literature, where multivariate studies tend to have modest dimension. Practically, this raises two main issues. First, if $p \approx m$ for dimension p and bandwidth m, then the approximation of the test statistic distribution by its asymptotic limit will be of poor quality, and the resulting test is likely to be miscalibrated. Second, it becomes difficult to interpret the long memory vector d, particularly when the coordinates of the corresponding time series are not meaningful themselves.

We resolve both issues by considering the *total memory* statistic $\bar{d}$, defined as

$$\bar{d} = \frac{1}{p} \sum_{j=1}^{p} \hat{d}_j^{\mathrm{GSE}}. \tag{2.4}$$

Computation of the total memory is no more complex than that of the GSE, and it has an intuitive interpretation as the coordinate-wise aggregate strength of long memory in a multivariate time series, where "aggregate" here refers to the average instead of the sum.

The total memory is a simple linear functional of the GSE, and thus its consistency and asymptotic normality can be established by a simple argument. In particular, defining $g(d) = \frac{1}{p} \sum_{j=1}^{p} d_i$, we see that $\nabla g(d) = \mathbf{1}/p$, which is clearly nonzero at zero, so that by Eq.

(2.3) and the delta method we have

$$\sqrt{m}\left(\bar{d}-\bar{d}_0\right) \to_d \mathcal{N}\left(0, \frac{1}{p^2}\sum_{j=1}^{p}\sum_{k=1}^{p}\Omega_{jk}^{-1}\right), \tag{2.5}$$

where $\bar{d}_0 = \frac{1}{p}\sum_{j=1}^{p} d_j$ is the true total memory of the observed process. To validate this proposed estimator, we provide a "sanity check" on simulated high-dimensional data with known long memory in Appendix A.

Visualizing and testing for long memory in high dimensions. The visual time-domain summary of long memory in Figure 1.1 can be extended to the multivariate setting. In this case, the autocovariance $\Gamma(k) = \mathrm{Cov}(X_t, X_{t+k})$ is matrix-valued, which for the purpose of evaluating long memory can be summarized by the scalar $\mathbf{Tr}\,|\Gamma(k)|$, where the absolute value is taken element-wise. Recall that a sufficient condition for short memory is the absolute convergence of the autocovariance series, whereas this series diverges for long memory processes.

From a testing perspective, a statistical decision rule for the presence of long memory can be derived from the asymptotic distribution of the corresponding estimator. However, when the dimension p is large and we conservatively set the bandwidth $m = \sqrt{T}$, we may have $m \approx p$ even when the observed sequence is relatively long.

The classical approach to testing for the multivariate Gaussian mean is based on the Wald statistic

$$m(\hat{d}-d_0)^T\Omega(\hat{d}-d_0),$$

which has a $\chi^2(p)$ distribution under the null hypothesis $\mathcal{H}_0 : d = d_0$. In Appendix A, we give a demonstration that the standard Wald test can be seriously miscalibrated when $m \approx p$, whereas testing for long memory with the total memory statistic remains well-calibrated in this setting. These results are consistent with previous observations that the Wald test for long memory can have poor finite-sample performance even in low dimensions (Shimotsu,

2007; Hurvich and Chen, 2000), though these studies suggest no alternative.

Testing for long memory. For all experiments, we test the null hypothesis

$$\mathcal{H}_0 : \bar{d}_0 = 0$$

against the one-sided alternative of long memory,

$$\mathcal{H}_1 : \bar{d}_0 > 0.$$

We set the level of the test to be $\alpha = 0.05$ and compute the corresponding critical value c_α from the asymptotic distribution of the total memory estimator.

Given an estimate of the total memory $\bar{d}(x_{1:T})$, a p-value is computed as $P(\bar{d} > \bar{d}(x_{1:T})|\bar{d}_0 = 0)$; note that a p-value less than $\alpha = 0.05$ corresponds to rejection of the null hypothesis in favor of the long memory alternative. In all tables, we indicate that that a test rejects null hypothesis $\mathcal{H}_0$ with a checkmark (✓), and that a test fails to reject $\mathcal{H}_0$ by a cross (✗).

Experiments

A full summary of results is given in Table 2.1, and autocovariance partial sums are visualized in Figure 2.1. We note that the total memory, as an average over the components of a multivariate time series, is naturally suited for comparison of the estimated long memory across multiple time series of different dimension.

Natural language data. We evaluate long memory in three different sources of English language text data: the Penn TreeBank training corpus (Marcus et al., 1993), the training set of the Children's Book Test from Facebook's bAbI tasks (Weston et al., 2016), and the King James Bible. The Penn TreeBank corpus and King James Bible are considered as single sequences, while the Children's Book Test data consists of 98 books, which are considered as separate sequences. We require that each sequence be of length at least $T = 2^{14}$, which

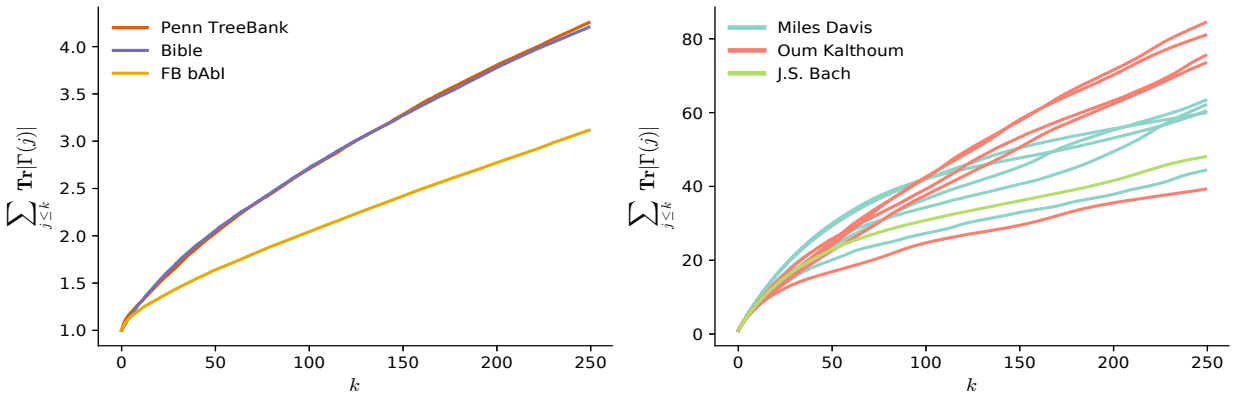

Figure 2.1: Partial sum of the autocovariance trace for embedded natural language and music data. *Left*: Natural language data. For clarity we include only the longest of the 98 books in the Facebook bAbI training set. *Right:* Music data. Each of the five tracks from both Miles Davis and Oum Kalthoum is plotted separately, while the Bach cello suite is treated as a single sequence.

ensures that the periodogram can be estimated with reasonable density near the origin. Finally, we use GloVe embeddings (Pennington et al., 2014) to convert each sequence of word tokens to an equal-length sequence of real vectors of dimension $k = 200$.

The results show significant long memory in each of the text sources, despite their apparent differences. As might be expected, the children's book measured from the Facebook bAbI dataset demonstrates the weakest long-range dependencies, as is evident both from the value of the total memory statistic and the slope of the autocovariance partial sum.

Music data. Modeling and generation of music has recently gained significant visibility in the deep learning community as a challenging set of tasks involving sequence data. As in the natural language experiments, we seek to evaluate long memory in a broad selection of representative data. To this end, we select a complete Bach cello suite consisting of 6 pieces from the MusicNet dataset (Thickstun et al., 2017), the jazz recordings from Miles Davis' *Kind of Blue*, and a collection of the most popular works of famous Egyptian singer Oum Kalthoum. For the Bach cello suite, we embed the data from its raw scalar wav file format using a reduced version of a deep convolutional model that has recently achieved near

Table 2.1: Total memory in natural language and music data.

	Data	Total memory	p-value	Reject $\mathcal{H}_0$?
Natural Language	Penn TreeBank	0.163	$<1 \times 10^{-16}$	✓
	Facebook CBT	0.0636	$<1 \times 10^{-16}$	✓
	King James Bible	0.192	$<1 \times 10^{-16}$	✓
Music	J.S. Bach	0.0997	$<1 \times 10^{-16}$	✓
	Miles Davis	0.322	$<1 \times 10^{-16}$	✓
	Oum Kalthoum	0.343	$<1 \times 10^{-16}$	✓

state-of-the-art prediction accuracy on the MusicNet collection of classical music (Thickstun et al., 2018). Details of the model training, including performance benchmarks, are provided in Appendix A.

We are not aware of a prominent deep learning model for either jazz music or vocal performances. Therefore, for the recordings of Miles Davis and Oum Kalthoum, we revert to a standard method and extract mel-frequency cepstral coefficients (MFCC) from the raw wav files at a sample rate of 32000 Hz (Logan, 2000). A study of the impact of embedding choice on estimated long memory, including a long memory analysis of the Bach data under MFCC features, is provided in Appendix A. The results show that long memory appears to be even more strongly represented in music than in text. We find evidence of particularly strong long-range dependence in the recordings of Miles Davis and Oum Kalthoum, consistent with their reputation for repetition and self-reference in their music.

Overall, while the results of this section are unlikely to surprise practitioners familiar with the modeling of language and music data, they are scientifically useful for two main reasons: first, they show that our long memory analysis is able to identify well-known instances of long-range dependence in real-world data; second, they establish quantitative criteria for the successful representation of this dependency structure by RNNs trained on such data.

2.4 A Statistical Criterion for Recurrent Neural Networks

The results of the previous section establish that long memory is present, and often strong, across a broad variety of natural language and music data sources. We now address the question of whether RNN models trained on these data are able to capture this property. It is worth emphasizing that nonlinear transitions of the previous state alone are no guarantee that the dependence structure is any more complex than a simple linear autoregressive or Markov model; this is demonstrated in the proposition below.

Proposition 2.4.1. *Define the scalar nonlinear autoregressive process*

$$X_{t+1} = f(X_t) + \varepsilon_t,$$

where $\{\varepsilon_t\}$ *is a white noise sequence with positive density with respect to Lebesgue measure and satisfying* $\mathbb{E}|\varepsilon_t| < \infty$*, while* $f : \mathbb{R} \to \mathbb{R}$ *is bounded on compact sets and satisfies*

$$\sup_{|x|>r} \left| \frac{f(x)}{x} \right| < 1$$

for some $r > 0$*. Then* X_t *has a unique stationary distribution* π*, and the sequence of random variables* $\{X_t, t \geq 0\}$ *initialized with* $X_0 \sim \pi$ *is strictly stationary and geometrically ergodic.*

Furthermore, if

$$\mathbb{E}|X_t|^{2+\delta} < \infty$$

for some $\delta > 0$*, then* $\{X_t\}$ *is a short memory process.*

Proof. The proof proceeds by analysis of X_t as a Markov chain on a general state space $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the standard Borel sigma algebra on the real line. Define the transition kernel $P(x, B) = P(X_t \in B | X_{t-1} = x)$ for any $x \in \mathbb{R}$ and $B \in \mathcal{B}$.

We first establish that X_t is aperiodic. A d-cycle is defined by a collection of disjoint sets $\{D_i\}, 0 = 1, ..., d-1$ such that

1. For $x \in D_i$, $P(x, D_{i+1}) = 1$, $i = 0, ..., d-1 \mod d$.

2. The set $[\cup_i D_i]^C$ has measure zero.

The period is defined as the largest d for which $\{X_t\}$ has a d-cycle (Meyn and Tweedie, 2012). Clearly, however, since ε_t has positive density with respect to Lebesgue measure, $p(x, D) = 1$ only if $D = \mathbb{R}$ up to null sets. Thus the period is $d = 1$, so $\{X_t\}$ is aperiodic.

Strict stationarity and geometric ergodicity are established by showing that the aperiodic chain $\{X_t\}$ satisfies a strengthened version of the Tweedie criterion (Meyn and Tweedie, 2012), which requires the existence of a measurable non-negative function $g : \mathbb{R} \to \mathbb{R}$, $\epsilon > 0$, $R > 1$ and $M < \infty$ such that

$$\mathbb{E}[g(X_{t+1})|X_t = x] \leq \frac{g(x) - \epsilon}{R}, \quad x \in K^c$$
$$\mathbb{E}[g(X_{t+1})\mathbb{1}\{X_{t+1} \in K^c\}|X_t = x] \leq M, \quad x \in K$$

for some set K satisfying

$$\inf_{x \in K} \sum_{n=1}^{m} P^n(x, B) > 0$$

Under the conditions of f and ε_t assumed above, this criterion is established for the process X_t by Tjøstheim (1990, Thm 4.1), with $g(x) = |x|$.

Geometric ergodicity implies that the

$$\|\lambda P^n - \pi\|_{TV} \leq C\rho^n,$$

with $C < \infty$, $\rho \in (0, 1)$, and where $\|\cdot\|_{TV}$ denotes the total variation distance between measures. A well-known result in the theory of Markov chains (Nummelin and Tuominen, 1982) establishes that geometric ergodicity is equivalent to absolute regularity, which is parameterized by

$$\beta(k) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cup B_j) - P(A_i)P(B_j)|,$$

where the supremum is taken over all finite partitions $\{A_1, ..., A_I\}$ and $\{B_1, ..., B_J\}$ of the sigma fields $\mathcal{A} = \sigma(X_t)$ and $\mathcal{B} = \sigma(X_{t+k})$. In particular, $\beta(k)$ decays at least exponentially fast. Furthermore, for any two sigma fields $\mathcal{A}$ and $\mathcal{B}$ we have (Bradley et al., 2005, §2.1)

$$\begin{aligned}\beta(\mathcal{A}, \mathcal{B}) &= \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cup B_j) - P(A_i)P(B_j)| \\ &\geq \sup \frac{1}{2} |P(A \cup B) - P(A)P(B)|, \quad A \in \mathcal{A}, B \in \mathcal{B} \\ &= 2\alpha(\mathcal{A}, \mathcal{B}),\end{aligned}$$

so that the α-mixing parameter is also bounded by an exponentially decaying sequence.

Finally, if $\mathbb{E}|X|^{2+\delta}$ for some $\delta > 0$, then the absolute covariance obeys (Ibragimov and Linnik, 1971, Thm. 17.2.2)

$$|\gamma(k)| = \sigma^{-2}|\rho(k)| \leq C\alpha(k)^{\delta/(2+\delta)},$$

which completes the proof. □

RNN hidden state as a nonlinear model for a long memory process.

The standard tool for statistical modeling of multivariate long memory processes is the vector autoregressive fractionally integrated moving average (VARFIMA) model, which represents the process $X_t \in \mathbb{R}^p$ with long memory parameter d as

$$\Phi(B)(1 - B)^d X_t = \theta(B)Z_t,$$

where Z_t is a white noise process and $(1 - B)^d = \text{diag}((1 - B)^{d_i})$, $i = 1, ..., p$ (Lobato, 1997; Sowell, 1992). Under the standard stationarity and invertibility conditions on the matrix polynomials $\Phi(B)$ and $\Theta(B)$, respectively, the process can be represented as

$$X_t = (1 - B)^{-d}\Phi^{-1}(B)\Theta(B)Z_t,$$

which shows that the X_t has a composite representation in terms of linear "features" of the input sequence and an explicit fractional integration step ensuring that it satisfies the definition of multivariate long memory in Eq. (1.1).

We extend this view to deep network models for sequences with long range dependencies. The key difference is that RNN models are not constrained to work with a linear representation of the data, nor do they explicitly contain a step that guarantees the long memory of X_t. To evaluate long memory in an RNN model, we study the stochastic process

$$X_t = \Psi(Z_{-\infty:t}), \tag{2.6}$$

where Z_t is again a white noise, and the nonlinear transformation Ψ describes the RNN transformation of inputs to the hidden state. In a typical RNN model, a decision rule is learned by linear modeling of the hidden state; this framework thus aligns with a broader theoretical characterization of deep learning as approximate linearization of complex decision boundaries in input space by means of a learned nonlinear feature representation (Bruna and Mallat, 2013; Mairal et al., 2014; Jones et al., 2019; Bietti and Mairal, 2019).

Testable criteria for RNN capture of long-range dependence.

The complexity of $\Psi(\cdot)$ corresponding to even the most basic RNN sequence models precludes a fully theoretical treatment of long memory in processes described by Eq. (2.6). Nonetheless, this characterization suggests an approach for the statistical evaluation of long memory in RNNs, as it establishes testable criteria under which a model of the form Eq. (2.6) describes a process X_t with long memory. In particular, to satisfy the definition in Eq. (1.1) we must have

$$X_t = \Psi(Z_{-\infty:t}) = (1-B)^{-d}\tilde{\Psi}(Z_{-\infty:t})$$

for some $d \neq 0$ and process $Y_t = \tilde{\Psi}(Z_{-\infty:t})$ with bounded and nonzero spectral density at zero frequency. Semiparametric estimation of d in the frequency domain provides a means

to evaluate this condition such that the results are agnostic to the behavior of $\tilde{\Psi}(Z_{-\infty:t})$ at higher frequencies. If $\Psi(Z_{-\infty:t})$ admits a representation in terms of an explicit fractional integration step, then this can be investigated in two complementary experiments:

1. **Integration of fractionally differenced input.** Define

$$\tilde{X}_t = (1 - B)^d Z_t,$$

where Z_t is a standard Gaussian white noise and d is the long memory parameter corresponding to the source X_t on which the model was trained. If the sequence $\tilde{x}_{1:T}$ is drawn from $\tilde{X}_t$, then we expect to find that

$$\hat{d}^{\mathrm{GSE}}(\tilde{h}_{1:T}) \approx 0,$$

where $\tilde{h}_{1:T} = (\Psi(\tilde{x}_1), \Psi(\tilde{x}_{1:2}) \ldots, \Psi(\tilde{x}_{1:T}))$ is the RNN hidden representation of the simulated input. On the other hand, nonzero long memory in the hidden state indicates a mismatch between fractional integration learned by the RNN and long memory of the data X_t.

2. **Long memory transformation of white noise.** Conversely, we expect to find that the RNN hidden representation of a white noise sequence has a nonzero long memory parameter. White noise has a constant spectrum and thus a long memory parameter equal to zero. If $\Psi(\cdot)$ performs both the feature representation and fractional integration functions that are handled separately and explicitly in the VARFIMA model, then a zero-memory input will be transformed to a nonzero-memory sequence of hidden states.

Table 2.2: Language model performance by RNN type

Model	Test Perplexity
Zaremba et al.	114.5
LSTM	114.5
Memory cell	119.0
SCRN	124.3

2.4.1 Long memory analysis of language model RNNs

We now turn to the question of whether RNNs trained on one of the datasets evaluated above are able to represent the long-range dependencies that we know to be present. We evaluate the criteria for long memory on three different RNN architectures: long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997a), memory cells (Levy et al., 2018), and structurally constrained recurrent networks (SCRN) (Mikolov et al., 2015). Each network is trained on the Penn TreeBank corpus as part of a language model that includes a learned word embedding and linear decoder of the hidden states; the architecture is identical to the "small" LSTM model of Zaremba et al. (2014), which is preferred for the tractable dimension of the hidden state. Model performance is evaluated in terms of the *perplexity*, defined as the exponentiated average cross-entropy between the predicted and true next-word distributions, on a held-out test set; see Table 2.2. Note that our objective is not to achieve state-of-the-art results, but rather to reproduce benchmark performance in a well-known deep learning task. Finally, for comparison, we will also include an untrained LSTM in our experiments; the parameters of this model are simply set by random initialization.

RNN integration of fractionally differenced input. Having estimated the long memory parameter d corresponding to the Penn TreeBank training data in the previous section, we simulate inputs $\tilde{x}_{1:T}$ with $T = 2^{16}$ from by fractional differencing of a standard Gaussian white noise and evaluate the total memory of the corresponding hidden representation $\Psi(\tilde{x}_{1:T})$ for each RNN. Results from $n = 100$ trials are compiled in Table 2.3, with standard

Table 2.3: Residual total memory in RNN representations of fractionally differenced input.

Model	**Total memory**	**p-value**	**Reject $\mathcal{H}_0$?**
LSTM (trained)	-8.36×10^{-3} (0.00475)	4.07×10^{-2}	✓
LSTM (untrained)	-6.20×10^{-2} (0.00387)	$<1 \times 10^{-16}$	✓
Memory cell	-1.18×10^{-2} (0.00539)	1.52×10^{-2}	✓
SCRN	-2.62×10^{-2} (0.00631)	3.32×10^{-5}	✓

errors reported in parentheses. We test the null hypothesis $\mathcal{H}_0 : \bar{d} = 0$ against the one-sided alternative $\mathcal{H}_1 : \bar{d} < 0$, which corresponds to the model's failure to represent the full strength of fractional integration observed in the data.

RNN transformation of white noise. For a complementary analysis, we evaluate whether the RNNs can impart nontrivial long-range dependency structure to white noise inputs. In this case, the input sequence $z_{1:T}$ is drawn from a standard Gaussian white noise process, and we test the corresponding hidden representation $\Psi(z_{1:T})$ for nonzero total memory. As in the previous experiment, we select $T = 2^{16}$, choose the bandwidth parameter $m = \sqrt{T}$, and simulate $n = 100$ trials for each RNN. Results are detailed in Table 2.4. We test $\mathcal{H}_0 : \bar{d}_0 = 0$ against $\mathcal{H}_1 : \bar{d}_0 > 0$; here, the alternative corresponds to successful transformation of white noise input to long memory hidden state.

We summarize the main experimental result as follows: there is a statistically well-

Table 2.4: Total memory in RNN representations of white noise input.

Model	**Total memory**	**p-value**	**Reject $\mathcal{H}_0$?**
LSTM (trained)	-8.59×10^{-4} (0.00405)	0.583	X
LSTM (untrained)	-4.17×10^{-4} (0.00223)	0.572	X
Memory cell	-5.96×10^{-4} (0.00452)	0.552	X
SCRN	2.37×10^{-3} (0.00522)	0.324	X

defined and practically identifiable property, relevant for prediction and broadly represented in language and music data, that is not present according to two fractional integration criteria in a collection of RNNs trained to benchmark performance. Tables 2.3 and 2.4 show that each evaluated RNN fails both criteria for representation of the long-range dependency structure of the data on which it was trained. The result holds despite a training protocol that reproduces benchmark performance, and for RNN architectures specifically engineered to alleviate the gradient issues typically implicated in the learning of long-range dependencies.

2.5 Modeling Multivariate Long Memory Time Series in the Frequency Domain

In the second part of this chapter, we define, analyze, and numerically illustrate a model for the spectral density function of a multivariate long-range dependent process. This topic arises naturally as a positive methodological contribution to complement the negative result of the previous section. Before defining the model, we briefly review the definitions and conditions required for valid specification of such a process via its spectral density function.

Conditions for short-range spectral components. The structure of the short-range dependent spectrum can be used to identify a broad function space from which to draw flexible representations of spectral behavior away from the origin.

It follows from basic properties of the complex exponential that the spectral density function $f_U(\lambda)$ corresponding to a short-range dependent process U_t is periodic and Hermitian, that is, $f_U(\lambda) = f_U(\lambda + 2\pi)$ and $f_U(\lambda) = f_U(-\lambda)^*$. By Parseval's theorem, the short-range dependence condition $\sum_{n=-\infty}^{\infty} \|\gamma_U(n)\| < \infty$ implies $\int_{-\pi}^{\pi} \mathbf{Tr}(f_U(\lambda) f_U(\lambda)^*) d\lambda < \infty$ so that the component functions $(f_U)_{jk}(\lambda)$, and therefore their real and imaginary parts, are elements of $L_2(-\pi, \pi)$ for each $j, k \in \{1, \ldots, p\}$.

If each component also has a square-integrable derivative with respect to λ, then it is an element of the periodic Sobolev space

$$\mathcal{P}_1 \triangleq \{g \in \mathcal{P} : g' \in \mathcal{P}, \int_{-\pi}^{\pi} [g'(\lambda)]^2 d\lambda < \infty\},$$

where $\mathcal{P}$ is the set of real-valued functions of period 2π. Equipping this space with the inner product defined for any f and g in $\mathcal{P}_1$ as

$$\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\frac{\partial}{\partial \lambda} g(\lambda) \right) \left(\frac{\partial}{\partial \lambda} f(\lambda) \right) d\lambda$$

yields the reproducing kernel Hilbert space of Cogburn and Davis (1974) with kernel

$$\xi(\lambda, \nu) = \sum_{m \in \mathbb{Z} \setminus \{0\}} (2\pi m)^{-1} \exp(im(\lambda - \nu)).$$

Since $f_U(\lambda)$ is Hermitian, its real and imaginary components must be even and odd, respectively. Restriction of the reproducing kernel to these subspaces results in (Krafty and Collinge, 2013)

$$\xi^{(\mathrm{Re})}(\lambda, \nu) = \sum_{m=1}^{\infty} (2\pi m)^{-1} \cos(m\lambda) \cos(m\nu)$$
$$\xi^{(\mathrm{Im})}(\lambda, \nu) = \sum_{m=1}^{\infty} (2\pi m)^{-1} \sin(m\lambda) \sin(m\nu).$$

The real and imaginary parts of the short-range spectrum can thus be represented in cosine and sine bases of periodic functions, respectively.

Multivariate long-range dependence in the frequency domain. Here we recall the general framework for long memory in the multivariate and frequency-domain setting due to Kechagias and Pipiras (2015), which was briefly introduced in Chapter 1. A weakly stationary process $X_t \in \mathbb{R}^p$ is long-range dependent if for every $j, k \in \{1, ..., p\}$ the $(j,k)^{th}$ component $f_{jk}(\lambda)$ of its spectral density can be written as

$$f_{jk}(\lambda) = G_{jk}(\lambda) |\lambda|^{-(d_j + d_k)}$$

such that

$$G_{jk}(\lambda)|\lambda|^{-(d_j+d_k)} \sim g_{jk}e^{i\phi_{jk}\text{sign}(\lambda)}|\lambda|^{-(d_j+d_k)} \tag{2.7}$$

as $|\lambda| \to 0$ for some $d \in (0, 1/2)^p$.

The matrix $G(\lambda) \in \mathbb{C}^{p\times p}$ is Hermitian positive definite for each λ, and $g_{jk} \in \mathbb{R}$, $g_{jj} > 0$, $\phi_{jk} \in (-\pi, \pi]$ for each $j, k \in \{1, ..., p\}$. Moreover, the mapping $\lambda \mapsto G(\lambda)$ is Hermitian, that is, $G(-\lambda) = G(\lambda)^*$.

The multivariate LRD spectrum therefore consists of two components: a long memory vector d controlling the behavior of the spectral components near zero frequency, and a complex matrix-valued function $G(\lambda)$ that specifies the behavior of the spectrum at higher frequencies. The conditions on $G(\lambda)$ and d ensure the validity of $f(\lambda)$ as a spectral density function.

Decomposition of the fractional spectrum. We focus in particular on processes $X_t \in \mathbb{R}^p$ that can be represented as fractionally integrated. The main advantage of considering such processes is that fractional integration, and therefore long-range dependence, is defined as a convolution with a linear filter. This implies that the spectral density function of X_t can be written as

$$f_X(\lambda) = \Lambda(d)f_U(\lambda)\Lambda(d)^*,$$

where $\Lambda(d) = \text{diag}\left((1 - e^{i\lambda})^{-d_1}, ..., (1 - e^{i\lambda})^{-d_p}\right)$ and $f_U(\lambda)$ is the spectral density function of the short-range dependent process U_t.

A natural approach to modeling the long-range dependent spectrum $f_X(\lambda)$ is therefore to separate the short-range dependent (or high-frequency) behavior of the process from the long-range dependence exhibited at low frequencies and parameterized in the case of fractionally integrated processes by the long memory vector d.

Model definition. We propose to model the multivariate LRD spectrum $f(\lambda)$ as $f(\lambda) = [H(\lambda)H(\lambda)^*]^\dagger$, where $H(\lambda) \in \mathbb{C}^{p\times p}$ is a lower-triangular matrix with real, non-negative diag-

onals and † denotes the Moore-Penrose pseudoinverse. The product $H(\lambda)H(\lambda)^*$ is positive semidefinite, so its pseudoinverse is as well. The pseudoinverse is unique and equal to the inverse when $H(\lambda)$ is positive definite, which occurs when it has strictly positive diagonals. Parameterization of the spectral density through a decomposition of the inverse has been proposed for short memory spectra by Rosen and Stoffer (2007) and Krafty and Collinge (2013); here we extend this approach to the long memory case.

The nonzero entries of $H(\lambda)$ are parameterized as

$$\begin{aligned} H_{jj}(\lambda) &= |1 - e^{i\lambda}|^{d_j} S_{jj}^{\mathrm{Re}}(\lambda) \\ \Re\{H_{jk}(\lambda)\} &= \Re\{(1 - e^{i\lambda})^{d_j}\} S_{jk}^{\mathrm{Re}}(\lambda) \\ \Im\{H_{jk}(\lambda)\} &= \Im\{(1 - e^{i\lambda})^{d_j}\} S_{jk}^{\mathrm{Im}}(\lambda), \end{aligned}$$

with $j \in \{1, \ldots, p\}, k < j$.

The short-range components $S_{jk}^{\mathrm{Re}}(\lambda)$ and $S_{jk}^{\mathrm{Im}}(\lambda)$ have a simple additive form corresponding to a truncated basis expansion in the even and odd subspaces, respectively, of $\mathcal{P}_1$

$$\begin{aligned} S_{jk}^{\mathrm{Re}}(\lambda) &= \sum_{\ell=0}^{L} \alpha_{jk\ell} \cos(\ell\lambda) \quad j = 1, \ldots, p; k \leq j \\ S_{jk}^{\mathrm{Im}}(\lambda) &= \sum_{\ell=1}^{L} \beta_{jk\ell} \sin(\ell\lambda) \quad j = 1, \ldots, p; k < j, \end{aligned}$$

with $j \in \{1, \ldots, p\}, k < j$.

The model parameters are thus $\theta = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{d}\}$ with

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha_{jk0}, \ldots, \alpha_{jkL})_{j=1,\ldots,p;k\leq j} \\ \boldsymbol{\beta} &= (\beta_{jk1}, \ldots, \beta_{jkL})_{j=1,\ldots,p;k<j} \\ \mathbf{d} &= (d_1, \ldots, d_p). \end{aligned}$$

The diagonal of $H(\lambda)$ is strictly positive if, for each $j = 1, \ldots, p$, the polynomial $r_j(\lambda) =$

$\sum_{\ell=0}^{L} \alpha_{jj\ell} \cos(\ell\lambda)$ does not have a real root in $(0, 2\pi]$. This is not guaranteed, though the penalized estimation scheme discussed below, which guarantees hierarchical sparsity in the estimate of the vector $(\alpha_{jj0}, \ldots, \alpha_{jjL})$, eliminates some simple cases in which positive definiteness may not be achieved. Regardless, we emphasize that by construction the spectral density is guaranteed to be positive semidefinite at every frequency.

Moreover, the short-range component of the inverse Cholesky factor $H(\lambda)$ is Hermitian and periodic, as required for a valid model of the fractionally integrated multivariate spectrum. This can be viewed as implicit enforcement of a constraint on the gradient of the real and complex components of the short-range spectrum with respect to the frequency λ at integer multiples of $1/2$. The term $(1 - e^{i\lambda})^d$ controls the behavior of the model at low frequencies and thus constitutes the long-range component.

Proposition 2.5.1. *Let $f(\lambda) = \tilde{\Lambda}(d)G(\lambda)\tilde{\Lambda}(d)^*$ be the spectral density matrix function of a process obtained by fractional differencing of a short-range dependent process with positive definite spectral density $G(\lambda)$. Then the function $G(\lambda)^{-1/2}$, defined as the lower-triangular Cholesky factor of $G(\lambda)^{-1}$, is periodic with even real part and odd complex part, and the lower-triangular Cholesky factor of $f(\lambda)^{-1}$ has the form $f(\lambda)^{-1/2} = [\tilde{\Lambda}(d)^*]^{-1}G(\lambda)^{-1/2}$ with $[\tilde{\Lambda}(d)^*]^{-1} = diag\left((1 - e^{i\lambda})^{d_1}, ..., (1 - e^{i\lambda})^{d_p}\right)$.*

Proof. By definition of $f(\lambda)$, $G(\lambda)$ is the spectral density matrix function of some second-order stationary process U_t. Its periodicity then follows from a basic property of the complex sinusoid

$$G(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-ik\lambda}\gamma(k) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-i}e^{-ik\lambda}\gamma(k) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{-ik(\lambda+1)}\gamma(k) = G(\lambda + 1).$$

Periodicity of $G(\lambda)$ implies periodicity of $G(\lambda)^{-1/2}$, as the inverse and the Cholesky decomposition are well-defined and unique.

Furthermore, $G(\lambda) = G(-\lambda)^*$ implies

$$\Re\{G(\lambda)\} = \Re\{G(-\lambda)\}$$
$$\Im\{G(\lambda)\} = -\Im\{G(-\lambda)\}.$$

Finally,

$$f(\lambda)^{-1} = [\tilde{\Lambda}(d)^*]^{-1} G(\lambda)^{-1} \tilde{\Lambda}(d)^{-1}$$

so that $[\tilde{\Lambda}(d)^*]^{-1} G(\lambda)^{-1/2}$ is the unique lower-triangular Cholesky factor of $f(\lambda)^{-1}$.

Moreover,

$$\left[(1 - e^{-i\lambda})^{-d}\right]^{-1} = |1 - e^{-i\lambda}|^d e^{id \arg(1 - e^{-i\lambda})} = (1 - e^{-i\lambda})^d$$

and $\left[(1 - e^{-i\lambda})^{-d}\right]^* = (1 - e^{i\lambda})^{-d}$; see Appendix A for details. Thus

$$[\tilde{\Lambda}(d)^*]^{-1} = \operatorname{diag}\left((1 - e^{i\lambda})^{d_1}, ..., (1 - e^{i\lambda})^{d_p}\right).$$

□

Proposition 2.5.2. *Let $f(\lambda) = \tilde{\Lambda}(d) G(\lambda) \tilde{\Lambda}(d)^*$, where $G(\lambda)$ is such that $G(\lambda)_{jk} \sim g_{jk} \in \mathbb{R}$ for every $j = 1, ..., p$, $k \leq j$ as $\lambda \to 0$. Then*

$$f(\lambda)_{jk} \sim \lambda^{-(d_j - d_k)} e^{-i \ sign(\lambda) \frac{\pi}{2}(d_j - d_k)} g_{jk} \quad as \ \lambda \to 0$$

so that $f(\lambda)$ has the asymptotic behavior of a multivariate LRD spectrum under the specific setting of the phase parameters $\phi_{jk} = -\frac{\pi}{2}(d_j - d_k)$.

Proof. We have

$$1 - e^{-i\lambda} = 1 - \cos(\lambda) + i \sin(\lambda) = 2 \sin\left(\frac{\lambda}{2}\right) e^{i(\pi - \lambda)/2},$$

where the second equality follows from the application of basic product-sum and translation

identities for trigonometric functions. Then given $f(\lambda) = \tilde{\Lambda}(d)G(\lambda)\tilde{\Lambda}(d)^*$ we have

$$\begin{aligned} f(\lambda)_{jk} &= (1 - e^{-i\lambda})^{-d_j}(1 - e^{i\lambda})^{-d_k} G(\lambda)_{jk} \\ &= \left[2 \sin\left(\frac{\lambda}{2}\right)\right]^{-(d_j+d_k)} e^{-i(\pi-\lambda)(d_j-d_k)/2} G(\lambda)_{jk}, \end{aligned}$$

so that

$$f(\lambda)_{jk} \sim \lambda^{-(d_j+d_k)} e^{-i\pi(d_j-d_k)/2} g_{jk}, \text{ as } \lambda \to 0^+$$

and

$$f(\lambda)_{jk} \sim -\lambda^{-(d_j+d_k)} e^{i\pi(d_j-d_k)/2} g_{jk}, \text{ as } \lambda \to 0^-,$$

where we have used $\lim_{x\to 0} \sin(x)/x = 1$.

□

2.5.1 Penalized estimation

Tree-structured penalization of a Whittle likelihood-type objective. Let $x_{1:T}$ be an observed sequence with corresponding discrete Fourier transform

$$y_j = T^{-1/2} \sum_{t=1}^{T} x_t e^{-i\lambda_j t}$$

and periodogram $I(\lambda_j) = y_j y_j^*$ evaluated at the Fourier frequencies $\lambda_j = 2\pi j/T$, $j = 1, \ldots, \lfloor T/2 \rfloor$.

The model parameters are estimated by minimization of a hierarchically penalized objective function related to the multivariate Whittle likelihood

$$J_\rho(\theta) = \mathcal{L}(\theta) + \Omega_\rho(\theta). \tag{2.8}$$

The loss function $L(\theta)$ is defined as

$$\mathcal{L}(\theta) = \sum_{j=1}^{\lfloor T/2 \rfloor} \mathbf{Tr}(I_j f(\lambda_j)^{-1}) - \log |f(\lambda_j)^{-1}|$$

and can be seen as a version multivariate negative Whittle log-likelihood.

Meanwhile, we penalize the short-range spectral components $S_{\text{Re}}^{jk}(\lambda)$ and $S_{\text{Im}}^{jk}(\lambda)$ individually via a sum of hierarchically ordered ℓ_2-norm penalties on their coefficients in the relevant basis expansion:

$$\Omega_\rho(\theta) = \rho \left[\sum_{j \geq k=1}^{p} \sum_{\ell=1}^{L} w_\ell \left\| \alpha_{j,k,\ell:L} \right\|_2 + \sum_{j > k=1}^{p} \sum_{\ell=1}^{L} w_\ell \left\| \beta_{j,k,\ell:L} \right\|_2 \right]. \tag{2.9}$$

Here we use the notation $\boldsymbol{\alpha}_{j,k,\ell:L}$ and $\boldsymbol{\beta}_{j,k,\ell:L}$ to denote the sub-vectors

$$\boldsymbol{\alpha}_{j,k,\ell:L} = (\boldsymbol{\alpha}_{j,k,\ell}, \boldsymbol{\alpha}_{j,k,\ell+1}, \ldots, \boldsymbol{\alpha}_{j,k,L})^T,$$
$$\boldsymbol{\beta}_{j,k,\ell:L} = (\boldsymbol{\beta}_{j,k,\ell}, \boldsymbol{\beta}_{j,k,\ell+1}, \ldots, \boldsymbol{\beta}_{j,k,L})^T,$$

respectively, for every $\ell \leq L$. The weights w_ℓ are given by $w_\ell = \ell^3 - (\ell - 1)^3$, consistent with the penalization scheme of Haris et al. (2019).

For each coordinate index (j, k) of the multivariate spectrum, the penalty Ω_ρ enforces the regularization of a tree-structured grouping of the corresponding real and imaginary basis coefficients. The proximal operator corresponding to such penalization schemes is studied by Jenatton et al. (2010), who also provide efficient implementations in the `SPAMS` toolbox.

Proposition 2.5.3. *The penalized Whittle objective $J_\rho(\theta)$ is jointly convex in the short-range parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ of the multivariate LRD spectral model over the set on which $S_{jj}(\lambda) > 0$ for each $j = 1, \ldots, p$.*

Proof. The penalized Whittle objective is written as the sum of multivariate Whittle likeli-

hood and roughness penalty terms

$$J_\rho(\theta) = \mathcal{L}(\theta) + \Omega_\rho(\theta).$$

First consider the likelihood term

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{j=1}^{\lfloor T/2 \rfloor} \mathbf{Tr}(I_j f(\lambda_j)^{-1}) - \log|f(\lambda_j)^{-1}| \\ &= \sum_{j=1}^{\lfloor T/2 \rfloor} \mathbf{Tr}(H(\lambda_j)^* w_j w_j^* H(\lambda_j)) - 2\log|f(\lambda_j)^{-1}|\end{aligned}$$

where parameterize the unique lower-triangular matrix $H(\lambda)$ satisfying $f(\lambda)^{-1} = H(\lambda)H(\lambda)^*$ via

$$\begin{aligned}H_{jj}(\lambda) &= |1-e^{i\lambda}|^{2d_j} S_{jj}^{\mathrm{Re}}(\lambda) \\ \Re\{H_{jk}(\lambda)\} &= \Re\{(1-e^{i\lambda})^{d_j}(1-e^{-i\lambda})^{d_k}\} S_{jk}^{\mathrm{Re}}(\lambda) \\ \Im\{H_{jk}(\lambda)\} &= \Im\{(1-e^{i\lambda})^{d_j}(1-e^{-i\lambda})^{d_k}\} S_{jk}^{\mathrm{Im}}(\lambda).\end{aligned}$$

Then

$$H(\lambda) = \tilde{\Phi}(\lambda, d)^* S(\lambda),$$

with $\tilde{\Phi}(\lambda, d) = \mathrm{diag}\left((1-e^{i\lambda})^{d_1}, ..., (1-e^{i\lambda})^{d_p}\right)$. The real and complex elements of $S(\lambda)$ are linear as a function of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, so the mapping $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto (\Re\{S(\lambda)\}, \Im\{S(\lambda)\})$ is convex in each component.

To show convexity of $\mathcal{L}(\theta)$ in (α, β) it suffices to show convexity of the term $\mathcal{L}_j(\theta) \triangleq \mathbf{Tr}(I_j f(\lambda_j)^{-1}) - \log|f(\lambda_j)^{-1}|$ for a fixed Fourier frequency λ_j.

The log-determinant $-\log|f(\lambda)^{-1}|$ decomposes as

$$\begin{aligned}
-\log|f(\lambda)^{-1}| &= -\log|\tilde{\Phi}(\lambda,d)^* S(\lambda)S(\lambda)^*\tilde{\Phi}(\lambda,d)| \\
&= -\log|\tilde{\Phi}(\lambda,d)^*| - \log|\tilde{\Phi}(\lambda,d)| - \log|S(\lambda)S(\lambda)^*| \\
&= -2\log|1-e^{i\lambda}|\sum_{j=1}^{p} d_j - 2\sum_{j=1}^{p}\log S_{jj}(\lambda), \qquad (2.10)
\end{aligned}$$

where we have used the fact that $S(\lambda)$ is triangular and positive along the diagonal. This term is thus convex in $(\boldsymbol{\alpha},\boldsymbol{\beta})$ as the composition of a linear and convex function. Let $I = yy^*$ be a periodogram ordinate. The term $\mathbf{Tr}(If(\lambda)^{-1})$ can be written as

$$\begin{aligned}
\mathbf{Tr}(If(\lambda)^{-1}) &= \mathbf{Tr}(I_j\tilde{\Phi}(\lambda,d)^* S(\lambda)S(\lambda)^*\tilde{\Phi}(\lambda,d)) \\
&= y^*\tilde{\Phi}(\lambda,d)^* S(\lambda)S(\lambda)^*\tilde{\Phi}(\lambda,d)y \\
&= \left\| y^*\tilde{\Phi}(\lambda,d)^* S(\lambda)\right\|_F^2
\end{aligned}$$

and is again convex in $(\boldsymbol{\alpha},\boldsymbol{\beta})$ by composition rules. It follows that $\mathcal{L}(\theta)$ is convex in $(\boldsymbol{\alpha},\boldsymbol{\beta})$. Finally, the penalty $\Omega_\rho(\theta)$ is convex as the sum of convex functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Each term in the sum is convex by composition rules, as the composition of a linear operation selecting the appropriate sub-vector of $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, and a norm. □

We include some further considerations of the conditions under which the objective is also convex in the long memory parameter d. Recall from Eq. (2.10) that the log-determinant term is linear and thus convex in d. It remains to consider the trace term. Writing

$$\tilde{\Phi}(\lambda,d)^* S(\lambda)S(\lambda)^*\tilde{\Phi}(\lambda,d) = S(\lambda)S(\lambda)^* \odot \phi_d\phi_d^*$$

with $\phi_d = \text{diag}(\tilde{\Phi}(\lambda, d)) \in \mathbb{C}^p$ and where $\odot$ denotes the Hadamard product, we have

$$\begin{aligned}
\mathbf{Tr}(I_j f(\lambda_j)^{-1}) &= y_j^* \tilde{\Phi}(\lambda, d)^* S(\lambda) S(\lambda)^* \tilde{\Phi}(\lambda, d) y_j \\
&= y_j^* [S(\lambda) S(\lambda)^* \odot \phi_d \phi_d^*] y_j \\
&= \mathbf{Tr}(Y_j^* S(\lambda) S(\lambda)^* Y_j \phi_d \phi_d^*) \\
&= \phi_d^* Y_j^* S(\lambda) S(\lambda)^* Y_j \phi_d \\
&= \sum_{m=1}^{p} \sum_{n=1}^{p} (1 - e^{i\lambda})^{d_m} (1 - e^{-i\lambda})^{d_n} [Y_j^* S(\lambda) S(\lambda)^* Y_j]_{mn} \\
&= \sum_{m=1}^{p} |1 - e^{i\lambda}|^{2d_m} [Y_j^* S(\lambda) S(\lambda)^* Y_j]_{mm} \\
&\qquad + \sum_{m=1}^{p} \sum_{n<m} 2\Re\{(1 - e^{i\lambda})^{d_m} (1 - e^{-i\lambda})^{d_n} [Y_j^* S(\lambda) S(\lambda)^* Y_j]_{mn}\}.
\end{aligned}$$

Analysis of the convexity of $\mathcal{L}(\theta)$ in d reduces to analysis of the conditions rendering this sum convex. Since the matrix $M_j \triangleq Y_j^* S(\lambda) S(\lambda)^* Y_j$ is Hermitian, it has positive diagonals, and thus the sum over diagonal elements is always convex. One condition for convexity is thus $(M_j)_{mn} = 0 \;\; \forall m \neq n$. If all components of the long memory vector d are equal, then this can be relaxed to the condition $\sum_{n<m} \Re\{(M_j)_{mn}\} > 0 \;\; \forall m$.

Estimation via alternating minimization. The convexity of the objective in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and (under additional conditions) d suggests an alternating minimization scheme. Following recent work on the convergence of alternating minimization to second-order stationary points (Li et al., 2019), we regularize the alternating minimization problems with a proximal term.

The minimization over the short-range parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is solved by accelerated proximal gradient descent, with the step-size selected via backtracking line search (Nesterov, 1983; Beck and Teboulle, 2009), while the minimization over the long memory parameter d is solved by L-BFGS (Byrd et al., 1995). Since the objective is convex in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we enforce a stopping criterion in terms of the norm of the gradient, whereas for the long memory parameter we stop when the change in objective falls below the threshold 1×10^{-2} or when a

maximum number of iterations is reached. The stopping criterion for the outer loop of the alternating minimization is given by $\left\|\theta^{(t)} - \theta^{(t-1)}\right\|_2 < \varepsilon$. For the simulation study below, we set $\varepsilon = 1 \times 10^{-6}$ and observe convergence within 150 iterations.

Algorithm 1 Alternating minimization for the spectral-LRD model.

Require: Data $x_{1:T}$
Require: Initial values $\boldsymbol{\alpha}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $d^{(0)}$
Require: Hierarchical penalization parameter $\rho > 0$
Require: Tolerance $\varepsilon > 0$
Compute periodogram $I(\lambda_j)$ for $j = 1, \ldots, \lfloor T/2 \rfloor$
Set $t = 0$
while not converged **do**
    Update $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, d^{(t)}) + \Omega_\rho(\boldsymbol{\alpha}, \boldsymbol{\beta})$
    Update $d^{(t+1)} = \arg\min_d \mathcal{L}(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, d)$
    Set $t = t + 1$
end while
Return $(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}), d^{(t)}$

As the objective is not jointly convex in the parameters, the estimator will depend on our choice of initialization. We use the following scheme:

1. For each component $i = 1, \ldots, p$, compute the (scalar) local Whittle estimator of d_i:

$$\hat{d}_i^{LW} = \underset{d \in (-\frac{1}{2}, \frac{1}{2})}{\arg\min} \left[\log \left(\frac{1}{m} \sum_{j=1}^{m} \frac{I(\lambda_j)_{ii}}{\lambda_j^{-2d}} \right) - d \left(\frac{2}{m} \sum_{j=1}^{m} \log \lambda_j \right) \right],$$

   where the bandwidth parameter m is typically set as $m = T^{1/2}$. Note that this scalar optimization problem is convex in d and so the estimator can be computed efficiently and to high accuracy.

