ABSTRACT

Title of dissertation: BAYESIAN MODELING AND ESTIMATION TECHNIQUES FOR THE ANALYSIS OF NEUROIMAGING DATA

Proloy Das, Doctor of Philosophy, 2020

Dissertation directed by: Professor Behtash Babadi
Department of Electrical & Computer Engineering

Brain function is hallmarked by its adaptivity and robustness, arising from underlying neural activity that admits well-structured representations in the temporal, spatial, or spectral domains. While neuroimaging techniques such as Electroencephalography (EEG) and magnetoencephalography (MEG) can record rapid neural dynamics at high temporal resolutions, they face several signal processing challenges that hinder their full utilization in capturing these characteristics of neural activity. The objective of this dissertation is to devise statistical modeling and estimation methodologies that account for the dynamic and structured representations of neural activity and to demonstrate their utility in application to experimentally-recorded data.

The first part of this dissertation concerns spectral analysis of neural data. In order to capture the non-stationarities involved in neural oscillations, we integrate multitaper spectral analysis and state-space modeling in a Bayesian estimation setting. We also present a multitaper spectral analysis method tailored for spike trains that captures the non-linearities involved in neuronal spiking. We apply our proposed algorithms to both EEG and spike recordings, which reveal significant gains in spectral resolution and noise reduction.

In the second part, we investigate cortical encoding of speech as manifested in MEG responses. These responses are often modeled via a linear filter, referred to as the temporal response function (TRF). While the TRFs estimated from the sensor-level MEG data have been widely studied, their cortical origins are not fully understood. We define the new notion of Neuro-Current Response Functions (NCRFs) for simultaneously determining the TRFs and their cortical distribution. We develop an efficient algorithm for NCRF estimation and apply it to MEG data, which provides new insights into the cortical dynamics underlying speech processing.

Finally, in the third part, we consider the inference of Granger causal (GC) influences in high-dimensional time series models with sparse coupling. We consider a canonical sparse bivariate autoregressive model and define a new statistic for inferring GC influences, which we refer to as the LASSO-based Granger Causal (LGC) statistic. We establish non-asymptotic guarantees for robust identification of GC influences via the LGC statistic. Applications to simulated and real data demonstrate the utility of the LGC statistic in robust GC identification.

BAYESIAN MODELING AND ESTIMATION TECHNIQUES FOR THE ANALYSIS OF NEUROIMAGING DATA

by

Proloy Das

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Behtash Babadi, Chair
Professor Jonathan Z. Simon
Professor Shihab Shamma
Professor Steve Marcus
Professor Tom Goldstein
Professor Daniel A. Butts, Dean's Representative

Acknowledgments

I would like to take this opportunity to express my profound gratitude to my advisor, Professor Behtash Babadi, for his valuable guidance and friendly support throughout my PhD period. This dissertation would not have been possible without his constant encouragement and untiring supervision. Thanks for enlightening my mindset toward research, through which I learned that the sky is the limit to the knowledge one can gain, and the contribution one can make to science. Thanks for the countless hours of brainstorming and scientific discussions we had through these years, through which I have broadened my horizons in science, and developed a researcher mindset. It has been, and will be an honor to have the opportunity to work under his supervision, as one of his earlier PhD students. As a supervisor, he has always inspired me to acquire a never-give-up mindset, and encouraged me to push the boundaries

I would like to thank Prof. Jonathan Z. Simon, Prof. Shihab Shamma, Prof. Steve Marcus, Prof. Tom Goldstein and Prof. Daniel A. Butts, who kindly agreed to serve on my dissertation committee, and generously shared their invaluable time and expertise to enrich this dissertation.

I would also like to thank my colleagues, Alireza Sheikhattar, Sayyed Sina Miran, Abbas Kazemipour, Shoutik Mukherjee, Anuththara Rupasinghe, Behrad Soleimani and Sahar Khosravi for their support, friendship and the countless hours of brainstorming, exchanging ideas and insightful discussions. Moreover, I would like to express my gratitude to all my collaborators, Christian Brodbeck, Alessandro Presacco, Joshua P. Kulasingham, Dushyanthi Karunathilake, for sharing their broad knowledge, extensive expertise and rich data recordings with me. I owe Christian Brodbeck a special thanks for his tremendous help to start-off my scientific Python journey.