2. Initialize $d_i = \hat{d}_i^{LW}$ for each component $i = 1, \ldots, p$ of the LRD spectrum.

3. Set $\alpha_{ii0} = 1$ for each $i = 1, \ldots, p$, and set the remaining elements of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to zero. The initial setting for the short-range spectrum is thus real, constant, and diagonal. We observe that it lies within the nullspace of the penalty $\Omega_\rho(\theta)$.

Numerical illustration

We illustrate the automatic model selection and support recovery properties of the penalized estimator through a simulation study. Data of length $T = 2^{12}$ are generated from a true model of dimension $p = 3$ with hierarchically sparse parameterization of the real and imaginary spectral components. The components vary in order, from a minimum of 1 to a maximum of 6; details are provided in Appendix A. We then estimate a spectral LRD model given the generated data, with an order upper bound of $L = 10$.

Support recovery as a function of ρ The experiment in this section is modeled on the support recovery experiments of Bach (2008). We compute the spectral LRD estimate over

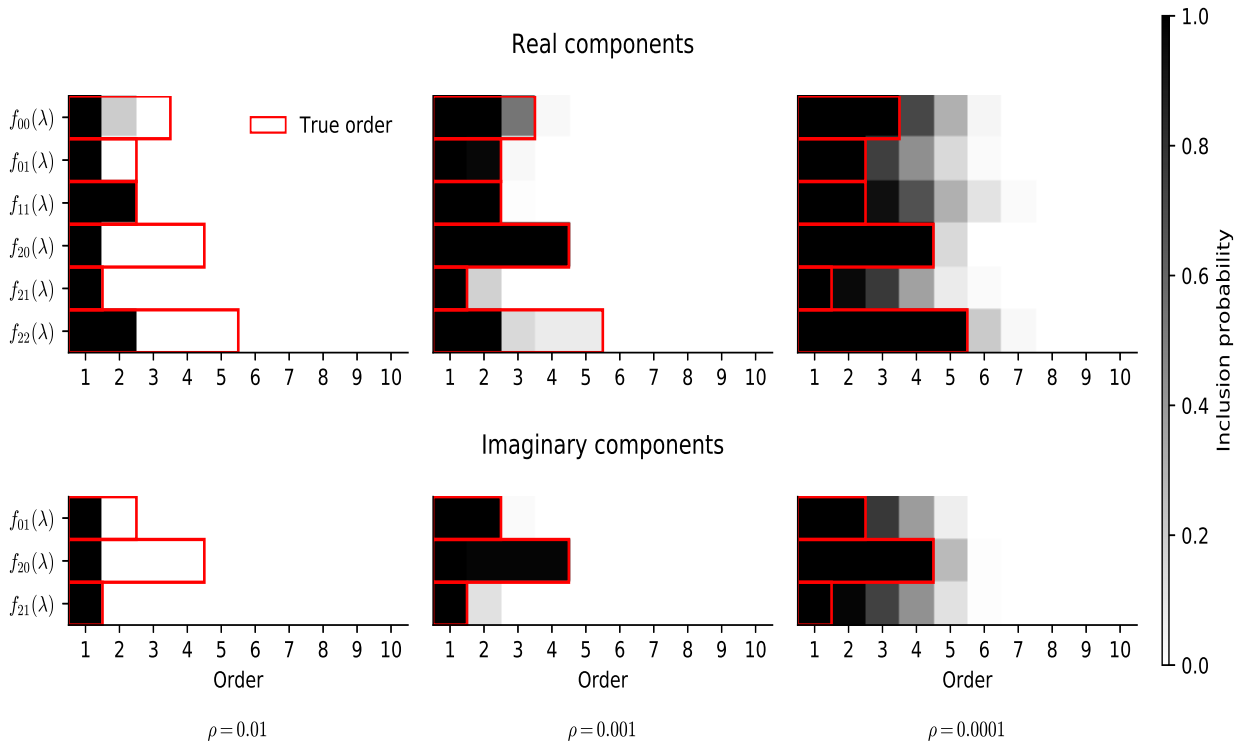

Figure 2.2: Probability of component selection over $n = 100$ trials for the spectral LRD support recovery experiment. The true order of each component is indicated by a red outline of the heatmap cells corresponding to nonzero coefficients.

a log-linearly spaced grid of values for the regularization parameter ρ, for each of $n = 100$ sample datasets drawn from the same hierarchically sparse model. For each value of ρ, we plot a heatmap of the model parameters where the intensity at a given grid location represents the proportion of the n experiments in which the parameter was included in the estimated model; see Figure 2.2. For reference, we also show the true order of each spectral component. The results demonstrate the component-wise hierarchical structure that results from penalization as in Eq. (2.9) and indicate the potential for accurate support recovery when the generating process is itself hierarchically sparse.

Selection of the regularization parameter. Next we demonstrate a holdout validation procedure for selection of the regularization parameter ρ. Again generating data of length $T = 2^{12}$, we use the first half of the sequence to train the model while reserving the second half as a holdout set on which to estimate out-of-sample spectral prediction performance. By construction, the validation set will be correlated with the training set, which is of nontrivial

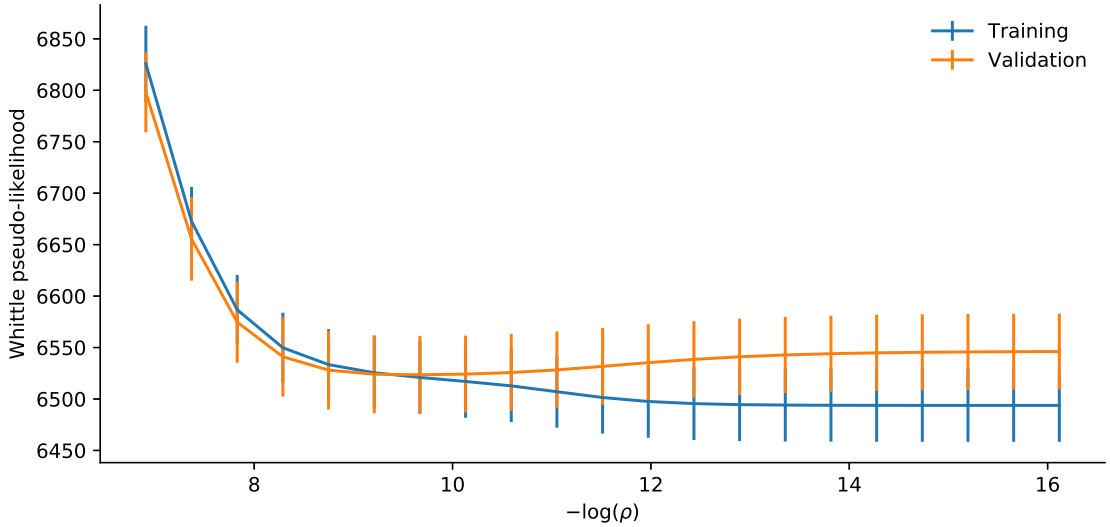

Figure 2.3: Whittle pseudo-likelihood of the training (blue) and held-out validation data (orange) as a function of the regularization parameter ρ. Error bars indicate standard error over $n = 5$ replicate datasets.

concern for a potentially long memory time series; nonetheless we accept some level of bias here in order to better estimate the long memory parameter. We estimate the spectral LRD model over a range of values for ρ and compute the Whittle pseudo-likelihood of both the train and validation sets under the model. The parameter ρ is selected over a grid of 21 logarithmically spaced values between $\rho = 1 \times 10^{-7}$ and $\rho = 1 \times 10^{-3}$. We repeat the procedure over $n = 5$ simulated datasets and plot the training and validation results in Figure 2.3.

We select the largest regularization parameter ρ such that the mean validation loss over the $n = 5$ trials is within one standard error of the minimum mean validation loss over all values of ρ in the grid. This "one standard error" rule is a standard heuristic that aligns with our preference for parsimonious model estimates (Friedman et al., 2001). For the simulated data, the selected value is $\hat{\rho} = 2.51 \times 10^{-4}$.

Model estimation. We estimate a spectral LRD model using the selected value of $\hat{\rho}$ and again setting the order upper bound $L = 10$. The final estimated sparsity pattern is plotted in Figure 2.4. We observe that the validation procedure yields a level of penalization that results in accurate recovery of the component-wise support, despite their varying orders. The selected value of $\hat{\rho}$ agrees with the support recovery results shown in Fig. 2.2, which suggest that a value of $1 \times 10^{-3} \leq \rho \leq 1 \times 10^{-4}$ will yield an estimate close to the truth.

At high frequencies, the asymptotic distribution of the the normalized periodogram $I(\lambda_j)_{kk}/f(\lambda_j)_{kk}$ is well-approximated by the standard exponential distribution, which is the exact asymptotic distribution in the short-memory case (Brockwell and Davis, 1991; Beran et al., 2013). Therefore, a measure of goodness-of-fit at high frequencies can be obtained visually by Q-Q plots of the component-wise normalized periodogram against the quantiles of the standard exponential distribution. This is plotted in Figure 2.5. Finally, the components of the estimated spectral density are plotted against both the true model and the raw periodogram computed from the simulated data in Figure 2.6. We note that the magnitude of the auto and cross-spectra varies over several orders of magnitude, and the model estimates

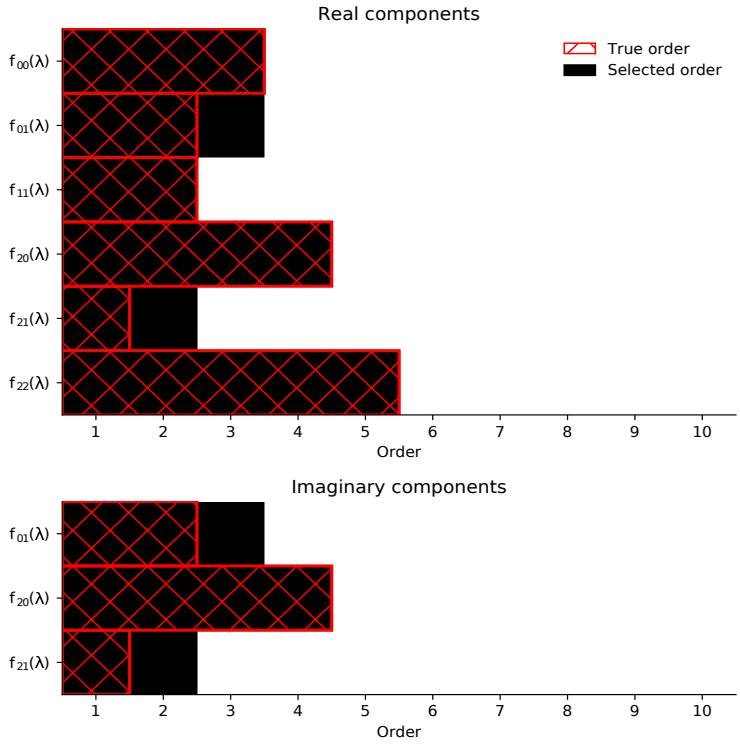

Figure 2.4: Sparsity pattern of the estimated model. Selected components are indicated in black, with the red outlines showing the true order of each component.

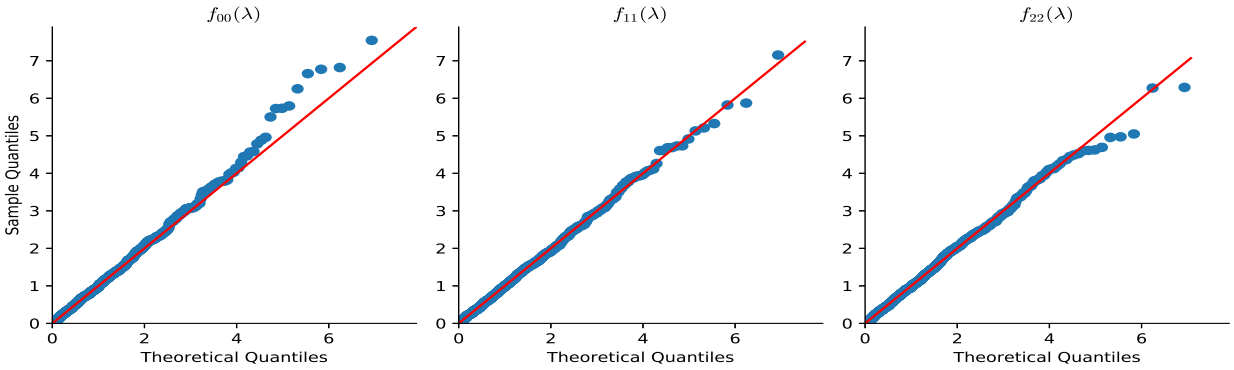

Figure 2.5: Quantile-quantile plots of the component-wise normalized spectral density estimates against the quantiles of the standard exponential distribution.

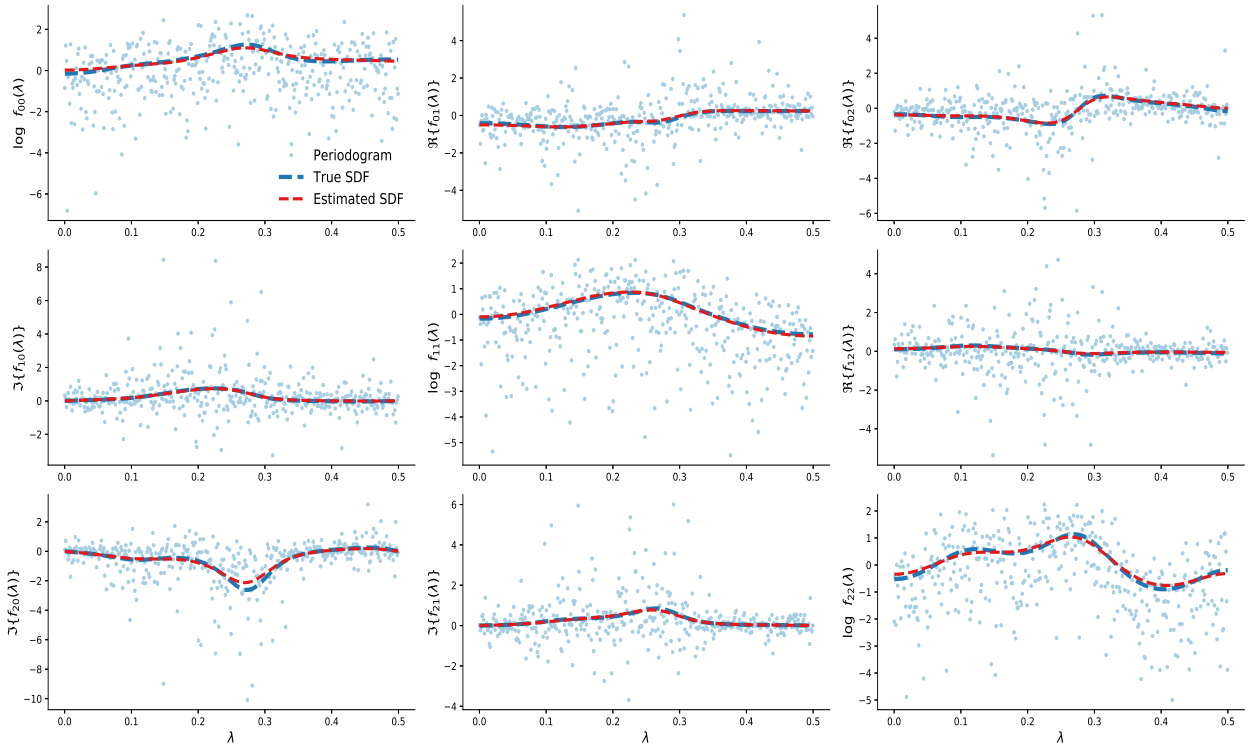

Figure 2.6: Components of the estimated spectral density function, with the true spectral density and raw periodogram values for comparison. Marginal components of the spectral density are plotted on the main diagonal. The real components of the cross-spectrum are plotted on the upper triangle, with the corresponding imaginary components on the lower triangle.

closely track this variation.

2.6 Analysis of Macaque ECoG Recordings

Electrophysiological recordings of the brain are observed in a naturally multivariate format, as a collection of waveforms corresponding to local field potentials measured by a number of electrodes extending over some area of interest. There is growing recognition that these measurements, in addition to exhibiting well-known oscillatory behavior across certain power bands, also tend to feature a non-oscillatory component with power-law behavior in the frequency domain. Evidence from a variety of experimental settings and data modalities indicates that the latter component may be a useful signal for classifying states of conscious-

ness or distinguishing between healthy and pathological resting-state behavior (Maxim et al., 2005; Timmermann et al., 2019; Wen and Liu, 2016).

In practice, these "oscillatory" and "non-oscillatory" components are identified as distinct features of the estimated spectral density function. The "oscillatory" components correspond to relatively well-separated peaks in the spectral density function, particularly in the range of approximately 1-200 Hz consistent with plausible biological neural activity. The "non-oscillatory" component refers to a broadband pattern that tends to be well-modeled by a power law as the frequency approaches zero. The model proposed in the previous section decomposes the spectral density function into short and long-range terms that can represent each of these respective behaviors. Here, we demonstrate its potential to aid modern investigations in neuroscience by reproducing and extending recent results from the spectral analysis of an ECoG dataset.

Experimental data and baseline results. The data consists of ECoG measurements from a rhesus macaque recorded during different states of consciousness (Yanagawa et al., 2013). It is published online and publicly available as part of the Project Tycho initiative (`http://www.neurotycho.org/`). We focus on the data derived from the "natural sleep" experiment, in which ECoG signals were recorded at 1kHz as the subject transitioned between two states of consciousness: sleep and awake with eyes closed (AEC). In addition to the ECoG signals themselves, the data contains temporal labels for the subject's state of consciousness. The authors defined the sleep state by the degree of spatial synchronization in slow-wave oscillations across electrodes in the ECoG array. The spatial layout of the ECoG electrodes is shown in Appendix A. Our analysis is focused on the temporal lobe, which was covered by 22 electrodes.

The data was collected and initially analyzed by Yanagawa et al. (2013), then subsequently analyzed by Wen and Liu (2016) specifically in terms of the long-range dependence of the ECoG signal. Yanagawa et al. (2013) train a support vector machine classifier on the tapered periodogram to distinguish between awake and unconscious states, then examine

the learned weights to identify biologically plausible aspects of the decision rule. Wen and Liu (2016) propose a shifting and averaging scheme for the periodogram to estimate the long-range dependent component of the spectrum. They implicitly estimate the short-range dependent component of the spectrum as the difference between this long-range dependent component and the observed (smoothed) periodogram.

Both studies analyze the ECoG data as a collection of univariate time series, averaging over electrodes within an anatomical region. They each find evidence of increased power in the alpha frequency band (8-13 Hz) during the awake state versus unconscious states. Furthermore, Wen and Liu (2016) estimate a decrease in the slope of the estimated "fractal" or "non-oscillatory" component in log-log coordinates for the AEC versus sleep condition, suggesting that long-range dependence of ECoG signals decreases in strength upon the transition from sleep to AEC.

Objective. In light of these previous findings, along with the limitations of the methods used, our data analysis has two major objectives:

1. To show that estimation of the multivariate spectrum with our model can recover the same qualitative and scientifically validated conclusions as reported previously in the literature (Yanagawa et al., 2013; Wen and Liu, 2016). In particular, this requires that the analysis simultaneously identify both a short-range dependent phenomenon in the differential behavior between AEC and sleep in the alpha band and a long-range dependent phenonmenon in the long memory decrease between sleep and AEC.

2. To demonstrate an extension beyond the methodological limits of these studies by generating an explicit estimate of the multivariate spectral density function. We use our estimate of the spectral density to visualize the differences in functional connectivity across the temporal lobe in the transition from sleep to AEC via the partial coherence.

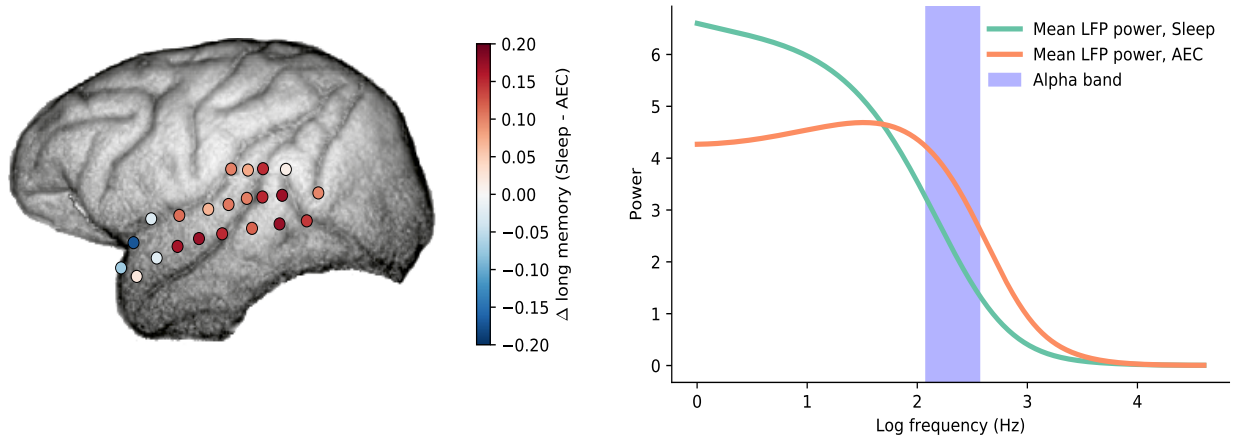

Figure 2.7: Difference in estimated long memory between sleep and awake-eyes open state for the activity recorded at each electrode (left) and marginal estimates of the spectral density, averaged across all electrodes, (right) with alpha band highlighed.

Modeling details. We isolate two segments of the data, each corresponding to a 10 second window, within a continuous temporal sequence in which the subject transitioned from sleep to AEC state. A one minute buffer is included on either side of this transition to avoid measurement of any transition effects between states.

As is standard practice in the analysis of ECoG data, we fit our models to the spectral window corresponding to the 1-200 Hz range, which is the relevant region for analyzing neural activity. We fit one model for each state, setting $L = 20$. The models are trained by batch proximal alternating minimization. Training curves and goodness of fit visualizations are reported in Appendix A.

Results. We first analyze the results in terms of the marginal characteristics of the estimated spectra. In particular, we are interested in contrasting both long-range (as parameterized by the long memory estimate) and short-range (as measured by estimated power in the alpha band 8-13 Hz) components between states. Results are plotted in Figure 2.7. We successfully reproduce the previously reported results that long memory strength decreases and alpha power increases on average in the transition from sleep to AEC states.

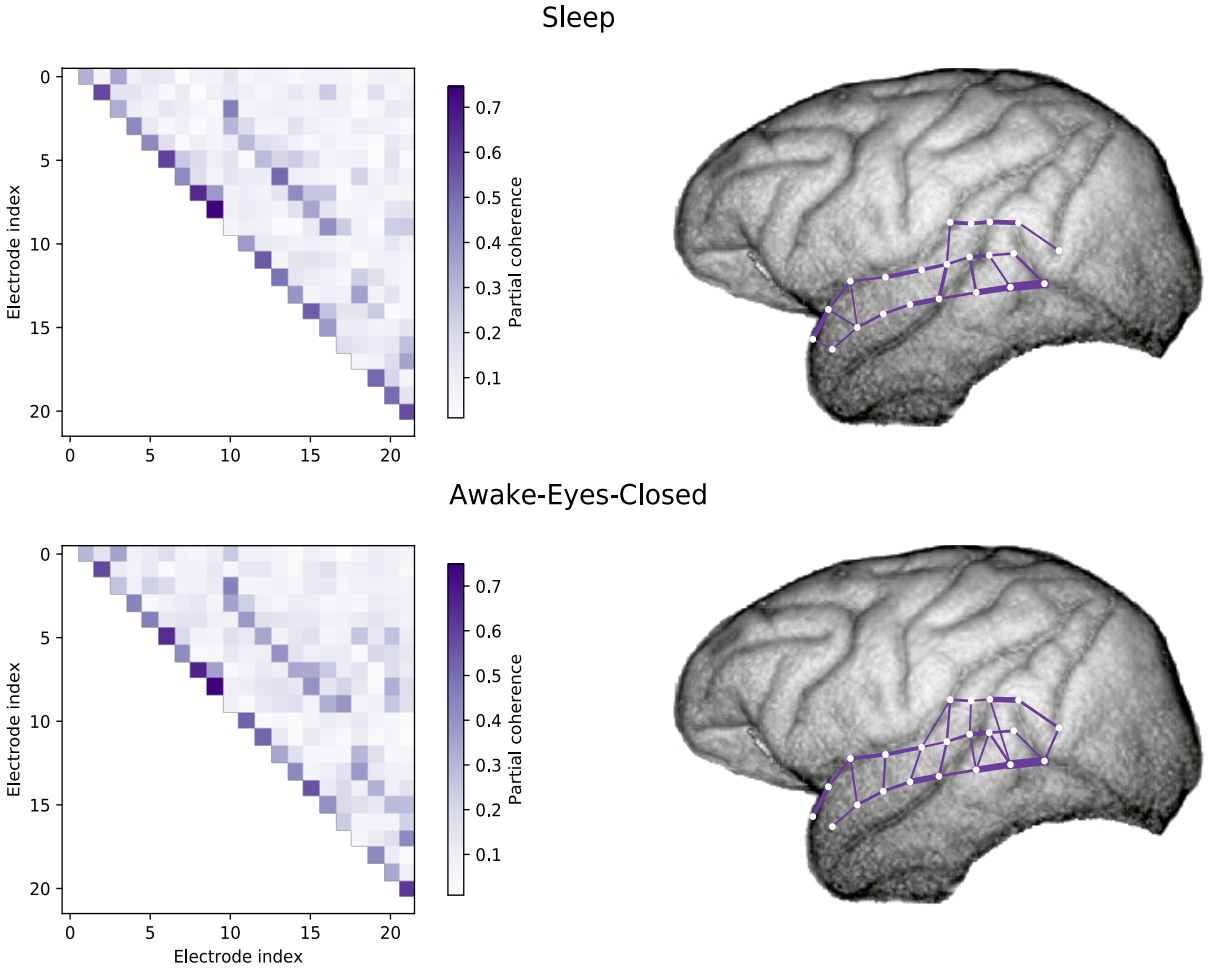

Figure 2.8: Partial coherence (left) and connectivity networks obtained from the thresholded mean partial coherence over frequencies in the alpha band 8-13 Hz (right). Results are compared for sleep (top row) and awake-eyes closed (bottom) states. An edge is plotted in the connectivity network at left if the corresponding mean partial coherence exceeds a threshold of 0.3. The network is visualized over the spatial layout of the ECoG electrodes in the right column; edges are proportional in width to the magnitude of the mean partial coherence.

Second, we go beyond marginal analyses (and consequently the method of Wen and Liu (2016)) to investigate the network behavior of the recordings in both sleep and AEC states. We use our fitted models to estimate the partial coherence in the alpha range 8-13 Hz and plot the results in Figure 2.8. We make two main observations about the results. First, in both

states, the estimated functional connectivity in the alpha band has a structure that closely reflects patterns of spatial proximity in the electrodes. This reproduces a known biological result, namely that functional connectivity reflects underlying anatomical structure (Wang et al., 2013), and thus functions as a positive sanity check. Second, functional connectivity is increased across the network in the AEC state versus sleep.

2.7 Discussion

In this chapter we have introduced a framework for evaluation of long memory in RNN models of natural language and music. Application of this framework to a variety of RNN architectures yielded evidence that these models fail to capture the long memory in the data on which they are trained. We subsequently proposed a model for the spectral density function of a multivariate long memory process, along with a penalized estimation framework to automatically select the smoothness of the real and imaginary components. We showed that the model reproduces previous findings on the marginal spectral behavior of macaque ECoG data while offering new insights in terms of the functional connectivity between the regions measured in the ECoG array.

The testing framework proposed for long memory evaluation in RNNs is easily adapted to recurrent architectures beyond those studied in this chapter. The idea of studying embedded sequences, rather than the raw input, may also be of value in extending this framework to more general data sources, such as images or graphs. At the same time, developments in deep learning for modeling natural language and music data continue apace, including the advent of non-recurrent architectures that pose challenges for our long memory analysis, which takes as its object a learned representation with explicit temporal ordering.

In particular, transformer networks (Vaswani et al., 2017) have achieved state-of-the-art performance tasks on a variety of language tasks (Radford et al., 2019; Brown et al., 2020), attaining in some cases a quality of conditionally generated text indistinguishable from human writing (Köbis and Mossink, 2021). Music modeling and generation has benefitted significantly from the incorporation of multi-scale convolutional structures, often as a sub-

stitute for recurrent hidden representations (Oord et al., 2016). It remains for future work to identify if and how the long memory perspective taken in this chapter may be extended to study these models.

Chapter 3

PREDICTION OF STIMULATION-INDUCED FUNCTIONAL CONNECTIVITY CHANGES IN LARGE SCALE NEURAL RECORDINGS

3.1 Introduction

Complex patterns of activity in the brain are coordinated across large, spatially distinct regions that contain millions of neurons. The network structure of coordinated activity at this scale has become a major focus of modern neuroscience, motivated in part by the fact that certain signatures in network-level neural activity are reliable indicators of pathological conditions. Remarkably, medical interventions involving the implantation of electrodes to provide *deep brain stimulation* have been found to have significant therapeutic effect, often when other approaches have failed (Lozano and Lipsman, 2013). While this suggests that stimulation may encourage the reorganization of cortical networks, there remains relatively little understanding of the exact mechanism underlying these results, or indeed a clear view of what factors mediate the network-level response to stimulation in general (Kringelbach et al., 2010; Saenger et al., 2017). There is thus significant motivation to investigate this phenomenon in greater detail, both from the perspective of basic science and with a view towards improved clinical technologies.

Until recently, the main challenge to this line of research lay in the inadequacy of available technology for experimentation or measurement. However, advances in bioengineering have provided solutions in the form of two complementary tools, *optogenetics* and *electrocorticography*, that together enable large-scale, high-quality recordings from neural populations while precise stimulation protocols are applied. Optogenetics (Boyden et al., 2005) renders the neuronal machinery responsible for its spiking activity sensitive to light, so that it can be

activated with high spatial and temporal precision by a laser. Electrocorticography (ECoG) offers a means to record the aggregated activities in a population of neurons via the local field potential (LFP), typically measured in microvolts, induced by their spiking behavior (Leuthardt et al., 2006). In contrast to other electrophysiological measurement techniques, ECoG signals are recorded from an electrode array placed directly on the surface of the brain, yielding relatively higher-quality signals and spatial resolutions. These technologies have been recently combined to measure network-wide response to stimulation in the primary sensorimotor cortex of the rhesus macaque, a non-human primate model species highly similar to humans (Yazdan-Shahmorad et al., 2016, 2018).

Measurement of neural activity as a multivariate time series enables formalization of the scientific investigation using signal processing and statistical techniques. In particular, the notion of connectivity between groups of neurons is extended from a literal, anatomical definition to a statistical definition, whereby groups of neurons are considered connected if they exhibit correlated activity. In computational neuroscience this concept is known as *functional connectivity* and is associated with multiple, non-equivalent mathematical formulations. Functional connectivity is commonly defined in the frequency domain, as there is particular scientific interest in understanding the behavior of neuronal activity within specific frequency bands (Fornito et al., 2016).

In this chapter, we present a framework to quantify, model, and predict stimulation-induced changes in functional connectivity observed in a novel dataset of ECoG recordings. We detail a method for signal processing by which this change is evaluated in terms of differences in the band-limited coherence between pairs of electrodes in the ECoG array. Guided by the main scientific objectives of the analysis, we propose a nonlinear additive model that primarily aims for accurate out-of-sample prediction yet remains amenable to feature-wise investigation. We detail further tools to investigate the stability and feature-wise predictive importance of the model estimates, and to quantify the similarity between the additive components of two different model estimates.

Our results show that the proposed model achieves good predictive accuracy on unseen

data. Moreover, we provide quantitative evidence addressing important scientific questions on the relative importance of various factors hypothesized to affect changes in functional connectivity, leading to novel insights not previously reported in the literature. A hierarchical penalization scheme yields component functions that are often simple and interpretable when they correspond to basic features. The framework is extended to account for potential evolution of the conditional mean after repeated stimulation.

3.2 Related Work

The problem of modeling LFP time series can be assessed both in relation to previous works in computational neuroscience and from a broader statistical perspective.

3.2.1 Network responses to stimulation and functional connectivity

The advent of ECoG recording technology has led to the development of numerous experimental platforms designed to investigate responses to stimulation over relatively large areas of the brain surface. Modeling approaches tend to be simple and closely tied to scientific hypotheses of interest. For example, Keller et al. (2018) use a linear model to show that the stimulation-induced activity at a specific location can be predicted by a small set of features describing the stimulation protocol. Yang et al. (2021) extend this perspective to a time series context, estimating a linear state space model to predict the evolution of LFP power from a controllable sequence of stimulation inputs. While there is considerable focus on aspects of the stimulation protocol as the salient features driving induced responses, deriving in part from classical models of stimulation response at the cellular level, a growing amount of evidence suggests that the network structure of the brain plays a strong mediating role at larger scales (Keller et al., 2011; Huang et al., 2019).

The linear models described above model activity at a single location on the cortical surface. A key advantage of ECoG recording, however, is that the network-wide response to stimulation is measured simultaneously, which provides the opportunity to study the change in *connections* between the regions measured by each electrode. In this context,

such "connections" do not necessarily refer to anatomical structure linking two regions but rather to the notion of *functional connectivity*, which measures the statistical association between their recorded activities. The precise statistical feature used to define functional connectivity varies somewhat in the literature, from cross-correlation in the time domain (Chu et al., 2012), to coherence in the frequency domain (Betzel et al., 2019), to directed measures such as Granger causality (Seth et al., 2015). Frequency-domain measures are particularly useful as it is often of scientific interest to separate the study of neural activity across distinct frequency bands. Band-limited ECoG coherence has been shown to change significantly after stimulation (Huang et al., 2019) and has been predicted in the absence of stimulation by a linear combination of anatomical and genetic factors (Betzel et al., 2019).

3.2.2 Stochastic models of multiple time series with structured covariance

The LFP data is a multiple time series whose components correspond to spatially distinct regions of the brain surface. Conditional independence relations among Gaussian time series components correspond to sparsity patterns in the partial autocovariance sequence or partial coherence (Dahlhaus, 2000). Frameworks for estimation of the dependence structure therefore include model selection approaches among simplified families of spectral densities (Bach and Jordan, 2004) or vector autoregressions with sparsity-inducing penalties (Davis et al., 2016). Spatially stationary models such as Gaussian Markov random fields (Rue and Held, 2005) offer an alternative approach. Abrupt changes in time series structure, for example due to the onset of stimulation, are commonly modeled by switching processes or changepoint models with stationary components.

Alternatively, a graph-valued measure of connectivity can be modeled directly as it evolves over time or in response to stimulation. Prediction of structured objects such as graphs can be formulated in the context of max-margin learning (Taskar et al., 2005), where smoothing approaches can accelerate the otherwise combinatorial optimization problem (Pillutla et al., 2018). Spatiotemporal sequences have also been predicted with neural networks combining convolutional and recurrent architectures (Shi et al., 2015).

3.2.3 Sparse additive modeling

Among related works in the neuroscience literature analyzing ECoG responses to stimulation, there is almost exclusive focus on linear modeling of the response. While these models are conceptually straightforward and may offer reasonable approximations for simple features, there is little reason to believe they adequately describe the influence of complex quantities such as genetic factors or structural aspects of the underlying network. However, many "black-box" nonlinear methods currently popular in machine learning either do not correspond to an explicit stochastic model or lack rigorous tools to quantify influence, uncertainty, or similarity on a feature-wise basis. In the context of scientific data analysis, these properties are not merely helpful but rather priorities in themselves.

Additive modeling offers a middle ground by constraining the structure of the conditional mean such that $\mathbb{E}[Y|X] = \sum_{j=1}^{p} f_j(X_j)$, where f_j, $j = 1, \ldots, p$ are univariate continuous functions (Hastie and Tibshirani, 2017; Gu, 2013). The development of additive modeling in the statistical literature grew out of interest in nonparametric smoothing techniques for regression; early algorithms for estimation included projection pursuit and backfitting (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1987). The advantage of flexible representation of the conditional mean comes at the expense of efficiency or in the worst case identifiability of the model due to *concurvity*, a nonlinear analogue of linear dependence in the design matrix (Donnell et al., 1994).

Convex penalization of the additive modeling objective enabled the use of fast iterative algorithms and extension to the high-dimensional setting (Ravikumar et al., 2009). Practically, the component functions f_j can be represented by truncation of a representation in some fixed basis of functions. Haris et al. (2019) propose a hierarchical penalty that controls the order of this representation and establish minimax convergence rates. The penalty is an example of a tree-structured regularization function, the fast computation of which is demonstrated by Jenatton et al. (2011) and implemented in the `SPAMS` toolbox.

3.3 Signal Processing and Feature Representation of ECoG LFP Time Series

The data consist of ECoG recordings collected across 33 sessions on 17 unique days, spanning a total of 9 months (Yazdan-Shahmorad et al., 2018). Experiments were conducted on two awake macaque monkeys, identified as subjects `G` and `J`. Each session consisted of exactly 5 *experimental blocks*. All blocks within a session featured both a baseline period of passive recording and a stimulation protocol involving alternating optogenetic stimulation at two laser sites in the primary sensorimotor cortex. We refer to the baseline periods as *resting state blocks* and the stimulation periods as *stimulation blocks*. Both the baseline and stimulation blocks varied in length, with the baseline periods typically 3-5 minutes and stimulation roughly 7-10 minutes.

The raw data are time series measurements of the local field potential in micro-volts (μV), recorded at a frequency of 10kHz across 96 electrodes. Within each block, the raw data can therefore be represented as a $96 \times T$ matrix with rows corresponding to electrodes and columns corresponding to time-ordered observations of the LFP. The frequency of the recordings far exceeds the maximum relevant rate of roughly 200Hz for neural activity. The time series are therefore decimated by convolution with an order 8 Chebychev type I filter and downsampling by a factor of 10.

Within each block, the data are partitioned into non-overlapping 20-second windows, corresponding to an assumption of local stationarity in the LFP on this timescale (see Fig. 3.1). We compute an estimate of the coherence between all non-faulty electrodes for each 20-second window in every block. Denote by E the number of electrodes and denote the multivariate LFP time series for a given window by the two-dimensional array $\boldsymbol{x}_{1:T} \in \mathbb{R}^{E \times T}$, which consists of the time-ordered concatenation of the LFP recordings $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{Et})^T$ for each electrode at time t. We estimate the multivariate spectral density function $S(\lambda) \in \mathbb{C}^{E \times E}$ for a the time series $\boldsymbol{x}_{1:T}$ by Welch's method (Welch, 1967), whereby a time series is partitioned into overlapping blocks, a periodogram is computed from tapered data in each block, and a final estimate is produced by averaging the periodogram ordinates at each

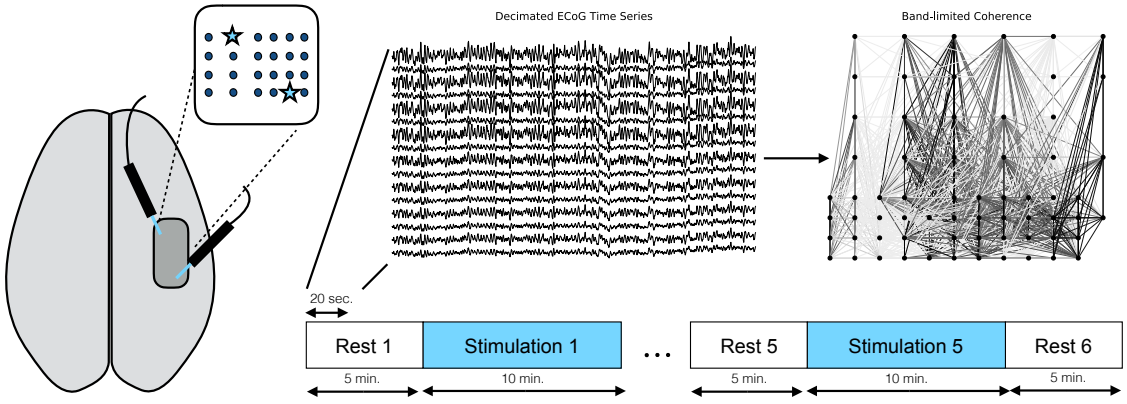

Figure 3.1: Overview of experimental session structure and ECoG signal processing. Data are collected from an electrode array on the cortical surface. Two sites are selected for laser stimulation. The stimulation protocol is applied in five blocks, which are interleaved with rest periods. The raw LFP time series is decimated and band-limited coherence is computed across nonoverlapping 20 second windows.

frequency over all blocks. We set the block length to correspond to 10 seconds of recording time and use a standard 50% overlap between blocks and Hann window.

The coherence between electrodes i and j at frequency λ is denoted as $C_{ij}(\lambda)$ and computed from the estimated spectral density via $C_{ij}(\lambda) = |S_{ij}(\lambda)|^2/S_{ii}(\lambda)S_{jj}(\lambda)$. We compute the coherence at 400 linearly spaced frequencies in the interval $(0, 200)$ Hz.

Summaries over four frequency bands, 4-7 Hz (Theta), 12-30 Hz (Beta), 30-70 Hz (Gamma), and 70-200 Hz (High Gamma), are computed by averaging the coherence estimates within each band. Thus, for each session, the raw LFP time series is transformed into a sequence of matrix-valued quantities representing the average band-limited coherence in each 20-second interval. Since the coherence is symmetric and the diagonal conveys no information on pairwise behavior of the LFP electrode signals, we retain only values in the upper triangle above the diagonal for modeling and prediction.

The stimulation-induced functional coherence change (SIFCC) is defined as the difference in mean coherence between two electrodes relative to their coherence measured during the

previous resting state block. We consider coherence changes over two timescales: changes observed during a stimulation block and changes observed between successive resting state blocks. Let $C_{ij}^{(w)}(\lambda_k)$ be the coherence between electrodes i and j at frequency λ_k in the 20-second window w, as estimated by the procedure described above. The band-limited coherence is obtained by averaging over the indices $k \in K_b$ corresponding to frequency band b: $C_{bij}^{(w)} = |K_b|^{-1} \sum_{k \in K_b} C_{bij}^{(w)}(\lambda_k)$. Within an experimental session, we denote by W_{B_ℓ}, $\ell \in \{1, 2, 3, 4, 5, 6\}$ and W_{S_ℓ}, $\ell \in \{1, 2, 3, 4, 5\}$, the collections of 20-second windows belonging to resting state block ℓ or stimulation block ℓ, respectively. Note that the additional resting state block corresponds to a final recording period after the 5th stimulation block. The electrode coherences are summarized for each of these blocks by their mean values:

$$\tilde{C}_{bij}^{(B_\ell)} = \frac{1}{|W_{B_\ell}|} \sum_{w \in W_{B_\ell}} C_{bij}^{(w)} \tag{3.1}$$

$$\tilde{C}_{bij}^{(S_\ell)} = \frac{1}{|W_{B_\ell}|} \sum_{w \in W_{S_\ell}} C_{bij}^{(w)} \tag{3.2}$$

The SIFCC during stimulation is then computed as

$$y_{bij\ell} = \tilde{C}_{bij}^{(S_\ell)} - \tilde{C}_{bij}^{(B_\ell)} \tag{3.3}$$

and the SIFCC after stimulation is

$$y'_{bij\ell} = \tilde{C}_{bij}^{(B_{\ell+1})} - \tilde{C}_{bij}^{(B_\ell)}. \tag{3.4}$$

3.3.1 Feature representation of processed data

In the next section, we adopt a nonlinear regression framework for prediction of the SIFCC. Under this approach, a single observation is denoted by the pair $(y_n, \boldsymbol{x}_n)$, $n = 1, \ldots, N$, with $y_n \in \mathbb{R}$ an SIFCC measurement and $\boldsymbol{x}_n \in \mathbb{R}^p$ the corresponding features. In switching to the single index n we indicate that the data is aggregated over all unique electrode pairs

and all experimental blocks (except when modeling the whole-session SIFCC, defined in the next section, which is not recorded per-block). The dependence on the band b remains but is dropped for notational simplicity; the same analysis is repeated for the data in each frequency band.

The features are constructed to satisfy two objectives: first, that the subsequent analysis can separate the influence of the protocol parameters from that of the network structure of the functional connectivity; second, that all information in the features is available prior to stimulation. To satisfy the first objective, we partition the features into two groups: *protocol* features, which summarize aspects of the experimental setting and protocol; and *network* features, which summarize information in the electrode coherences during baseline recording. The designation "network features" derives from consideration of the estimated band-limited coherence as the adjacency matrix of an undirected, edge-weighted graph. The protocol features describe the key parameters of the experimental framework of Yazdan-Shahmorad et al. (2018). The network features are intended to serve as simple but informative summaries of the connectivity information available in recordings of baseline activity prior to stimulation. They correspond to summary statistics computed over spatial or temporal ranges, basic quantities pertaining to the estimated spectrum, or network features that have found previous application in graph analysis (Barrat et al., 2004; Salton and McGill, 1983).

Network features

All features in this category (apart from the phase, Eq. (3.5)) are computed from the resting state coherences $C_{bij}^{(w)}$, $w \in W_{B_\ell}$. Most features in this section and the next are computed on a per-window basis, and then summarized by their mean across the total collection of windows B_ℓ in the resting state period corresponding to the given observation. Precisely, let $f : \mathbb{C}^{96\times 96} \to \mathbb{R}$ map the window spectral estimate $\{C_{bij}^{(w)}\}_{i,j=1}^{96}$ to a given feature. Then unless otherwise noted, each quantity in this section and the next is reported both in terms

of the mean feature

$$f(\tilde{\boldsymbol{C}}_b^{(B_\ell)}) = \frac{1}{|W_{B_\ell}|} \sum_{w' \in W_{B_\ell}} f(\tilde{\boldsymbol{C}}_b^{(w')}), \quad [\tilde{\boldsymbol{C}}_b^{(B_\ell)}]_{ij} = \tilde{C}_{bij}^{(B_\ell)}.$$

The spectral features are enumerated below. As the features are computed in the same way for each band, we drop the indexing by band b.

1. The `initial coherence` is given by $\tilde{C}_{ij}^{(B_\ell)}$.

2. The `L=2 path strength`. Define the vector

$$\overrightarrow{C}_i^{(\ell)} = (\tilde{C}_{i1}^{(B_\ell)}, \tilde{C}_{i2}^{(B_\ell)}, \ldots, \tilde{C}_{iN}^{(B_\ell)})^T$$

   for each electrode $i = 1, \ldots, N$ and block ℓ. Then the $L = 2$ path strength for the observation $Y_{ij\ell}$ is $\langle \overrightarrow{C}_i^{(\ell)}, \overrightarrow{C}_j^{(\ell)} \rangle$. This is an inner product, and as such a measure of similarity. It also has an interpretation from the graph perspective as the sum of weights of all length-2 paths between electrodes i and j (Onnela et al., 2005).

3. The `pair coherence to network` summarizes the average coherence between electrodes i and j and the remaining electrodes in the array:

$$\frac{\overrightarrow{C}_i^{(\ell)} + \overrightarrow{C}_j^{(\ell)}}{N}$$

   This corresponds to the normalized sum of vertex strengths of nodes i and j (Barrat et al., 2004).

4. The `pair MAD (mean absolute difference) to network` measures the average difference in connectivity between electrodes i and j to other electrodes in the network:

$$\frac{|\overrightarrow{C}_i^{(\ell)} - \overrightarrow{C}_j^{(\ell)}|}{N}.$$

5. The `mean electrode coherence to stim` is the average coherence between electrodes i and j to stimulation sites A and B. The indices a and b below refer to the electrode sites corresponding to the laser locations for a given experiment.

$$\frac{1}{4}[\tilde{C}_{ia}^{(B_\ell)} + \tilde{C}_{ib}^{(B_\ell)} + \tilde{C}_{ja}^{(B_\ell)} + \tilde{C}_{jb}^{(B_\ell)}]$$

6. The `pair electrode covariance` and `pair time covariance` capture the variability of the coherence measurements over the electrode array and over the total number of 20-second time windows in a baseline period, respectively. We define the vectors

$$C_{\text{elec}(i)}^{(w)} = (C_{i1}^{(w)}, ..., C_{iN}^{(w)})$$

and

$$C_{\text{time}(i,j)}^{(B_\ell)} = (C_{ij}^{(w_1)}, ..., C_{iN}^{(w_{|B_\ell|})}).$$

Then the pair electrode and time covariance are given by

$$\texttt{TimeCov}(i,j,\ell) = \frac{1}{N}\sum_{k=1}^{N} \text{Cov}(C_{\text{time}(i,k)}^{(B_\ell)}, C_{\text{time}(j,k)}^{(B_\ell)})$$
$$\texttt{ElecCov}(i,j,\ell) = \frac{1}{|B_\ell|}\sum_{w\in B_\ell} \text{Cov}(C_{\text{elec}(i)}^{(w)}, C_{\text{elec}(j)}^{(w)}).$$

7. Finally, the `phase` feature captures spectral information not available in the coherence. Writing the $(i,j)^{th}$ cross-spectral component computed within window w as

$$S_{ij}^{(w)}(\lambda_k) = |S_{ij}^{(w)}(\lambda_k)|e^{i\theta_{ijk}^{(w)}},$$

the phase feature θ_{ij} is given by the average

$$\tilde{\theta}_{ij}^{(B_\ell)} = \frac{1}{K|W_{B_\ell}|} \sum_{w \in W_{B_\ell}} \sum_{k=1}^{K} \theta_{ijk}^{(w)}. \tag{3.5}$$

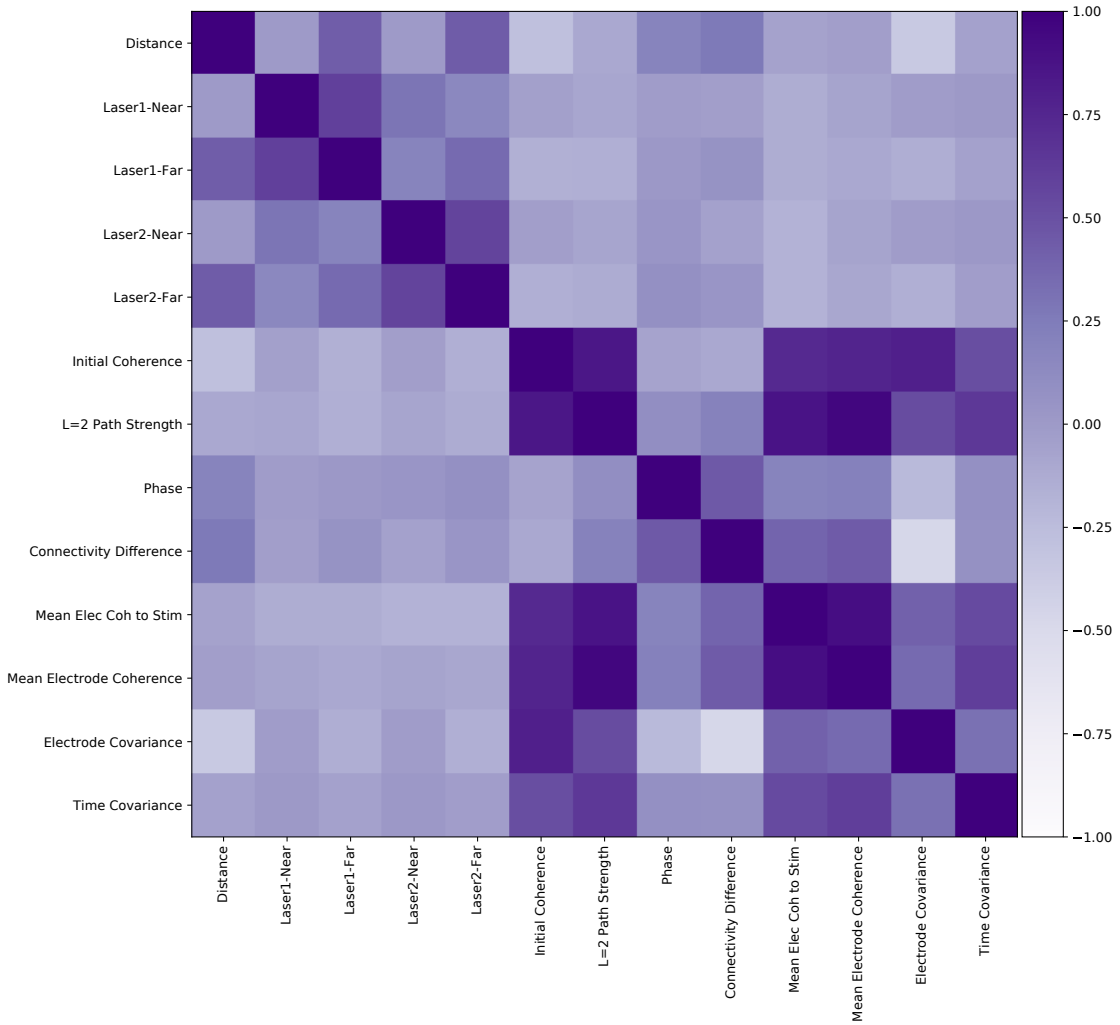

Figure 3.2: Correlation heatmap between the real-valued features of the high gamma band of the ECoG coherence dataset. Correlations within the protocol and network feature groups are higher on average than correlations between the feature groups.

Protocol features

The protocol features first include the categorical measurements `Subject`, `Region`, `Delay`, and `Block number`. `Subject` is a binary variable encoding the identity of the experimental subject. `Region` indicates whether the two electrodes corresponding to a given measurement are both in region M1, both in region S1, or if one is in each region. `Delay` encodes three levels of time-delay in the pulse of the paired laser stimulation: 10, 30, or 100 ms. While the delay parameter in the stimulation protocol is real-valued, in the context of our data it is only measured at three distinct settings. We therefore choose to estimate an effect for each setting individually rather than to estimate a nonlinear function of the delay given observations at only three points. `Block number` indicates the time-order position of the experimental block in which the observation was recorded. In the "block analysis" regression design, we remove this feature and instead allow all feature mappings to vary with the block number, thus investigating the predictive impact of allowing the entire model to evolve dynamically over the discrete time-stages of an experimental session. In Appendix B we provide summaries of the data by the level of each categorical feature.

Physical distances between the relevant components of the stimulation-recording setup are incorporated in five additional, real-valued features. `Distance` measures the distance between the two electrodes. The four remaining features, `Laser1-to-Nearest`, `Laser1-to-Farthest`, `Laser2-to-Nearest`, and `Laser2-to-Farthest`, encode distances between the electrodes and the locations of the two lasers.

3.4 Additive Modeling of Stimulus-Induced Functional Connectivity Changes

Let $(y_i, \boldsymbol{x}_i), i \in \{1, \ldots, N\}$ denote a single observation in the data, so that $y_i \in \mathbb{R}$ represents the change in coherence for a given electrode pair and experimental block, while $\boldsymbol{x}_i \in \mathbb{R}^p$ represents the corresponding protocol and graph features defined above.

We model the relationship between features and response as

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i. \tag{3.6}$$

The model consists of a fixed intercept β_0, p nonlinear functions f_j controlling the impact of each feature $j = 1, ..., p$ on the response, and an error term ε_i.

The utility of the model derives from the nonlinearity of the component functions and the additive procedure through which they are combined to yield a prediction for the change in coherence. The additive structure of the model allows us to visualize and interpret individual feature-response relationships. For each feature $j = 1, ..., p$, this is represented by the component function f_j. The nonlinearity of these component functions enables the identification of more complex relationships between coherence changes and the experimental features than can be captured by a linear model. Linearity is a strong modeling assumption that is both difficult to justify scientifically and, as we show in the results, significantly harmful in terms of predictive accuracy.

Each nonlinear component function is itself modeled as a sum of polynomial basis functions, up to maximum order K:

$$f_j(x) = \beta_{j1}x + \beta_{j2}x^2 + \ldots + \beta_{jK}x^K \tag{3.7}$$

Automatic order selection. We leverage the recent proposal of Haris et al. (2019) to automatically select the order of each component function in the model. This approach augments the standard square loss with a penalty term $\Omega_j(\beta_j)$ for the coefficient vector $\beta_j \in \mathbb{R}^K$ corresponding to each component j, defined via

$$\Omega_j(\beta_j) = \sum_{k=1}^{K} w_k \left(\sum_{i=1}^{N} (\beta_{jk}x_{ij}^k)^2 + ... + (\beta_{jK}x_{ij}^K)^2 \right)^{\frac{1}{2}} \tag{3.8}$$

The penalty induces hierarchical sparsity in the parameter estimates $\beta_{j1}, ..., \beta_{jK}$ while shrinking the magnitudes of the non-sparse components towards zero. Hierarchical sparsity guarantees that if an estimated coefficient $\hat{\beta}_{jk} = 0$, then all higher-order coefficients for that component $\hat{\beta}_{jk'} = 0$, with $k < k' \leq K$; this is equivalent to selection of an order $k - 1$ representation for the component function f_j.

Handling of categorical features. All categorical features are dummy-coded such that the C categories for each feature are represented by $C - 1$ indicator variables. Polynomial expansion of these variables generates identical columns and thus rank-deficiency in the design matrix, which leads to instability in the optimization routine. This problem is avoided by truncating the polynomial expansion of all categorical variables to order 1 rather than order K.

Objective function. Let $\Psi_K \in \mathbb{R}^{N \times (Kp+1)}$ denote the design matrix (with intercept) induced by this choice of basis and maximum order, with $\Psi_K^j \in \mathbb{R}^{N \times K}$ the columns corresponding to feature j. The estimation problem is given by

$$\hat{\beta}_0, \hat{\beta}_1^{\text{hier}}, \ldots, \hat{\beta}_p^{\text{hier}} = \underset{\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}^K}{\arg\min} \frac{1}{2N} \left\| y - \beta_0 - \sum_{j=1}^{p} \Psi_K^j \beta_j \right\|_2^2 + \text{Pen}_{\lambda,\alpha}(\beta), \tag{3.9}$$

with

$$\text{Pen}_{\lambda,\alpha}(\beta) = \lambda\alpha \sum_{j=1}^{p} \Omega_j(\beta_j) + \lambda(1-\alpha) N^{-1/2} \sum_{j=1}^{p} \left\| \Psi_K^j \beta_j \right\|_2. \tag{3.10}$$

The penalty is a convex combination of $\sum_{j=1}^{p} \Omega_j(\beta_j)$ and a group lasso penalty on the coefficients of each additive component. The overall degree of penalization is controlled by λ, while α controls the tradeoff between these penalty terms. It is straightforward to see that the objective function is convex in the parameters of the nonlinear additive model. Numerically, we solve Eq. (3.9) by proximal gradient descent (Bertsekas, 2015, §6.3).

Experimental designs. We apply the nonlinear modeling framework to three different configurations of the design matrix, each corresponding to a specific hypothesis explored in the results.

Full data: First we investigate the performance of the nonlinear model on the full data, which includes all experimental blocks from all sessions. This yields 481505 observations of 17 features, expanded to 22 features after dummy-coding of the categorical variables `Subject`, `Delay`, `Region`, and `Block number`.

Block interactions: In the full data analysis, the influence of the block number on the prediction is restricted to a constant shift. We subsequently investigate the impact of allowing the shape of the component functions to vary with the block number. This constitutes an interaction design, whereby the categorical feature `Block number` is removed, and the remaining features are each augmented with interaction terms for binary indicators that denote whether the measurement was made in each of blocks 2, 3, 4, and 5. The number of observations remains 481505, while the number of features expands to 94.

Whole session: Finally, we investigate whether the nonlinear model can predict the net coherence change for a given pair of electrodes across an entire session consisting of 5 consecutive experimental blocks. Here, the features correspond to the protocol and graph features as calculated in the baseline period before the first experimental block. As the net coherence change is calculated per-session and not per-block, the design matrix is reduced to 96301 observations of 16 features. The decrease from 17 features in the full setting is due to the removal of the `Block number` feature, which doesn't apply for whole-session observations. Whole session results for each frequency band are reported in Appendix B; for the remainder of this chapter, we will focus on the full data and block interaction designs.

Selection of regularization parameters. The penalized loss function requires specification of two regularization parameters $\lambda \in \mathbb{R}_+$ and $\alpha \in [0, 1]$, which control the overall regularization strength and tradeoff between terms inducing hierarchical and component-wise sparsity, respectively. The parameter α is selected over the range from 0 to 1, inclusive,

in increments of 0.1. The parameter λ is selected over 100 log-linearly-spaced values between $\lambda_{\max} \times 10^{-5}$ and $\lambda_{\max}$.

The value $\lambda_{\max}$ is selected such that all estimated components are equal to zero for every value of investigated α. It is obtained by backtracking line search from an initial value λ_0, defined as any value of $\lambda > 0$ such that $\widehat{\boldsymbol{\beta}} = 0$ for all grid values of α. Starting from $\lambda = \lambda_0$, we compute $\widehat{\boldsymbol{\beta}}$ for each α; if in each case $\widehat{\boldsymbol{\beta}} = 0$, we reduce λ by a factor of 1.2 and repeat the procedure. We find that $\lambda_0 = 0.1$ suffices for all designs and frequency bands.

We select the pair (α^*, λ^*) by first finding the (α, λ) pair minimizing the average R^2 on the validation set over 5-fold cross-validation on the training data. The details of this procedure are provided in Alg. 2. We select α^* as the α-coordinate of this pair, and select λ^* as the largest value of λ such that the mean validation R^2 of (α^*, λ) is within one standard error of the mean validation R^2 at (α^*, λ^*). This mimics the "one standard error" strategy for one-dimensional cross-validation (Friedman et al., 2001) and represents a conservative approach to regularization corresponding to our preference for smoother component estimates.

Algorithm 2 Hierarchically penalized estimation with k-fold cross-validation.

Require: Training data $(\boldsymbol{y}, \boldsymbol{x}) \in \mathbb{R}^{n,p+1}$
Require: Model order upper bound K
Require: Cross-validation folds C
Require: Regularization parameter grid (λ_ℓ, α_m), $\ell = 1, \ldots, L$, $m = 1, \ldots, M$
for $c = 1, \ldots, C$ **do**
    Define c^{th} fold as $(\boldsymbol{x}^{\text{valid}}, \boldsymbol{y}^{\text{valid}})$, remainder $(\boldsymbol{x}^{\text{train}}, \boldsymbol{y}^{\text{train}})$
    for $\ell = 1, \ldots, L$ **do**
        for $m = 1, \ldots, M$ **do**
            Set $(\lambda, \alpha) = (\lambda_\ell, \alpha_m)$
            Estimate parameters $\hat{\boldsymbol{\beta}}_1^{(c)}, \ldots, \hat{\boldsymbol{\beta}}_p^{(c)}$ from $(\boldsymbol{x}^{\text{train}}, \boldsymbol{y}^{\text{train}})$ as in (3.9)
            Evaluate model prediction performance on $(\boldsymbol{x}^{\text{valid}}, \boldsymbol{y}^{\text{valid}})$
        end for
    end for
end for
Select $(\lambda, \alpha) = (\lambda_{\text{CV}}^*, \alpha_{\text{CV}}^*)$ via the "one standard error" rule
Compute $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$ by solving (3.9) over the entire training data $(\boldsymbol{y}, \boldsymbol{x})$
Return $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_p$

Component function stability and feature importance via resampling

The main advantage of an additive modeling approach is that it preserves feature-wise analysis of the estimated model even as we extend to a nonlinear conditional mean. Our collaborators are interested in two further aspects of this feature-wise analysis: equipping the parameter estimates, and consequently the estimated nonlinear component functions themselves, with a notion of uncertainty; and ranking the features in terms of their importance for prediction.

We address the first question through subsampling of the training data. The rationale for this procedure derives from the analysis of Meinshausen and Bühlmann (2010), who show that subsampling may have distinct advantages with respect to model selection in the high-dimensional regime. For $S = 100$ trials, we generate a dataset $\{y_s, X_s\}_{s=1}^{S}$ by subsampling $\lfloor N/2 \rfloor$ observations uniformly at random and compute point estimates by solving Eq. (3.9) on this data. Stability of the estimated component functions is assessed by pointwise visualization of upper and lower quantiles of the S component functions.

To measure feature importance, we solve the reduced regression problem

$$\hat{\beta}_0^{(-j)}, \ldots, \hat{\beta}_p^{(-j)} = \underset{\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}^K}{\arg\min} \frac{1}{2N} \left\| y - \beta_0 - \sum_{j' \neq j} \Psi_K^{j'} \beta_{j'} \right\|_2^2 + \mathrm{Pen}_{\lambda,\alpha}^{(-j)}(\beta), \tag{3.11}$$

with

$$\mathrm{Pen}_{\lambda,\alpha}^{(-j)}(\beta) = \lambda\alpha \sum_{j' \neq j} \Omega_{j'}(\beta_{j'}) + \lambda(1-\alpha) N^{-1/2} \sum_{j' \neq j} \left\| \Psi_K^{j'} \beta_{j'} \right\|_2 .$$

The procedure is repeated for each feature $j = 1, ..., p$. The importance of feature j is calculated as the change in square loss on the training data,

$$\left\| Y_{\text{train}} - \Psi_{\text{train}} \widehat{\boldsymbol{\beta}} \right\|_2^2 - \left\| Y_{\text{train}} - \Psi_{\text{train}}^{(-j)} \widehat{\boldsymbol{\beta}}^{(-j)} \right\|_2^2 ,$$

between the full model and the model estimated without feature j. Furthermore, since this procedure generalizes to any partition of the features, it is also used to investigate the relative

predictive importance of the network features versus the protocol and distance features. We apply the same subsampling procedure as above to estimate the uncertainty in our estimates of the feature (or feature group) importance.

Comparison of component functions across frequency bands

Let $\hat{f}_j^a$ and $\hat{f}_j^b$ be the estimated feature mappings for feature j on the data corresponding to frequency bands a and b, respectively. We introduce a quantitative measure of their similarity,

$$s_j^{ab} = \frac{\langle f_j^a, f_j^b \rangle}{\left\| f_j^a \right\| \left\| f_j^b \right\|}. \tag{3.12}$$

The quantity s_j^{ab} is the cosine similarity of the estimated feature mappings, considered as elements of a common Hilbert space of functions. By definition, $s_j^{ab} \in [-1, 1]$. Similarity increases as s_j^{ab} approaches 1 or -1 (which indicates f_j^a and $-f_j^b$ are identical up to a positive constant) and is minimized at 0.

The inner product $\langle f_j^a, f_j^b \rangle = \int_{\mathcal{X}_j} f_j^a(x) f_j^b(x) dx$ and norm $\|f_j\| = \langle f_j, f_j \rangle$ are given by integrals whose domain $\mathcal{X}_j$ depends on the feature j. For real-valued features, we take $\mathcal{X}_j$ to be the interval $[-5, 5]$, which after standardization corresponds to the range of all observed measurements within 5 standard deviations of the mean. Due to the polynomial representation of the nonlinear feature mappings, these can be computed exactly.

3.5 Experimental Results

We evaluate the method in terms of the scientific objectives stated at the beginning of this chapter. The primary goal is to obtain good prediction accuracy on held-out test data, a proxy for the clinical prediction setting that motivates this line of research. This is balanced by the need to evaluate specific scientific hypotheses regarding the factors that contribute to SIFCC, their relative importance, and their variation across frequency bands or prediction settings. In contrast to typical black-box nonlinear modeling of neuroscience data, we further

equip our estimates of component functions and feature importances with a notion of stability via resampling.

As the dataset is unique and obtained from a novel experimental method, we must also provide a means of calibrating expectations with respect to prediction performance. We address this in two ways: first, by estimating a linear model as a baseline for prediction performance; second, by reference to recent analyses involving prediction of ECoG data in similar settings. The linear model offers a direct comparison for our results and establishes a minimum condition for success.

For recent analyses, we refer to Yang et al. (2021) and Betzel et al. (2019). In Yang et al. (2021), the authors use a linear state-space model to predict the response to stimulation in a univariate time series corresponding to LFP power measurements; thus whereas we predict the *edges* of a coherence network, they predict the (non-normalized) nodes themselves. Their model achieves a test R^2 of 0.58 on step-ahead prediction corresponding to a $0.5 - 1$s interval, but performance degrades rapidly ($R^2 = 0.20$) over timescales on the order of our experimental blocks. Betzel et al. (2019) model ECoG coherence in the absence of stimulation, achieving prediction accuracy of up to $R^2 = 0.43$ in a linear model that combines anatomical, distance, and genetic predictors. While neither study's results are directly comparable to our own, they at least provide some relevant context: Yang et al. (2021) suggest that prediction of the response to stimulation over longer timescales is expected to be more challenging, while Betzel et al. (2019), who predicts ECoG coherence in a simpler setting (i.e. without stimulation) and from a broader set of predictors, establish an optimistic standard for coherence prediction in our setting.

3.5.1 Experimental details

In all experiments we set the polynomial order upper bound $K = 10$ and the lower bound for the regularization parameter $\lambda_{\min} = \lambda_{\max}/10^5$. We provide evidence in Appendix B that these settings are reasonable for our data, as the order and hyperparameter selection procedures automatically select values in the interior of their respective ranges. The training

set is obtained by selecting 70% of the data uniformly at random; the remainder is withheld as the test set. The test set is held fixed across resampling trials, that is, resampled datasets are drawn from the training set only. The real-valued features are standardized to have zero mean and unit standard deviation prior to estimation of the model. The linear transformation corresponding to this standardization is computed from the training set only, and applied to the test set during prediction.

3.5.2 Results

Additive modeling with hierarchical penalization outperforms the linear baseline. We begin by comparing prediction results for the additive model against the linear baseline; these are summarized in Table 3.1. The nonlinear model outperforms the linear model in each frequency band. Moreover, the nonlinear model fit to network features only achieves a relatively larger fraction of the full-model prediction accuracy than in the linear case, suggesting that nonlinear modeling is particularly advantageous for deriving useful prediction rules from these features.

These results are not merely a consequence of estimation within a strictly larger class of functions, as they report error on held-out test data while estimation and hyperparameter selection are performed entirely on the test set. While they are not particularly surprising in view of the well-known advantages of nonlinear modeling for prediction in other applications, this has not yet been demonstrated for ECoG data, where linear modeling remains the dominant paradigm and negative results for nonlinear extensions have been recently reported (Yang et al., 2021).

Network-derived features are important for prediction. The results in Table 3.1 also show that the set of network-derived features are particularly important for prediction. This result is scientifically important as it provides strong evidence that the "baseline" network structure prior to stimulation mediates the response to stimulation. The relative inadequacy of the protocol-only models shows that the traditionally emphasized features

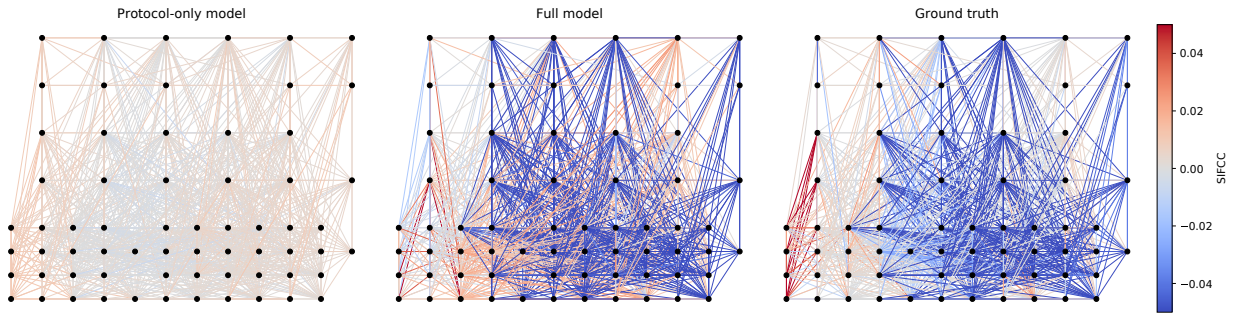

Figure 3.3: Comparison of nonlinear model predictions of SIFCC during stimulation with and without network features for a single representative experimental block. Black dots indicate the locations of electrodes in the ECoG array. Model predictions are plotted as edges between the electrodes in the left and center panels. The true SIFCC for the experimental block is plotted on the right.

of stimulation studies in neuroscience are insufficient on their own to explain the brain's response to stimulation at this scale. We emphasize this point in Figure 3.3 by contrasting protocol-only and full nonlinear model predictions across the entire ECoG array in Block 1 of the session `MonkeyG_20150925_Session1_S1`. Whereas the protocol-only model largely fails to capture meaningful variation in the SIFCC, incorporating network features yields expressive and network-wide predictions that more closely correspond to the ground truth.

Beyond demonstrating the aggregate importance of the network features, we apply the proposed methods to study the importance, stability, and similarity of the individual com-

Table 3.1: Prediction performance (R^2) of linear baseline and the proposed nonlinear additive model on held-out test data.

	Linear			Nonlinear		
Band	Full model	Network-only	Protocol-only	Full model	Network-only	Protocol-only
Theta	0.164	0.059	0.027	0.240	0.139	0.045
Beta	0.104	0.034	0.013	0.130	0.092	0.057
Gamma	0.157	0.043	0.020	0.221	0.162	0.026
High gamma	0.244	0.074	0.014	0.323	0.255	0.045

ponent functions in the estimated nonlinear models; results are presented in Figure 3.4. Consistent with the results for protocol-only and network-only models, we note generally larger importances for the network features. No individual network feature dominates, suggesting that it is really the full collection of these features providing relevant, if possibly somewhat redundant, information with respect to SIFCC. The single most important feature appears to be the Subject ID, consistent with the significant inter-subject variation observed broadly in neuroscience.

In the middle row of Figure 3.4, we show the estimated component functions for five of these features. The learned component functions appear generally stable. Network features tend to have more complex functions, while simple features such as the distance have correspondingly low-order components. The model reproduces some basic expected relationships, such as the smooth decay of the SIFCC with the distance between the two electrodes. Finally, we compare the similarity of the component functions across different frequency bands (Fig. 3.4, bottom row). We show the average similarity over all components in addition to group averages for the protocol and graph features. Overall, the estimated components are largely similar according to our measure. Disaggregating over protocol and network features reveals that the protocol features are highly similar, while there is greater variation in network features. Adjacent frequency bands tend to have learned more similar component functions than more distant bands.

Block-interaction modeling provides evidence for a time-dynamic conditional mean. Properly considered, the SIFCC is a graph-structured time series whose mean conditional on the protocol and baseline network features could reasonably be expected to evolve over the course of repeated stimulation periods in an experimental session. The method pursued thus far does not allow for variation in this conditional mean beyond the block-specific intercept estimated by the protocol feature `block number`. On the other hand, time series modeling of this high-dimensional quantity at the frequency of its recording would involve significant modeling effort to capture structure highly unlikely to be relevant for prediction

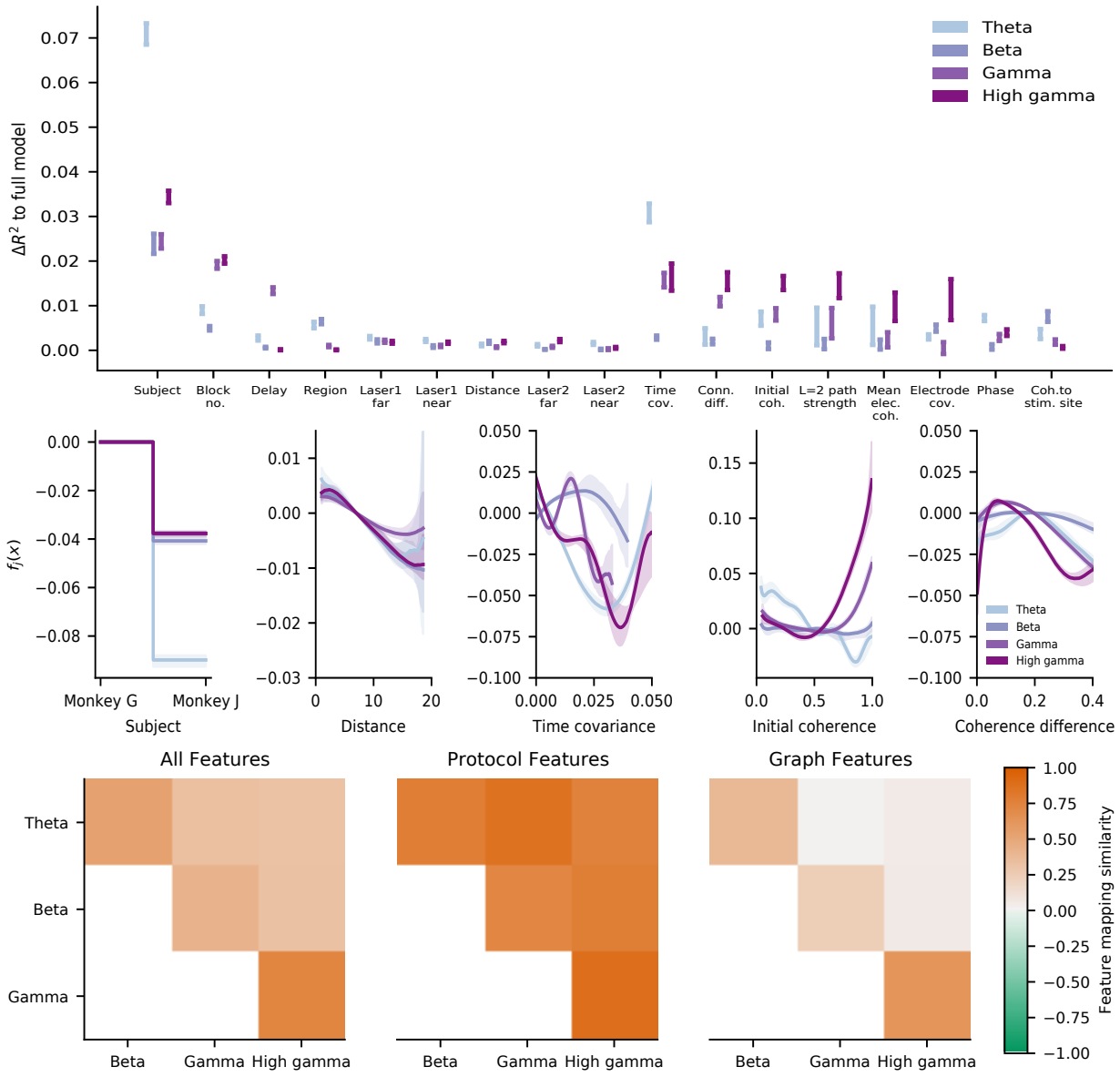

Figure 3.4: Feature-wise investigation of the nonlinear additive model. Top row: Feature importances, clustered by group (protocol or network) and sorted by median. Intervals span the $2.5 - 97.5^{th}$ quantiles of the resampling distribution. Middle row: Example component functions for each frequency band. Bands indicate the pointwise $2.5 - 97.5^{th}$ quantiles of the resampled component functions. Bottom row: Average feature similarities across bands, and average feature similarities within the protocol and network groups of features.

over the timescales of interest in this work, as seen for example in the poor forward prediction results obtained in a relatively simpler context by Yang et al. (2021).

Here we consider an intermediate solution corresponding to a coarse discretization of the experimental session into five time steps consisting of successive baseline and stimulation blocks, and we extend the model to estimate a distinct nonlinear additive conditional mean per block. As the baseline coherence is available to the model as the `initial coherence` feature, the model can be thought of as a nonlinear time-inhomogeneous autoregression with exogenous variables. In practice, it is estimated by extending the polynomial design matrix to include interaction terms for the block number.

Results for this modeling approach are shown in Figure 3.5. The top row shows prediction accuracy on the test set by block number. Prediction accuracy is uniformly improved over the static model and relatively more consistent. The significant variability in prediction error of the static model across blocks suggests that a more expressive model would be more suitable to capture the time variation of the conditional mean. Meanwhile, the hierarchical penalization method extends naturally to this setting by discouraging block-variation of the component functions unless they significantly contribute to improving the validation error. We plot the estimated block-varying component functions for two features in the second and third rows of Figure 3.5. Finally, we compute the average component function similarities between each pair of frequency bands over the five blocks. The result, shown in the bottom row of Figure 3.5, indicates that while the SIFCC conditional mean is relatively similar across all frequency bands at the start of a session, these similarities decrease with repeated stimulation. As in the static modeling case, adjacent frequency bands continue to demonstrate higher similarity than those farther apart.

SIFCC can be accurately predicted after stimulation has ended. We consider nonlinear additive modeling of the SIFCC over longer timescales. Whereas the SIFCC between electrodes i and j in frequency band b and block ℓ was originally defined as $y_{bij\ell} = \tilde{C}_{bij}^{(S_\ell)} - \tilde{C}_{bij}^{(B_\ell)}$, we further consider modeling the response $y'_{bij\ell} = \tilde{C}_{bij}^{(B_{\ell+1})} - \tilde{C}_{bij}^{(B_\ell)}$.

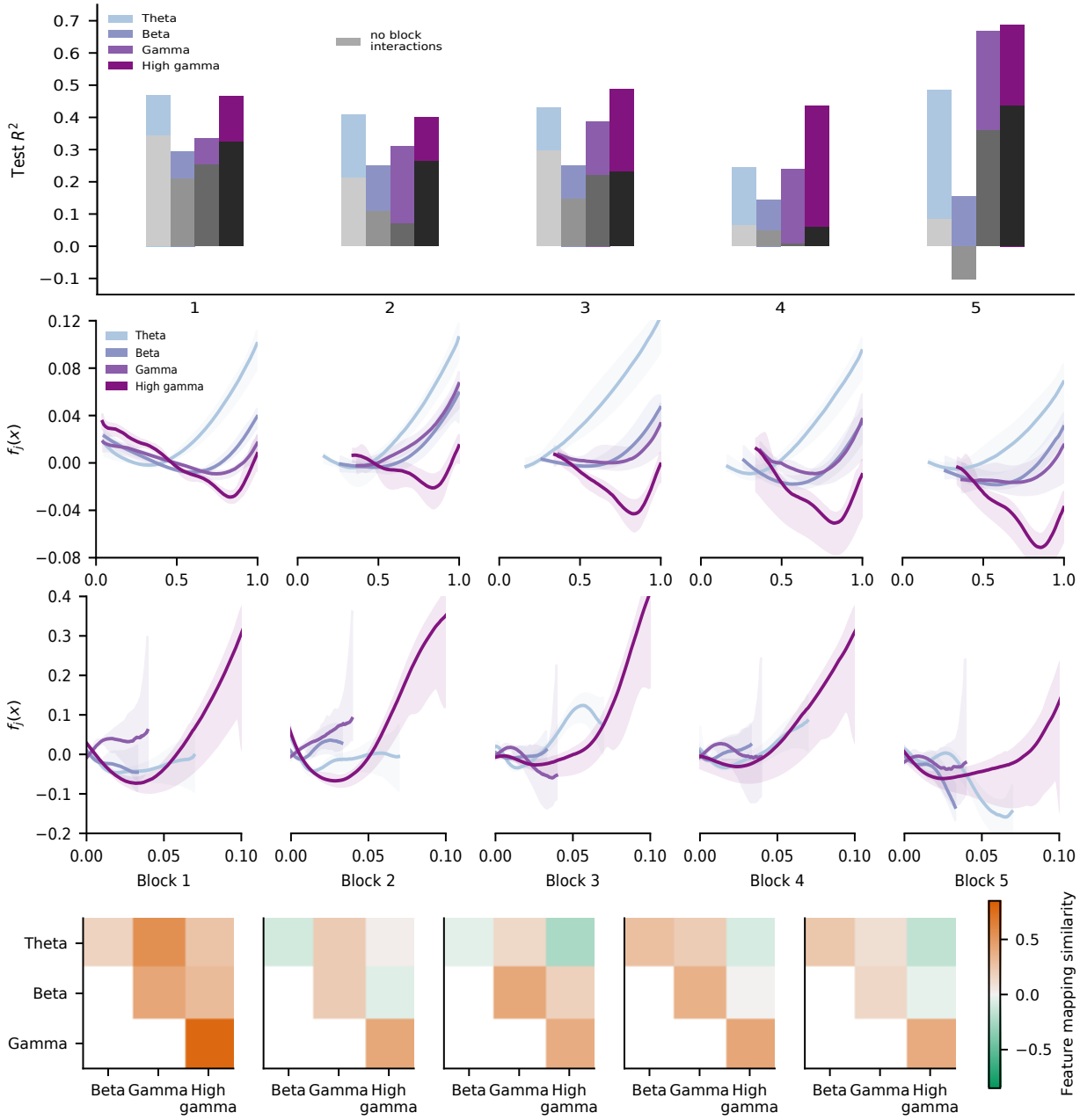

Figure 3.5: Results for nonlinear additive modeling with block interactions. Top row: Prediction accuracy of the block-interaction and static models by frequency band and block number. Second and third rows: Variation over blocks of the component functions estimated for the `initial coherence` and `time covariance` features. Bottom row: Average component similarities by block.

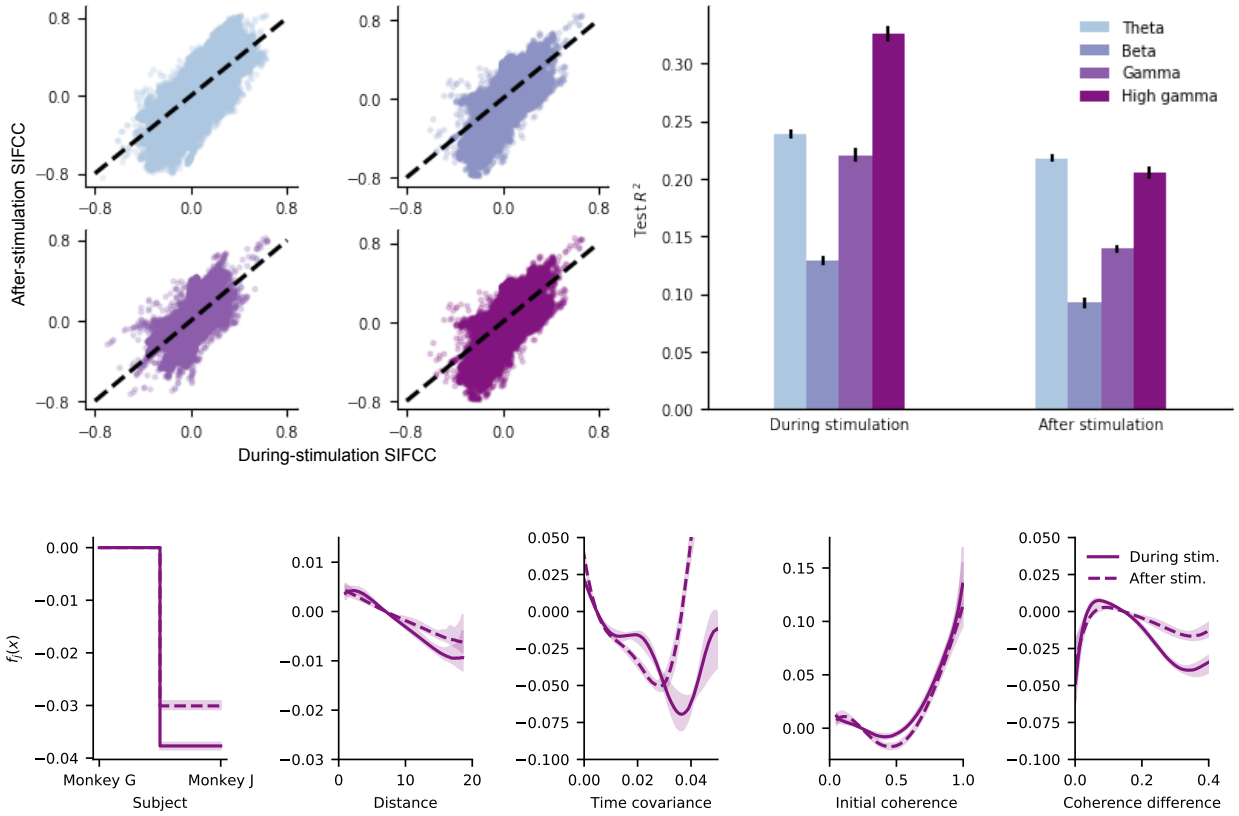

Figure 3.6: Results for prediction of SIFCC computed during versus after stimulation. Top left: Scatter plots of during and after SIFCC in each frequency band. Top right: Comparison of nonlinear prediction results on the test set. Bottom row: Example component functions in the high gamma band.

This corresponds to the coherence change between two successive baseline periods, rather than the coherence change between stimulation and baseline, and as such represents a significantly longer timescale over which to predict the evolution of functional connectivity. We distinguish between $y_{bij\ell}$ and $y'_{bij\ell}$ in this section by referring to them as the *during stimulation* and *after stimulation* SIFCC, respectively.

We estimate the nonlinear additive model using the same penalization and cross-validation framework as above. We find that the after simulation SIFCC can still be predicted with relative accuracy on held-out data, particularly in the theta and high gamma bands, though

in each band the performance is diminished with respect to the shorter-timescale during stimulation SIFCC prediction task; results are plotted in the top row of Figure 3.6. It is of scientific interest to understand what factors contribute to the changes in functional connectivity over different timescales. We thus compare the estimated component functions of models fit to both versions of SIFCC for a given frequency band; examples corresponding to the high gamma band are plotted in the bottom row of Figure 3.6. The close correspondence of these component functions shows that the models not only achieve similar prediction accuracy, but also that they do so by means of an estimate for the conditional mean in which each feature plays a very similar role. This suggests that the underlying mechanisms between SIFCC at these timescales may be largely similar.

Concurvity of additive features in the high gamma band. As a final and complementary analysis, we implement the additive principal component (APC) method of Donnell et al. (1994) to investigate nonlinear dependence in the regression features. This method is concerned with identifying *concurvity* in the features, which extends the usual notion of linear dependence to the additive setting. Exact concurvity corresponds to the existence of component functions $\phi_j \in \mathcal{H}_j, j = 1, \ldots, p$, with $\mathcal{H}_j$ a closed subspace of centered L_2 variables (e.g. the linear span of some basis functions), such that

$$\sum_{j=1}^{p} \phi_j(X_j) = 0.$$

In this case, the data lies on an additive manifold and the model proposed in Eq. 3.6 is not identifiable, as for any $f_j \in \mathcal{H}_j, j = 1, \ldots, p$, we have

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon = \beta_0 + \sum_{j=1}^{p} [f_j(X_j) + \phi_j(X_j)] + \varepsilon.$$

Concurvity is diagnosed by investigation of the smallest additive principal component of

the training data, which is given by

$$\begin{aligned} \text{argmin}_{\Phi \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_d} \quad & \text{Var}\left(\sum \phi_j(X_j)\right) \\ \text{s.t.} \quad & \sum_{j=1}^{p} \text{Var}(\phi_j(X_j)) = 1, \end{aligned}$$

where $\Phi = (\phi_1, \ldots, \phi_p)$. Subsequent APCs $\Phi^{(k)}$ are subject to the additional orthogonality condition

$$\langle \Phi^{(\ell)}, \Phi^{(k)} \rangle = \sum_{j=1}^{p} \text{Cov}\left(\phi_j^{(\ell)}(X_j), \phi_j^{(k)}(X_j)\right) = 0, \quad \forall \ell < k.$$

Exact concurvity obtains (up to a shift in the intercept term) when the eigenvalue $\text{Var}\left(\sum \phi_i(X_i)\right) = 0$, while *approximate* concurvity indicates that caution should be exercised in the interpretation of nonlinear component functions estimated from these features. We compute the complete set of additive principal components and their corresponding eigenvalues for the additive model features on the training data for the high gamma band. The minimum eigenvalue is $\lambda_{\min} \approx 1 \times 10^{-4}$ and significantly separated from the second-smallest eigenvalue at 4×10^{-3}. Investigation of the APC corresponding to the minimum eigenvalue reveals approximate concurvity among two main features, `L=2 path strength` and `mean coherence to network`. Results are plotted in Figure 3.7.

The additive principal component analysis provides a cautionary perspective against overinterpretation of the specific form of the estimated component functions in the additive model of SIFCC, particularly for the `L=2 path strength` and `pair coherence to network` features. Aspects of the foregoing analysis related to prediction, including the improvement in prediction quality on the held-out set and demonstration of the network feature importance, are not affected. While convex penalization of the additive modeling objective function guarantees the existence of a unique solution even under exact concurvity of the features, the concurvity analysis may be useful in clarifying nonlinear dependence relations between features and in guiding future iterations of feature development.

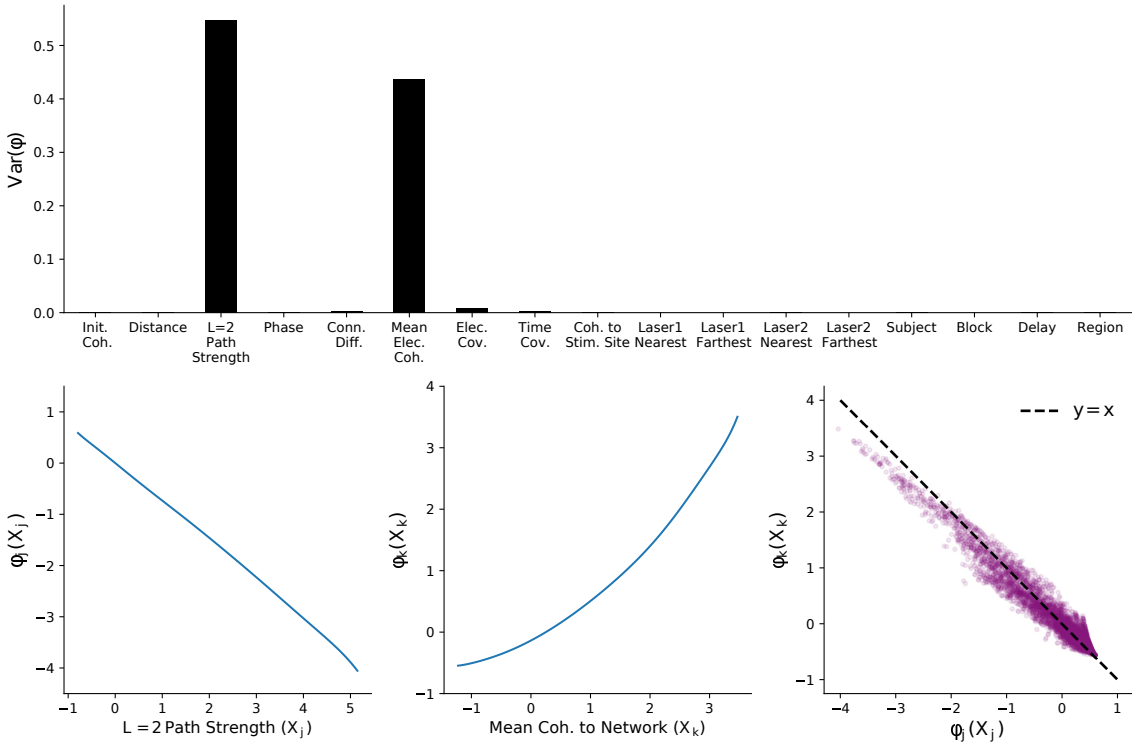

Figure 3.7: Concurvity analysis for additive modeling of the ECoG features. Top: relative contribution of each feature to the smallest APC. Bottom: nonlinear components of the APC corresponding to the features `L=2 path strength` (left) and `mean coherence to network` (middle). A scatterplot (right) shows strong (negative) linear association between these nonlinear transformations of the original features, indicating approximate concurvity.