My housemates at my places of residence have been a crucial factor in my finishing smoothly. I'd like to express my gratitude to Abhishek Chakraborti, Ayan Mallik, Nitish Mehta, Ravin Singh, Srijan Kumar for their friendship and support. I would also like to thank all my wonderful friends who made this journey a memorable and pleasant one, Amit, Ankit, Jooik, Josh, Mike, Udit, Prithviraj, Rajeev, Rajdeep, Rajrupa, Samnhita, Sohini, Udit, Upamanyu, Yuntao, and all other dear friends.

Lastly and most importantly, I owe my deepest thanks to my family - my parents, Angana and Provat Das, my brother, Alloy and my girlfriend, Lisa Barman for their lifetime love and endless support. They have always believed in me, and encouraged me to follow my dreams always and have pulled me through against impossible odds at times. They have lost a lot due to my studies abroad. Words cannot express the gratitude I owe them. I would also like to express special thanks to Sayantan Sarkar and Ananya Mukherjee who are like my elder siblings from another mother, because of them my graduate experience has been one that I will cherish forever.

August 2020

Proloy Das

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AC	Auditory Cortex
AIC	Akaike Information Criterion
AR	Auto-Regressive
ARMA	Auto-Regressive Moving-Average
BIC	Bayesian Information Criterion
BOLD	Blood Oxygen Level Dependant
BVAR	Bi-variate auto-regressive
CI	Confidence Interval
dpss	Discrete prolate spheroidal sequences
DBMT	Dynamic Bayesian Multitaper
EAF	Expected Ambiguity Function
ECoG	Electrocorticography
EEG	Electroencephalography
EM	Expectation Maximization
EMD	Empirical Mode Decomposition
ENF	Electrical Network Frequency
ERP	Event-Related Potentials
FASTA	Fast Adaptive Shrinkage/Thresholding Algorithm
FBS	Forward-Backward Splitting
FIR	Finite Impulse Response
(f)MR(I)	(functional) Magnetic Resonance (Imaging)
GC	Granger Causal(ity)
GES	Generalized Evolutionary Spectra
IFG	Inferior Frontal Gyrus
iid	Independent and identically distributed
LASSO	Least Absolute Shrinkage and Selection Operator
LFP	Local Field Potential
LGC	Lasso-based Granger Causal(ity)
LTI	Linear Time Invariant
MAP	Maximum A Posteriori
MEG	Magnetoencephalography
ML	Maximum Likelihood
MNE	Minimum Norm Estimate
MT	Multitaper
MVAR	Multi-variate auto-regressive
NCRF	Neuro-Current Response Function
OLS	Ordinary Least Squares
PET	Positron Emission Tomography
PMC	Primary Motor Cortex

PSD	Power Spectral Density
PSWF	Prolate Spheroidal Wave Functions
RE	Restricted Eigenvalue
SNR	Signal-to-Noise Ratio
SPARLS	Sparse Least Squares
STS	Superior Temporal Sulcus
TFAR(MA)	Time-Frequency Auto-Regressive (Moving-Average)
TFCE	Threshold-Free Cluster-Enhancement
TRF	Temporal Response Function

Chapter 1: Introduction

In recent years, a wide range of experimental methodologies for recording the structure and function of the human brain have produced abundant neural datasets in various spatiotemporal resolutions and modalities. Modalities such as electrocorticography (ECoG), local field potential (LFP) recording, and single/multi-unit depth electrode recordings acquire electrophysiological activity directly from the cortex in an invasive fashion. On the contrary, noninvasive approaches such electroencephalography (EEG) and magnetoencephalography (MEG), record electromagnetic fields generated by cortical activity, at the scalp or around the head. Other examples of noninvasive brain-imaging modalities include positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), which measure the changes in cerebral blood flow as correlates of neural activity. While the invasive and noninvasive approaches have specific benefits as well as limitations, noninvasive modalities have become increasingly popular in research settings, and particularly in studying the human brain, due to the convenience of their adoption in non-clinical environments.

This convenience, however, comes at the cost of losing either the temporal or spatial resolution due to indirect measurement of neural activity. For example, while fMRI can distinguish between brain sources separated by $\sim 1.5\,\mathrm{mm}$ [1], it has a temporal resolution of the order of seconds. M/EEG recordings, on the other hand, can capture neural activity with millisecond resolution, but suffer from spatial mixing of the activity over the sensor space. A key advantage of such high temporal resolution is the ability to track the rapid temporal dynamics of coordinated neuronal population activity under sensory or cognitive tasks [2, 3, 4]. However, the indirect measurement of brain activity via M/EEG makes the inference of its spatial characteristics heavily dependent on modeling and computational tools. While invasive electrophysilogy recordings do not suffer from loss of spatial resolution, the come with other sets of challenges. For example, spike train recordings of individual neurons' action potentials result in *binary* time series, which require the usage of statistical inference techniques capable of capturing the characteristics of such binary data.