3.6 Discussion

In this chapter we have presented a data analysis of ECoG recordings emphasizing prediction of stimulation-induced functional connectivity changes and investigation as to the aspects of the experimental protocol or initial connectivity structure that mediate the network-wide response. We define and estimate the SIFCC via the band-limited coherence between electrode LFP time series, then propose a nonlinear model with additive contributions from a diverse range of features. We apply a hierarchically penalized estimation scheme to control the complexity of the model components and identify additional tools to address their stability, importance, and similarity. Our results demonstrate good prediction accuracy that

significantly improves over a linear baseline, and they highlight in particular the importance of the baseline network structure as summarized by the network features. We show that extension to a coarse time-varying model over blocks yields further improvements in prediction, and that the SIFCC can be accurately predicted even after the end of stimulation.

The complexity of the raw time series data and the underlying process that generated it leave significant room for development of more sophisticated modeling approaches. One opportunity lies in explicit modeling of the LFP data as a time series, with stimulation details modeled as exogenous inputs. Challenges for such an approach would include the high dimension of the time series, the inadequacy of linear modeling for prediction, and the challenging forms of nonstationarity in the data, which likely involve both global drift and structural breaks during stimulation. A separate line of extension might consider more clinically realistic prediction settings, in which entire blocks, sessions, or even subjects are held out for test evaluation. This may lead to consideration of distributionally robust estimation schemes that can handle differences in the data generating distribution between training and test data.

Chapter 4

TOWARDS CLINICAL READINESS: STATISTICAL TOOLS FOR DATA HETEROGENEITY AND COUNTERFACTUAL SIMULATION

4.1 Introduction

In the previous chapter we proposed and implemented a data analysis for prediction of coherence changes between regions spanned by an ECoG electrode array using features derived from both the stimulation protocol and the baseline coherence network prior to stimulation. This analysis represents a first effort to formalize and estimate a statistical relationship between the coherence change and the raw data collected in the experiment of Yazdan-Shahmorad et al. (2018), namely a large collection of multivariate time series recordings of the local field potential on the cortical surface. The methods applied or developed reflect the main scientific questions that arise from comparison of this novel experimental protocol to the computational approaches adopted in the recent literature (Betzel et al., 2019; Keller et al., 2018; Huang et al., 2019; Yang et al., 2021): whether coherence changes could be accurately predicted in this experimental setting, how to quantify the relative impact of features relating to either the baseline network or the experimental protocol, and if nonlinear modeling could improve prediction quality under an appropriate scheme for model selection.

One weakness of this approach is that the data collection process itself is treated in a relatively abstract manner. Implicit in Eq. 3.6 is an assumption of independence between the observed coherence changes conditional on the features, but this is an idealistic assumption that masks the complexity of modern experimentation in neuroscience. The novelty of the experimental protocol in Yazdan-Shahmorad et al. (2018) heightens the importance of careful exploratory data analysis; while the biological validity of the experiments is well-

established, there is notably less emphasis, both for this data and in related studies, on statistical evaluation of the assumptions implicit in assembly and analysis of the processed dataset.

At the same time, the ultimate goal of brain stimulation research from the clinical perspective is to develop data-driven approaches to *intervention* on a cortical network to achieve a target connectivity associated with some therapeutic outcome (Jackson et al., 2006; Edwardson et al., 2013). Our results from the previous chapter show that the state of functional connectivity at baseline is highly predictive of subsequent changes due to stimulation. The challenge is then to identify a configuration of the protocol variables (i.e. a stimulation protocol within some parameterized set) likely to yield a target outcome, given the current state of the network. The availability of relevant data is limited by the massive cost of *in vivo* experimentation. One potential alternative is simulation of neuronal networks under various stimulation conditions, which can be used to investigate outcomes under alternative interventions.

In this chapter we present two studies that, while methodologically distinct, are related by their motivation to develop statistical perspective on the complex observational realities and inferential goals that will arise in translating brain stimulation research to the clinical setting. First, we conduct a careful exploration of the ECoG data in the high gamma frequency band, with an emphasis on potential sources of heterogeneity in the data generating process. Our results motivate a survey of regression techniques accounting for heterogeneity across levels of a (known) partition of the observed data, with an emphasis on prediction for entirely unseen sessions or subjects. Second, we review the recent literature on simulation of stochastic spiking neurons. We implement a Poisson spiking network and show how it can be used to identify an optimal stimulation protocol to achieve a target connectivity.

4.2 Evidence for Session and Subject-Level Heterogeneity

4.2.1 Details of the optogenetic ECoG data collection process

Yazdan-Shahmorad et al. (2018) introduced not only a novel dataset but an entirely new approach to stimulation of and recording from the macaque cortex. While the details of the stimulation protocol are described in the previous chapter, here we emphasize the technological differences that distinguish this experiment from others in the literature, and we document some challenging aspects of the data collection process that are often omitted in published work.

Technologically, there are three aspects of the experiment that together distinguish it from other contemporary studies: the use of macaque subjects, recording via ECoG array, and localized stimulation of the cortex via optogenetics. Macaque subjects are relatively uncommon for ECoG studies of stimulation-induced response; the standard source of ECoG data is human subjects whose severe epilepsy warrants the invasive neurosurgery required for direct access to the cortex (Rao, 2013; Leuthardt et al., 2006). Macaque subjects allow for broader and longer-term experimentation, and they are not restricted to a sub-population with severe neurological issues. One challenge, however, is that ethical considerations and constrained resources lead nearly all macaque studies in this area to evaluate just two individuals (Yanagawa et al., 2013; Yazdan-Shahmorad et al., 2018; Yang et al., 2021), so that quantification of inter-subject variation is very limited from a statistical standpoint. The combination of optogenetic stimulation and ECoG recording renders the protocol of Yazdan-Shahmorad et al. (2018) significantly less invasive than comparable studies of stimulation-induced response in macaque subjects, such as Yang et al. (2021). The result is longer-running trials, greater flexibility to troubleshoot and modify the experiment in progress, and ultimately more experimental sessions per subject.

The reality of data collection in a novel experimental paradigm in neuroscience is that it occurs to some extent in parallel with procedural fine-tuning and equipment malfunction. Yazdan-Shahmorad et al. (2018) collected data on 15 separate days spanning roughly 9

Table 4.1: Summary of ECoG data collection details by subject.

Subject	Sessions	Timespan (days)	Total obs.	Missing electrodes	Pct. missing data
G	15	10	266125	11.9 (10.7)	22.2
J	18	67	215380	28.0 (15.2)	47.5

months. The two subjects were studied sequentially rather than in parallel, and over notably different timespans. For each subject, the ECoG array was set to cover roughly the same regions - the somatosensory and motor cortex - but its placement was not fixed across sessions, effectively removing a "geographic" interpretation of the electrode indices. Moreover, electrode failure was common; in every session fewer than the 96 total electrodes in the array provided a viable signal, and by the chronologically latest sessions over half the electrodes were missing. We show in Appendix C that the nonlinear additive model fit separately to each subject varies significantly in its predictive accuracy at every frequency band.

A simple summary of the data collection process is provided in Table 4.1. For each subject, we report the number of sessions, timespan covering these sessions, total observations of coherence change, mean and standard deviation of missing electrodes per session, and overall percent missing data (as a fraction of the total possible observations with 96 working electrodes). Throughout, we will assume that electrode malfunction is independent of the stimulation and network features, and thus that the unobserved coherence changes are missing completely at random.

4.2.2 Exploratory analysis reveals heterogeneity at the subject and session levels

We begin with an exploratory analysis in which the pairwise distance between elements of the session or subject-level partitions are quantified and compared against a permutation-based null distribution. For each partition, denote by N_G the number of partition elements and assign to each a label $1, \ldots, N_G$. For $g \in \{1, \ldots, N_G\}$, denote by $\mathbb{P}_g$ the empirical distribution

of the ECoG features in partition element g, and let P_g refer to the corresponding population distribution. We compute and summarize the pairwise distances between these subgroup distributions using four different distance metrics:

1. The mean distance

$$D_{\text{mean}}(g, g') = \|\mu_g - \mu_{g'}\|,$$

where $\mu_g = \int x dP_g$. The corresponding estimator is $\widehat{D}_{\text{mean}}(s, s') = \|\hat{\mu}_g - \hat{\mu}_{g'}\|$, with $\hat{\mu}_g = \frac{1}{n_g} \sum_{i:G_i=g} X_i$.

2. The Bhattacharya distance

$$D_{\text{B}}(g, g') = -\log\left(\int \sqrt{p_g(x) p_{g'}(x)} dx\right),$$

where p_g is the density of P_g. We approximate each P_g by a Gaussian distribution with mean $\hat{\mu}_g$ and covariance $\hat{\Sigma}_g = \frac{1}{n_g - 1} \sum_{i:G_i=g} (X_i - \hat{\mu}_g)(X_i - \hat{\mu}_g)^T$, such that the Bhattacharya distance can be estimated as

$$\widehat{D}_{\text{B}}(g, g') = \frac{1}{8}(\hat{\mu}_g - \hat{\mu}_{g'})^T \Sigma^{-1} (\hat{\mu}_g - \hat{\mu}_{g'}) + \frac{1}{2} \log\left(\frac{\det \Sigma}{\sqrt{\det \hat{\Sigma}_g \det \hat{\Sigma}_{g'}}}\right),$$

where $\Sigma = \frac{1}{2}\left(\hat{\Sigma}_g + \hat{\Sigma}_{g'}\right)$. Under these assumptions the Bhattacharya distance is equivalent to the Mahalanobis distance between the distribution centers, penalized by the difference in distribution covariances.

3. The Jensen-Shannon distance

$$D_{\text{JS}} = \frac{1}{2}\left(D_{\text{KL}}(P_g || P) + D_{\text{KL}}(P_{g'} || P)\right),$$

with $P = \frac{1}{2}(P_g + P_{g'})$, where we use the Gaussian approximation to P_g and $P_{g'}$ as above.

Table 4.2: Pairwise distance summary for session and summary-level partitions of the ECoG data.

	Session			*Subject*		
Metric	**Distance**	**p-value**	**Reject $\mathcal{H}_0$?**	**Distance**	**p-value**	**Reject $\mathcal{H}_0$?**
Mean	3.85×10^{0}	$<1 \times 10^{-2}$	✓	7.48×10^{-1}	$<1 \times 10^{-2}$	✓
Bhattacharya	1.01×10^{1}	$<1 \times 10^{-2}$	✓	5.07×10^{0}	$<1 \times 10^{-2}$	✓
Jensen-Shannon	4.81×10^{0}	$<1 \times 10^{-2}$	✓	1.97×10^{0}	$<1 \times 10^{-2}$	✓
MMD	8.73×10^{-2}	$<1 \times 10^{-2}$	✓	5.42×10^{-3}	$<1 \times 10^{-2}$	✓

4. The maximum mean discrepancy (MMD)

$$D_{\text{MMD}}(g, g') = \left\| \mathbb{E}_{X \sim P_g}[\varphi(X)] - \mathbb{E}_{Y \sim P_{g'}}[\varphi(Y)] \right\|_{\mathcal{H}}$$

with feature map $\varphi : \mathcal{X} \to \mathcal{H}$ and reproducing kernel Hilbert space $\mathcal{H}$. We take φ corresponding to the Gaussian radial basis function kernel

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} = \exp(- \|x - y\|_2^2 / p),$$

where p is the dimension of $\mathcal{X}$. The mean distance is a special case of the MMD, with identity feature map $\varphi(x) = x$. Somewhat less trivially, the L_2 distance between kernel density estimates $\hat{p}_g(x) = \int_t \kappa(t - x) d\mathbb{P}_g$ and $\hat{p}_{g'}(y) = \int_t \kappa(t - y) d\mathbb{P}_{g'}$ is a special case of the MMD with kernel $k(x, y) = \int_z \kappa(x - z)\kappa(y - z) dz$ (Muandet et al., 2017, §5.1).

There are only two subjects and thus a single pairwise distance for the subject partition. For the session partition, the $\binom{33}{2} = 578$ pairwise distances are summarized by their mean. We compare these distance measurements to their distribution under a permutation-based null hypothesis. For $n = 100$ trials, we randomly partition the ECoG data into subgroups of the same size as the original subject or session partitions, then summarize the distance between the empirical distributions of the permuted partition elements. We compare the distance summary computed for the subject and session partitions to their permutation-based

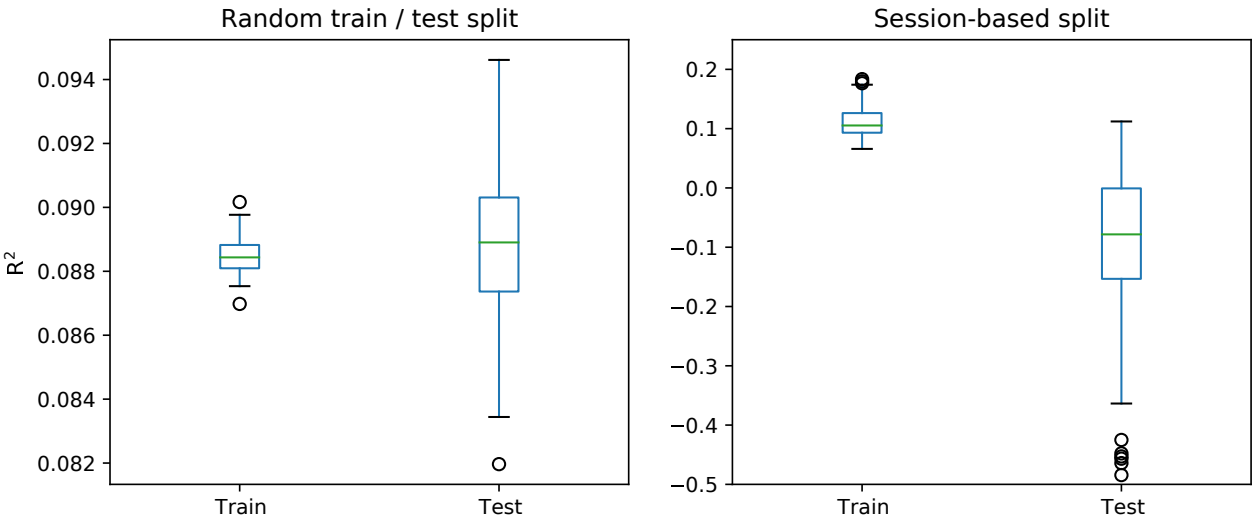

Figure 4.1: Prediction performance of the least squares estimator under naive and session-stratified randomization.

null distributions. Results are summarized in Table 4.2; we observe significant heterogeneity at both the subject and session levels.

4.2.3 Predictive failure of the least squares estimator

The observed heterogeneity across sessions motivates a session-stratified analysis in which the observations in the train and test sets have disjoint session membership. Importantly, this setting also corresponds to a much more realistic framework for application in clinical practice: given a trained model for stimulation-induced changes of functional connectivity, the objective is to generate useful predictions in future settings that will necessarily have been absent from the training data. We evaluate results from a simple linear model estimated by ordinary least squares under this stratified split; results are plotted in Figure 4.1 and compared to the "naive" method of randomizing train and test data across all splits. Results show the distribution in predictive accuracy, as measured by R^2, over 100 train-test splits computed by both methods.

Together, the exploratory results in this section provide evidence for significant hetero-

geneity in the ECoG data across sessions, and demonstrate that naive approaches to prediction on unseen sessions are likely to fail, regardless of their goodness-of-fit on the session-aggregated training data. Therefore, in applying the regression methods to be discussed below to the ECoG data, we will consider how to account for group heterogeneity at the session or subject levels and evaluate prediction results on held-out sessions corresponding to either chronological or subject-based splits.

4.3 A Survey of Likelihood-based Methods for Heterogeneous Data

We conduct a survey of modern regression methods for estimation and inference of the conditional mean parameters for data exhibiting group-heterogeneous structure. The motivating context and focus of the applied analysis is prediction of stimulation-induced functional connectivity changes from the data introduced in the previous chapter. We restrict the analysis to the after-stimulation coherence change observed in the high gamma frequency band.

4.3.1 General modeling framework

We begin by introducing notation for a general framework encompassing the range of methods investigated below. We posit the following hierarchical data generating process for the stimulation-induced coherence change Y_{sr} for subject s in session r, given observed features X_{sr}:

$$\varphi_s \sim P_{\text{subject}} \tag{4.1}$$

$$\alpha_{sr} \sim P_{\text{session}}(\cdot; \varphi_s) \tag{4.2}$$

$$Y_{sr} = \mu(X_{sr}; \beta) + \eta(X_{sr}; \alpha_{sr}) + \varepsilon_{sr}. \tag{4.3}$$

At the highest level, experimental subjects are drawn from a hypothetical superpopulation. Session-specific parameters are drawn from a distribution that may depend on the subject parameter. Finally, the observed data are generated as the sum of three terms: the population mean $\mu(X_{sr}; \beta)$, a session and subject-specific term $\eta(X_{sr}; \alpha_{sr})$, and the session

error ε_{sr}.

For each subject s and recording session r, $Y_{sr} \in \mathbb{R}^n$ and $X_{sr} \in \mathbb{R}^{p \times n}$, where $n = 5E(E-1)/2$ and $E = 96$ is the number of electrodes in the ECoG array. While n is fixed across sessions, not all data is observed for each session; we assume the missingness pattern is independent of Y_{sr} and X_{sr}. The session errors ε_{sr} are drawn independently from the distribution P_ε across all sessions and subjects. The covariance of the session noise is assumed to be diagonal; it remains for future work to investigate richer models for within-session covariance that may arise from the spatiotemporal structure in the observations.

The functional form of μ and η is known and may depend on the session features X_{sr}. Throughout this chapter we will emphasize the case of the linear population mean $\mu(X_{sr}; \beta) = X_{sr}\beta$, but in many cases extensions to the nonlinear setting are straightforward. The term $\eta(X_{sr}; \alpha_{sr})$ will be specified for each regression method. Finally, while the full data generating process specifies a three-level hierarchy, in reality we have data from only two subjects. We therefore focus mainly on heterogeneity at the session level.

4.3.2 Feature-dependent heteroskedasticity

We first investigate two likelihood-based procedures for estimating β, both of which arise from modeling the covariance of the observations Y_{sr}. The first of these models the observation variance as a function of the features, so that heteroskedasticity is realized at the level of individual observations. In particular we assume a linear population mean and random session-specific term $\eta(X_{sr}; \alpha_{sr})$ satisfying

$$\begin{aligned} \mathbb{E}[\eta(X_{sr}; \alpha_{sr})] &= 0 \\ \mathrm{Var}(\eta(X_{sr}; \alpha_{sr})) &= \mathrm{diag}(\exp(X_{sr}\alpha)), \end{aligned}$$

where the exponential is applied element-wise. Note that we have dropped the indices on the variance parameter α, indicating that they are the same across sessions, and that there is no additional session error under this model, i.e. $\varepsilon_{sr} = 0$. In the context of the ECoG

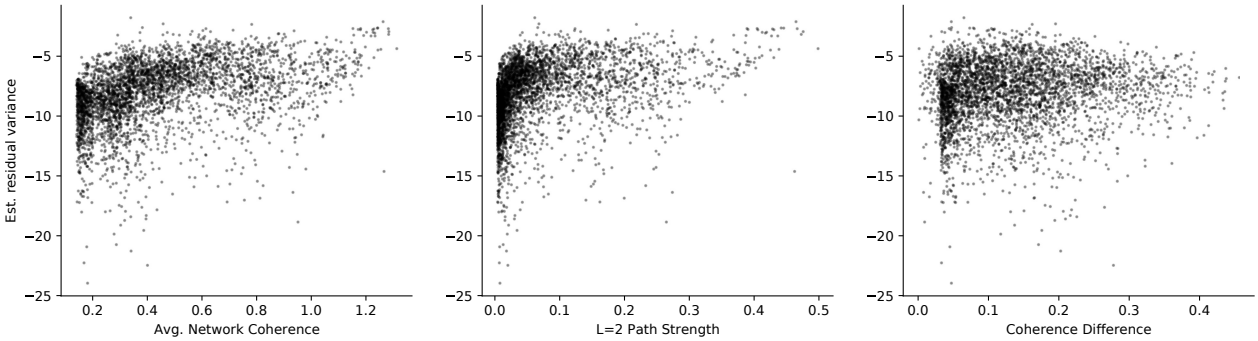

Figure 4.2: Log residual variance plotted against the three features with largest estimated effect size under the conditional variance model.

data, the model structure is motivated by an exploratory analysis of the residual variance of the ordinary least squares estimator, which exhibits visible trends with respect to several features; see Fig. 4.2.

We compute estimates $\hat{\beta}$ and $\hat{\alpha}$ of the mean and covariance parameters, respectively, in alternating fashion; see Alg. 3 for details. Estimation of these parameters via a single update after initialization is known in econometrics as "feasible weighted least squares" (Baltagi, 2008, §10.2) and is equivalent in the context of Alg. 3 to enforcing a stopping criterion based on iteration number. The results we report are for this estimation strategy. We also investigated the convergence criterion $\left\|\hat{\beta}^{(t+1)} - \hat{\beta}^{(t)}\right\|_2^2 < \tau$ with $\tau = 1 \times 10^{-6}$; for the ECoG data in the high-gamma band, we find that this criterion is met within 20 iterations. Our implementation uses the Python package `statsmodels` (Seabold and Perktold, 2010).

The estimator $\hat{\beta}$ computed above is an instance of a general class of "weighted least squares" estimators (Shao, 2003, §3.5). It is consistent for β and asymptotically efficient in the class of linear unbiased estimators only if $\widehat{\Sigma}_{sr} = \mathrm{diag}(\exp(X_{sr}\hat{\alpha}))$ is consistent for the true session-wise covariance Σ_{sr}. However, $\hat{\beta}$ may still be consistent, though not necessarily asymptotically efficient relative to the OLS estimator, so long as $\widehat{\Sigma}_{sr}$ converges to some $S \in \mathbb{R}^{n_{sr} \times n_{sr}}$, with convergence defined as $\left\|\widehat{\Sigma}_{sr}^{-1} S - I\right\|_{\max} \to_p 0$ (Chen and Shao, 1993). Moreover, unbiasedness of $\hat{\beta}$ is only guaranteed under the assumptions that ε_{sr} is symmetric

around zero and that $\widehat{\Sigma}_{sr}$ is unchanged by changing ε_{sr} to $-\varepsilon_{sr}$ (Shao, 2003, Examples 3.29-30). Note that neither of these conditions is necessarily satisfied in the general formulation for simultaneous estimation of the mean and variance parameters in §2.5.2 of Wakefield (2013).

Meanwhile, the estimator $\hat{\alpha}$ can be understood as the result of a particular approach to variance function regression (Carroll and Ruppert, 1988, Chapter 3), where the standard log-linear form for the variance function guarantees positivity for any setting of the features. Particularly in an applied setting in which predictions and not estimates are of primary importance, such as the clinical application hypothesized at the start of this chapter, estimation of the variance function is itself of major interest as this controls the probabilistic behavior of predictions under the model. Under the setting in which observations are Gaussian and the variance parameters depend on the features but not the mean, as posited above, the asymptotic variance of the generalized least squares estimator for the mean – that is, the estimator $\hat{\beta}^{(t)}$ from Alg. 3 – depends additively on the asymptotic variance of the variance parameters; thus improved estimation of the variance positively impacts estimation of the mean (Rothenberg, 1984).

Algorithm 3 Parameter estimation under feature-dependent heteroskedasticity.

Require: Data (Y_{sr}, X_{sr})
Require: Tolerance level τ
Initialize $\hat{\beta}^{(0)} = \arg\min_{\beta} \sum_{s,r} \|Y_{sr} - X_{sr}\beta\|_2^2$
Initialize $\hat{\alpha}^{(0)} = 0$
Set $t = 0$
while not converged **do**
    Compute residuals $\epsilon_{sr} = Y_{sr} - X_{sr}\hat{\beta}^{(t)}$
    Form covariance estimates $\widehat{\Sigma}_{sr} = \epsilon_{sr}\epsilon_{sr}^T$
    Update covariance parameters $\hat{\alpha}^{(t+1)} = \arg\min_{\alpha} \sum_{s,r} \left\| \log\left(\mathrm{diag}(\widehat{\Sigma}_{sr})\right) - X_{sr}\alpha \right\|_2^2$
    Update mean parameters $\hat{\beta}^{(t+1)} = \arg\min_{\beta} \sum_{s,r} \left\| \widehat{\Sigma}_{sr}^{-1}(Y_{sr} - X_{sr}\beta) \right\|_2^2$
    Set $t = t + 1$
end while
Return $\hat{\beta}^{(t)}, \hat{\alpha}^{(t)}$

As an aside, we note that this modeling approach is related to, but more general than, the classical quasi-likelihood framework for generalized linear models. The difference lies in the parameterization for the conditional variance. In quasi-likelihood, the conditional variance is modeled as the product

$$\mathrm{Var}(Y_{sr}) = \phi g(\mu(X_{sr}; \beta)),$$

with $\phi > 0$ an unknown dispersion parameter and g a known function of the conditional mean. This form implies that the estimating equation

$$X_{sr}^T \mathrm{Var}(Y_{sr})^{-1}(Y_{sr} - X_{sr}\beta) = 0$$

depends only on the mean parameter β. In particular, this leads to two consequences that do not necessarily hold under the more general modeling framework proposed here: first, the estimator $\hat{\beta}$ does not depend on the variance parameter α; second, consistency of $\hat{\beta}$ requires only that the mean model be correctly specified.

4.3.3 Session-weighted least squares

As an alternative to parameterization of the conditional variance in terms of the features, we consider instead an assumption of session-wise heteroskedasticity, such that

$$\mathbb{E}[\eta(X_{sr}; \alpha_{sr})] = 0$$
$$\mathrm{Var}(\eta(X_{sr}; \alpha_{sr})) = \alpha_{sr} I_n.$$

Heteroskedasticity is thus realized at the session level, parameterized per session by the scalar $\alpha_{sr} > 0$, rather than at the observation level as above. As before, $\varepsilon_{sr} = 0$.

The estimator is computed by the same method as in Alg. 3, where we instead initialize each $\hat{\alpha}_{sr}^{(0)} = 1$ and update as

$$\hat{\alpha}_{sr}^{(t+1)} = \frac{1}{n_{sr} - p - 1} \epsilon_{sr}^T \epsilon_{sr}$$

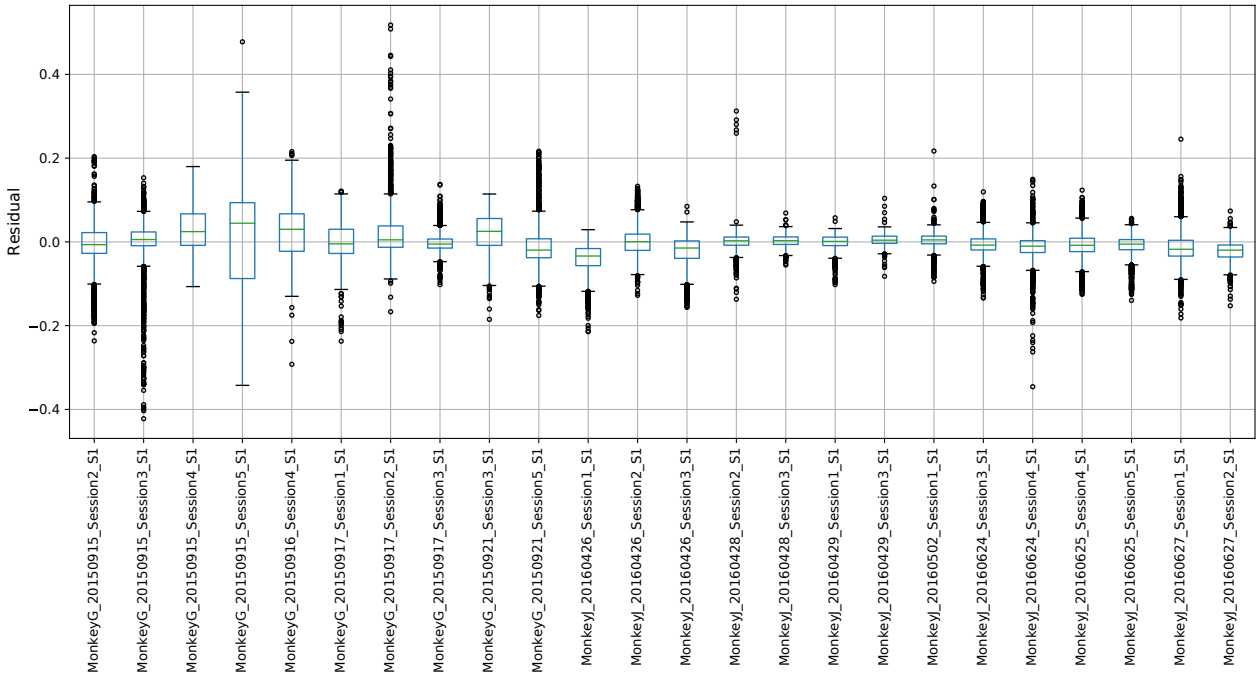

Figure 4.3: Session-wise distribution of residuals from a linear model of the after-stimulation coherence change.

with ϵ_{sr} defined as in the section above and n_{sr} the number of observations in session r for subject s. As with the feature-dependent heteroskedastic model, this approach investigates a hypothesis obtained from exploration of the ordinary least squares residuals. Plotting the residuals by session reveals significant heterogeneity in the residual variance at the session level; see Fig. 4.3. We note, however, that visualization is insufficient to draw statistical conclusions about this heterogeneity, especially as this corresponds to a fourth-order property of the data generating process. We therefore apply Levene's test for equality of variance among the session residuals, using the median as the measure of centrality for robustness against skewness of the residual distribution (Levine, 1960; Brown and Forsythe, 1974). The test is rejected with $p < 1 \times 10^{-16}$ in favor of the alternative that at least two sessions have unequal residual variance.

4.3.4 Anchor regression

Rothenhäusler et al. (2021) suppose a setting in which the data (Y_{sr}, X_{sr}) are observed across heterogeneous environments, here indexed by s and r, where the heterogeneity is driven by α_{sr}. In contrast to the other methods investigated in this section, it is assumed that $\alpha_{sr} \in \mathbb{R}^{n_{sr} \times q}$ is *observed*. The model for the data is

$$Y_{sr} = X_{sr}\beta + \sum_{k=1}^{q} \alpha_{rs} m_k + \varepsilon_{sr}, \tag{4.4}$$

with β the parameter of interest, $M = (m_1, \ldots, m_q) \in \mathbb{R}^{q \times q}$ is a fixed, unknown shift matrix, $\mathbb{E}[\varepsilon_{sr}] = 0$, and $\mathrm{Var}(\varepsilon_{sr}) = \sigma^2 I_{n_{sr}}$. The model is interpretable as a linear structural equation model in which α_{sr} may influence the distribution of X_{sr}, Y_{sr}, or both, but is itself uncaused by these variables. It thus has the interpretation as the driver of environment or experiment-based heterogeneity and is termed the *anchor variable*.

The proposed anchor regression estimator is given by

$$\hat{\beta}^{\gamma} = \arg\min_{\beta} \sum_{s,r} (1-\gamma) \left\| (I - \Pi_{\alpha_{sr}})(Y_{sr} - X_{sr}\beta) \right\|_2^2 + \gamma \left\| \Pi_{\alpha_{sr}} (Y_{sr} - X_{sr}\beta) \right\|_2^2$$

where $\Pi_{\alpha_{sr}}$ denotes the orthogonal projection matrix $\alpha_{sr}(\alpha_{sr}^T \alpha_{sr})^{-1} \alpha_{sr}^T$; again, this can be computed because we assume that α_{sr} is observed in this setting. The parameter $\gamma \in [0, 1]$ controls a tradeoff between projection of the residuals of the linear model onto the columnspace of α_{sr} versus its orthogonal complement.

At $\gamma = 0$, the $\hat{\beta}^{\gamma}$ coincides with least squares "adjusted for" α_{sr} by replacing X_{sr} and Y_{sr} with their residuals after regressing each against α_{sr}; at $\gamma = 1/2$ it coincides with ordinary least squares; and at $\gamma = 1$ it coincides with the two-stage least squares estimator in an instrumental variables framework, with α_{sr} as the instrumental variable. Informally, γ can be understood to represent the degree to which the information available in α_{sr} is a nuisance source of variation to be removed ($\gamma < 1/2$), irrelevant for the estimation problem at hand ($\gamma = 1/2$), or a useful source of variation for determining the effect of X_{sr} on Y_{sr} in the

presence of potential confounders ($\gamma > 1/2$).

Simple linear algebra shows that, for $\gamma \in (0, 1)$, $\hat{\beta}^\gamma$ can be computed analytically as a session-weighted generalized least squares estimator. In particular, define

$$W_{sr} = (\sqrt{1-\gamma})I + (\sqrt{\gamma} - \sqrt{1-\gamma})\Pi_{\alpha_{sr}}.$$

Then

$$\hat{\beta}^\gamma = \arg\min_\beta \|W_{sr}(Y_{sr} - X_{sr}\beta)\|_2^2 .$$

We will be particularly interested in the case where α_{sr} is a dummy-coded observation of a categorical variable indicating the session. Then $\Pi_{\alpha_{sr}} = n_{sr}^{-1}\mathbb{1}\mathbb{1}^T$,

$$W_{sr}^T W_{sr} = (1-\gamma)I_{n_{sr}} + n_{sr}^{-1}(2\gamma - 1)\mathbf{1}_{n_{sr}}\mathbf{1}_{n_{sr}}^T,$$

and by the Sherman-Morrison-Woodbury formula,

$$(W_{sr}^T W_{sr})^{-1} = (1-\gamma)^{-1}I_{n_{sr}} - \frac{1}{n_{sr}}\left(\frac{2\gamma - 1}{\gamma(1-\gamma)}\right)\mathbf{1}_{n_{sr}}\mathbf{1}_{n_{sr}}^T.$$

As the setting $\gamma = 1/2$ reduces to ordinary least squares estimation, and correspondingly to irrelevance of the anchor variable A, a model selection procedure can be applied to identify whether and how the anchor variable can be used to improve prediction over an OLS baseline. For the ECoG data, we apply a leave-one-session-out validation procedure, evaluating for each γ the goodness-of-fit in terms of the coefficient of determination R^2 on the held-out session. Results are shown in Fig. 4.4; we find that a nontrivial level of anchor penalization yields a slight but consistent improvement in predictive accuracy over the OLS estimator.

Theoretically, the anchor parameter β^γ is characterized as a *distributionally robust* estimator, that is, one optimizing the worst-case risk over a certain class of distributions.

Theorem 4.3.1. *(Rothenhäusler et al., 2021, Theorem 2) Suppose the data are generated according to Eq. 4.4, with Y_{sr} and X_{sr} centered. Denote by P_{train} the joint distribution*

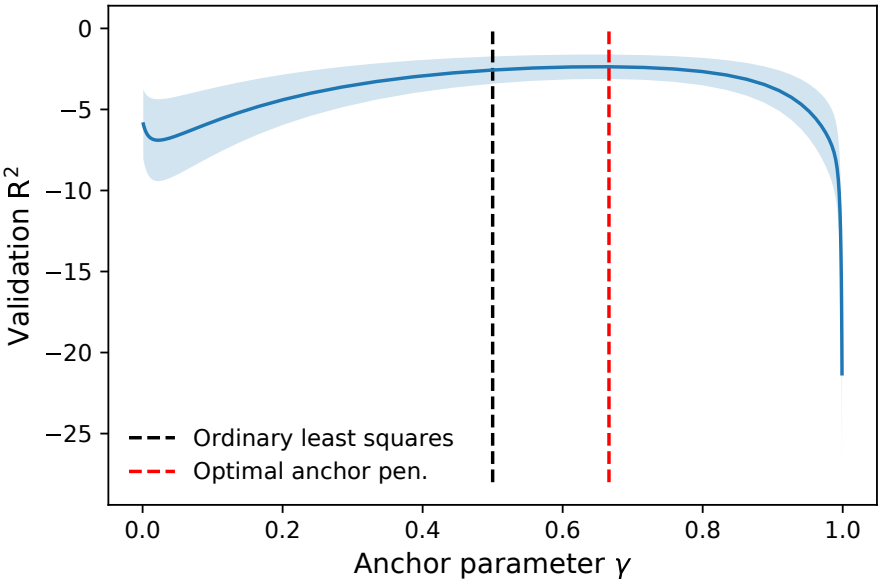

Figure 4.4: Validation results for selection of the anchor regression penalty γ on the ECoG data. The blue line indicates mean R^2 and bands show standard error over all held-out sessions. The selected value of γ is indicated in red, while the value corresponding to ordinary least squares is in black.

of (Y_{sr}, X_{sr}), $\mathbb{E}_{train}$ *the corresponding expectation. Define* $P_{\alpha_{sr}}(X_{sr}) = \mathbb{E}_{train}[X_{sr}|\alpha_{sr}]$ *and* $P_{\alpha_{sr}}(Y_{sr}) = \mathbb{E}_{train}[Y_{sr}|\alpha_{sr}]$. *Let* $\gamma \in (0, 1)$. *Then*

$$\mathbb{E}_{train}\left[\sum_{s,r} \|(1-\gamma)(I - P_{\alpha_{sr}})(Y_{sr} - X_{sr}\beta)\|_2^2 + \gamma \|P_{\alpha_{sr}}(Y_{sr} - X_{sr}\beta)\|_2^2\right] = \sup_{\nu \in C^\gamma} \mathbb{E}_\nu[(Y_{sr} - X_{sr}\beta)^2],$$

where the expectation $\mathbb{E}_\nu$ *is with respect to the shifted distribution obtained by replacing* $\sum_{k=1}^q \alpha_{rs} m_k$ *with* $\nu \in \mathbb{R}^{n_{sr}}$ *in Eq. 4.4, and we define*

$$C^\gamma = \{\nu \in \mathbb{R}^{n_{sr}} : \nu\nu^T \preccurlyeq \frac{\gamma}{1-\gamma} M\mathbb{E}[\alpha_{sr}\alpha_{sr}^T]M^T\}.$$

Relation to linear mixed effects. The generalized least squares formulation for the anchor regression estimator shows that the quantity W_{sr} implicitly functions to specify a model for the within-session covariance depending on the anchor variables α_{sr}; specifically,

$\widehat{\Sigma}_{sr} = (W_{sr}^T W_{sr})^{-1}$. Here we show that $\widehat{\Sigma}_{sr}$ corresponds to the same clustered-dependence structure that is induced by a random intercept term in a linear mixed-effects model. The linear mixed effects model writes the conditional mean of Y_{sr} given features X_{sr} and *random effect* b_{sr} as

$$\mathbb{E}[Y_{sr}|b_{sr}] = X_{sr}\beta + Z_{sr}b_{sr} + \varepsilon_{sr},$$

where $Z_{sr} \in \mathbb{R}^{n_{sr}\times q}$ are observed features that may coincide with the X_{sr}. It is typical to assume that the random effect and noise terms and independent and distributed as $\varepsilon_{sr} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, E(\alpha))$ and $b_{sr} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, D(\alpha))$, so that marginalizing over the random effect yields

$$\begin{aligned}\mathbb{E}[Y_{sr}|X_{sr}] &= X_{sr}^T\beta \\ \text{Var}(Y_{sr}|X_{sr}) \triangleq V_{sr}(\alpha) &= Z_{sr}D(\alpha)Z_{sr}^T + E(\alpha).\end{aligned}$$

Maximum likelihood estimators for the fixed effects β and variance components α are then obtained as

$$(\widehat{\beta}, \widehat{\alpha}) = \arg\min_{\beta,\alpha} -\frac{1}{2}\sum_{s,r}\log|V_{sr}(\alpha)| - \frac{1}{2}(Y_{sr} - X_{sr}^T\beta)^T V_{sr}(\alpha)^{-1}(Y_{sr} - X_{sr}^T\beta).$$

In the random intercept model, $Z_{sr} = \mathbf{1}_{n_{sr}}$ is simply a constant, $b_{sr} \in \mathbb{R}$, $E(\alpha) = \sigma_\varepsilon^2 I_{n_{sr}}$, $D(\alpha) = \sigma_b^2$, and thus the covariance parameters are given by $\alpha_{sr} = (\sigma_\varepsilon, \sigma_b)$. The resulting model for the session-wise covariance is *exchangeable* (Wakefield, 2013, §8.4):

$$\text{Var}(Y_{sr}) = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix},$$

where $\sigma^2 = \sigma_b^2 + \sigma_\varepsilon^2$ and $\rho = \sigma_b^2/\sigma^2$. Matching terms to $\widehat{\Sigma}_{sr} = (W_{sr}^T W_{sr})^{-1}$ as computed above,

we find an equivalence between the log-likelihood function for the linear mixed effects model with random session intercepts and the anchor regression objective with $\sigma_b^2 = (1-\gamma)^{-1}$ and $\sigma_\varepsilon^2 = (1-2\gamma)(n_{sr}\gamma(1-\gamma))^{-1}$. This is limited to the setting $\gamma < 1/2$, as we must have $\sigma_\varepsilon^2 > 0$.

4.3.5 Estimating equations with cluster dependence

The relation between anchor regression and estimation in the linear mixed effects framework suggests a comparison to a model in which the within-session dependence structure is explicitly represented and estimated. As the focus of estimation is on the population-level parameters β, we adopt the generalized estimation equations approach with linear mean and

$$\mathbb{E}[\eta(X_{sr}; \alpha_{sr})] = 0$$
$$\mathrm{Var}(\eta(X_{sr}; \alpha_{sr})) = \sigma R(\rho).$$

The covariance parameters are thus $\alpha = (\sigma, \rho)$; we set $\varepsilon_{sr} = 0$. The matrix $R(\rho) \in \mathbb{R}^{n_{sr} \times n_{sr}}$ encodes the structure of the working covariance model. We make the exchangeable assumption for within-session data such that

$$R(\rho) = (1-\rho) I_{n_{sr}} + \rho \mathbf{1}_{n_{sr}} \mathbf{1}_{n_{sr}}^T;$$

this is selected to correspond to the implicit within-session covariance of the anchor regression estimator derived above. Exchangeability is the simplest non-trivial model for within-session covariance, specifying a constant correlation between all pairs of observed coherence changes in a given session. The spatial layout of the ECoG electrodes and anatomical structure of the region they cover suggest that more complex covariance models may be worth exploring in future work.

The generalized estimating equation framework bares some similarity to the likelihood-based methods described above, and more generally to quasi-likelihood methods for estimation, in that assumptions are limited to the first two moments of the observation. Here,

however, we investigate the case of dependence between observations within a single session.

Computation of estimators $\hat{\beta}$ and $\hat{\alpha}$ again proceeds iteratively, as in Alg. 3. Here we initialize $\hat{\alpha}^{(0)} = (\hat{\sigma}^{(0)}, \hat{\rho}^{(0)}) = (1, 0)$ and at iteration t estimate the session covariances as $\widehat{\Sigma}_{sr} = \hat{\sigma}^{(t)} R(\hat{\rho}^{(t)})$. Given $\hat{\beta}^{(t)}$, we form the residuals $\epsilon_{sr} = Y_{sr} - X_{sr}\hat{\beta}^{(t)}$ and estimate $\hat{\alpha}^{(t+1)}$ by a second estimating equation

$$\sum_{s,r} E_{sr}^T (T_{sr} - C(\hat{\alpha}^{(t+1)})) = 0,$$

where $T_{sr} \in \mathbb{R}^{n_{sr}+n_{sr}(n_{sr}-1)/2}$ are covariance "data" obtained from the residuals, $C(\alpha) \in \mathbb{R}^{n_{sr}+n_{sr}(n_{sr}-1)/2}$ are their corresponding expectation under the working covariance model, and $E_{sr} \in \mathbb{R}^{n_{sr}+n_{sr}(n_{sr}-1)/2 \times 2} = \nabla_\alpha C(\alpha)$ is the matrix of partial derivatives with respect to the variance components:

$$\begin{aligned} T_{sr} &= [\epsilon_{sr1}\epsilon_{sr2}, \ldots, \epsilon_{sr(n_{sr}-1)}\epsilon_{srn_{sr}}, \epsilon_{sr1}^2, \ldots, \epsilon_{srn_{sr}}^2]^T, \\ C(\alpha) &= [\alpha\rho, \ldots, \alpha\rho, \alpha, \ldots, \alpha]^T, \\ E_{sr} &= \begin{bmatrix} \rho & \ldots & \rho & 1 & \ldots & 1 \\ \sigma & \ldots & \sigma & 0 & \ldots & 0 \end{bmatrix}. \end{aligned}$$

The estimator $\hat{\beta}$ is consistent so long as $\hat{\alpha}$ converges to some fixed value as $n_{sr} \to \infty$, while consistent estimation of the variance further requires that the covariance model is correct, so that $\mathbb{E}[T_{sr}] = C(\alpha_{sr})$ (Wakefield, 2013, §8.7). The sandwich estimator of the variance

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \sum_{s,r} \left(X_{sr}^T \widehat{\Sigma}_{sr}^{-1} X_{sr}\right)^{-1} \left(X_{sr}^T \widehat{\Sigma}_{sr}^{-1} \widehat{\mathrm{Var}}(Y_{sr}) \widehat{\Sigma}_{sr}^{-1} X_{sr}\right) \left(X_{sr}^T \widehat{\Sigma}_{sr}^{-1} X_{sr}\right)^{-1},$$

with $\widehat{\Sigma}_{sr} = \hat{\sigma} R(\hat{\rho})$ and $\widehat{\mathrm{Var}}(Y_{sr}) = (Y_{sr} - X_{sr}\hat{\beta})(Y_{sr} - X_{sr}\hat{\beta})^T$, is consistent under the assumption that the cluster units, i.e. sessions, are independent.

4.3.6 Surrogate variable analysis

Surrogate variable analysis (SVA) (Leek and Storey, 2007) attempts to account for observation heterogeneity due to unmodeled factors across experimental trials. The method was originally developed in the context of gene expression analysis, in which the fluctuations of expression measurements across arrays are colloquially known as "batch effects." It can be situated within a broader literature in biostatistics concerned with estimation and removal of latent sources of variation that might otherwise bias the regression analysis of interest; an alternative method assuming access to data from a negative control experiment is developed by Gagnon-Bartsch et al. (2013).

In surrogate variable analysis, batch effects are assumed to arise when outcomes at a collection of sites of interest (corresponding, for example, to genes or microarray locations) are observed repeatedly across potentially heterogeneous experimental settings. Adapting this to the ECoG data requires that the indices of the observed vector Y_{sr} be scientifically interpretable across subjects s and sessions r. While this is not the case for observations at the session level, we can recover a level of interpretability by further partitioning Y_{sr} according to the stimulation block $b \in \{1, 2, 3, 4, 5\}$ during which the observation was recorded. At the block level, the indices of $Y_{srb} \in \mathbb{R}^{n_{sr}/5}$ correspond to unique pairs of electrodes in the ECoG array and can thus be identified across subjects s, sessions r, and blocks b. The SVA model is then

$$Y_{srb} = X_{srb}\beta + \alpha_{srb} + \varepsilon_{srb}, \tag{4.5}$$

with $\varepsilon_{srb} \sim \mathcal{N}(0, \sigma^2 I_{n_{srb}})$. The "batch effects" $\alpha_{srb} \in \mathbb{R}^{n_{sr}/5}$ are considered a source of unwanted variation and modeled as $\alpha_{srb} = \sum_{\ell=1}^{L} \gamma_\ell g_{srb}^{(\ell)}$, where $g_{srb} = (g_{srb}^{(1)}, \ldots, g_{srb}^{(L)}) \in \mathbb{R}^L$ is a vector of latent factors realized per experiment (in this case, each unique combination of subject s, session r, and block b).

The SVA algorithm computes a set of orthogonal *surrogate variables* $\{\hat{h}^{(k)}\}_{k=1}^{K}$ that ap-

proximately span the linear subspace defined by the latent factors $\{g^{(\ell)}\}_{\ell=1}^{L}$,

$$\sum_{\ell=1}^{L} \gamma_\ell g_{srb}^{(\ell)} \approx \sum_{k=1}^{K} \lambda_k \hat{h}_{srb}^{(k)},$$

The number of surrogate variables K is selected by the user. Following Leek and Storey (2007), we adapt the SVA algorithm to the ECoG data via the procedure detailed in Alg. 4.

"Consistency" for SVA refers to convergence of the column space of the estimates $\{\hat{h}^{(k)}\}_{k=1}^{K}$ to that of the true latent variables $\{g_\ell\}_{\ell=1}^{L}$. In general, consistency of the surrogate variables is not guaranteed, but it is established in the special case of mutual orthogonality between the features and latent variables.

Theorem 4.3.2. *(Leek, 2007, Theorem 9) Let Y_{srb} be generated according to the model in Eq. 4.5 and suppose that for every subject s, session r, and block b the features X_{srb} are orthogonal to the vector g_{srb} of latent factors. Suppose the true number K^* of latent factors is known.*

Then the column space of the estimated surrogate variables $\{\hat{h}^{(k)}\}_{k=1}^{K^*}$ converges almost surely to that of the true latent factors $\{g_\ell\}_{\ell=1}^{L}$ as the number of observed experiments $m = s \times r \times b \to \infty$.

The requirement that the number of true latent factors K^* be known can be relaxed under additional assumptions; the authors use the algorithm of Bai and Ng (2002) to estimate K^* prior to estimation of the surrogate variables.

Algorithm 4 Estimation of surrogate variables for ECoG data.

Require: Data (Y_{srb}, X_{srb})
Require: Number of surrogate variables K
Require: Family-wise error threshold α
Compute the naive estimator $\hat{\beta}_0 = \arg\min_\beta \sum_{s,r,b} \|Y_{srb} - X_{srb}\beta\|_2^2$
Form the residuals $\epsilon_{srb} = Y_{srb} - X_{srb}\hat{\beta}_0$
Concatenate ϵ_{srb} over s, r, b to obtain matrix E
Compute the singular value decomposition $E = U\Sigma V^T$
Concatenate X_{srb} over s, r, b to obtain matrix X
for $k = 1, \ldots, K$ **do**
    Let V_k be the k^{th} column of V, sorted by magnitude of the singular values
    for all s, r, b **do**
        Regress V_k against X_{srb}
        Store the p-value resulting from the two-sided t-test of the estimated slope
    end for
    Apply Bonferroni correction at level α to all p-values
    Form X_R from columns of X that remain significant after correction
    Compute the singular value decomposition $X_R = U_R\Sigma_R V_R^T$
    Set $\hat{h}^{(k)} = V_{Rj^*}$, with $j^* = \arg\max_j \mathrm{Corr}(V_k, V_{Rj})$
end for
Return surrogate variables $(\hat{h}^{(1)}, \ldots, \hat{h}^{(K)})$

After estimation of the surrogate variables, estimates $\hat{\beta}, \hat{\lambda}$ of the parameters in the adjusted model

$$Y_{srb} = X_{srb}\beta + \sum_{k=1}^{K} \lambda_\ell \hat{h}_{srb}^{(k)} + \varepsilon_{srb}$$

are obtained by ordinary least squares. The estimator $\hat{\beta}$ of the population mean parameter is considered the "final" estimate after correction for batch effects. We implement this procedure for the ECoG data, setting $K = 5$. Unsurprisingly given the construction of the surrogate variables and the observed session-level variation in the data, the final model explains much more variation in the training data than the naive linear model estimated by OLS: the R^2 value is 0.649 versus 0.204, and the median absolute error is 1.25×10^{-2} versus 1.83×10^{-2}.

Finally, we note that SVA is not naturally suited to prediction on new "experiments" as

the surrogate variables $\hat{h}^{(k)}$ are defined only on observed subjects s, sessions r, and blocks b. Given features $\tilde{X}$ from a new setting, we compute predictions simply as $\hat{Y} = \tilde{X}\hat{\beta}$, implicitly setting $\lambda = 0$. This corresponds to an assumption of negligible batch effects.

4.3.7 Maximin estimation

The *maximin* framework of Meinshausen and Bühlmann (2015) is proposed for estimation of a common effect over data generated from the mixture model

$$Y_{sr} = X_{sr}\alpha_{sr} + \varepsilon_{sr}, \tag{4.6}$$

with session-level coefficients α_{sr} and additive errors $\varepsilon_{sr} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_{n_{sr}})$. Under the maximin model, $\mu(X_{sr}; \beta) = 0$, so that there is no true population-level effect. Nonetheless, the objective is to compute a single statistic that in some sense captures the part of the session-level effects that is common across the observed sessions. This is formalized by the "maximin" criterion

$$\beta_{\text{maximin}} = \text{argmax}_\beta \min_{s,r} V(\beta, \alpha_{sr}),$$

where

$$\begin{aligned} V(\beta, \alpha_{sr}) &= \mathbb{E}_{X_{sr},\varepsilon_{sr}}[\|Y_{sr}\|_2^2] - \mathbb{E}_{X_{sr},\varepsilon_{sr}}[\|Y_{sr} - X_{sr}\beta\|_2^2] \\ &= 2\beta^T \Sigma^X \alpha_{sr} - \beta^T \Sigma^X \beta, \end{aligned}$$

with $\Sigma^X = \mathbb{E}[X_{sr}^T X_{sr}]$ the feature covariance; Σ^X is not indexed by s or r, reflecting the assumption that the features X_{sr} are drawn independently from a common distribution P_X in each session. The scalar quantity $V(\beta, \alpha_{sr})$ can be interpreted as the variance explained by β for the data from subject s and session r with true parameter α_{sr}, and the maximin parameter β_{maximin} as the coefficient maximizing this quantity under the most adversarial setting observed in the data.

Estimates of α_{sr} for each subject s and session r are computed by ordinary least squares on

the session-specific data (Y_{sr}, X_{sr}). The covariance Σ^X is estimated by $\widehat{\Sigma}^X = (n_s n_r)^{-1} \sum_{s,r} n_{sr}^{-1} X_{sr}^T X_{sr}$. From these, we estimate $V(\beta, \alpha_{sr})$ as $\widehat{V}_{sr}(\beta) = 2\beta^T \widehat{\Sigma}^X \hat{\alpha}_{sr} - \beta^T \widehat{\Sigma}^X \beta$. Finally, the minimum in the definition of β_{maximin} can be smoothed via

$$\min_{s,r} V(\beta, \alpha_{sr}) \approx \sum_{s,r} \left(V(\beta, \alpha_{sr})^{\xi} \right)^{\frac{1}{\xi}}$$

for small values $\xi > 0$, which yields a weighted least squares estimator for β_{maximin} as

$$\hat{\beta}_{\text{maximin}} = \arg\min_{\beta} \sum_{s,r} V(\beta, \alpha_{sr})^{\xi - 1} (Y_{sr} - X_{sr}\beta).$$

Theoretical exposition relies on the following alternative characterization of the maximin parameter.

Theorem 4.3.3. *(Meinshausen and Bühlmann, 2015, Theorem 1) For each session and subject, let $X_{sr} \sim P_X$ with covariance Σ^X. Define*

$$M(B, \Sigma) = \arg\min_{\beta \in CVX(B)} \beta^T \Sigma \beta,$$

where $\Sigma \succcurlyeq 0$, $B = (b_1, \ldots, b_n)$ is a collection of vectors in $\mathbb{R}^p$, and $CVX(B)$ denotes the convex hull of the set B. Then

$$\beta_{maximin} = M(\alpha, \Sigma^X),$$

where $\alpha = (\alpha_{sr})_{sr}$ is the collection of true regression coefficients across all sessions and subjects.

Rothenhäusler et al. (2016) show that $M(\hat{\alpha}, \widehat{\Sigma}^X)$ is consistent for $\beta_{\text{maximin}} = M(\alpha, \Sigma^X)$ and asypmtotically normal. In applying this definition of the maximin estimator to the

ECoG data, we obtain a negative result. It follows from this characterization that

$$\exists (s,r),(s',r') : \operatorname{sgn}([\hat{\alpha}_{sr}]_j) \neq \operatorname{sgn}([\hat{\alpha}_{s'r'}]_j) \implies (\beta_{\text{maximin}})_j = 0,$$

where $[\hat{\alpha}_{sr}]_j$ is the j^{th} component of the OLS estimator $\hat{\alpha}_{sr}$ computed from (Y_{sr}, X_{sr}). This is indeed the case for the ECoG data, for every feature; see Appendix C. In this application, therefore, conservatism to the degree of maximin estimation results in a pessimistic evaluation of the population-level effect.

4.3.8 Superquantile regression

An alternative approach to robust estimation of the conditional mean that does not yield a trivial estimate for the ECoG data can be obtained by *superquantile regression*. As with the maximin formulation, the superquantile estimator does not seek to optimize a likelihood or generalized likelihood criterion. Instead, this procedure aims to minimize the superquantile, also known as the conditional value at risk (CVaR) (Rockafellar et al., 2014) of a given loss function L. Under a stochastic model in which $(Y, X) \sim \mathbb{P}$, the loss function $L = \ell(Y, f(X))$ is itself a random variable with distribution F_L, and the superquantile is given by

$$\bar{Q}_\rho(L) = \frac{1}{1-\rho} \int_{\rho'=\rho}^{1} Q_{\rho'}(L) d\rho',$$

with $Q_\rho(R) = \min\{x \in \mathbb{R} : F_L(x) \geq \rho\}$ the ρ^{th} quantile of the loss. Denote by $\bar{Q}_{n,\rho}(L)$ the empirical version of $\bar{Q}_\rho(L)$, obtained by taking $(Y, X) \sim \mathbb{P}_n$. Furthermore, suppose that the loss function L is parameterized by $\beta \in \mathbb{R}^p$. The superquantile equivalent of risk consistency is then established in the following theorem.

Theorem 4.3.4. *(Laguel et al., 2021, Theorem 1) Fix $\rho \in (0,1)$. Let L be parameterized by $\beta \in B$, where B is bounded and admits an ϵ-cover. Furthermore, suppose L is P-almost*

surely bounded for each $\beta \in B$ *and* M*-Lipschitz as a function of the features* X*. Denote by*

$$\beta_\rho^* \in \arg\min_{\beta \in B} \bar{Q}_\rho(L) \quad \textit{and} \quad \beta_{n,\rho}^* \in \arg\min_{\beta \in B} \bar{Q}_{n,\rho}(L)$$

exact minimizers of the population and empirical superquantiles of the loss, respectively. Then $\bar{Q}_{n,\rho}(L(\beta_{n,\rho}^*)) \to \bar{Q}_\rho(L(\beta_\rho^*))$ *almost surely.*

The superquantile admits a dual formulation as the supremum over a set of measures absolutely continuous with respect to $\mathbb{P}$. In the case of the discrete measure $\mathbb{P}_n$ and square loss $L = (Y - X\beta)^2$, it can be written as

$$\bar{Q}_{n,\rho}(L) = \sup_{w \in K_\rho} \sum_{i=1}^{n} w_i (Y_i - X_i^T \beta)^2,$$

with $K_\rho = \{w \in \mathbb{R}^n : \sum_{i=1}^n w_i = 1, w_i \in [0, (n(1-\rho))^{-1}] \forall i\}$. The superquantile regression estimator $\hat{\beta}_\rho = \text{argmin}_\beta \bar{Q}_{n,\rho}(L)$ thus takes the form of a weighted least squares estimator, with weights selected adversarially from K_ρ. Laguel et al. (2020) show that $\hat{\beta}_\rho$ can be efficiently computed by first-order optimization methods using a smoothed version of $\bar{Q}_{n,\rho}(L)$.

The superquantile approach optimizes over the worst-case reweighting of data on an individual level, but can be modified in a straightforward manner to address reweighting on a session level. Specifically, the per-observation loss $L_i = (Y_i - X_i^T \beta)^2$ is replaced by the per-group loss $L_{sr} = n_{sr}^{-1} \|Y_{sr} - X_{sr}\beta\|_2^2$, and the session-superquantile estimator is given by

$$\hat{\beta}_\rho = \arg\min_\beta \sup_{w \in K_\rho^{sr}} \sum_{s,r} \frac{w_{sr}}{n_{sr}} \|Y_{sr} - X_{sr}\beta\|_2^2,$$

with $K_\rho^{sr} = \{w \in \mathbb{R}^{n_s n_r} : \sum_{s,r} w_{rs} = 1, w_{rs} \in [0, (n_r n_s (1-\rho))^{-1}] \forall s, r\}$. We use the package `SPQR` of Laguel et al. (2020) to compute this estimator and select L-BFGS for the optimizer as the loss is smooth; the algorithm converges within 500 iterations.

4.3.9 Results

We apply each of the preceding frameworks to estimation of a linear conditional mean for the ECoG data. Throughout, we maintain a focus on heterogeneity at the level of recording sessions. In keeping with both the statistical objective of generalization from heterogeneous data and the clinical objective of accurate coherence prediction from future ECoG recordings, the methods are evaluated on hold-out sets consisting of entire sessions.

Hold-out settings. We investigate two hold-out settings in particular. In the *chronological* hold-out setting, the test set consists of all recording sessions that occurred on the chronologically latest two days of experiments for each subject. In the *subject* hold-out setting, the test set consists of all sessions recorded from the second of two macaque experimental subjects ("Subject J").

Evaluation metrics. Denote by $\hat{\beta}$, $\hat{Y} = X\hat{\beta}$, and $R = Y - \hat{Y}$ the parameter estimate, predictions given features X, and residuals given true outcomes Y, respectively, obtained by a given regression method. We report the following metrics for prediction on the hold-out set $\{(Y_i, X_i)\}_{i=1}^{n'}$:

- The coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n'}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n'}(Y_i - \bar{Y})^2},$$

  with $\bar{Y} = \frac{1}{n'}\sum_{i=1}^{n'} Y_i$.

- The 50^{th} and 90^{th} quantiles of the absolute error $E = (|R_1|, \ldots, |R_{n'}|)^T$

$$Q(E)_p = \min\{x \in \mathbb{R} : F_E(x) \geq p\}, \quad p \in \{0.5, 0.9\},$$

  with $F_E(x) = \frac{1}{n'}\sum_{i=1}^{n'} \mathbb{1}\{E_i \leq x\}$.

- The balanced accuracy of a binary classifier predicting $S = \text{sign}(Y)$ by $\hat{S} = \text{sign}(\hat{Y})$

$$B = \frac{1}{2}\left(\frac{\sum_{i=1}^{n'} \mathbb{1}\{S_i = 1, \hat{S} = 1\}}{\sum_{i=1}^{n'} \mathbb{1}\{S_i = 1\}} + \frac{\sum_{i=1}^{n'} \mathbb{1}\{S_i = 0, \hat{S} = 0\}}{\sum_{i=1}^{n'} \mathbb{1}\{S_i = 0\}}\right).$$

The variety of metrics is intended to provide a robust summary of prediction performance: the R^2 and median absolute errors as summaries of average performance, the 90^{th} quantile of the absolute error as a measure of the upper tail, and the balanced accuracy as an assessment of the model's capacity to predict the sign of the coherence change, adjusted for class imbalance. Finally, we estimate the variance of each metric computed by a session-wise jackknife procedure.

Chronological hold-out

Train and test set residuals for the chronological hold-out setting are plotted in Fig. 4.5, and evaluation metrics on the test set are reported in Table 4.3. We observe a significant range in the prediction metrics across methods. The best performance is achieved by the likelihood-based methods accounting for heteroskedasticity and the surrogate variable analysis, which estimates the effect of unmeasured confounders across sessions.

Subject hold-out

Train and test set residuals for the subject hold-out setting are plotted in Fig. 4.6, and evaluation metrics on the test set are reported in Table 4.4. In contrast to the chronological hold-out setting, the prediction results in this application are uniformly of low quality. This suggests that the subject-level heterogeneity in the ECoG data may be even more extreme than the session-level heterogeneity that is the main focus of this analysis. The distribution of the residuals in Fig. 4.6 show that *every* method investigated exhibits significant bias in its predictions on the new subject.

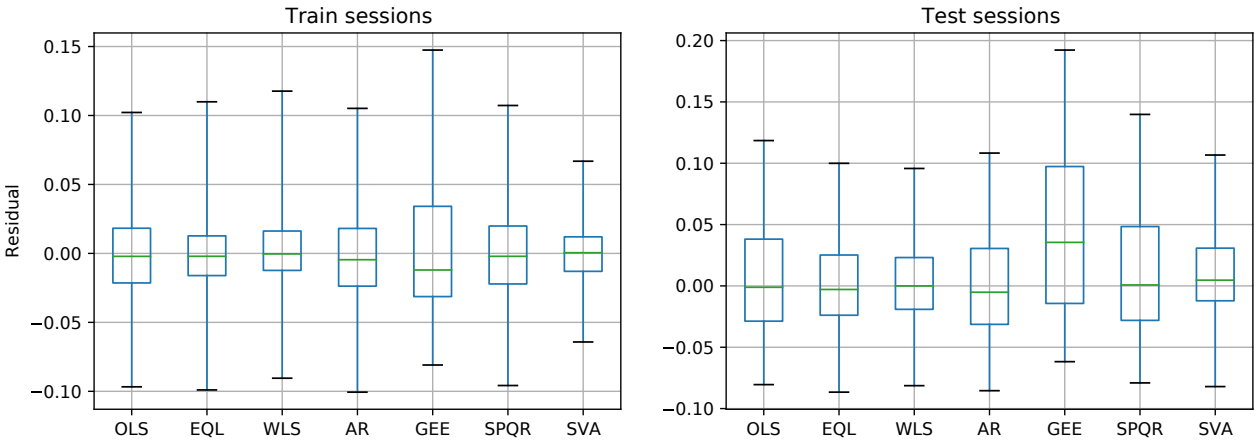

Figure 4.5: Residuals on training and held-out test sessions for each regression method.

	Method	**Test R^2**	**Median abs. error**	**Q90 abs. error**	**Balanced accuracy**
Baseline	Least squares	-0.170	0.032	0.090	0.590
		(0.169)	(0.002)	(0.009)	(0.009)
Modeled Heteroskedasticity	Feature-dependent heteroskedasticity	0.070	0.025	0.078	0.590
		(0.079)	(0.001)	(0.005)	(0.005)
	Session-weighted least squares	0.162	0.021	0.072	0.623
		(0.033)	(0.001)	(0.002)	(0.006)
Modeled Confounding	Surrogate variable analysis	0.057	0.022	0.076	0.567
		(0.025)	(0.001)	(0.001)	(0.005)
Clustered Dependence Structure	Anchor regression	-0.062	0.031	0.082	0.568
		(0.073)	(0.001)	(0.004)	(0.008)
	Linear mixed effects	-1.751	0.047	0.151	0.512
		(0.456)	(0.003)	(0.016)	(0.014)
Sample Reweighting	Superquantile regression	-0.411	0.035	0.102	0.576
		(0.206)	(0.002)	(0.010)	(0.010)

Table 4.3: Prediction results on held-out sessions in the chronological hold-out setting. Numbers in parentheses indicate jackknife estimates of the standard error.

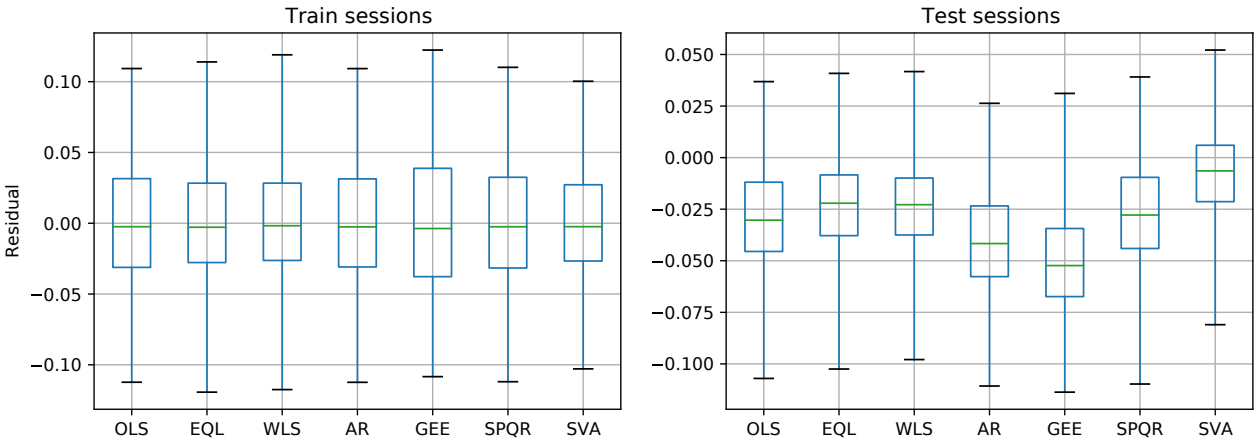

Figure 4.6: Residuals on training and held-out test subjects for each regression method.

	Method	**Test R^2**	**Median abs. error**	**Q90 abs. error**	**Balanced accuracy**
Baseline	Least squares	-2.258 (0.322)	0.031 (0.003)	0.071 (0.003)	0.496 (0.002)
Modeled Heteroskedasticity	Feature-dependent heteroskedasticity	-1.640 (0.229)	0.024 (0.003)	0.066 (0.002)	0.492 (0.004)
	Session-weighted least squares	-1.581 (0.378)	0.025 (0.004)	0.064 (0.003)	0.498 (0.004)
Modeled Confounding	Surrogate variable analysis	-1.036 (0.426)	0.017 (0.005)	0.057 (0.006)	0.490 (0.006)
Clustered Dependence Structure	Anchor regression	-3.322 (0.326)	0.042 (0.002)	0.078 (0.003)	0.499 (3.51×10^{-4})
	Linear mixed effects	-4.681 (0.920)	0.053 (0.006)	0.084 (0.006)	0.480 (0.002)
Sample Reweighting	Superquantile regression	-2.221 (0.332)	0.030 (0.003)	0.071 (0.003)	0.490 (0.004)

Table 4.4: Prediction results on held-out sessions in the subject hold-out setting. Numbers in parentheses indicate jackknife estimates of the standard error.

Interpretation of the prediction results

The contrast between the largely negative results reported here and the predictive success of the nonlinear model in the previous section merits some discussion. It is clear that the difference in prediction results is due to the construction of the training and test sets. When the test set is drawn uniformly at random from the data, it is guaranteed to be close in distribution to that of the training set when both are of sufficient size. The resulting nonlinear model is interpretable as an explanation for the variation in the observed data, though not as a prediction rule expected to generalize to new sessions or subjects. The major scientific conclusions of the chapter, including the relative importance of the graph features and similarity of nonlinear features across frequency bands, are not affected by this qualification and may be useful to the neuroscience community as previously unreported insights derived from a novel dataset.

On the other hand, when the test set is obtained by subject or session-level stratification, then it can differ significantly from the training set, regardless of their sizes, if the features or response are not both independent of the stratification factor. This more closely reflects the reality of clinical deployment of a prediction model, which will necessarily be trained on past sessions or even different individuals. Our approach in this chapter is to document the experiment-level variation in ECoG recordings and implement a survey of modern regression methods that may serve as a reference for explicitly accounting for this phenomenon in future analyses. We find in particular that accounting for session-level heterogeneity either as heteroskedastic noise or as an unobserved confounder yields promising results in the session hold-out setting. Robust methods such as maximin or superquantile regression perform poorly and are likely too conservative, though this is in itself informative as to the severe dissimilarity across sessions and subjects. Accounting for within-group dependence, as in linear mixed effects and implicitly through anchor regression, is ineffective with respect to prediction, though such dependence surely exists. More sophisticated models for the spatial and temporal structure of this dependence may be warranted.

4.4 Statistical Approaches to Simulation of Spiking Neurons

In the second part of this chapter, we review stochastic methods for simulation of neural activity and implement an approach tailored to the problem of identifying optimal stimulation interventions. The broad term "neural activity" refers to some quantitative characterization of the specialized, coordinated activity of neuronal populations by which they exert influence on each others' cell states through electrochemical signaling. Mathematically, this activity - the result of highly complex biological mechanisms - is summarized by the membrane potential $v(t) \in \mathbb{R}$, typically measured in millivolts (mV), which describes the gradient in electric potential between the interior and exterior of the neuron. Of particular interest are discrete events known as action potentials or *spikes*, where the cell membrane rapidly depolarizes and the membrane voltage is consequently reset. Spikes form the basis of an extensively developed and tested theory of neuronal communication (Gerstner and Kistler, 2002). This theory is fundamentally statistical: spike patterns are characterized in information-theoretic terms (Dayan and Abbott, 2001, §4.3) and considered as realizations from an underlying process that can be estimated under standard Bayesian or frequentist frameworks (Gerstner et al., 2014, §10.2). The spiking neuron model has itself motivated certain lines of investigation in applied probability, including superpositions (Srinivasan and Sampath, 2013, §4.2) and information capacity (Johnson, 1996, §3.1) of point processes.

Given a set of spike times $\mathbb{S} = \{t_1, ..., t_k\} \in [0, T]^k$ for an individual neuron, its activity can be represented by the *spike train* $S(t) = \sum_{t_s \in \mathbb{S}} \delta(t - t_s)$. Where appropriate, we will index the membrane potential $v_i(t)$, spike set $\mathbb{S}_i$, and spike train $S_i(t)$ by i to denote measurements for neuron i within a population of size N. The spike train of an individual neuron carries information to other neurons in a population by virtue of their anatomical connection, known as a synapse. The neuron that spikes is denoted as *presynaptic*, while the neuron receiving the spike is *postsynaptic*. Anatomical connectivity is not necessarily (or even typically) symmetric; neuron i can synapse onto neuron j without the reverse being true. For a population of N neurons, we denote by $C(t) \in \mathbb{R}^{n \times n}$ the connectivity matrix at time t, where

$C_{ij}(t) \neq 0$ indicates a connection from neuron i to j, and $\text{sign}(C_{ij}(t))$ indicates whether a spike in neuron i increases or decreases the chance of a subsequent spike in neuron j. For all $i = 1, \ldots, N$ and $t \in [0, T]$, we set $C_{ii}(t) = 0$.

Classical approaches to simulation of neural activity hewed closely to biophysical models for the electrochemical state of a neuron in continuous time, most prominently the Hodgkin-Huxley model (Hodgkin and Huxley, 1952), a system of partial differential equations for the membrane potential within each spatial compartment of an individual neuron. The simplified *integrate-and-fire* (IF) model (Lapique, 1907; Hill, 1936) reduces the cell to a single point, and thus to a single value for the membrane potential, which evolves according to

$$C_m \frac{\partial v_i(t)}{\partial t} = I_{\text{network}}(t) + I_{\text{input}}(t) + I_{\text{leak}}(t), \tag{4.7}$$

where C_m is the membrane capacitance in farads (F) and $I(t)$ denotes the current, in amps (A), from various sources at time t. In particular, the leak current $I_{\text{leak}}(t) = -C_m/\tau_m(v_i(t) - v_0)$, $\tau_m > 0$, models a background decrease in voltage towards a resting potential v_0; the input current $I_{\text{input}}(t)$ represents injected current exogenous to the network, for example from experimental stimulation; and the network current $I_{\text{network}}(t) = C_m \tau_s^{-1} \sum_{j=1}^{N} \sum_{t_s \in \mathbb{S}_j} C_{ji}(t) \exp((t - t_s)/\tau_s)\Theta(t - t_s)$ captures the influence of presynaptic network spikes. The Heaviside function $\Theta(t)$ ensures that a spike at time t_s only affects the postsynaptic membrane potential at $t \geq t_s$, or equivalently, that the filter relating presynaptic spikes to membrane potential is *causal*. Finally, a spike is generated ($S_i(t) = 1$) from the IF model at time t if $v_i(t) > v_{\text{spike}}$, after which $v_i(t)$ is reset to v_0; thus under the IF model a neuron "integrates" presynaptic and external stimulation in its membrane potential and "fires" when this quantity exceeds a fixed threshold, after which the process repeats.

A taxonomy of stochastic models for spiking neurons can be constructed in terms of increasing abstraction from the deterministic IF setup. At the most basic level, a stochastic spiking model is obtained by combining the deterministic IF mechanism with stochastic spike inputs. These are almost universally modeled as a Poisson process, which is considered a

reasonable approximation to the superposition of incoming spike trains $\sum_{j:C_{ji}\neq 0} S_j(t)$ even if not for an individual neuron's spiking activity. Under the simple assumption of homogeneous Poisson spiking input, certain properties of the IF model are analytically tractable, including the spiking rate and interspike interval length (Burkitt, 2006a). Empirically, however, more complex stochastic input is required to reproduce observed neuronal behavior. For example, Jackson (2004) demonstrates that superposed doubly stochastic Poisson spike trains with fractional Gaussian noise intensities induce long-range dependence in the spiking output of an IF model, and that both the input and output processes better match the observed statistics of spike trains from cortical neurons.

At a further level of abstraction, the membrane potential $v(t)$ is itself modeled as a stochastic process. This approach can be thought of as modeling the combined noise from both intrinsic (e.g. biological mechanisms) and extrinsic (e.g. spike arrival times) sources directly in terms of their influence on $v(t)$. One common choice is the Ornstein-Uhlenbeck model, which replaces the right hand side of Eq. 4.7 can be replaced with $-(v(t)-v_0)+\mu(t)+\sigma(t)\sqrt{2C_m}\xi(t)$, where $\xi(t)$ is a Gaussian white noise. The drift term $\mu(t)$ can be used to model input to the neuron; simple models such as the periodic input $\mu(t) = \mu_0 + \mu_1 \cos(\omega t + \phi_0)$ can be studied analytically in terms of their spiking behavior (Burkitt, 2006b). A final level of abstraction is obtained by ignoring the IF model's mechanistic approach to spike generation via membrane potential thresholding, and instead directly modeling a neuron's instantaneous propensity to fire, or *firing rate*. This is identified with the intensity $\lambda(t)$ of an inhomogeneous Poisson process. A typical parameterization (Gilson et al., 2010) is

$$\lambda_i(t) = \nu_i(t) + \sum_{j\neq i} \sum_{t_s\in\mathbb{S}_j} C_{ji}(t)\Theta(t-t_s)\frac{1}{\tau}\exp\left(\frac{t-t_s}{\tau}\right), \tag{4.8}$$

where $\nu_i(t)$ represents a contribution to the firing rate from sources exogenous to the network (e.g. stimulation) and $\tau > 0$ is a time constant, in milliseconds, controlling the time decay of excitement from incoming spikes.

A major topic of study in modern neuroscience, including some work presented in this

thesis, is that of *network plasticity*, which refers to the evolution of the connectivity matrix $C(t)$ and its consequent impact on the behavior of the spiking network. At sufficiently long timescales, the main biological mechanism driving this evolution is *spike timing dependent plasticity* (STDP), first postulated by Hebb (1949) and by now validated through extensive experimentation (Bi and Poo, 1998). STDP specifies a mechanism by which changes to the connectivity component $C_{ij}(t)$ depend on the relative timing of spike pairs from neurons i and j. Let $t \in \mathbb{S}_i$, $t' \in \mathbb{S}_j$ and define $\Delta t = t - t'$, $t^* = \max\{t, t'\}$. Then for a given time step δt and learning rate $\eta > 0$, the STDP rule can be expressed as

$$C_{ij}(t + \delta t) = C_{ij}(t) + \eta \left[\alpha f(C_{ij}(t)) \frac{|\Delta t|}{\tau_p} \exp\left(-\frac{|\Delta t|}{\tau_p} \right) \right], \tag{4.9}$$

where $f(C_{ij}(t)) = (1 - C_{ij}(t)/C_{\max})^\gamma$ if $\Delta t > 0$ and $f(C_{ij}(t)) = (C_{ij}(t)/C_{\max})^\gamma$ if $\Delta t \leq 0$ (Gilson et al., 2010; Lajoie et al., 2017). Thus for a single pair of spikes in neurons i and j, $C_{ij}(t)$ increases if j spikes before i and decreases otherwise. The magnitude of this change depends on Δt and $f(C_{ij}(t))$, the latter of which models saturation of connectivity as it approaches some value $C_{\max} > 0$.

The theory of neuronal communication and network plasticity is developed almost exclusively in terms of the spiking model, but it is rare for spikes to be directly recorded in modern experimental neuroscience. Most often, the observed quantity is a (multivariate) waveform measuring electromagnetic fluctuation associated with neural activity at specified locations on the cortical surface, as for example in electrocorticography (ECoG), electroencephalography (EEG), and magnetoencephalography (MEG). This motivates a model and set of associated statistical procedures for converting between spikes and waveforms. Here, it is important to note the transition from characterizing the spiking behavior of individual neurons to that of collections of neurons within the spatial region recorded by a single electrode. Denote by $\{A_r\}_{r=1}^R$ a partition of N recorded neurons into R regions. Let $y_r(t)$, $r = 1, \ldots, R$ indicate LFP measurements at region r and time t, with $\boldsymbol{y}(t) = (y_1(t), \ldots, y_R(t)) \in \mathbb{R}^R$ and

$\boldsymbol{S}(t) = (S_1(t), \ldots, S_N(t)) \in \mathbb{R}^N$. Then the multivariate LFP recording can be modeled as

$$\boldsymbol{y}(t) = \int_{\tau_1}^{\tau_2} H(\tau)\boldsymbol{S}(t+\tau)d\tau, \tag{4.10}$$

where $H(\tau) \in \mathbb{R}^{R\times N}$ is a matrix of convolution filters. Hall et al. (2014) estimate $H(\tau)$ by least squares and propose a method to estimate $\boldsymbol{S}(t)$ given $\boldsymbol{y}(t)$ based on Weiner deconvolution. More recent proposals for simulation of LFPs from spiking neurons attempt to account for biological details such as neuron morphology and the spatial layout of recording electrodes with respect to neuron cell bodies (Telenczuk et al., 2020). A proposal for jointly modeling observed waveforms and the latent activity of the underlying neurons was proposed by Friston et al. (2003) under the name "dynamic causal modeling;" however, estimation in this nonlinear, continuous-time state space framework is challenging and the method has not been widely used for simulation (Daunizeau et al., 2011). Despite clear scientific interest in relating spiking neurons to the waveform measurements of modern recording devices, methodological development and validation is limited by the lack of publicly available data.

Finally, we consider the experimental evidence justifying simulation-based approaches to understanding neural responses to stimulation. Interestingly, this literature is particularly well-developed for human subjects, where deep-brain stimulation (DBS) is already used to treat some neurodegenerative diseases. Computational models of pathological activity and degeneration at the neuronal level have become standard for testing hypotheses regarding the DBS mechanism (Wang et al., 2015), which remains unknown, and point the way towards closed-loop systems in which stimulation is tailored to specific patterns of activity observed in the patient (Widge et al., 2018). For nonhuman primate subjects, Lajoie et al. (2017) demonstrate that a Poisson spiking model based on Eq. 4.8 reproduces the long-term plasticity reported experimentally by Jackson et al. (2006) upon stimulation with an implanted electrode. Subsequent work based on a network of stochastic IF neurons predicts the network activity resulting from four different stimulation paradigms, without the need to individually tune simulation parameters for each trial (Shupe and Fetz, 2021)

4.5 *Identification of Optimal Stimulation Protocol in a Poisson Spiking Network*

We propose and implement an algorithm for identifying the optimal parameters for the stimulation protocol of Yazdan-Shahmorad et al. (2018), as applied to a simulated population of Poisson spiking neurons. A major clinical objective of research into stimulation-induced neuroplasticity can be stated as follows: given a current connectivity structure $C \in \mathbb{R}^{N\times N}$, target connectivity $\tilde{C} \in \mathbb{R}^{N\times N}$, and intervention protocol f_θ parameterized by controllable inputs $\theta \in \mathbb{R}^p$, identify an intervention θ^* such that $f_{\theta^*}(C) \approx \tilde{C}$. Ethical and resource constraints heavily limit the amount of data that can be gathered from live experimentation. Simulation-based models of network activity, studied under conditions of exogenous stimulation and incorporating network plasticity via STDP rules, thus emerge as a useful alternative.

Our simulation framework adopts the Poisson spiking model introduced in the previous section. Each neuron in our simulated network generates spikes as an inhomogeneous Poisson process with intensity given by Eq. 4.8; we set the time constant $\tau = 5$ms. In the context of an optogenetic stimulation protocol (Yazdan-Shahmorad et al., 2018), the exogenous input term $\nu_i(t)$ represents the instantaneous modification to the firing rate induced by laser stimulation at a given site. We implement a generalized version of the STDP update rule in Eq. 4.9, under which $C_{ij}(t+\delta t) = C_{ij}(t) + \eta W(\Delta t, C_{ij}(t))$ with

$$W(\Delta t, C_{ij}(t)) = \begin{cases} \left(1 - \frac{C_{ij}(t)}{C_{\max}}\right) \alpha^+ \frac{|\Delta t|}{\tau^+} \exp(-|\Delta t|/\tau^+) & \Delta t < 0 \\ \left(\frac{C_{ij}(t)}{C_{\max}}\right) \alpha^- \frac{|\Delta t|}{\tau^-} \exp(-|\Delta t|/\tau^-) & \Delta t > 0 \end{cases}.$$

This allows the parameterization of the connectivity change to depend on the sign of Δt, in line with experimental evidence that STDP is asymmetric around zero in the time delay (Kempter et al., 1999; Gilson et al., 2010); we follow these references in setting $\alpha^+ = 30, \alpha^- = 10, \tau^+ = 8.5$ms, $\tau^- = 16$ms.

4.5.1 Illustration of model components

We illustrate the simulation framework through an example based on the "functional subgroup" study of Lajoie et al. (2017). This computational study proposes to simulate from $N = 60$ neurons, initially connected via a random realization of the sparse connectivity matrix $C(0)$ with edges

$$C_{jk}(0) = \begin{cases} c & U_{jk} < p \\ 0 & \text{otherwise} \end{cases},$$

where U_{jk}, $j, k = 1, \ldots, N$ are i.i.d. standard uniform random variables, $c = 0.005$ is the initial connection strength, and $p = 0.3$ is the expected sparsity level. The neurons are divided into three equal groups $g \in \{1, 2, 3\}$, each driven by a truncated sinusoidal pulse $\nu_(t)$

$$\nu_g(t) = \max\{0, a \sin(2\pi\lambda t + \omega_g)\},$$

with amplitude $a = 100$, frequency $\lambda = 2$Hz. The pulses are offset from one another by the phase shift $\omega_g = \frac{2}{3}(g-1)\pi$. This setup is designed to illustrate the effect of the STDP update rule on the connectivity matrix: despite the random, group-unaware initialization of the connectivity, repeated synchronous activation of each subgroup is expected to result in greater within-group connections and relatively diminished between-group connections in the final connectivity matrix. We simulate from this system for $T = 60$s, at a discretization interval of $\delta t = 0.5$ms. A summary of the experiment and outcome is shown in Figure 4.7.

4.5.2 Identification of optimal stimulation protocols

We apply the simulator to the problem of identifying the optimal parameters for the optogenetic stimulation protocol of Yazdan-Shahmorad et al. (2018) to achieve a desired connectivity structure $\tilde{C}$. The optogenetic stimulation protocol consists of short (5 ms) pulses of

stimulation at 5Hz at each of two network locations:

$$\nu_\ell(t) = \begin{cases} s\mathbb{1}\{0 \leq t \mod 200 \leq 5\} & \ell = j \\ s\mathbb{1}\{0 \leq (t-d) \mod 200 \leq 5\} & \ell = k \\ 0 & \text{otherwise} \end{cases} .$$

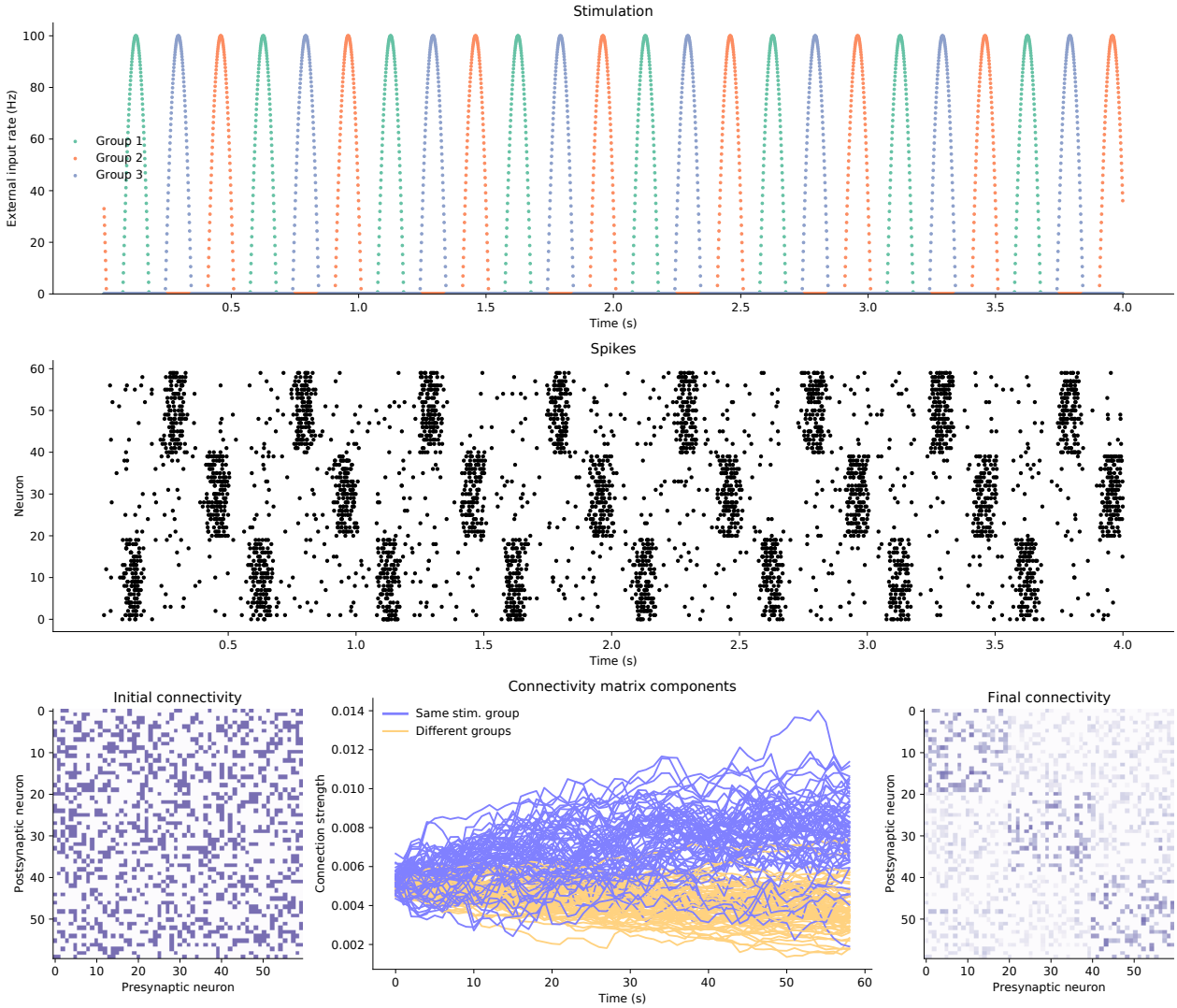

Figure 4.7: Simulator output under the "functional subgroup" stimulation protocol. Top: visualization of the exogenous stimulation signal $\eta_g(t)$ for each group $g \in \{1, 2, 3\}$. Middle: Simulated spike trains for each neuron under the stimulation protocol. Bottom left: Initial connectivity matrix, with connections selected uniformly at random and initialized to a constant value. Bottom middle: Evolution of the nonzero connectivities over the simulated stimulation trial. Bottom right: Final connectivity matrix. The within-group connections are relatively enhanced, while the between-group connections are diminished.

The parameters are the location indices $j, k \in \{1, \ldots, N\}$ and the delay $d > 0$ between pulse trains. The stimulation size s, which models the instantaneous effect of optogenetic stimulation on the neural spike rate, is currently unknown but could be estimated from spike train data. We set $s = 200$. Note that in contrast to the example above, stimulation is sparse both spatially (i.e. in terms of neurons directly stimulated) and in time.

Given an initial connectivity $C(0)$ and target connectivity $\tilde{C}$, we propose to identify a stimulation protocol by brute-force search over the parameter space. For each spatial configuration of the laser locations $j \neq k$ and delay level d, we simulate from the spiking Poisson population model for T seconds with a discretization interval of δt. The stimulation protocol parameters minimizing the Frobenius norm of $C(T) - \tilde{C}$ are selected as the optimal intervention. Multiple trials for each setting can be run in parallel to characterize the uncertainty induced by the stochastic model for spike generation.

4.5.3 Results

We apply the method to a network of $N = 6$ neural regions. The target connectivity $\tilde{C}$ is generated from a single simulation run, initialized at $C(0)$ and with true stimulation

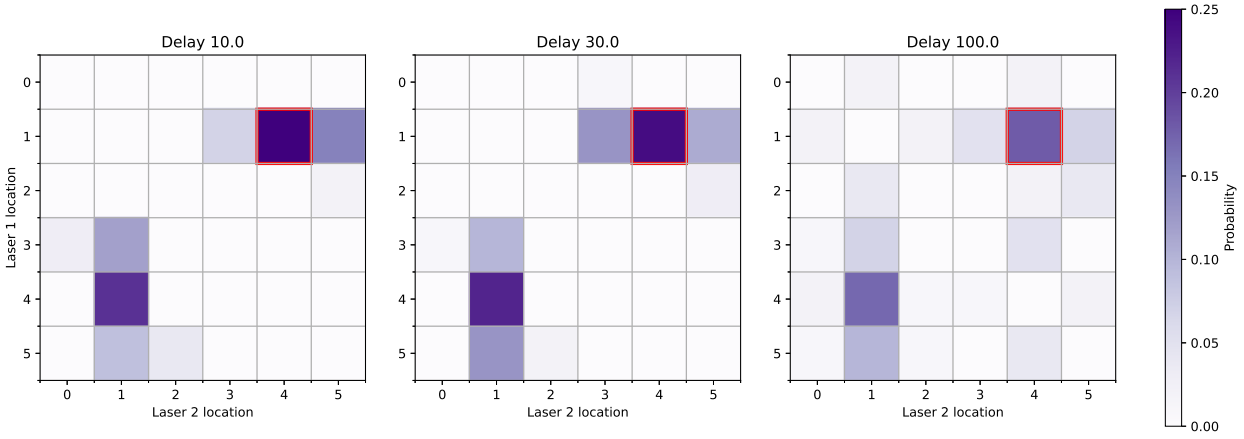

Figure 4.8: Heatmap of best-peforming protocol parameters $\hat{\theta}$ over 100 trials. The "true" laser location generating the target matrix $\tilde{C}$ is highlighted in red.

parameters $\theta^* = (j^*, k^*, d^*)$, with $(j^*, k, ^*) = (1, 4)$ and $d^* = 30$ms. For $n = 100$ simulation trials, we simulate the trajectory of the connectivity matrix $C(t)$, initialized at $C(0)$, under the stimulation protocol with parameters $\theta = (j, k, d)$ for each $j \neq k \in \{1, ..., N\}$ and $d \in \{10, 30, 100\}$ms. Runs are simulated for $T = 300$s and discretized at $\delta t = 0.5$ms. For each trial, we record the best-performing parameter $\hat{\theta}$; the result over many trials yields an empirical distribution for $\hat{\theta}$ under the simulation mechanism. A heatmap of these results, per delay value, is shown in Figure 4.8.

We observe that the method recovers the true arrangement of the laser stimulation sites as the configuration selected with greatest frequency across trials. There is relatively less capacity to distinguish the correct delay, though we note that for $d = 100$ms the results are more diffuse, as the delay-dependent effects of STDP begin to decay in magnitude. We also observe that the second-most-likely configuration in each setting involves reversing the laser placements. Finally, we compare the average final connectivity under the selected protocol to that of the target connectivity in Figure 4.9.

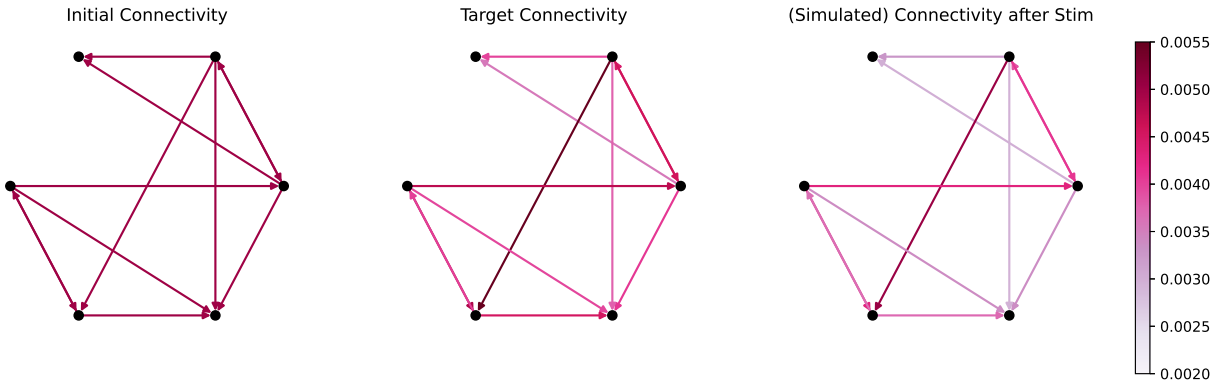

Figure 4.9: Initial (left), target (center), and predicted final connectivity under the selected protocol (right) for the protocol search method.

4.6 Discussion

In this chapter we have taken a statistical perspective on two major statistical challenges for real-world deployment of brain stimulation technologies. First, we identified and addressed the problem of group-level heterogeneity in a novel dataset derived from ECoG recordings of the macaque cortex. We have shown that the average pairwise distance between the elements of both session and subject-level partitions of the data are significant with respect to a permutation-based null distribution. This is of practical importance from the prediction standpoint, as session-level stratification of the train-test split dramatically decreases prediction performance of the least squares estimator. We survey and implement a broad range of statistical proposals to account for known, group-level heterogeneity in the estimation of a linear conditional mean. Our results demonstrate that extension to a heteroskedastic variance model or an estimation scheme accounting for unmodeled confounders results in a significantly improved linear prediction rule on unseen sessions. On the other hand, prediction results for the subject-level hold-out are uniformly negative.

As a consideration for future work, we note that the regression methods surveyed involved either trivial or crude models of the within-session dependence structure; in reality, there is likely to be both spatial (across electrodes) and temporal (across stimulation blocks) dependence among the observed coherence changes. Parameterization and estimation of this structure is of scientific interest in its own right. Additional "control" data from the experiments of Yazdan-Shahmorad et al. (2018), in which recordings were made while no stimulation occurred, is available but unused in the analyses thus far. While this data is itself significantly heterogeneous across control sessions, it may be useful for estimating session-level variation, as in Gagnon-Bartsch et al. (2013).

In the second part of the chapter, we offer a statistically oriented overview of spiking models and implement a simulator that combines inhomogeneous Poisson neurons, spike timing dependent plasticity in the connectivity matrix, and an exogenous stimulation signal based on recent experimental work. We demonstrate a simple search procedure by which the

simulator is used to identify optimal interventions given an initial and target connectivity state.

There are several considerations for subsequent development of this approach. First, the brute-force search method is computationally infeasible for large-scale networks or more complex parameterizations of the stimulation protocol. Control-based approaches based on deterministic models have been proposed for both spike train (Li et al., 2012) and functional connectivity (Menara et al., 2019) targets, but like our approach these have only been validated in synthetic settings. Second, the simulation-based approach can be developed to further distinguish between processes that occur at the cellular versus regional scale. In the optogenetic protocol, stimulation and recording takes place at the regional scale, while STDP occurs at the level of individual cells. One solution is to simulate a collection of artificial neurons for each region, each of which receives roughly the same stimulation input. Lajoie et al. (2017) find that the regionally-averaged connectivity dynamics under such a setup closely reproduce what would be expected at the cellular level by STDP.

Finally, further effort is required to reconcile spike simulations with waveform data. As the data itself is obtained as a waveform, it is tempting from a statistical perspective to ignore the spiking paradigm entirely and simply estimate a model for stimulation-induced structural change in a multivariate ECoG time series. Unfortunately this is unlikely to be satisfactory to neuroscientists, as the scientific interest lies in exploring the consequences of stimulation under a known (though flexibly parameterized) biological mechanism specified at the spiking level, namely STDP. Using the ECoG data analyzed previously, a method to convert waveforms to spikes could be used to estimate the magnitude of $\nu_i(t)$, i.e. the instantaneous increase in spike intensity due to optogenetic stimulation. On the other hand, converting from spikes to waveforms would allow us to simulate artificial ECoG data and explore the connection between (simulated) anatomical connectivity among spiking neurons and statistical notions of connectivity between cortical regions.

Chapter 5

CONCLUSION

This dissertation presents research contributions in two areas: long-range dependent multivariate time series and statistical analysis of neural recording data. At a high level, these contributions are unified in their focus on the development and implementation of statistical tools for complex, sequentially dependent data, ranging from natural language and music to high-frequency voltage measurements in the cerebral cortex. We contribute a statistical perspective on the fundamental question of memory in neural networks; introduce a new model for the spectrum of a multivariate long memory time series; develop a signal processing and nonlinear modeling pipeline for a novel dataset at the forefront of brain stimulation research; address major sources of heterogeneity in the data with an emphasis on prediction in clinically relevant settings; and develop a simulation-based approach for targeted reorganization of the connectivity between spiking neurons.

The applied areas spanned by this research continue to develop rapidly and thus motivate several directions for further investigation. In Chapter 2, our analysis of long memory in deep neural networks focuses on recurrent architectures. The largest and most successful language models now make use of the transformer architecture (Vaswani et al., 2017; Brown et al., 2020). Unlike the RNN or LSTM architectures, the hidden state of a transformer does not admit a causal representation in terms of the input sequence, nor is it inherently sensitive to the order of the inputs. Analogous to our analysis of recurrent architectures, recent analysis of self-attention reveals representational deficiencies that challenge the currently popular notion of its sufficiency in practice (Dong et al., 2021). In light of the remarkable generative capabilities of these models, a next step for our work would be to investigate long memory in the hidden representations or (embedded) output of transformer-based language models.

Aside from investigation of the models themselves, there remains the linguistic question of how the statistical property of long memory relates to qualitative aspects of language as evaluated by human listeners. Many simple statistical summaries of text have thus far demonstrated only weak correlation with human evaluations of their naturalness or quality (Novikova et al., 2017; Chaganty et al., 2018). In developing a framework for long memory analysis of language and music data, we have enabled the extension of this descriptive work to more sophisticated statistical features. It is of course unlikely that language "quality" is reducible to any scalar summary, but statistical perspectives may be useful in summarizing or diagnosing specific aspects that are qualitatively relevant to a human audience. For example, long memory may be worth investigating as an indicator of long-range coherence in text, which continues to be a limitation distinguishing algorithmically generated text from the capabilities of human writers (Holtzman et al., 2020). Looking farther ahead, advances in brain-computer interface technology demonstrate the possibility of natural language generation directly from real-time human ECoG signals (Moses et al., 2021), directly connecting two applied topics in this dissertation and raising new methodological challenges for decoding of sentence and paragraph-length text.

In the frequency-domain model, we have contributed a positive solution for long memory time series to complement the largely negative results for recurrent neural networks demonstrated earlier in Chapter 2. This model may be suitable in particular for the analysis of neural recording data, for which both frequency-domain and long memory analysis are of ongoing scientific interest. The next step will be to extend the model to the nonstationary case; as a simple example, we could rephrase the study in Section 2.6 as a problem of changepoint estimation that may affect both low frequency (via the long memory) and higher frequency (particularly in terms of the cross-spectrum) components of the spectrum. A nonstationary extension to the spectral model could be defined in terms of the wavelet transform, which offers a principled decomposition of the second-order structure of a time series, and which naturally gives rise to decorrelated representations in the presence of long memory (Percival and Walden, 2006; Achard and Gannaz, 2016).

The analysis in Chapter 3 advances the state of the art for prediction of the large-scale changes in functional connectivity that result from local stimulation of the cortex. Nonetheless, it represents just a preliminary step towards comprehensive understanding of the brain's response to stimulation at a network level. The scientific directions for future investigation often bear useful formulation as time series problems. For example, as an immediate next step it will be worth investigating the question of nonstationarity in the ECoG data, which must be addressed to distinguish stimulation-induced connectivity change from potential background drift. Some "control" data is available from this experiment, but it is itself heterogeneous. Alternatively, we could use recently published recordings from a large-scale study involving passive or "naturalistic" ECoG observation in humans (Singh et al., 2021). Beyond nonstationarity, there is longstanding interest in distinguishing the effects of cortical stimulation over various timescales (Hariz, 2017; Zarzycki and Domitrz, 2020), towards which our simple distinction between during-stimulation and resting-state SIFCC can be considered an initial effort.

In Chapter 4 we show that even greater care in the statistical analysis of brain stimulation data will be required to translate these advances to the clinical setting. The main challenge lies in the significant heterogeneity of the data across experimental sessions and subjects; the origin of this heterogeneity is unknown but may include variations in the noise level, confounding due to unobserved factors, covariate shift, or other sources. While the statistical evidence we show is specific to ECoG recordings of brain stimulation, where it has not been addressed previously, this is a problem widely encountered in the life sciences (Leek et al., 2010). Given the success of the heteroskedastic or estimated-confounder approaches for linear prediction, a clear next step is to extend this work to the nonlinear setting. The methodological alternatives that we survey could also be augmented with the additional availability of "control" sessions to estimate session variation (Gagnon-Bartsch et al., 2013), or with more recent statistical tools that relax assumptions on the residual covariance (McKennan and Nicolae, 2020) or exploit invariance to estimate causal relationships (Rojas-Carulla et al., 2018).

Finally, research on brain stimulation is motivated in part by the clinical objective of designing stimulation interventions to achieve a desired connectivity pattern. This suggests a framing as control of a dynamical cortical network by a sequence of stimulation inputs that may depend on the present state of the network; this "closed-loop" framework stands in contrast to the always-on, state-agnostic "open-loop" implementation of current deep brain stimulation systems (Widge et al., 2018). At present, such control-based approaches seem confined to demonstration in simplified and simulated settings, as for example in the single-location linear state space approach of Yang et al. (2021) or the deterministic dynamical system of Menara et al. (2019). In our approach, simulation is used as an alternative to explicit specification of system evolution, with network-level dynamics emerging from the collective behavior of stochastic spiking neurons whose connectivity evolves in accordance with biologically validated rules. The main drawback of this approach is its computational expense, including the naive search over the stimulation parameter space. Further improvements to spiking neuron simulators in terms of their biological realism and reproduction of known responses to stimulation (Shupe and Fetz, 2021) can easily be incorporated into this framework for identification of stimulation interventions.

BIBLIOGRAPHY

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In {*USENIX*} *Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

Patrice Abry and Darryl Veitch. Wavelet analysis of long-range-dependent traffic. *IEEE Transactions on Information Theory*, 44(1):2–15, 1998.

Sophie Achard and Irène Gannaz. Multivariate wavelet Whittle estimation in long-range dependence. *Journal of Time Series Analysis*, 37(4):476–512, 2016.

Francis R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *International Conference on Machine Learning*, pages 33–40. PMLR, 2008.

Francis R Bach and Michael I Jordan. Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, 52(8):2189–2199, 2004.

Sohail Bahmani and Justin Romberg. Convex programming for estimation in nonlinear recurrent models. *arXiv preprint arXiv:1908.09915*, 2019.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Badi H. Baltagi. *Econometrics*. Springer, 4th edition, 2008.

Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004.

André M Bastos and Jan-Mathijs Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in Systems Neuroscience*, 9: 175, 2016.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Francois Belletti, Alex Beutel, Sagar Jain, and Ed Chi. Factorized recurrent neural architectures for longer range dependence. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR, 2018.

Yoshua Bengio and Paolo Frasconi. Credit assignment through time: Alternatives to backpropagation. In *Neural Information Processing Systems*, 1994.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Jan Beran. Fitting long-memory models by generalized linear regression. *Biometrika*, 80(4): 817–822, 1993.

Jan Beran. *Statistics in Musicology*. CRC Press, 2003.

Jan Beran, Yuanhua Feng, Sucharita Ghosh, and Rafal Kulik. *Long-Memory Processes: Probabilistic Properties and Statistical Methods*. Springer, 2013.

Stefano Bertelli and Massimiliano Caporin. A note on calculating autocovariances of long-memory processes. *Journal of Time Series Analysis*, 23(5):503–508, 2002.

Dimitri Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, 2015.

Julian Besag and Charles Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995.

Richard F Betzel, John D Medaglia, Ari E Kahn, Jonathan Soffer, Daniel R Schonhaut, and Danielle S Bassett. Structural, geometric and genetic factors predict interregional brain connectivity patterns probed by electrocorticography. *Nature Biomedical Engineering*, 3 (11):902–916, 2019.

Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18(24):10464–10472, 1998.

Alberto Bietti and Julien Mairal. Group invariance, stability to deformations, and complexity of deep convolutional representations. *The Journal of Machine Learning Research*, 20(1): 876–924, 2019.

Julien Bloch, Alexander Greaves-Tunnell, Eric Shea-Brown, Zaid Harchaoui, Ali Shojaie, and Azadeh Yazdan-Shahmorad. Cortical network structure mediates response to stimulation: an optogenetic study in non-human primates. *bioRxiv doi:10.1101/2021.05.17.444526*, 2021.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. In *International Coference on International Conference on Machine Learning*, pages 1881–1888. PMLR, 2012.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, 8(9):1263–1268, 2005.

Richard C Bradley et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. Calibration, entropy rates, and memory in language models. In *International Conference on Machine Learning*, pages 1089–1099. PMLR, 2020.

David R Brillinger. *Time Series: Data Analysis and Theory*. SIAM, 2001.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer, 1991.

Julia Brodsky and Clifford M Hurvich. Multi-step forecasting for long-memory processes. *Journal of Forecasting*, 18(1):59–75, 1999.

Morton B Brown and Alan B Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

Peter Bühlmann et al. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Anthony N Burkitt. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biological Cybernetics*, 95(1):1–19, 2006a.

Anthony N Burkitt. A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. *Biological Cybernetics*, 95(2):97–112, 2006b.

György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012.

Richard H Byrd, Jorge Nocedal, and Robert B Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming*, 63(1-3): 129–156, 1994.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

T Tony Cai, Cun-Hui Zhang, and Harrison H Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

David J Caldwell, Jeffrey G Ojemann, and Rajesh PN Rao. Direct electrical stimulation in electrocorticographic brain–computer interfaces: Enabling technologies for input to cortex. *Frontiers in Neuroscience*, 13:804, 2019.

Raymond J Carroll and David Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, 1988.

Celemony. Melodyne, 2018. URL http://www.celemony.com/en/melodyne/what-is-melodyne.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. The price of debiasing automatic metrics in natural language evalaution. In *Association for Computational Linguistics*, pages 643–653, 2018.

Jiahua Chen and Jun Shao. Iterative weighted least squares estimators. *The Annals of Statistics*, pages 1071–1092, 1993.

Shizhe Chen, Ali Shojaie, and Daniela M Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112(520):1697–1707, 2017.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, 2014.

Catherine J Chu, Mark A Kramer, Jay Pathmanathan, Matt T Bianchi, M Brandon Westover, Lauren Wizon, and Sydney S Cash. Emergence of stable functional networks in long-term human electroencephalography. *Journal of Neuroscience*, 32(8):2703–2713, 2012.

Ching-Fan Chung. Sample means, sample autocovariances, and linear regression of stationary multivariate long memory processes. *Econometric Theory*, 18(1):51–78, 2002.

Robert Cogburn and Herbert T Davis. Periodic splines and spectral estimation. *The Annals of Statistics*, pages 1108–1126, 1974.

Mike X Cohen. *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014.

Rainer Dahlhaus. Graphical interaction models for multivariate time series. *Metrika*, 51(2): 157–172, 2000.

Ming Dai and Wensheng Guo. Multivariate spectral analysis using Cholesky decomposition. *Biometrika*, 91(3):629–643, 2004.

Jean Daunizeau, Olivier David, and Klaas E Stephan. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage*, 58(2):312–322, 2011.

Richard A Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, 25(4):1077–1096, 2016.

Peter Dayan and Laurence F Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

Persi Diaconis and David Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021.

Deborah J Donnell, Andreas Buja, and Werner Stuetzle. Analysis of additive dependencies and concurvities using smallest additive principal components. *The Annals of Statistics*, pages 1635–1668, 1994.

Randal Douc, Eric Moulines, and David Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Chapman and Hall/CRC, 2014.

MA Edwardson, TH Lucas, JR Carey, and EE Fetz. New modalities of brain stimulation for stroke rehabilitation. *Experimental Brain Research*, 224(3):335–358, 2013.

Mark Fiecas, Rainer von Sachs, et al. Data-driven shrinkage of the spectral density matrix of a high-dimensional time series. *Electronic Journal of Statistics*, 8(2):2975–3003, 2014.

Alex Fornito, Andrew Zalesky, and Edward Bullmore. *Fundamentals of Brain Network Analysis*. Academic Press, 2016.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19 (4):1273–1302, 2003.

Johann A Gagnon-Bartsch, Laurent Jacob, and Terence P Speed. Removing unwanted variation from high dimensional data with negative controls. Technical report, University of California, Berkeley, 2013.

Abigail G Garrity, Godfrey D Pearlson, Kristen McKiernan, Dan Lloyd, Kent A Kiehl, and Vince D Calhoun. Aberrant "default mode" functional connectivity in schizophrenia. *American Journal of Psychiatry*, 164(3):450–457, 2007.

Ramazan Gençay, Faruk Selçuk, and Brandon J Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*. Elsevier, 2001.

Wulfram Gerstner and Werner M Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

Wulfram Gerstner, Werner M Kistler, Richard Naud, and Liam Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982.

John Geweke and Susan Porter-Hudak. The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4):221–238, 1983.

Matthieu Gilson, Anthony N Burkitt, David B Grayden, Doreen A Thomas, and J Leo van Hemmen. Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks V: Self-organization schemes and weight dependence. *Biological Cybernetics*, 103(5):365–386, 2010.

Christian Gourieroux and Joann Jasiak. Nonlinear innovations and impulse responses with application to VAR sensitivity. *Annales d'Economie et de Statistique*, pages 1–31, 2005.

Clive WJ Granger. The typical spectral shape of an economic variable. *Econometrica: Journal of the Econometric Society*, pages 150–161, 1966.

Clive WJ Granger and Roselyne Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980.

Alexander Greaves-Tunnell and Zaid Harchaoui. A statistical investigation of long memory in language and music. In *International Conference on Machine Learning*, pages 2394–2403. PMLR, 2019.

Chong Gu. *Smoothing Spline ANOVA Models*, volume 297. Springer Science & Business Media, 2013.

Thomas M Hall, Kianoush Nazarpour, and Andrew Jackson. Real-time estimation and biofeedback of single-neuron firing rates using local field potentials. *Nature Communications*, 5(1):1–12, 2014.

Edward James Hannan. *Multiple Time Series*. John Wiley & Sons, 1970.

Asad Haris, Ali Shojaie, and Noah Simon. Nonparametric regression with adaptive truncation via a convex hierarchical penalty. *Biometrika*, 106(1):87–107, 2019.

Marwan Hariz. My 25 stimulating years with DBS in Parkinson's disease. *Journal of Parkinson's Disease*, 7(s1):S33–S41, 2017.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Trevor J Hastie and Robert J Tibshirani. *Generalized Additive Models*. Routledge, 2017.

Biyu J He, John M Zempel, Abraham Z Snyder, and Marcus E Raichle. The temporal structures and functional significance of scale-free brain activity. *Neuron*, 66(3):353–369, 2010.

Harold Stanley Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, 1978.

Donald O Hebb. *The Organization of Behavior*. Wiley, 1949.

Ann M Hermundstad, Danielle S Bassett, Kevin S Brown, Elissa M Aminoff, David Clewett, Scott Freeman, Amy Frithsen, Arianne Johnson, Christine M Tipper, Michael B Miller, et al. Structural foundations of resting-state and task-based functional connectivity in the human brain. *Proceedings of the National Academy of Sciences*, 110(15):6169–6174, 2013.

Archibald Vivian Hill. Excitation and accommodation in nerve. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 119(814):305–355, 1936.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997a.

Sepp Hochreiter and Jürgen Schmidhuber. LSTM can solve hard long time lag problems. In *Neural Information Processing Systems*, 1997b.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4): 500–544, 1952.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

Jonathan RM Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 1981.

Jonathan RM Hosking. Asymptotic distributions of the sample mean, autocovariances, and autocorrelations of long-memory time series. *Journal of Econometrics*, 73(1):261–284, 1996.

Yuhao Huang, Boglárka Hajnal, László Entz, Dániel Fabó, Jose L Herrero, Ashesh D Mehta, and Corey J Keller. Intracortical dynamics underlying repetitive stimulation predicts changes in network connectivity. *Journal of Neuroscience*, 39(31):6122–6135, 2019.

Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116(1):770–799, 1951.

Clifford M Hurvich and Willa W Chen. An efficient taper for potentially overdifferenced long-memory time series. *Journal of Time Series Analysis*, 21(2):155–180, 2000.

IA Ibragimov and Yu V Linnik. *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, 1971.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical sciences*. Cambridge University Press, 2015.

Andrew Jackson, Jaideep Mavoori, and Eberhard E Fetz. Long-term motor cortex plasticity induced by an electronic neural implant. *Nature*, 444(7115):56–60, 2006.

B Scott Jackson. Including long-range dependence in integrate-and-fire models of the high interspike-interval variability of cortical neurons. *Neural Computation*, 16(10):2125–2195, 2004.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for sparse hierarchical dictionary learning. In *International Conference on Machine Learning*, pages 487–494. PMLR, 2010.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12:2297–2334, 2011.

Don H Johnson. Point process models of single-neuron discharges. *Journal of Computational Neuroscience*, 3(4):275–299, 1996.

Corinne Jones, Vincent Roulet, and Zaid Harchaoui. Kernel-based translations of convolutional networks. *arXiv preprint arXiv:1903.08131*, 2019.

Anatoli Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization*. Princeton University Press, 2020.

Stefanos Kechagias and Vladas Pipiras. Definitions and representations of multivariate long-range dependent time series. *Journal of Time Series Analysis*, 36(1):1–25, 2015.

Corey J Keller, Stephan Bickel, László Entz, Istvan Ulbert, Michael P Milham, Clare Kelly, and Ashesh D Mehta. Intrinsic functional architecture predicts electrically evoked responses in the human brain. *Proceedings of the National Academy of Sciences*, 108(25): 10308–10313, 2011.

Corey J Keller, Yuhao Huang, Jose L Herrero, Maria E Fini, Victor Du, Fred A Lado, Christopher J Honey, and Ashesh D Mehta. Induction and quantification of excitability changes in human cortical networks. *Journal of Neuroscience*, 38(23):5384–5398, 2018.

Richard Kempter, Wulfram Gerstner, and J Leo Van Hemmen. Hebbian learning and spiking neurons. *Physical Review E*, 59(4):4498, 1999.

Nils Köbis and Luca D Mossink. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, 114:106553, 2021.

Eric D Kolaczyk. Empirical likelihood for generalized linear models. *Statistica Sinica*, pages 199–218, 1994.

Andrey N Kolmogorov. On the representation of continuous functions of several variables by superpositions of continuous functions of lesser variable count. In *Dokl. Akad. Nauk SSSR*, volume 108, page 2, 1956.

Robert T Krafty and William O Collinge. Penalized multivariate Whittle likelihood for power spectrum estimation. *Biometrika*, 100(2):447–458, 2013.

Morten L Kringelbach, Alexander L Green, Sarah LF Owen, Patrick M Schweder, and Tipu Z Aziz. Sing the mind electric–principles of deep brain stimulation. *European Journal of Neuroscience*, 32(7):1070–1079, 2010.

Hans Rudolph Kunsch. Statistical aspects of self-similar processes. In *Proceedings of the First World Congress of the Bernoulli Society, 1987*, volume 1, pages 67–74. VNU Science Press, 1987.

Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. First-order optimization for superquantile-based supervised learning. In *IEEE 30th International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2020.

Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 2021.

Guillaume Lajoie, Nedialko I Krouchev, John F Kalaska, Adrienne L Fairhall, and Eberhard E Fetz. Correlation-based model of artificially induced plasticity in motor cortex by a bidirectional brain-computer interface. *PLoS Computational Biology*, 13(2):e1005343, 2017.

Louis Lapique. Recherches quantitatives sur l'excitation electrique des nerfs traitee comme une polarization. *Journal of Physiology and Pathololgy*, 9:620–635, 1907.

Steffen L Lauritzen. *Graphical Models*. Clarendon Press, 1996.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

Jeffrey Tullis Leek. *Surrogate Variable Analysis*. PhD thesis, University of Washington, 2007.

Tao Lei, Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Deriving neural architectures from sequence and graph kernels. In *International Conference on Machine Learning*, 2017.

A.J. Lerner and C.E. Schenk. Cerebral cortex. In Michael J. Aminoff and Robert B. Daroff, editors, *Encyclopedia of the Neurological Sciences (Second Edition)*, pages 662–671. Academic Press, 2014.

Eric C Leuthardt, Kai J Miller, Gerwin Schalk, Rajesh PN Rao, and Jeffrey G Ojemann. Electrocorticography-based brain computer interface-the Seattle experience. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):194–198, 2006.

Howard Levine. Robust tests for equality of variances. In I. Olkin, editor, *Contributions to Probability and Statistics*, pages 278–92. Stanford University Press, 1960.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Neural Information Processing Systems*, volume 27, pages 2177–2185, 2014.

Omer Levy, Kenton Lee, Nicholas FitzGerald, and Luke Zettlemoyer. Long short-term memory as a dynamically computed element-wise weighted sum. In *Association for Computational Linguistics*, 2018.

Lin Li, Il Memming Park, Austin Brockmeier, Badong Chen, Sohan Seth, Joseph T Francis, Justin C Sanchez, and Jose C Principe. Adaptive inverse control of neural spatiotemporal spike patterns with a reproducing kernel Hilbert space (RKHS) framework. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(4):532–543, 2012.

Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Alternating minimizations converge to second-order optimal solutions. In *International Conference on Machine Learning*, pages 3935–3943, 2019.

Wai Keung Li and A Ian McLeod. Fractional time series modelling. *Biometrika*, 73(1): 217–221, 1986.

Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

Tsungnan Lin, Bill G Horne, Peter Tino, and C Lee Giles. Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, 7(6):1329–1338, 1996.

Ignacio N Lobato. Consistency of the averaged cross-periodogram in long memory series. *Journal of Time Series Analysis*, 18(2):137–155, 1997.

Beth Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

Andres M Lozano and Nir Lipsman. Probing and regulating dysfunctional circuits using deep brain stimulation. *Neuron*, 77(3):406–424, 2013.

Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Neural Information Processing Systems*, pages 2627–2635, 2014.

Stéphane Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 2016.

B Mandelbrot. Some mathematical questions arising in fractal geometry. In *Development of Mathematics*, pages 795–811. Birkhauser, 2000.

Benoit B Mandelbrot and John W Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Georges Matheron. The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468, 1973.

Voichiţa Maxim, Levent Şendur, Jalal Fadili, John Suckling, Rebecca Gould, Rob Howard, and Ed Bullmore. Fractional Gaussian noise, functional MRI and Alzheimer's disease. *Neuroimage*, 25(1):141–158, 2005.

Peter McCullagh. *Tensor Methods in Statistics*. Chapman and Hall/CRC, 2018.

Chris McKennan and Dan Nicolae. Estimating and accounting for unobserved covariates in high-dimensional correlated data. *Journal of the American Statistical Association*, pages 1–12, 2020.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.

Tommaso Menara, Giacomo Baggio, Danielle S Bassett, and Fabio Pasqualetti. A framework to control functional connectivity in the human brain. In *IEEE Conference on Decision and Control*, pages 4697–4704. IEEE, 2019.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tomas Mikolov, Armand Joulin, Sumit Chopra, Michael Mathieu, and Marc'Aurelio Ranzato. Learning longer memory in recurrent neural networks. In *International Conference on Learning Representations*, 2015.

John Miller and Moritz Hardt. Stable recurrent models. In *International Conference on Learning Representations*, 2019.

Kai J Miller, Pradeep Shenoy, John W Miller, Rajesh PN Rao, Jeffrey G Ojemann, et al. Real-time functional brain mapping using electrocorticography. *Neuroimage*, 37(2):504–507, 2007.

Rosaleena Mohanty, William A Sethares, Veena A Nair, and Vivek Prabhakaran. Rethinking measures of functional connectivity via feature extraction. *Scientific Reports*, 10(1):1–17, 2020.

Brian B Monson, Andrew J Lotto, and Brad H Story. Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives. *The Journal of the Acoustical Society of America*, 132(3):1754–1764, 2012.

Isabel Moreno-Sánchez, Francesc Font-Clos, and Álvaro Corral. Large-scale analysis of Zipf's law in english texts. *PloS One*, 11(1):e0147073, 2016.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Eric Moulines, François Roueff, and Murad S Taqqu. A wavelet Whittle estimator of the memory parameter of a nonstationary Gaussian time series. *The Annals of Statistics*, 36 (4):1925–1956, 2008.

K Muandet, K Fukumizu, B Sriperumbudur, and B Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10 (1-2):1–144, 2017.

Yasuo Nakai, Hiroki Nishibayashi, Tomohiro Donishi, Masaki Terada, Naoyuki Nakao, and Yoshiki Kaneoke. Regional abnormality of functional connectivity is associated with clinical manifestations in individuals with intractable focal epilepsy. *Scientific Reports*, 11(1): 1–10, 2021.

Stephen G Nash. A survey of truncated-newton methods. *Journal of Computational and Applied Mathematics*, 124(1-2):45–59, 2000.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Simon Newcomb. *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. US Government Printing Office, 1895.

Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors*, 12(2):1211–1279, 2012.

Stephanie Noble, Dustin Scheinost, and R Todd Constable. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, 203: 116157, 2019.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Empirical Methods in Natural Language Processing*, pages 2231–2242. ACL, 2017.

Esa Nummelin and Pekka Tuominen. Geometric ergodicity of harris recurrent marcov chains with applications to renewal theory. *Stochastic Processes and Their Applications*, 12(2): 187–202, 1982.

Gabriel Koch Ocker, Ashok Litwin-Kumar, and Brent Doiron. Self-organization of microcircuits in networks of spiking neurons with plastic synapses. *PLoS Computational Biology*, 11(8):e1004458, 2015.

Jukka-Pekka Onnela, Jari Saramäki, János Kertész, and Kimmo Kaski. Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, 71(6):065103, 2005.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Steven Orey. Recurrent Markov chains. *Pacific Journal of Mathematics*, 9(3):805–827, 1959.

Art Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1): 90–120, 1990.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6): 1191–1253, 2003.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318. PMLR, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, volume 32, pages 8026–8037, 2019.

Karl Pearson. On the methematical theory of errors of judgement, with special reference to the personal equation. *Philosophical Transactions of the Royal Society of London, Series A*, 198(300-311):235–299, 1902.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, 2014.

Donald B Percival and Peter Guttorp. Long-memory processes, the Allan variance and

wavelets. In *Wavelet Analysis and its Applications*, volume 4, pages 325–344. Elsevier, 1994.

Donald B Percival and Andrew T Walden. *Wavelet Methods for Time Series Analysis*, volume 4. Cambridge University Press, 2006.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

Venkata K Pillutla, Vincent Roulet, Sham M Kakade, and Zaid Harchaoui. A smoother way to train structured prediction models. *Advances in Neural Information Processing Systems*, 2018.

Vladas Pipiras and Murad S Taqqu. *Long-range Dependence and Self-Similarity*. Cambridge University Press, 2017.

Maurice Bertram Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, pages 300–325, 1994.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019.

Adrian E Raftery. A model for high-order Markov chains. *Journal of the Royal Statistical Society, Series B*, pages 528–539, 1985.

Rajesh PN Rao. *Brain-computer Interfacing: An Introduction*. Cambridge University Press, 2013.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B*, 71(5):1009–1030, 2009.

Valdério A Reisen, Céline Lévy-Leduc, and Murad S Taqqu. An M-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference*, 187:44–55, 2017.

Peter M Robinson. Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, 47(1):67–84, 1991.

Peter M Robinson. Semiparametric analysis of long-memory time series. *The Annals of Statistics*, pages 515–539, 1994.

Peter M Robinson. Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, 23(5):1630–1661, 1995.

Peter M Robinson. Multiple local whittle estimation in stationary systems. *The Annals of Statistics*, 36(5):2508–2530, 2008.

R Terry Rockafellar, Johannes O Royset, and Sofia I Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1): 1309–1342, 2018.

Ori Rosen and David S Stoffer. Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, 94(2):335–345, 2007.

Thomas J Rothenberg. Approximate normality of generalized least squares estimates. *Econometrica*, pages 811–825, 1984.

Dominik Rothenhäusler, Nicolai Meinshausen, and Peter Bühlmann. Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*, pages 255–277. Springer, 2016.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *Journal of the Royal Statistical Society, Series B*, 83(2):215–246, 2021.

Havard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC press, 2005.

Victor M Saenger, Joshua Kahan, Tom Foltynie, Karl Friston, Tipu Z Aziz, Alexander L Green, Tim J van Hartevelt, Joana Cabral, Angus BA Stevner, Henrique M Fernandes, et al. Uncovering the underlying mechanisms and whole-brain dynamics of deep brain stimulation for parkinson's disease. *Scientific Reports*, 7(1):1–14, 2017.

G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

Gennady Samorodnitsky. *Stochastic Processes and Long Range Dependence*. Springer, 2016.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61, 2010.

Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

Claude E Shannon. Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30(1):50–64, 1951.

Jun Shao. *Mathematical Statistics*. Springer, 2003.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit Yan Yeung, Wai Kin Wong, and Wangchun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Neural Information Processing Systems*, 2015.

Katsumi Shimotsu. Gaussian semiparametric estimation of multivariate fractionally integrated processes. *Journal of Econometrics*, 137(2):277–310, 2007.

Katsumi Shimotsu and Peter CB Phillips. Exact local Whittle estimation of fractional integration. *The Annals of Statistics*, 33(4):1890–1933, 2005.

Larry Shupe and Eberhard Fetz. An integrate-and-fire spiking neural network model simulating artificially induced cortical plasticity. *ENeuro*, 8(2), 2021.

Siddharth Sigtia, Emmanouil Benetos, Nicolas Boulanger-Lewandowski, Tillman Weyde, Artur S d'Avila Garcez, and Simon Dixon. A hybrid recurrent neural network for music transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2061–2065. IEEE, 2015.

Satpreet H Singh, Steven M Peterson, Rajesh PN Rao, and Bingni W Brunton. Mining naturalistic human behaviors in long-term video and neural recordings. *Journal of Neuroscience Methods*, 358:109199, 2021.

H Fairfield Smith. An empirical law describing heterogeneity in the yields of agricultural crops. *The Journal of Agricultural Science*, 28(1):1–23, 1938.

Fallaw Sowell. Maximum likelihood estimation of fractionally integrated time series models. *Journal of Econometrics*, 53(1-3):165–188, 1992.

SK Srinivasan and Gopalan Sampath. *Stochastic Models for Spike Trains of Single Neurons*. Springer Science & Business Media, 2013.

Elias M Stein and Rami Shakarchi. *Fourier Analysis: An Introduction*, volume 1. Princeton University Press, 2011.

Student. Errors of routine analysis. *Biometrika*, pages 151–164, 1927.

Felice T Sun, Lee M Miller, and Mark D'esposito. Measuring interregional functional connectivity using coherence and partial coherence analyses of fmri data. *Neuroimage*, 21(2): 647–658, 2004.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. Do neural nets learn statistical laws behind natural language? *PloS One*, 12(12):e0189326, 2017.

Kumiko Tanaka-Ishii and Armin Bunde. Long-range memory in literary texts: on the universal clustering of the rare words. *PLoS One*, 11(11):e0164658, 2016.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *International Conference on Machine Learning*, pages 896–903, 2005.

Bartosz Telenczuk, Maria Telenczuk, and Alain Destexhe. A kernel-based method to calculate local field potentials from networks of spiking neurons. *Journal of Neuroscience Methods*, 344:108871, 2020.

John Thickstun, Zaid Harchaoui, and Sham Kakade. Learning features of music from scratch. In *International Conference on Learning Representations*, 2017.

John Thickstun, Zaid Harchaoui, Dean P Foster, and Sham M Kakade. Invariances and data augmentation for supervised music transcription. In *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

Christopher Timmermann, Leor Roseman, Michael Schartner, Raphael Milliere, Luke TJ Williams, David Erritzoe, Suresh Muthukumaraswamy, Michael Ashton, Adam Bendrioua, Okdeep Kaur, et al. Neural correlates of the DMT experience assessed with multivariate EEG. *Nature Scientific Reports*, 9(1):1–13, 2019.

Dag Tjøstheim. Non-linear time series and Markov chains. *Advances in Applied Probability*, 22(3):587–611, 1990.

Masaru Tomita. Dynamic construction of finite-state automata from examples using hill-climbing. In *Proceedings of the Fourth Annual Conference of the Cognitive Science Society*, pages 105–108, 1982.

O Toussoun. Mémoire sur l'histoire du nil. *L'Institut Français D'Archaeologie Orientale (Cairo)*, 1925.

Wen-Jen Tsay. Maximum likelihood estimation of stationary multivariate ARFIMA processes. *Journal of Statistical Computation and Simulation*, 80(7):729–745, 2010.

Dick C van Leijenhorst and Th P Van der Weide. A formal derivation of heaps' law. *Information Sciences*, 170(2-4):263–272, 2005.

Mark CW Van Rossum, Guo Qiang Bi, and Gina G Turrigiano. Stable hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience*, 20(23):8812–8821, 2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, pages 5998–6008, 2017.

RF Voss and J Clarke. 1/f noise in music and speech. *Nature*, 258(5533), 1975.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer Science & Business Media, 2013.

Yujiang Wang, Frances Hutchings, and Marcus Kaiser. Computational modeling of neurostimulation in brain diseases. *Progress in Brain Research*, 222:191–228, 2015.

Zheng Wang, Li Min Chen, László Négyessy, Robert M Friedman, Arabinda Mishra, John C Gore, and Anna W Roe. The relationship of anatomical and functional connectivity to resting-state connectivity in primate somatosensory cortex. *Neuron*, 78(6):1116–1126, 2013.

Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

Haiguang Wen and Zhongming Liu. Separating fractal and oscillatory components in the power spectrum of neurophysiological signal. *Brain Topography*, 29(1):13–26, 2016.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*, 2016.

P Whittle. On the variation of yield variance with plot size. *Biometrika*, 43(3-4):337–343, 1956.

Peter Whittle. Estimation and information in stationary time series. *Arkiv för Matematik*, 2(5):423–434, 1953.

Alik S Widge, Donald A Malone Jr, and Darin D Dougherty. Closing the loop on deep brain stimulation for treatment-resistant depression. *Frontiers in Neuroscience*, 12:175, 2018.

Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Toru Yanagawa, Zenas C Chao, Naomi Hasegawa, and Naotaka Fujii. Large-scale information flow in conscious and unconscious states: an ECoG study in monkeys. *PloS One*, 8(11): e80845, 2013.

Yuxiao Yang, Shaoyu Qiao, Omid G Sani, J Isaac Sedillo, Breonna Ferrentino, Bijan Pesaran, and Maryam M Shanechi. Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. *Nature Biomedical Engineering*, pages 1–22, 2021.

Azadeh Yazdan-Shahmorad, Camilo Diaz-Botia, Timothy L Hanson, Viktor Kharazia, Peter Ledochowitsch, Michel M Maharbiz, and Philip N Sabes. A large-scale interface for optogenetic stimulation and recording in nonhuman primates. *Neuron*, 89(5):927–939, 2016.

Azadeh Yazdan-Shahmorad, Daniel B Silversmith, Viktor Kharazia, and Philip N Sabes. Targeted cortical reorganization using optogenetics in non-human primates. *Elife*, 7:e31034, 2018.

John R Younce, Meghan C Campbell, Tamara Hershey, Aaron B Tanenbaum, Mikhail Milchenko, Mwiza Ushe, Morvarid Karimi, Samer D Tabbal, Albert E Kim, Abraham Z Snyder, Joel S Perlmutter, and Scott A Norris. Resting-state functional connectivity predicts STN DBS clinical response. *Movement Disorders*, 36(3):662–671, 2021.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. CoRR abs/1409.2329, 2014.

Marcin Zygmunt Zarzycki and Izabela Domitrz. Stimulation-induced side effects after deep brain stimulation–a systematic review. *Acta Neuropsychiatrica*, 32(2):57–64, 2020.

Jingyu Zhao, Feiqing Huang, Jia Lv, Yanjie Duan, Zhen Qin, Guodong Li, and Guangjian Tian. Do RNN and LSTM have long memory? In *International Conference on Machine Learning*, pages 11365–11375. PMLR, 2020.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, 1949.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Gradient of the GSE Objective

The Gaussian semiparametric estimator is defined as

$$\hat{d}_{\text{GSE}} = \arg\min_{d\in\Theta} \log\det \widehat{G}(d) - 2\sum_{i=1}^{m} d_i \frac{1}{m}\sum_{j=1}^{m}\log\lambda_j,$$

with

$$\begin{aligned}
\widehat{G}(d) &= \frac{1}{m}\sum_{j=1}^{m} \text{Re}\left[\Lambda_j(d)^{-1} I_{T,X}(\lambda_j)\Lambda_j^*(d)^{-1}\right] \\
\Lambda_j(d) &= \text{diag}(\lambda_j^{-d} e^{i(\pi-\lambda_j)/2}) \\
I_{T,X}(\lambda_j) &= y_j y_j^*, \quad y_j = \frac{1}{\sqrt{2\pi T}}\sum_{t=1}^{T} x_t e^{-i\lambda_j t}, \quad \lambda_j = 2\pi j/T.
\end{aligned}$$

Denote by

$$\mathcal{L}_m(d) = \log\det \widehat{G}(d) - 2\sum_{i=1}^{m} d_i \frac{1}{m}\sum_{j=1}^{m}\log\lambda_j,$$

the objective to be minimized, given a fixed choice of the bandwidth parameter m.

The partial derivative of $\mathcal{L}_m(d)$ with respect to the element d_ℓ of the long memory vector d, for any $\ell = 1, ..., p$, is

$$\frac{\partial}{\partial d_\ell}\mathcal{L}_m(d) = \mathbf{Tr}\left[\widehat{G}(d)^{-1}\frac{\partial}{\partial d_\ell}\widehat{G}(d)\right] - \frac{2}{m}\sum_{j=1}^{m}\log\lambda_j.$$

Note that Fourier frequencies λ_j are strictly positive for $j \geq 1$, so that $\log \lambda_j$ is well defined.

For the term $\frac{\partial}{\partial d_\ell}\widehat{G}(d)$, note that the (h,k) element of the matrix $\widehat{G}(d)$ can be written as

$$\frac{1}{m}\sum_{j=1}^{m} \mathrm{Re}\left[I(\lambda_j)_{h,k} \exp\left((d_h + d_k)\log\lambda_j + \frac{i(\pi - \lambda_j)(d_h - d_k)}{2}\right)\right],$$

and therefore the derivative $\frac{\partial}{\partial d_\ell}\widehat{G}(d)$ is given by

$$\left(\frac{\partial}{\partial d_\ell}\widehat{G}(d)\right)_{h,k} = \begin{cases} \frac{1}{m}\sum_{j=1}^{m} \mathrm{Re}\left[I(\lambda_j)_{\ell,k} c_j^- \exp(c_j^+ d_k)\exp(c_j^- d_\ell)\right] & \text{for } h = \ell, h \neq k \\ \frac{1}{m}\sum_{j=1}^{m} \mathrm{Re}\left[I(\lambda_j)_{h,\ell} c_j^+ \exp(c_j^- d_h)\exp(c_j^+ d_\ell)\right] & \text{for } k = \ell, h \neq k \\ \frac{1}{m}\sum_{j=1}^{m} \mathrm{Re}\left[2I(\lambda_j)_{\ell,\ell}\log\lambda_j \exp(2d_\ell \log\lambda_j)\right] & \text{for } \ell = h = k \\ 0 & \text{otherwise} \end{cases},$$

where

$$c_j^- = \log\lambda_j - i\left(\frac{\pi - \lambda_j}{2}\right)$$
$$c_j^+ = \log\lambda_j + i\left(\frac{\pi - \lambda_j}{2}\right).$$

A.2 Bias Study for the Bandwidth Parameter

We demonstrate the potential for semiparametric estimation to incur bias when the bandwidth parameter m is set too high relative to the length T of the observed sequence. The bias results from inclusion of periodogram ordinates in the long memory estimator that capture behavior in the spectral density function not local to the origin.

We give an illustration for univariate time series, which allows us to take advantage of a convenient visual interpretation of the long memory as the slope of $\log I(\lambda_j)$ against $-2\log(\lambda_j)$ as $\lambda_j \to 0$. Figure A.1 shows the spectral density function corresponding to three

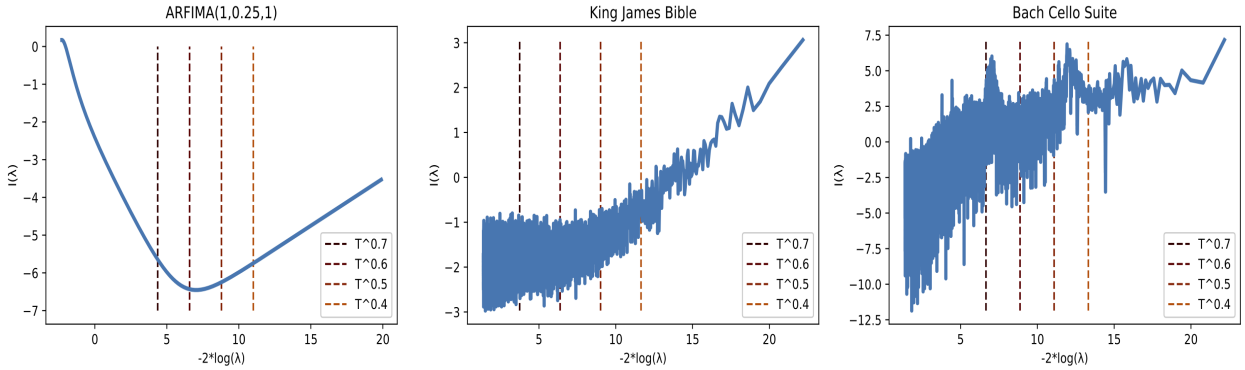

Figure A.1: Spectral density function of an ARFIMA(1,d,1) process (left) and smoothed estimates of the periodogram for the first coordinate of the embedded Bible text and Bach cello suite (center and right, respectively). Cutoff points associated with four choices of the bandwidth m are plotted as vertical dashed lines; the semiparametric estimate of the long memory for each sequence is essentially a measure of the slope based on the subset of points $(-2\log\lambda, \log I(\lambda))$ to the *right* of this line.

scalar processes: an ARFIMA(1,d,1) process with $d = 0.25$, a univariate projection of the embedded text from the King James Bible, and a univariate projection of the embedded Bach cello suite. For the ARFIMA process, the spectral density function can be computed exactly; for the other two sequences, it is estimated by the smoothed periodogram.

By marking the cutoff points $-2\log\lambda_m$ associated with different choices of m, we indicate the subset of points $(-2\log\lambda_j, \log I(\lambda_j))$ to the right of this cutoff used to compute the semparametric estimate of d. In the scalar case, this is essentially the slope of the SDF as λ approaches zero; thus it becomes clear that bias can be introduced when points sufficiently far from the origin are included. On the other hand, choosing m too small introduces the risk of high variance in the estimator; note for example that the estimate with $m = T^{0.4}$ for the Bach cello suite would be strongly influenced by a single point just to the left of $-2\log\lambda = 15$.

A.3 Simulation Study for the Total Memory Statistic

We compute the total memory statistic

$$\bar{d} = \mathbb{1}^T \hat{d}_{\text{GSE}}$$

for simulated fractionally differenced Gaussian white noise sequences of dimension $k = 200$. We simulate four different settings for the long memory parameter:

- **Zero**: Each coordinate of d is equal to zero.
- **Constant**: Each coordinate of d is set to the same value, $d = 0.25$.
- **Subset**: 90% of the coordinates are set to 0, while the remaining 10% are set to have strong long memory with $d = 0.4$.
- **Range**: The elements of d are drawn from a scaled Beta distribution with support on $(0, 0.25)$ and centered at 0.125.

For each setting, we simulate $n = 100$ sequences and compute the total memory. Results are plotted in Figure A.2, while in Table A.1 we compare the sample mean and variance of the estimator compared to the asymptotic value stated in Eq. (2.3).

Table A.1: Comparison of the sample mean and variance for the total memory estimator with the true total memory of the generating process and the asymptotic variance of the total memory estimator (both given in parentheses).

Setting	**Mean**	**Variance**
Zero	2.82×10^{-4} (0.0)	0.00801 (0.00698)
Constant	0.249 (0.25)	0.00793 (0.00698)
Subset	0.382 (0.04)	0.00804 (0.00698)
Range	0.101 (0.1029)	0.00696 (0.00698)

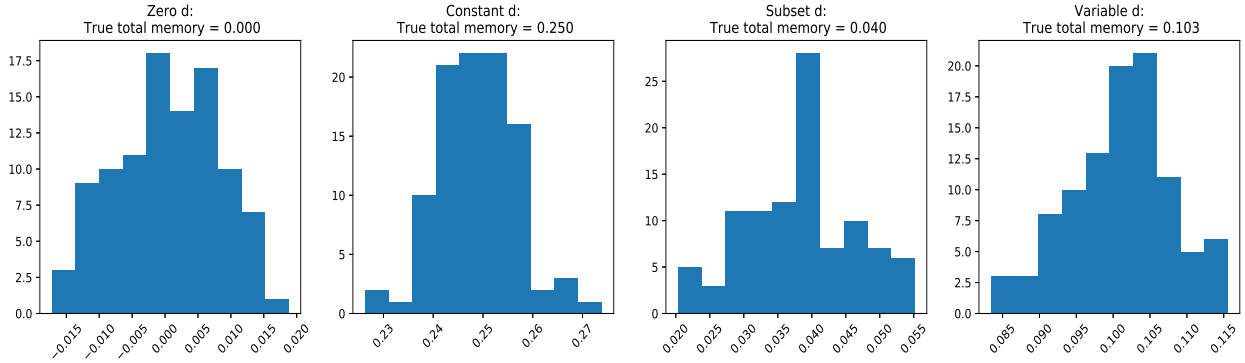

Figure A.2: Sample distribution of the total memory estimator $\bar{d}$ in four different simulation settings.

In each of these four diverse simulation settings, the total memory estimator accurately recovers the true underlying parameter of the data generating process.

A.4 Miscalibration of the Wald Test in High Dimension

Here we demonstrate that the standard Wald test can be badly miscalibrated in the high-dimensional regime, whereas testing for long memory with the total memory statistic remains well-calibrated. Recall that, given an estimate $\hat{d}_{\text{GSE}}$ of the multivariate long memory parameter, the Wald statistic for the null hypothesis $\mathcal{H}_0 : d = 0$ is computed as

$$t_{\text{Wald}} = \hat{d}_{\text{GSE}}^T (\Omega/m) \hat{d}_{\text{GSE}}.$$

This quantity is distributed as a $\chi^2(p)$ random variable under $\mathcal{H}_0$.

For the total memory, we compute

$$\bar{d} = \mathbb{1}^T \hat{d}_{\text{GSE}},$$

and in the main paper we have shown that this quantity is distributed as a $\mathcal{N}(0, \Omega/m)$ random variable when the true total memory $\bar{d}_0 = 0$.

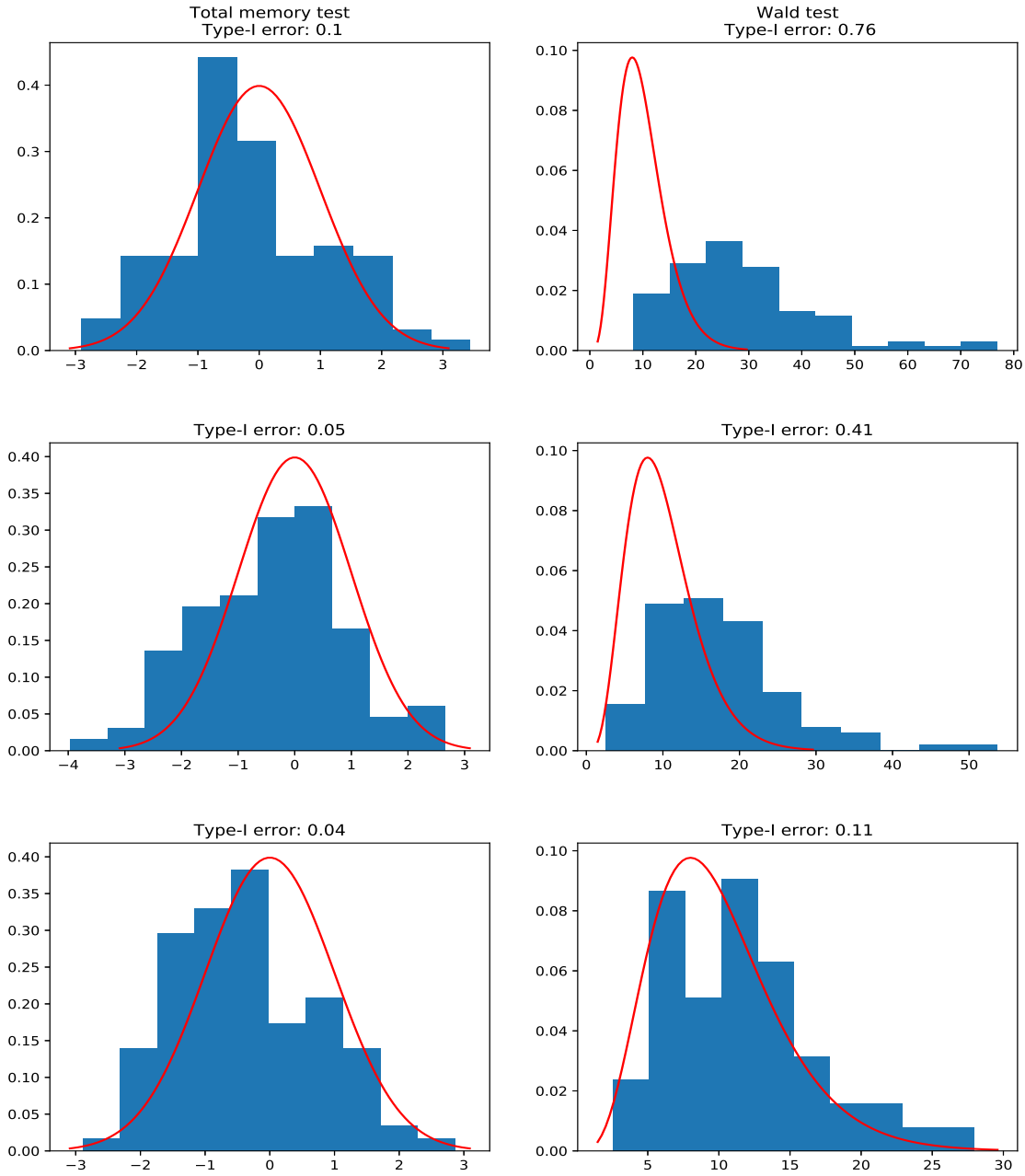

Figure A.3: Sample distribution of the test statistic over $n = 100$ trials for $m = \sqrt{T} = 256$ (top row), $m = 512$ (middle), and $m = 1280$ (bottom). Empirical type-I errors are computed using the critical value corresponding to a nominal type-I error of 0.05.

We simulate $n = 100$ realizations of length $T = 2^{16}$ from a standard Gaussian process (thus $d = 0$) of dimension $p = 200$, computing both the Wald and total memory test statistics. In Figure A.3, we plot the a comparison of the sample distribution of each test statistic against its asymptotic distribution over a range of values for m. For values of m close to p, we see that the empirical type-I error of the Wald test is severely inflated relative to the nominal level $\alpha = 0.05$; in other words, the test spuriously rejects the null and claims to find long memory when none exists at a rate much higher than accounted for. The total memory test, by contrast, largely avoids this issue, even in the case where there are barely more observations than dimensions.

Of course, with enough data, the Wald test becomes increasingly well-calibrated, but this is not at all an easy condition to satisfy while maintaining the integrity of the statistical analysis. We have already seen in Appendix A.2 that simply increasing m is not an option for real-world data, as this is likely to induce significant bias. On the other hand, the length T of the observed sequence would have to be enormous, even by machine learning standards, to achieve $m \gg p$ with the reasonable choice $m = \sqrt{T}$ when the dimension p is large. Finally, even if such data were available, we would likely prefer a method that allows valid inference at lower m for computational reasons.

A.5 Model Details for Music Data

The reduced version of the MusicNet model of Thickstun et al. (2018) used to obtain an embedding for the Bach cello suite is derived from the convolutional model implemented in `musicnet_module.ipynb`, a PyTorch interface to MusicNet available at https://github.com/jthickstun/pytorch_musicnet. We reduce the number of hidden states to 200, both for computational tractability in the optimization procedure and to achieve consistency with the embedding dimension for our natural language experiments.

The model is trained on the MusicNet training corpus with no further modification of the tutorial notebook. Successful training and an informative feature mapping are indicated by the competitive performance of the model, even despite the reduced dimension of the

hidden representation, in terms of the average precision of its predictions on the test set (see Table A.2). Results for our trained model (*longmem-embed*) are favorable in comparison to both short-time Fourier transform (STFT) and commercial software (Melodyne) baselines, while approaching the quality of the fully learned filterbank (Learned filterbank; Thickstun et al. (2017)) and state-of-the-art translation invariant network (Wide-translation-invariant; Thickstun et al. (2018)).

Table A.2: Performance Comparison for Models of MusicNet Data

Model	**Avg. Precision**
STFT	60.4
Melodyne	58.8
longmem-embed	65.1
Learned filterbank	67.8
Wide-translation-invariant	77.3

A.6 Impact of Embedding Choice on Long Memory Analysis

We evaluate the impact of embedding choice on estimated long memory from two perspectives. First, we include a re-analysis of the Bach cello suite data using the same MFCC features as used for the Miles Davis and Oum Kalthoum recordings. This allows us to state results for long memory estimation uniformly across a single choice of embedding, and to evaluate the impact of embedding choice on the long memory analysis across two very different but informative representations of the raw time series. The results (see Table A.3) show that the Bach data has long memory under both representations, though the average strength as measured by normalized total memory is somewhat variable.

Second, we consider a "negative control" experiment in which we re-estimate the long memory vector for the embedded Penn TreeBank training set after permuting the sequential ordering of the data. This addresses the question of whether our positive result truly captures

Table A.3: Long memory of Bach data by choice of embedding.

Embedding	**Norm. total memory**	**p-value**	**Reject $\mathcal{H}_0$?**
Mel-frequency cepstral coefficients	0.308	0.003	✓
Convolutional features	0.0997	$< 1 \times 10^{-16}$	✓

a sequence-dependent property of the data, or if it could have been produced spuriously as the consequence of other decisions related to the data analysis (including, for example, the choice of embedding). We compute the total memory statistic for $n = 100$ random permutations of the Penn TreeBank training data.

The results (see Figure A.4) show that the total memory of the permuted data is concentrated near zero, with a sample mean of 1.90×10^{-5} and standard error 0.00136; a one-sample test of the mean correspondingly fails to reject the null hypothesis $\mathcal{H}_0 : \bar{d} = 0$ with $p = 0.494$.

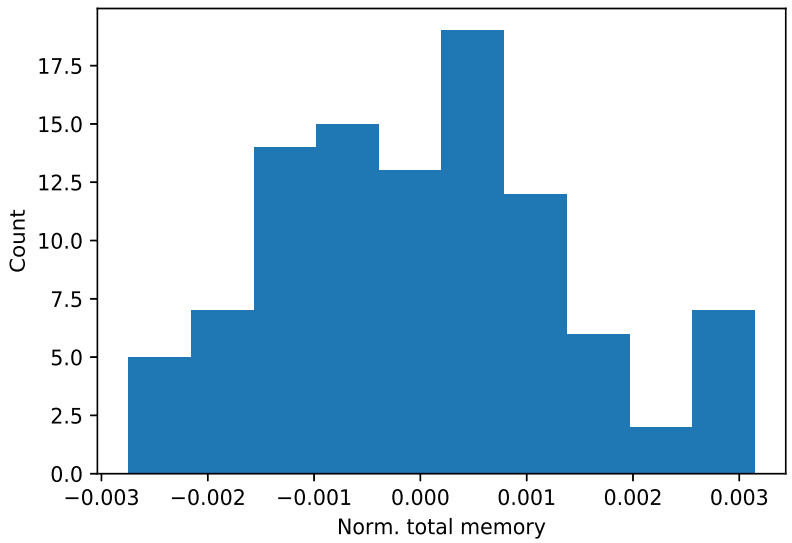

Figure A.4: Histogram of normalized total memory computed from $n = 100$ permutations of the Penn TreeBank training data.

A.7 Gradient Computations for the Spectral LRD Model

We compute the gradient of the penalized Whittle objective

$$J(\theta) = \mathcal{L}(\theta) + \Omega_\rho(\theta)$$

with respect to the model parameters $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, d)$. Let θ_k be an arbitrary element of the parameter vector θ. Then

$$\begin{aligned}\frac{\partial}{\partial\theta_k} J(\theta) &= \frac{\partial}{\partial\theta_k}\left[\left[\sum_{j=1}^{\lfloor T/2\rfloor} \mathbf{Tr}(I_j f(\lambda_j;\theta)^{-1}) - \log|f(\lambda_j;\theta)^{-1}|\right] + \Omega(\theta)\right] \\ &= \left[\sum_{j=1}^{\lfloor T/2\rfloor} \frac{\partial}{\partial\theta_k}\mathbf{Tr}(I_j f(\lambda_j;\theta)^{-1}) - \frac{\partial}{\partial\theta_k}\log|f(\lambda_j;\theta)^{-1}|\right] + \frac{\partial}{\partial\theta_k}\Omega(\theta),\end{aligned}$$

so that it suffices to compute the partial derivatives with respect to the penalty $\Omega(\theta)$ and to each of the trace and log-determinant terms in the pseudo-likelihood components

$$\mathcal{L}_j(\theta) \triangleq \mathbf{Tr}(I_j f(\lambda_j;\theta)^{-1}) - \log|f(\lambda_j;\theta)^{-1}|.$$

From the model definition, we have

$$f(\lambda_j;\theta)^{-1} = \Phi(\lambda_j;d) S(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta}) \Phi(\lambda_j;d)^*$$

so that

$$\begin{aligned}
\log|f(\lambda_j;\theta)^{-1}| &= \log|\Phi(\lambda_j;d)S(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})\Phi(\lambda_j;d)^*| \\
&= \log\left[|\Phi(\lambda_j;d)\Phi(\lambda_j;d)^*| \;\; |S(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})|\right] \\
&= \log|\Phi(\lambda_j;d)\Phi(\lambda_j;d)^*| + \log|S(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})| \\
&= \log\prod_{k=1}^{p}|1-e^{i\lambda_j}|^{2d_k} + \log|H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*| \\
&= \log\prod_{k=1}^{p}|1-e^{i\lambda_j}|^{2d_k} + 2\log|H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})| \\
&= \log\prod_{k=1}^{p}|1-e^{i\lambda_j}|^{2d_k} + 2\log\prod_{k=1}^{p}\left[\sum_{\ell=0}^{L}\alpha_{kk\ell}\cos(\ell\lambda_j)\right] \\
&= 2\sum_{k=1}^{p}\left[d_k|1-e^{i\lambda_j}| + \log\left[\sum_{\ell=0}^{L}\alpha_{kk\ell}\cos(\ell\lambda_j)\right]\right].
\end{aligned}$$

Next, considering the trace term, we write

$$\mathbf{Tr}(I_j f(\lambda_j;\theta)^{-1}) = y_j^*\Phi(\lambda_j;d)S(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})\Phi(\lambda_j;d)^* y_j$$

and define

$$z_j \triangleq y_j^*\Phi(\lambda_j;d)H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*$$

for ease of notation. Then the partial derivative is obtained as

$$\frac{\partial}{\partial\theta_k}\mathbf{Tr}(I_j f(\lambda_j;\theta)^{-1}) = \left[\frac{\partial}{\partial\theta_k}z_j\right]z_j^* + z_j\left[\frac{\partial}{\partial\theta_k}z_j^*\right].$$

For $\theta_k = d_k$ we have

$$\frac{\partial}{\partial d_k}z_j = \frac{\partial}{\partial d_k}\left[y_j^*\Phi(\lambda_j;d)H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*\right] = y_j^*\left[\frac{\partial}{\partial d_k}\Phi(\lambda_j;d)\right]H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*$$

Noting that

$$\frac{\partial}{\partial d_k}(1-e^{-i\lambda})^{d_k} = \frac{\partial}{\partial d_k}\left[2\sin\left(\frac{\lambda}{2}\right)^{d_k}\left[\cos\left(\frac{d_k}{2}(\pi-\lambda)\right)+i\sin\left(\frac{d_k}{2}(\pi-\lambda)\right)\right]\right],$$

we have

$$\frac{\partial}{\partial d_k}\Phi(\lambda_j;d)_{pq} = \begin{cases} 2\left[\operatorname{sign}\left(\frac{\lambda_j}{2}\right)\log\left|\frac{\lambda_j}{2}\right|\sin\left(\frac{\lambda_j}{2}\right)^{d_k}\left[\cos\left(\frac{d_k}{2}(\pi-\lambda_j)\right)+i\sin\left(\frac{d_k}{2}(\pi-\lambda_j)\right)\right]\right. & \\ \qquad \left. +\sin\left(\frac{\lambda_j}{2}\right)^{d_k}\left(\frac{\pi-\lambda_j}{2}\right)\left[\sin\left(\frac{\lambda_j}{2}\right)-i\cos\left(\frac{\lambda_j}{2}\right)\right]\right] & p=q=k \\ 0 & \text{otherwise.} \end{cases}$$

For $\theta_k = \alpha_{mn\ell}$ or $\theta_k = \beta mn\ell$ we have

$$\frac{\partial}{\partial\theta_k}z_j = \frac{\partial}{\partial\theta_k}\left[y_j^*\Phi(\lambda_j;d)H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*\right] = y_j^*\Phi(\lambda_j;d)\left[\frac{\partial}{\partial\theta_k}H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*\right]$$

with

$$\frac{\partial}{\partial\alpha_{mn\ell}}H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*_{pq} = \begin{cases}\cos(\ell\lambda_j) & p=m,q=n\\ 0 & \text{otherwise.}\end{cases}$$

and

$$\frac{\partial}{\partial\beta_{mn\ell}}H(\lambda_j;\boldsymbol{\alpha},\boldsymbol{\beta})^*_{pq} = \begin{cases}i\sin(\ell\lambda_j) & p=m,q=n\\ 0 & \text{otherwise.}\end{cases}$$

A.8 Details of Trigonometric and Complex Exponential Calculations

Here we provide the line-by-line arithmetic calculations that contribute to the proofs of Propositions 2.5.1 and 2.5.2. First, we calculate the inverse and complex conjugate of the quantity $(1-e^{-i\lambda})^{-d}$.

Let $z \in \mathbb{C}$ such that $z \neq 0$. The inverse can be written as $z^{-1} = \frac{z^*}{\|z\|^2}$. We first wish to show that we can write

$$(1 - e^{-i\lambda})^{-d} = |1 - e^{-i\lambda}|^{-d} e^{-id \arg(1-e^{-i\lambda})}.$$

When $z, w \in \mathbb{C}$, the expression z^w is defined as $e^{w \log z}$, where $\log z$ indicates the complex logarithm, which itself is defined either as a multivalued function or through a principal value. Here we shall take the principal value definition. Let $z = 1 - e^{-i\lambda}$ and $w = -d$. We write the polar form $z = |1 - e^{-i\lambda}| e^{i \arg(1-e^{-i\lambda})}$. The principal value of $\log z$ is $\log|1 - e^{-i\lambda}| + i \arg(1 - e^{-i\lambda})$, while the other values of the logarithm, considered as a multivalued function, are obtained by adding $2\pi ik$ to this quantity for any integer $k \neq 0$. Multiplying by $w = -d$ yields $w \log z = -d \log|1 - e^{-i\lambda}| - id \arg(1 - e^{-i\lambda})$. Finally, exponentiating, and using the property $e^{x+y} = e^x e^y$ along with the definition $z^w = e^{w \log z}$, yields

$$\begin{aligned} e^{w \log z} &= e^{-d \log|1-e^{-i\lambda}| - id \arg(1-e^{-i\lambda})} \\ &= e^{-d \log|1-e^{-i\lambda}|} e^{-id \arg(1-e^{-i\lambda})} \\ &= |1 - e^{-i\lambda}|^{-d} e^{-id \arg(1-e^{-i\lambda})} \end{aligned}$$

as desired. Therefore, for the inverse computation we can write

$$\begin{aligned} [(1 - e^{-i\lambda})^{-d}]^{-1} &= \frac{[|1 - e^{-i\lambda}|^{-d} e^{-id \arg(1-e^{-i\lambda})}]^*}{|1 - e^{-i\lambda}|^{-2d}} \\ &= |1 - e^{-i\lambda}|^{d} [e^{-id \arg(1-e^{-i\lambda})}]^* \\ &= |1 - e^{-i\lambda}|^{d} e^{id \arg(1-e^{-i\lambda})} \\ &= (1 - e^{-i\lambda})^{d}. \end{aligned}$$

For the complex conjugate, we have

$$\begin{aligned}
[(1-e^{-i\lambda})^{-d}]^* &= [|1-e^{-i\lambda}|^{-d}e^{-id\,\arg(1-e^{-i\lambda})}]^* \\
&= [|1-e^{i\lambda}|^{-d}e^{-id\,\arg(1-e^{-i\lambda})}]^* \\
&= [|1-e^{i\lambda}|^{-d}e^{-id(-\arg(1-e^{i\lambda}))}]^* \\
&= [|1-e^{i\lambda}|^{-d}e^{id\,\arg(1-e^{i\lambda})}]^* \\
&= |1-e^{i\lambda}|^{-d}e^{-id\,\arg(1-e^{i\lambda})} \\
&= (1-e^{i\lambda})^{-d}.
\end{aligned}$$

Next,

$$\begin{aligned}
1-e^{-i\lambda} &= 1-\cos(\lambda)+i\sin(\lambda) \\
&= 1-\sin\left(\frac{\pi}{2}-\lambda\right)+i\cos\left(\lambda-\frac{\pi}{2}\right) \\
&= \sin\left(\frac{\pi}{2}\right)+\sin\left(\frac{2\lambda-\pi}{2}\right)+i\left[\cos\left(\frac{\pi-2\lambda}{2}\right)-\cos\left(\frac{\pi}{2}\right)\right] \\
&= 2\sin\left(\frac{\lambda}{2}\right)\cos\left(\frac{\pi-\lambda}{2}\right)+2i\sin\left(\frac{\lambda}{2}\right)\sin\left(\frac{\pi-\lambda}{2}\right) \\
&= 2\sin\left(\frac{\lambda}{2}\right)\left[\cos\left(\frac{\pi-\lambda}{2}\right)+i\sin\left(\frac{\pi-\lambda}{2}\right)\right] \\
&= 2\sin\left(\frac{\lambda}{2}\right)e^{i(\pi-\lambda)/2},
\end{aligned}$$

which justifies the first line in the proof of Proposition 2.5.2.

A.9 Details of the Simulation Experiment for the Spectral LRD Model

We generate data from a multivariate time series of dimension $p = 3$ whose short-range spectral density function $S(\lambda)$ has components

$$
\begin{aligned}
S_{00}^{\mathrm{Re}}(\lambda) &= 0.2\cos(2\pi\lambda) + 0.1\cos(4\pi\lambda) + 0.2\cos(6\pi\lambda) \\
S_{11}^{\mathrm{Re}}(\lambda) &= -0.2\cos(2\pi\lambda) + 0.3\cos(4\pi\lambda) \\
S_{22}^{\mathrm{Re}}(\lambda) &= -0.2\cos(2\pi\lambda) + 0.3\cos(4\pi\lambda) + 0.1\cos(8\pi\lambda) - 0.1\cos(10\pi\lambda) + 0.2\cos(12\pi\lambda) \\
S_{10}^{\mathrm{Re}}(\lambda) &= 0.5\cos(2\pi\lambda) + 0.1\cos(4\pi\lambda) \\
S_{20}^{\mathrm{Re}}(\lambda) &= 0.5\cos(2\pi\lambda) + 0.1\cos(4\pi\lambda) - 0.2\cos(6\pi\lambda) + 0.3\cos(8\pi\lambda) \\
S_{21}^{\mathrm{Re}}(\lambda) &= -0.2\cos(2\pi\lambda) \\
S_{10}^{\mathrm{Im}}(\lambda) &= -0.2\sin(2\pi\lambda) - 0.1\sin(4\pi\lambda) \\
S_{20}^{\mathrm{Im}}(\lambda) &= 0.5\sin(2\pi\lambda) + 0.1\sin(4\pi\lambda) - 0.2\sin(6\pi\lambda) + 0.3\sin(8\pi\lambda) \\
S_{21}^{\mathrm{Im}}(\lambda) &= -0.2\sin(2\pi\lambda)
\end{aligned}
$$

To generate a simulated periodogram ordinate at Fourier frequency λ_j, we generate $Z_j = S^{1/2}(\lambda_j)Y$, where Y is a standard complex Gaussian random variable, and subsequently form the periodogram ordinate $I_j = Z_j Z_j^T$.

A.10 Details of the Rhesus Macaque ECoG Dataset

The ECoG data used in this numerical analysis are obtained from the "natural sleep" experiment of Yanagawa et al. (2013), in which ECoG recordings were obtained from a rhesus macaque subject during two states of consciousness: natural (i.e. achieved without the administration of sedatives) sleep while blindfolded in a dark room, and an awake state after the light was turned on but while the blindfold remained in place. ECoG recording occurred continuously throughout this transition; the total experiment lasted roughly 100 minutes.

The full data is available via the NeuroTycho project at http://www.neurotycho.org/

data/20120705slpanesthesiaandsleepchibitoruyanagawa. We focus on the transition between sleep and awake states over all 22 electrodes covering the temporal lobe; see Fig. A.5. In order to mitigate the impact of transition effects between the two states, we enforce a 60 second buffer on either side of the labeled transition time. For both states of consciousness, we obtain a time series of dimension $p = 22$ and length $T = 10000$.

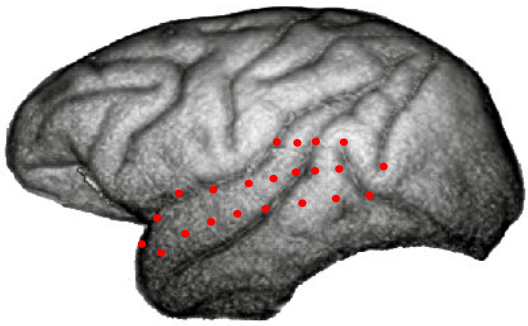

Figure A.5: Location of the ECoG electrodes on the macaque temporal lobe.

A.11 Further Modeling Results for the Multivariate ECoG Spectrum

We examine the marginal fit of the spectral LRD models for the sleep and AEC data, respectively, in Figures A.6 and A.7.

As in the numerical example of Section 2.5, we further investigate goodness-of-fit by boxplots of the log-residuals (Figs. A.8 and A.9), Q-Q plots for the normalized marginal periodogram at high frequencies (Figs. A.10 and A.11), and comparison to the local GPH model at low frequencies (Figs. A.12 and A.13).

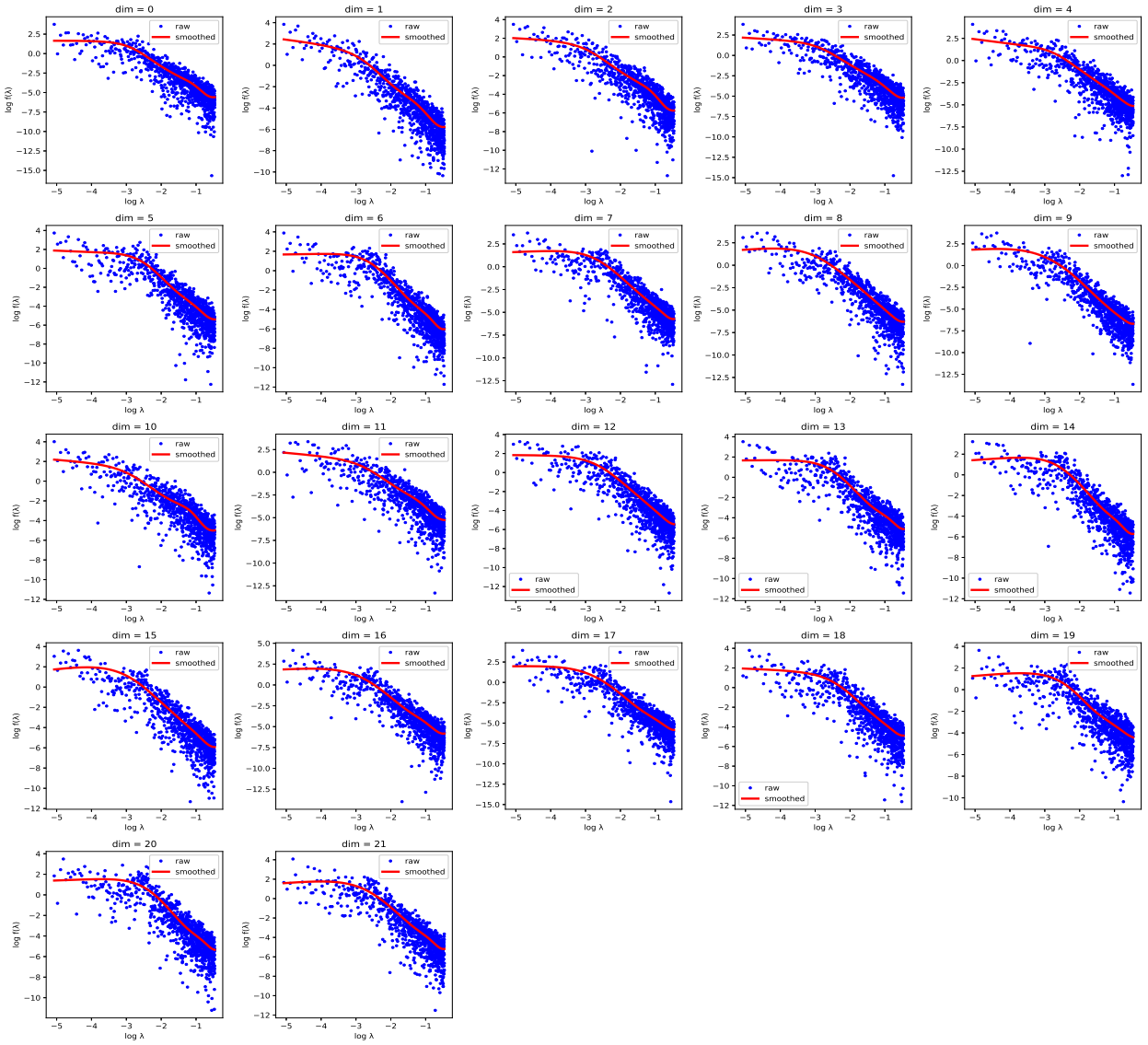

Figure A.6: Marginal spectral fits for temporal lobe electrodes during sleep.

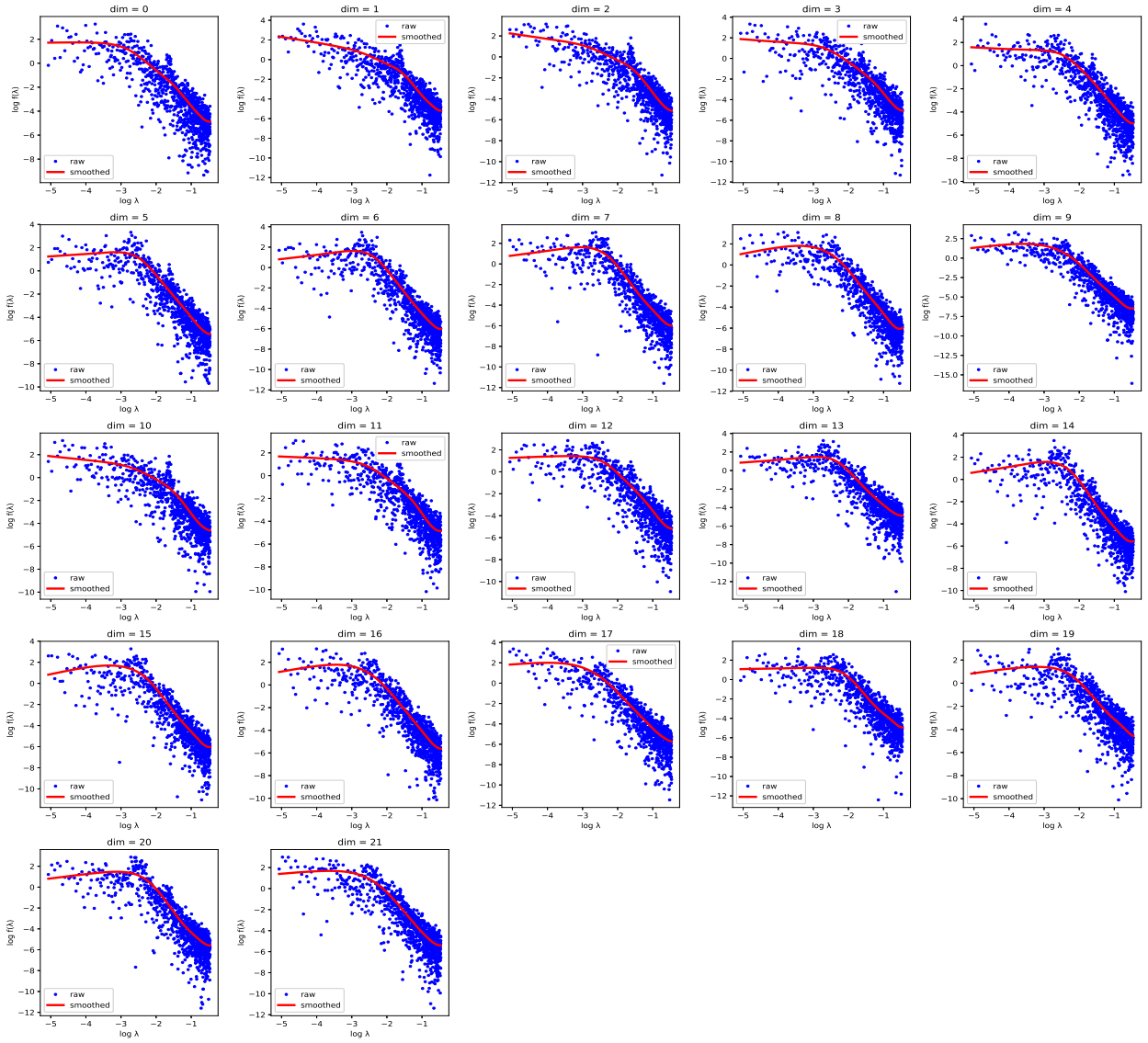

Figure A.7: Marginal spectral fits for temporal lobe electrodes during awake-eyes closed state.

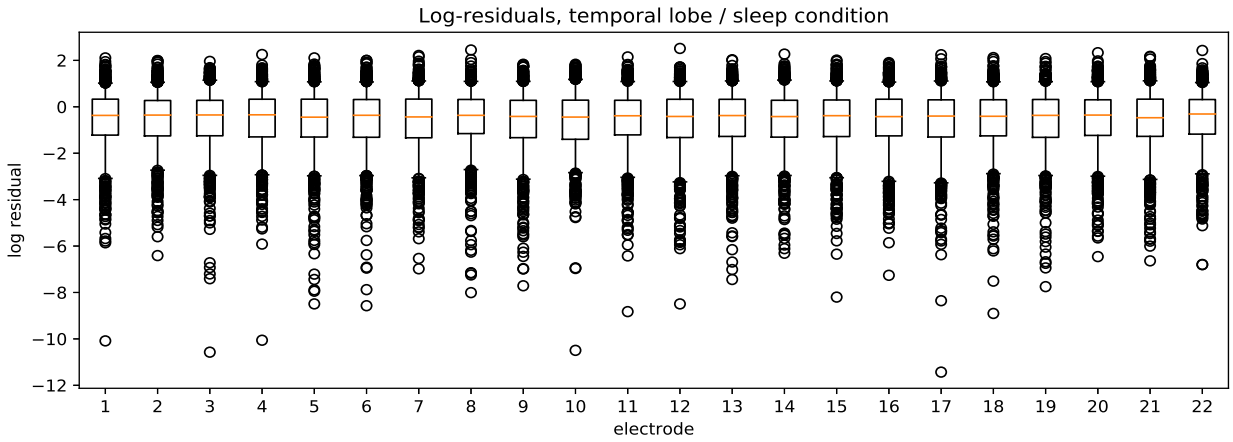

Figure A.8: Boxplot of marginal log-residuals for the model fit to macaque ECoG recordings in sleep state.

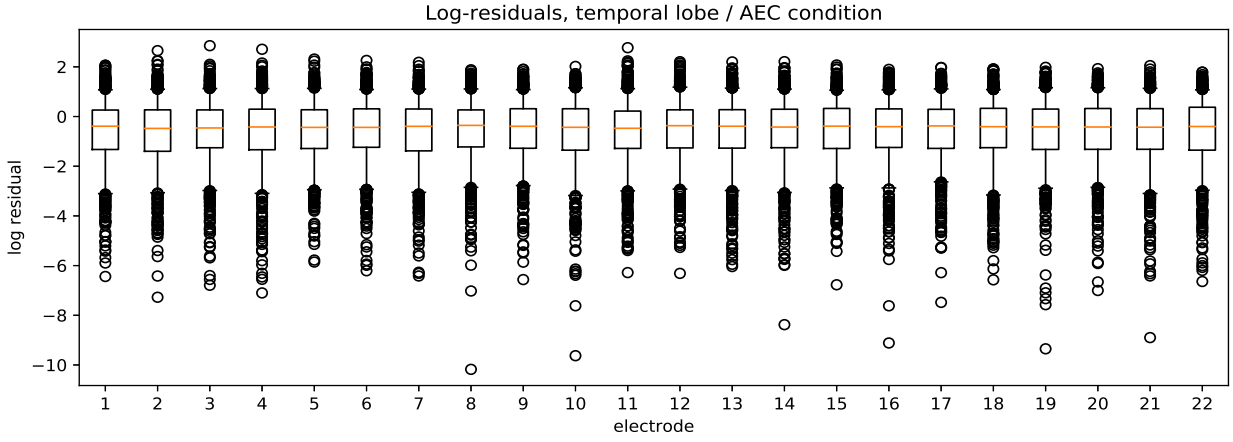

Figure A.9: Boxplot of marginal log-residuals for the model fit to macaque ECoG recordings in awake-eyes closed state.

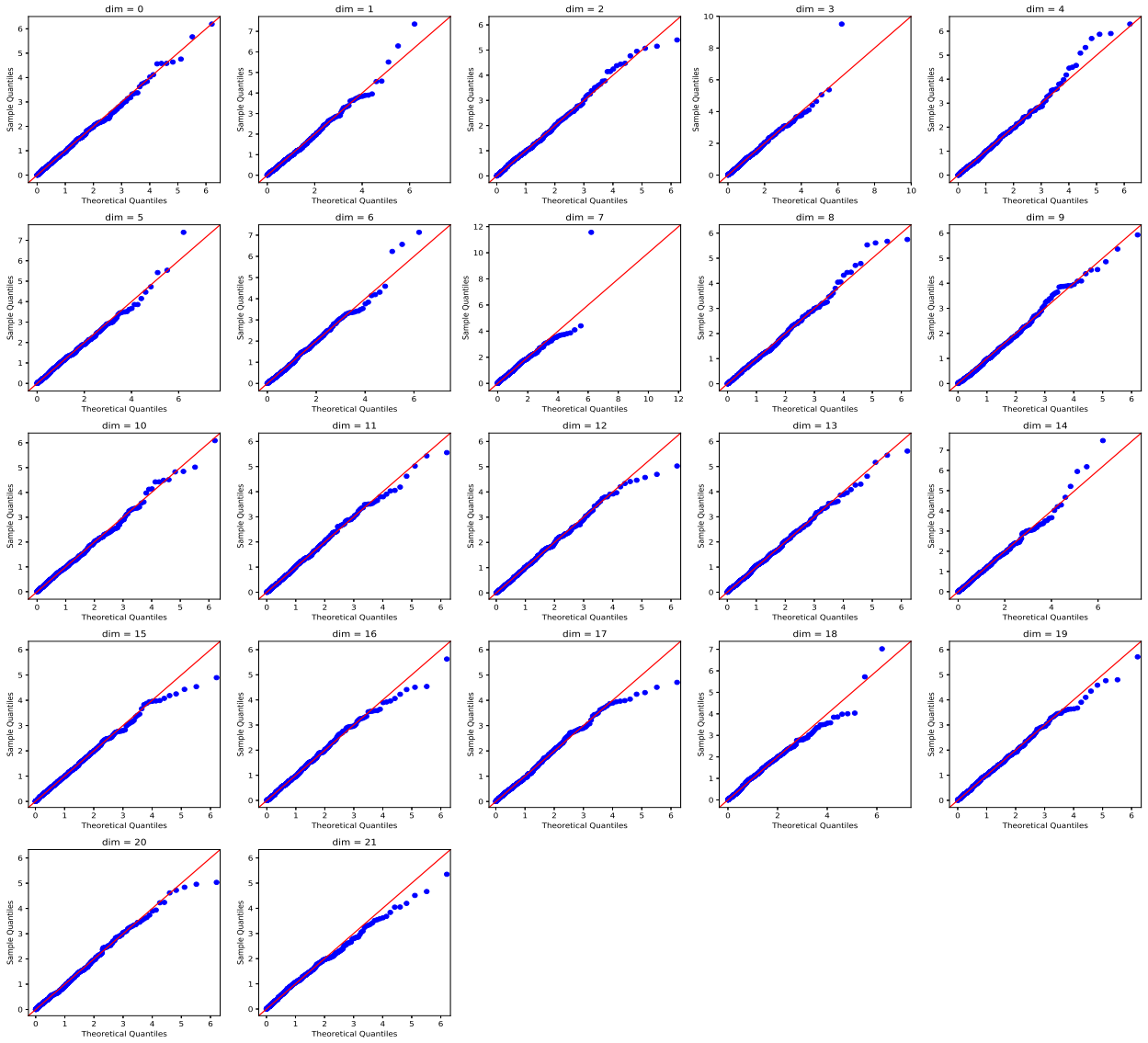

Figure A.10: Quantile-quantile plots of the component-wise normalized periodogram for macaque sleep data against the quantiles of the standard exponential distribution.

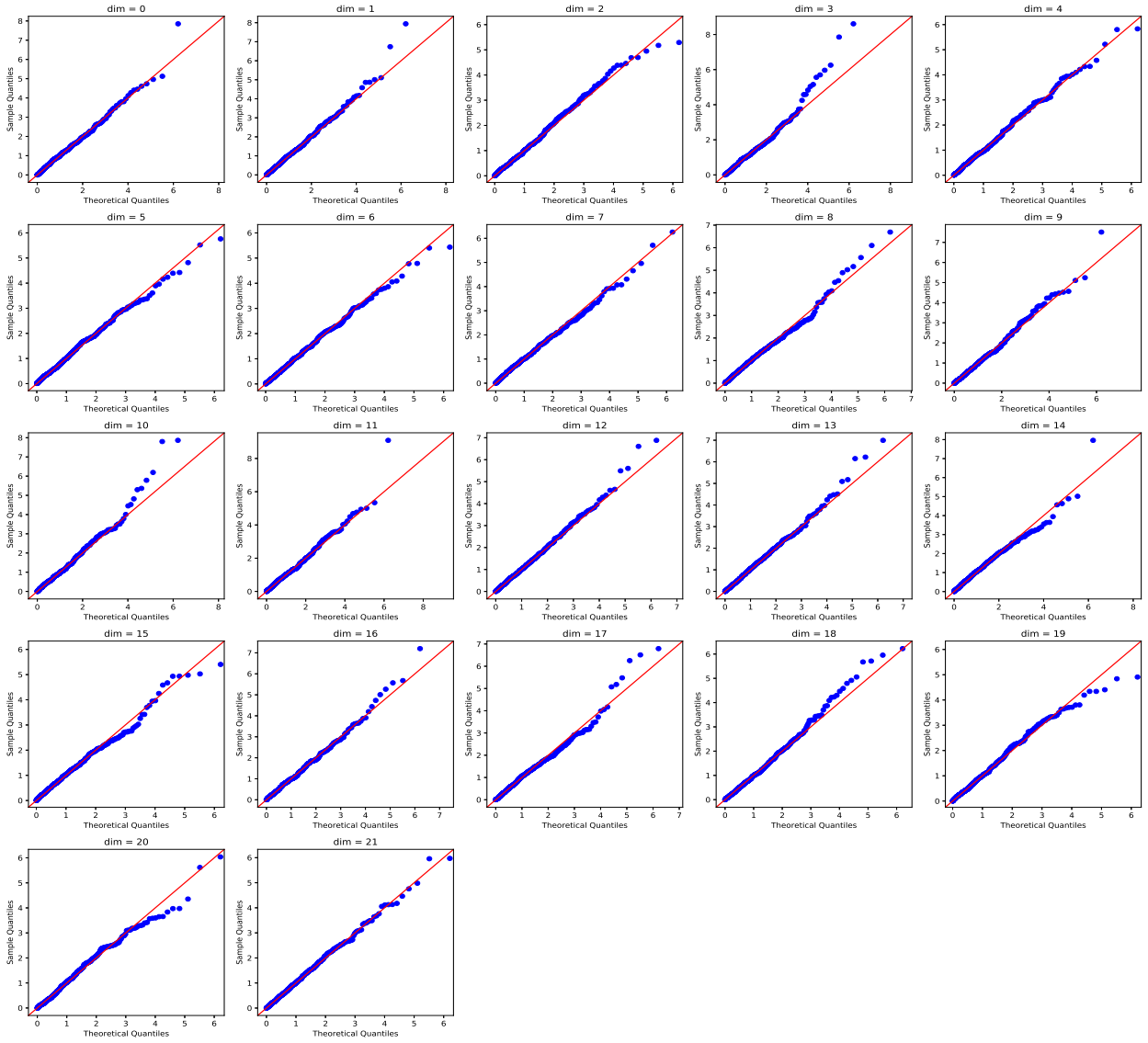

Figure A.11: Quantile-quantile plots of the component-wise normalized periodogram for macaque awake-eyes closed data against the quantiles of the standard exponential distribution.

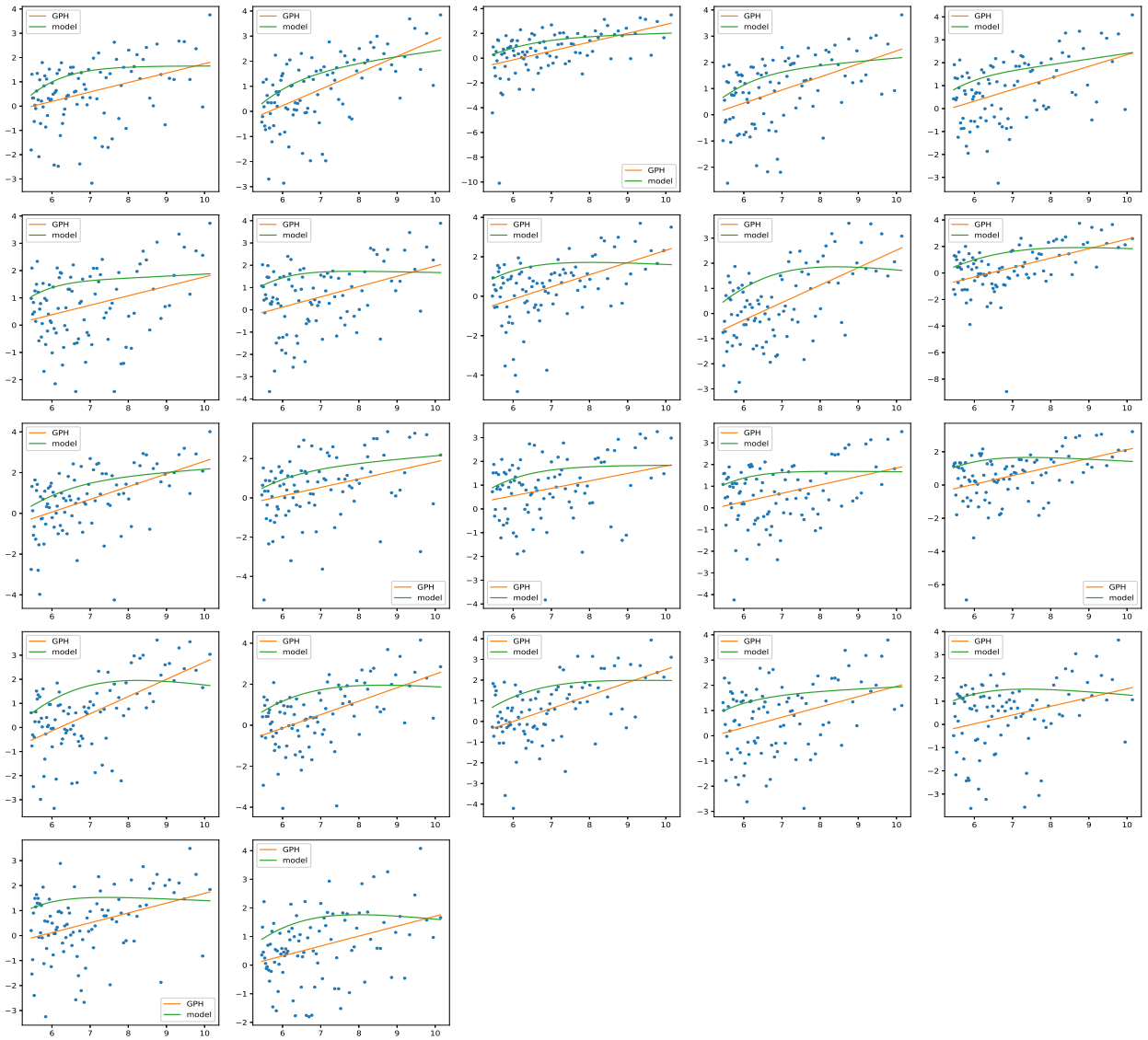

Figure A.12: Comparison of model fit to macaque sleep data at low frequencies to the local semiparametric model implied by the consistent GPH estimator.

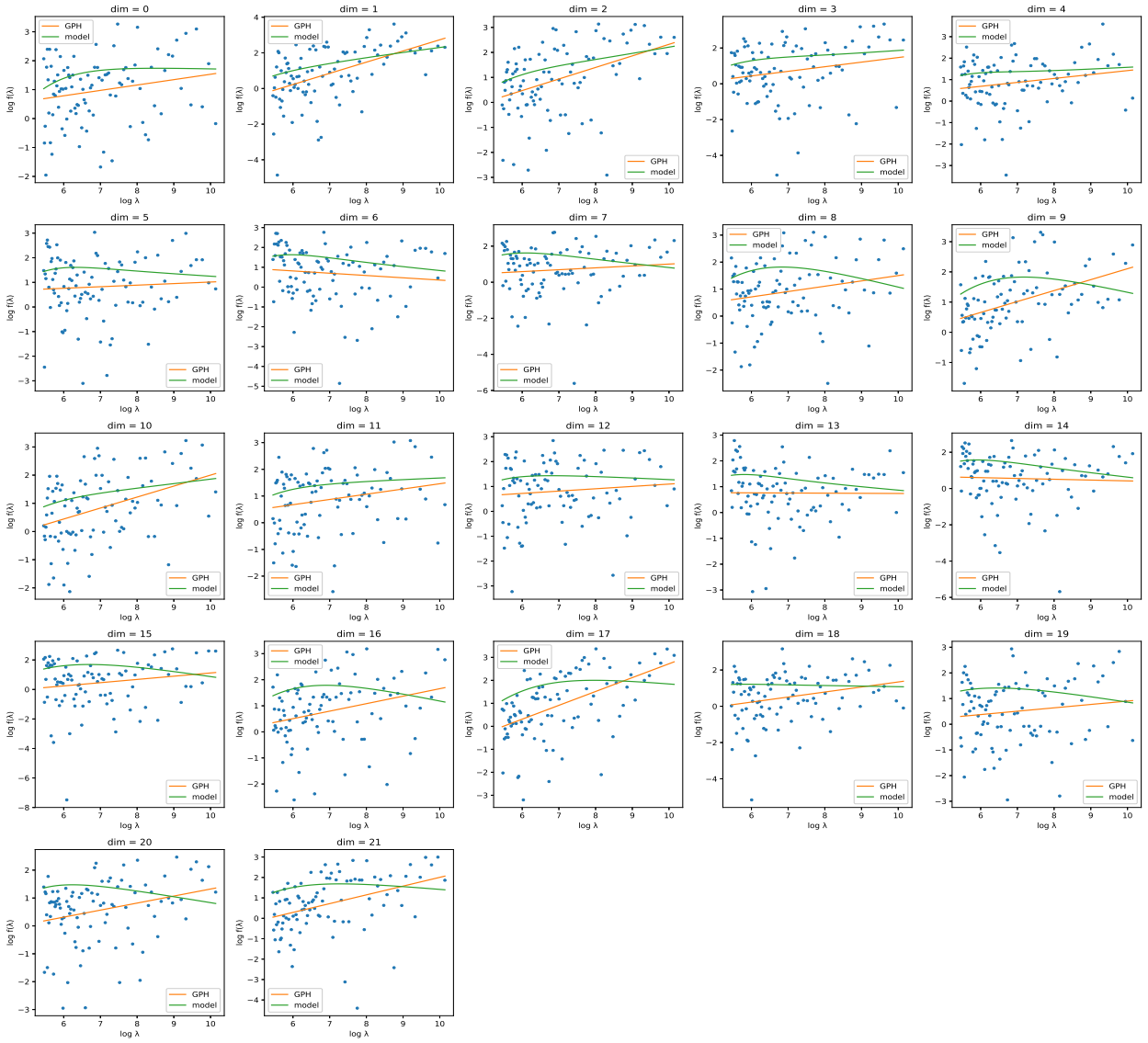

Figure A.13: Comparison of model fit to macaque awake-eyes closed data at low frequencies to the local semiparametric model implied by the consistent GPH estimator.

Appendix B

APPENDIX TO CHAPTER 3

B.1 Summary of the ECoG Data by Categorical Features

We report summaries of the number of observations in the ECoG data per level of the categorical features. The categorical features `Subject` and `Delay` vary at the session level. We report these values, along with total observation counts and information on the length of the stimulation and rest periods per block, in Table B.2. The variation in observations between sessions is due entirely to variation in the number of electrode failures in the ECoG array per session; see Chapter 4 for a detailed discussion of session-wise heterogeneity in the data collection process. For each session, there are an equal number of observations for all five levels of the categorical feature `Block number`.

The `Region` feature describes the location of the two electrodes corresponding to each observation of SIFCC; counts are summarized in Table B.1.

Finally, we note that the observation counts are equal across all frequency bands, as separation of the data by frequency band is downstream of the categorical features describing the experimental protocol.

Region	**Observations**
Both M1	134910
Both S1	159425
M1 / S1	187170

Table B.1: Summary of ECoG observation counts by electrode regions.

Date	**Session #**	**Subject**	**Delay** (ms)	**Stim.** (s)	**Rest** (s)	**# Blocks**	**# Obs.**
9/15/2015	2	G	30	310	610	5	18705
9/15/2015	3	G	10	310	610	5	20025
9/15/2015	4	G	30	310	610	5	21390
9/15/2015	5	G	10	310	610	5	17850
9/16/2015	4	G	10	310	610	5	19580
9/17/2015	1	G	10	310	610	5	13140
9/17/2015	2	G	10	310	610	5	13140
9/17/2015	3	G	100	310	610	5	20930
9/21/2015	3	G	10	310	610	5	21855
9/21/2015	5	G	10	310	610	5	21855
9/22/2015	1	G	10	310	610	5	21855
9/22/2015	2	G	10	310	610	5	18275
9/22/2015	3	G	100	310	610	5	18275
9/25/2015	1	G	10	310	610	5	10400
9/25/2015	2	G	10	310	610	5	8850
4/26/2016	1	J	30	310	610	5	13875
4/26/2016	2	J	10	310	610	5	14630
4/26/2016	3	J	10	310	610	5	13505
4/28/2016	2	J	100	310	610	5	12425
4/28/2016	3	J	10	310	610	5	12075
4/29/2016	1	J	10	310	610	5	11055
4/29/2016	3	J	100	310	610	5	10725
5/2/2016	1	J	10	310	610	5	10400
6/24/2016	3	J	10	310	610	5	20025
6/24/2016	4	J	100	310	610	5	20025
6/25/2016	4	J	10	310	610	5	20025
6/25/2016	5	J	10	310	610	5	17015
6/27/2016	1	J	10	310	610	5	10080
6/27/2016	2	J	100	310	610	5	9765
6/30/2016	1	J	100	310	610	5	4515
6/30/2016	3	J	10	310	610	5	4305
7/2/2016	2	J	10	310	610	5	4305
7/2/2016	4	J	100	310	610	5	6630

Table B.2: Breakdown of session-specific experimental details and total observations contributing to the full ECoG dataset.

B.2 Prediction Results for the Whole-session Design

We report prediction results for the nonlinear model on the whole-session design. These results represent a third, longer timescale for SIFCC prediction, as the regression target is the net connectivity change over an *entire* experimental session, roughly one hour in length. The whole-session analysis is secondary to the scientific objectives of Chapter 2 and was thus investigated in a preliminary way only, without resampling or similarity comparisons across frequency bands.

Nonetheless, we report results in Figure B.1 that warrant a more thorough investigation in future work. Two aspects deserve particular emphasis. First, prediction of connectivity changes over a full experimental session, given only information on the stimulation protocol and connectivity prior to *any* stimulation (i.e. resting block 1), is not only possible, it is in fact more accurate both for the linear and nonlinear methods than for the shorter timescale

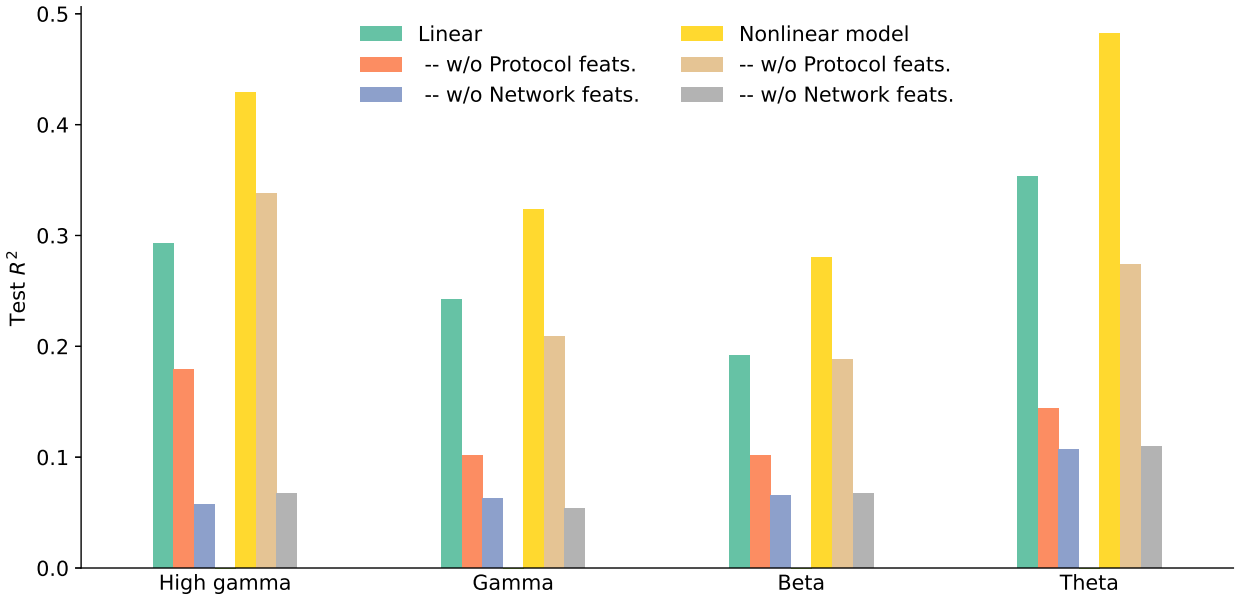

Figure B.1: Prediction accuracy, in terms of R^2 on the held-out test set, of the linear and nonlinear model on the whole-session design. Relative importance of the network and protocol features is measured by test prediction performance for a model estimated without these feature groups.

analyses reported in Chapter 2. Second, the relative significance of the network and protocol features remains the same as in the main chapter: the network features are much more important, in the sense that prediction quality is much worse in their absence than for the protocol features.

Taken together, these results are somewhat surprising, as it would typically be expected that prediction of changes over a shorter timescale, and in particular while stimulation is ongoing, should be easier than over longer timescales. One hypothesis for these results is that repeated stimulation reinforces connectivity changes in a manner still largely dependent on the initial connectivity, but with a larger effect size that is easier to distinguish from background variation.

B.3 Order Selection Results for the Hierarchically Penalized Additive Model

In Figures B.2 - B.5 we plot heatmaps showing the (hierarchically structured) estimated support for each additive model fit to the full data design in Chapter 2. White space indicates coefficient values equal to zero. For all components of the models fit to each frequency band and regression target, we observe that the coefficient at order $L = 10$ is either identically or approximately zero, indicating the suitability of this choice of upper bound.

B.4 Additive Principal Components in the High-gamma Band

We include further details on the calculation of additive principal components (APCs) and the APC analysis of the ECoG data in the high gamma band. The APCs are constructed from the design matrix $X \in \mathbb{R}^{n \times p}$ by the method outlined by (Donnell et al., 1994, §5); here, we reproduce the exact steps and results of the specific application to the ECoG data in greater detail. For each continuous feature, we expand the observation vector $X_j \in \mathbb{R}^n$ polynomially up to maximum order $K = 10$ and remove the column-wise mean. The polynomial features $\Psi_j \in \mathbb{R}^{n \times K}$ are orthogonalized via the QR decomposition $\Psi_j = Q_j R_j$, and we concatenate the matrices $\sqrt{n} Q_j$ column-wise to produce a matrix $\Psi_c \in \mathbb{R}^{n \times K p_c}$ of feature-wise orthogonalized polynomial features, where p_c is the number of continuous features. The discrete features

X_j are first dummy-coded to a binary matrix B_j whose columns are indicators for each level of the discrete feature. We compute the column-wise normalized matrix $\tilde{B}_j$ by dividing each column b_j of B_j by $\sqrt{n^{-1}\sum_{i=1}^{n} b_{ij}}$, and we concatenate the matrices $\tilde{B}_j$ column-wise across all discrete features. The final matrix Ψ is obtained by concatenating the discrete and continuous feature matrices; it has dimension 384095×143. The APCs are computed as the principal components of this design matrix, that is, as the eigenvectors of the positive-definite matrix $\frac{1}{n}\Psi^T\Psi$.

For each additive principle component, the relative contribution of the individual features is defined in terms of their empirical variance. Recall that for each APC Φ with nonlinear components $\{\phi_1, \ldots, \phi_p\}$, we must have $\sum_{j=1}^{p} \mathrm{Var}(\phi_j) = 1$. Let $v \in \mathbb{R}^{Kp_c+p_d}$ be the eigenvector corresponding to a given APC. For a continuous feature X_j, we compute the nonlinear APC component $\phi_j = \sqrt{n}Q_j v_j \in \mathbb{R}^n$, where $v_j \in \mathbb{R}^K$ is the sub-vector of v corresponding to the orthogonalized features Q_j. For a discrete feature X_j, we compute $\phi_j = v_j/\sqrt{n^{-1}\sum_{i=1}^{n} b_{ij}} \in \mathbb{R}^{l_j}$, where l_j is the number of levels taken by X_j. The relative contributions for each feature are then estimated by the empirical variance of ϕ_j.

The full set of components of the minimum APC is plotted in Figure B.6, expanding on the results partially visualized in Figure 3.7. As reported in Chapter 2, the contributions from all features other than the `L=2 Path Strength` and `mean coherence to network`, including the discrete features, are either identically or approximately zero. By contrast, the components of the maximum APC are shown in Figure B.7, and the maximum APC component weights are plotted in Figure B.8. These show that, in contrast to the minimum APC and thus the potential sources of concurvity, the maximum-variance APC contains significant contributions from most features.

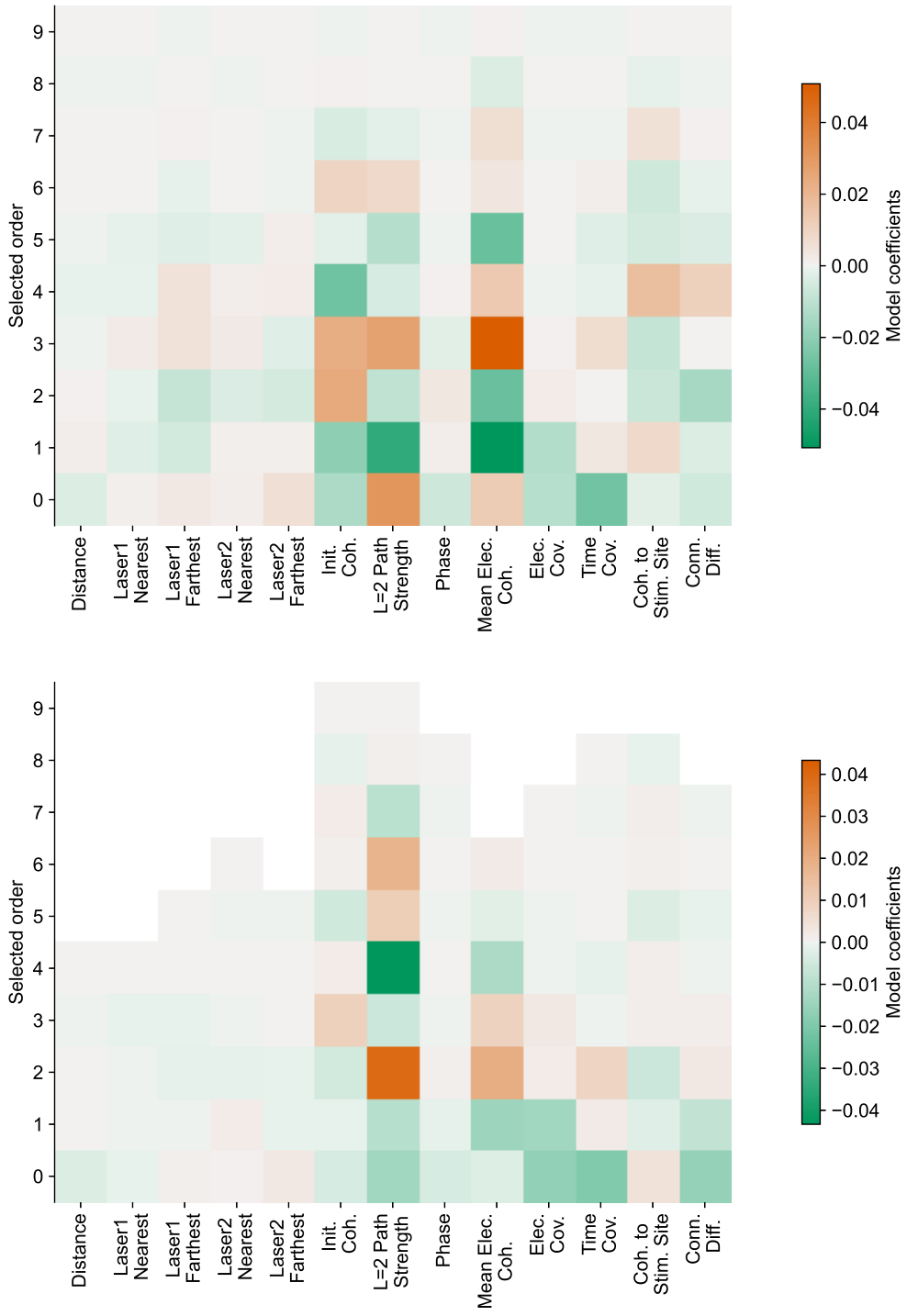

Figure B.2: Selected order for each continuous component function of the additive model fit to the full data design and SS-FCC target in high gamma band (top) and gamma band (bottom).

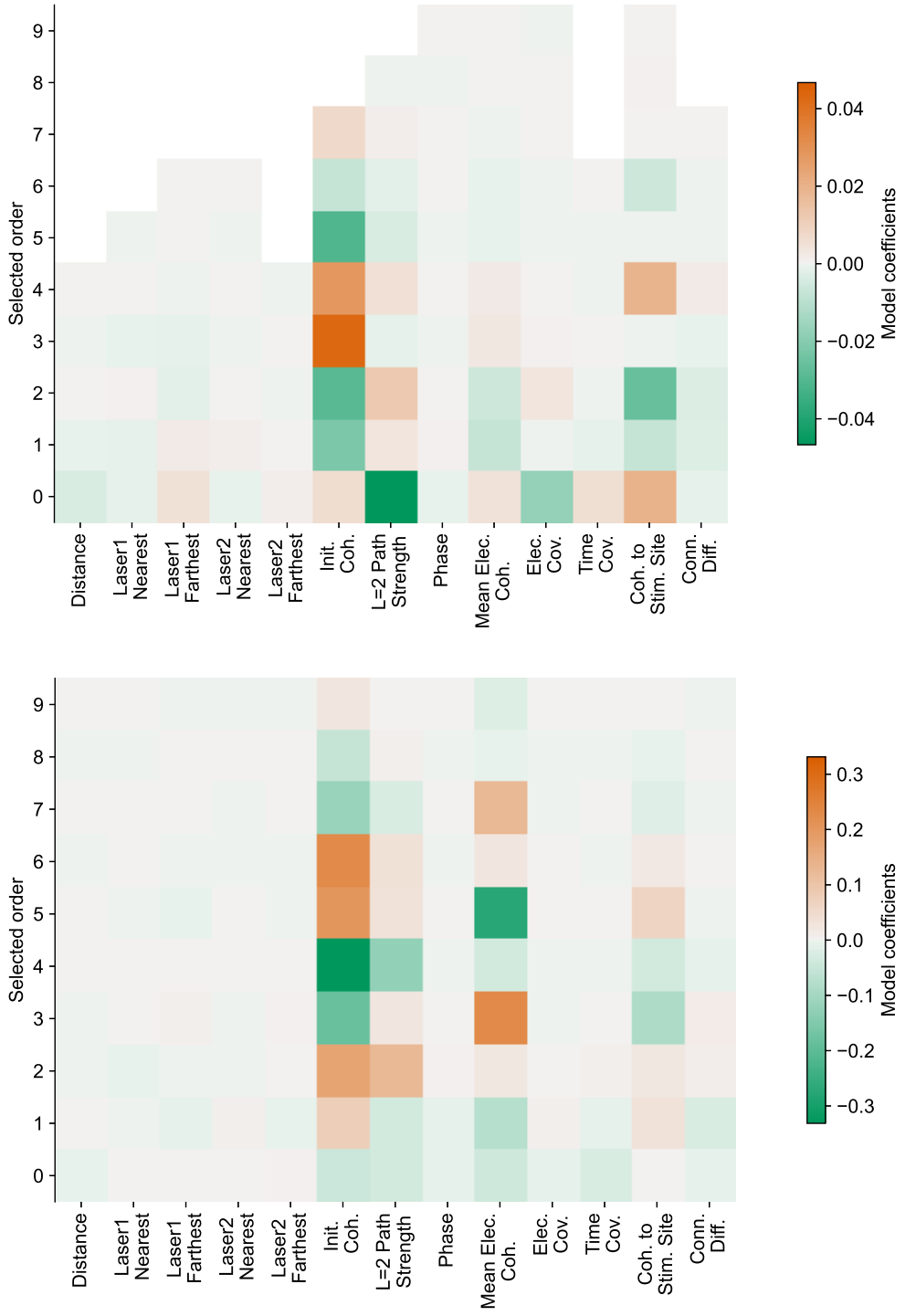

Figure B.3: Selected order for each continuous component function of the additive model fit to the full data design and SS-FCC target in beta band (top) and theta band (bottom).

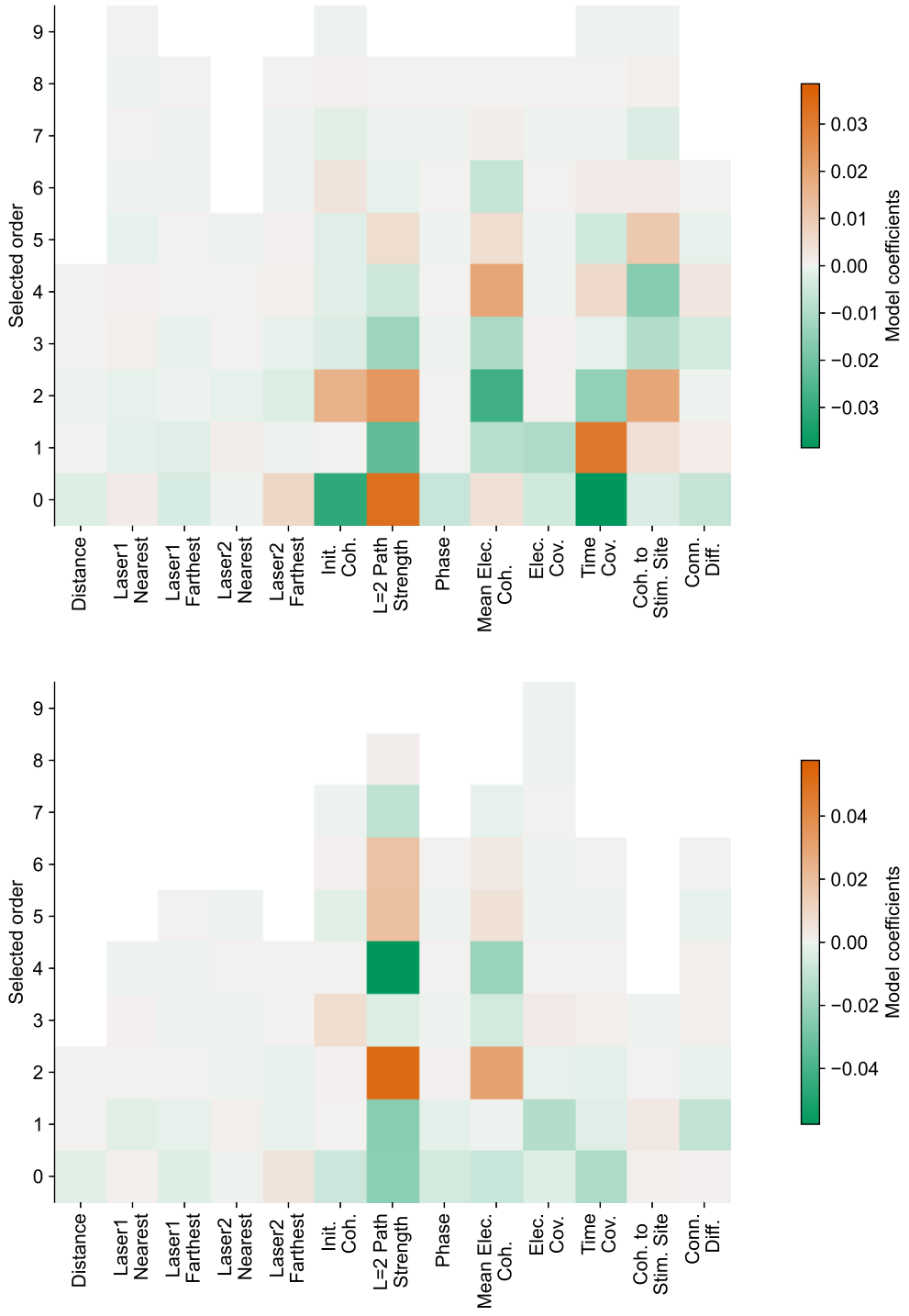

Figure B.4: Selected order for each continuous component function of the additive model fit to the full data design and RS-FCC target in high gamma band (top) and gamma band (bottom).

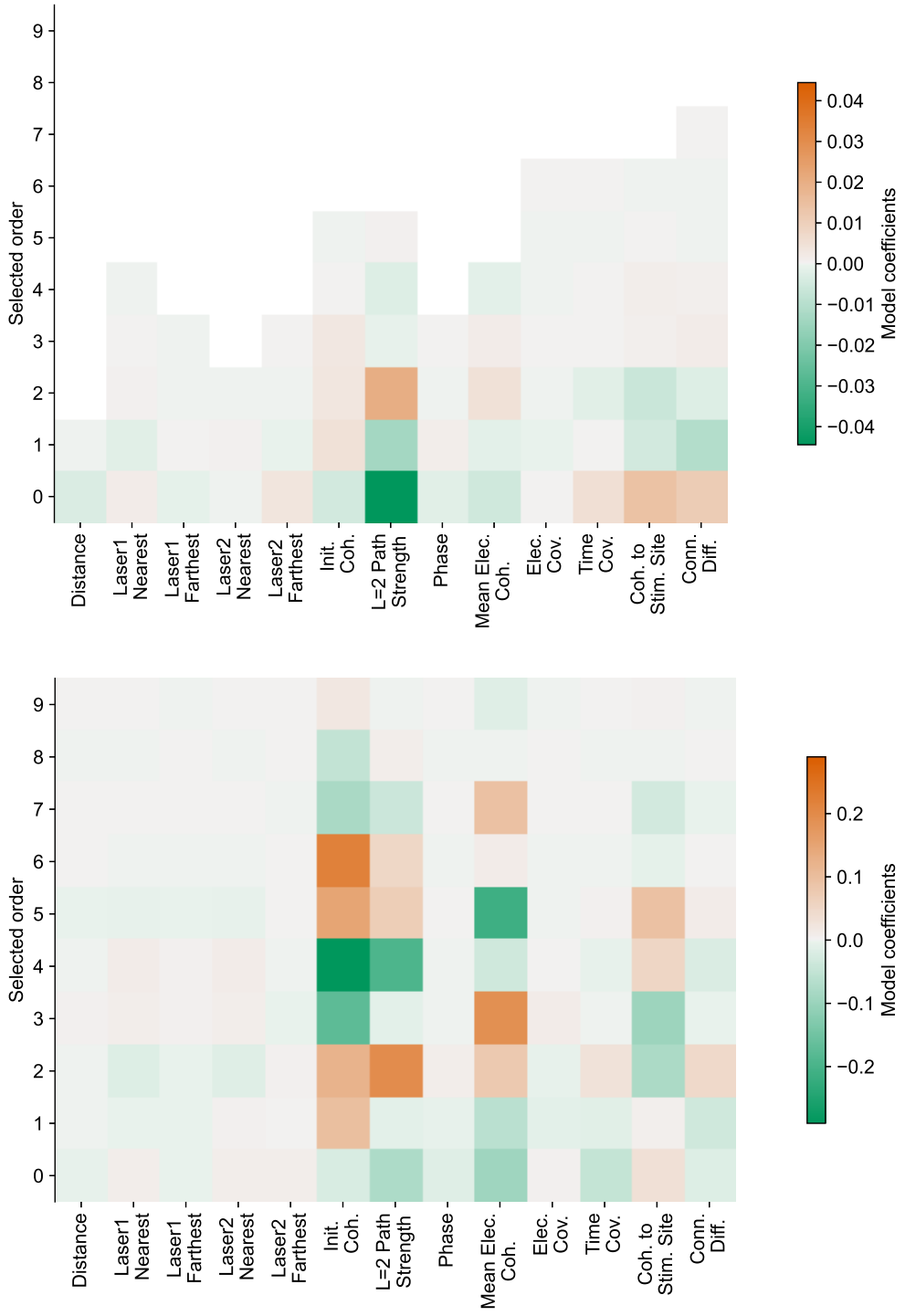

Figure B.5: Selected order for each continuous component function of the additive model fit to the full data design and RS-FCC target in beta band (top) and theta band (bottom).

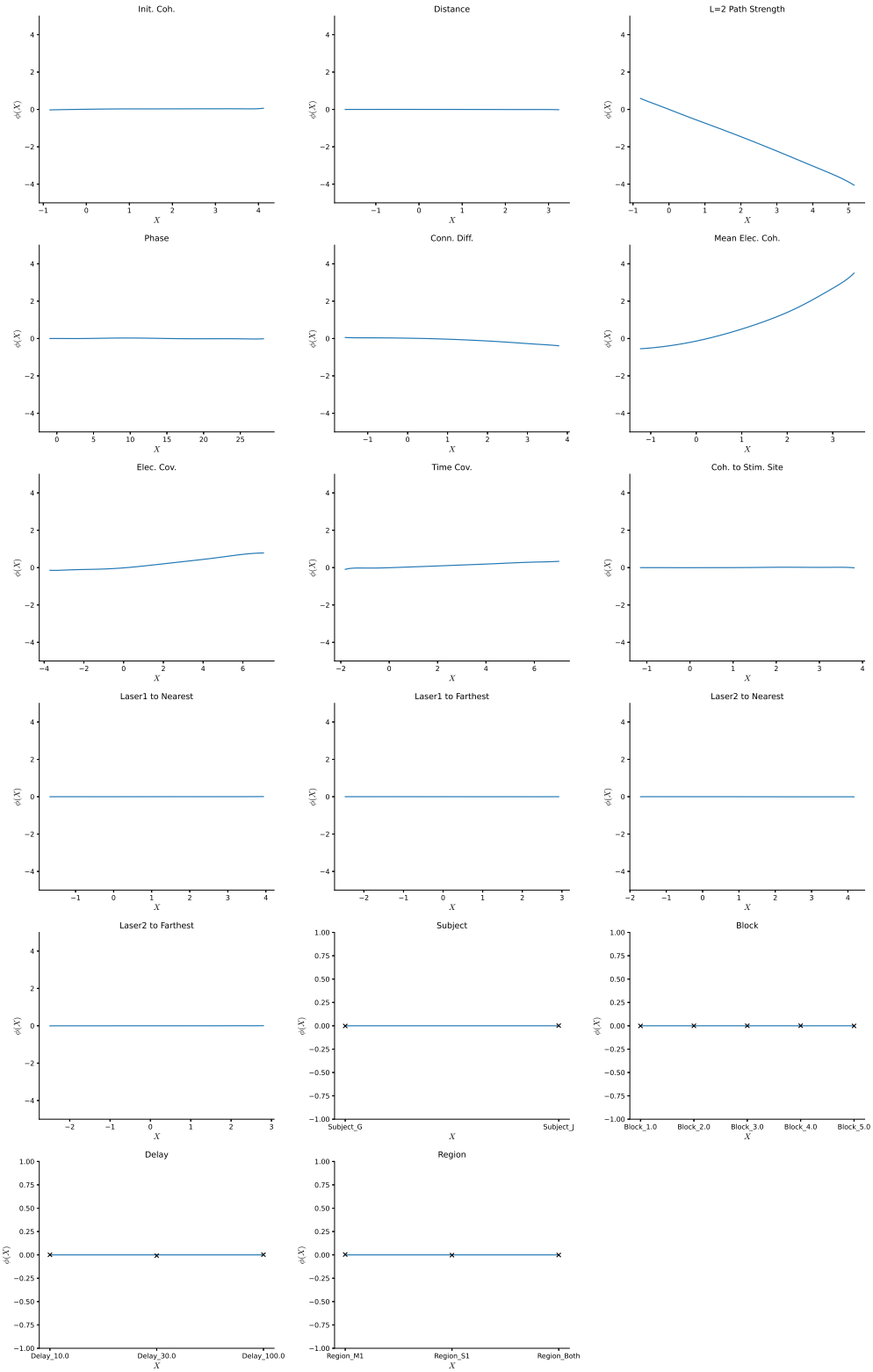

Figure B.6: Full set of APC component functions for the minimum APC of the ECoG data in high gamma band.

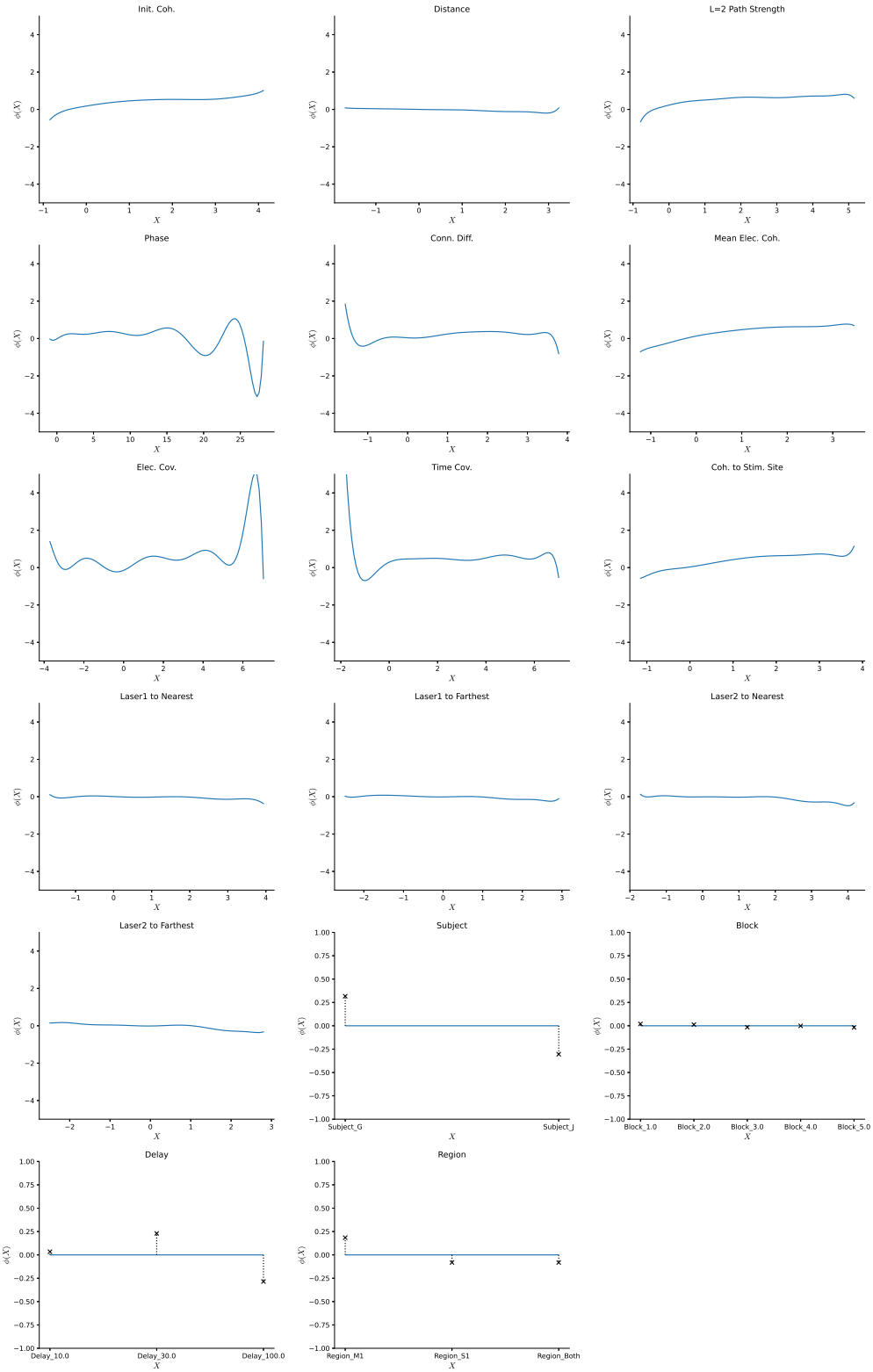

Figure B.7: Full set of APC component functions for the maximum APC of the ECoG data in high gamma band.

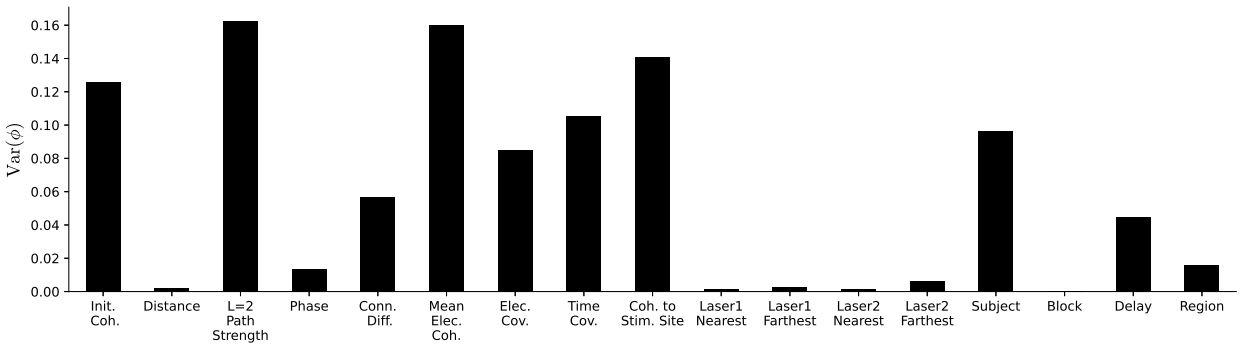

Figure B.8: Relative weights of the feature-wise contributions to the maximum APC.

Appendix C

APPENDIX TO CHAPTER 4

C.1 Subject-level Prediction Results for the Nonlinear Additive Model

We provide evidence for heterogeneity at the subject level by comparison of the prediction results for the nonlinear additive model of Chapter 3 when the model is estimated separately for each of the two subjects. We split the original ECoG data by subject, maintaining for each subject the same separation between train and test data as in the original experiment. We fit the nonlinear additive model for both during and after stimulation regression targets, over all frequency bands. Error bars are obtained from 100 resampling trials.

C.2 Session-wise Results for Maxmin Estimation

We confirm the negative result for maximin estimation of the linear model across sessions in the training data. The maximin coefficient of a given feature is zero if, when the linear model is fit separately to each group of data in a known partition, the range of coefficient values estimated for that feature includes zero. We plot this range for each feature across all of the 24 sessions in the training data of the chronological session hold-out experiment. Two features, `Subject` and `Delay`, do not vary within a session and thus are omitted from the analysis. Results are plotted in Figure C.2. We see that while at least some features *tend* to have linear coefficients of the same sign across sessions, there is no feature for which this holds in *all* sessions. Maximin estimation thus conservatively selects the zero vector for the high gamma band ECoG data.

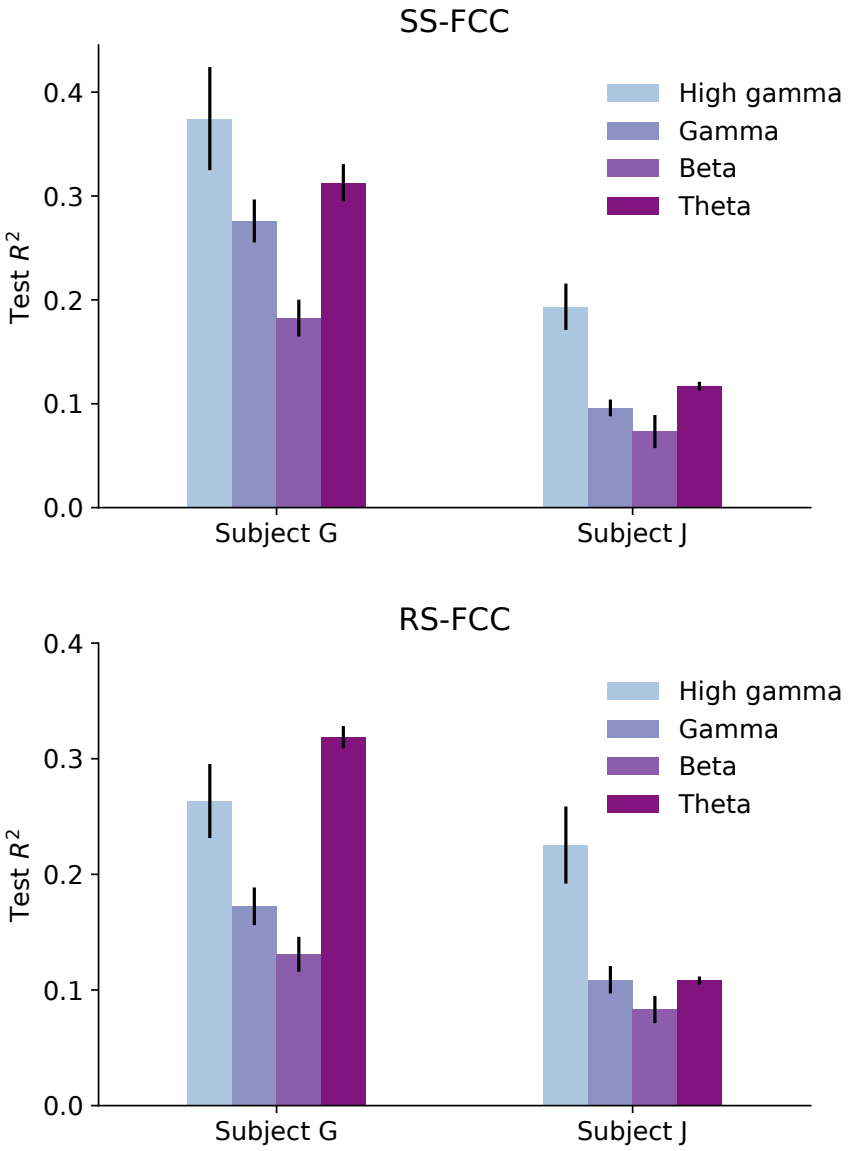

Figure C.1: Prediction accuracy on the test set for the nonlinear additive model estimated and evaluated separately for each subject. Results are shown for both for SIFCC during stimulation (top) and SIFCC after stimulation (bottom). Error bars are obtained from 100 trials of the subsampling procedure described in Chapter 3.

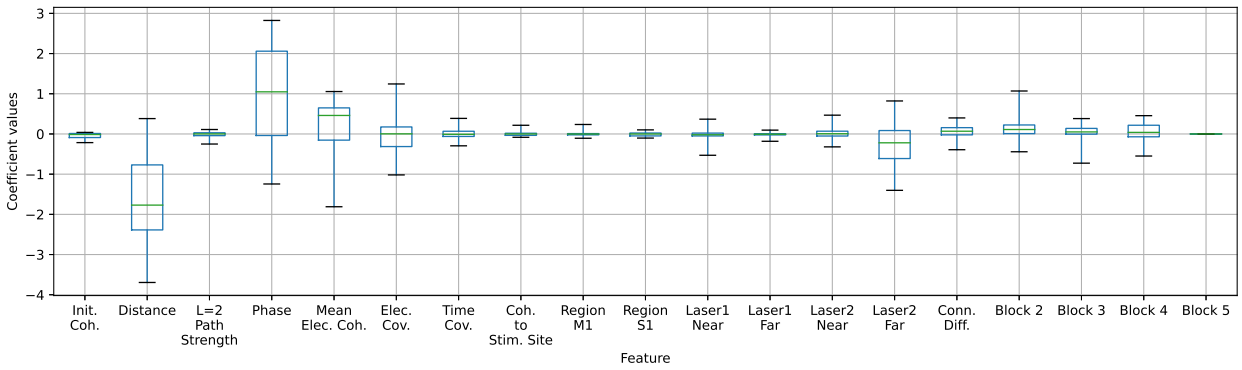

Figure C.2: Distribution of estimated coefficients of a linear model fit separately to each of the 24 sessions in the ECoG training data with high gamma frequency band and during-stimulation SIFCC regression target.