Existing inference algorithms face several key challenges in deciphering the brain dynamics that underlie neural data. First, under sensory and cognitive tasks, the underlying cortical activity exhibits representations that are well structured in the temporal, spatial, or spectral domains, or combinations thereof [5, 6, 7]. An important challenge in neural data analysis is therefore how to exploit the aforementioned structured representations towards more robust and interpretable inference procedures. Second, existing approaches often adopt linear models (specially for encoding of external stimuli), even though various nonlinearities are involved in cortical processing [8, 9]. Undermining these nonlinearities results in large biases in the inferred characteristics of neural activity. Third, neuronal activity often undergoes rapid changes, often referred to as neuronal plasticity [10], in order to adapt to changing stimulus salience and behavioral context. Most existing techniques for neuroimaging data analysis either assume stationarity of the data or resort to sliding window processing, which result in loss of temporal resolution. Fourth, the majority of existing data analysis techniques designed for continuous-time data such as M/EEG and LFP are not optimal for the analysis of binary time series obtained by single-/multi-unit recordings of neuronal activity. In this thesis, we aim at addressing these challenges for three classes of data analysis techniques, namely, spectral analysis, encoding models, and causal inference, primarily in context of M/EEG and spiking data analysis.

Part I: Spectral Analysis of Neural Data

The first part of this dissertation concerns the spectral analysis of neural data from EEG and neuronal spike recordings. Although spectral analysis of continuous time-series is well established, hallmarked by the multitaper method that offers optimal bias-variance trade-offs [11, 12, 13], some of the key characteristics of neural data cannot be readily captured by existing techniques. In particular, we aim at capturing the non-stationarity of EEG data and the non-linearities involved in spiking activity in the context of multitaper spectral analysis.

In Chapter 2, we develop a Bayesian framework for estimating time-varying spectra of non-stationary data at high spectrotemporal resolutions. Classically, spectral analysis techniques, such as the multitaper method, are used in conjunction with overlapping sliding windows in analyzing such non-stationary time series data. Although sliding window analysis is convenient to implement, the resulting estimates are sensitive to the window length and overlap size. In addition, it undermines the dynamics of the time series as the estimate associated to each window uses only the data within. Finally, the overlap between consecutive windows hinders a precise statistical assessment of the estimated spectra.

We address these shortcomings by explicitly modeling the spectral dynamics through augmenting the multitaper method with a state-space model within a Bayesian estimation framework [14, 15]. The underlying states pertaining to the eigen-spectral quantities arising in multitaper analysis are estimated using instances of the Expectation-Maximization algorithm, and are used to construct spectrograms and their respective confidence intervals. We propose spectral estimators that are robust to noise and are able to capture spectral dynamics at high spectrotemporal resolution. We provide a theoretical analysis of the bias-variance trade-off, which establishes performance gains over the standard overlapping multitaper method. We compare the spectrogram estimates from our algorithms to other state-of-the-art estimation techniques using synthetic data, the results of which also validate our theoretical analysis. We apply our algorithms to EEG data recorded during sleep, as well as electric network frequency recordings, to demonstrate their utility and versatility in real world data analysis.

Next, we consider the problem of inferring the spectra of latent neural covariates that underlie spiking activity. Access to such spectra is critical to understanding the role of brain rhythms in mediating cognitive functions. While the spectral estimation of continuous time-series is a well-established domain, computing the spectral representation of these latent neural covariates from spiking data sets forth various challenges due to the intrinsic non-linearities involved. In Chapter 3, we address this problem by proposing a variant of the multitaper method specifically tailored for neural spiking data [16]. To this end, we construct auxiliary spiking statistics akin to tapered data, from which the eigen-spectra of the underlying latent process can be directly inferred using maximum likelihood estimation, and thereby the multitaper estimate can be efficiently computed. Comparison of our proposed technique to existing methods using simulated spike trains and multi-unit recordings under general anesthesia reveals significant gains in terms of the bias-variance trade-off.

Part II: A Cortically-distributed Encoding Model of Speech Processing for M/EEG Analysis

The second part of this dissertation concerns the characterization of the neural dynamics that underlie speech processing at the cortical level. Neuroimaging techniques such as M/EEG have provided significant insights into the meso-scale neural processing of continuous stimuli, such as speech, thanks to their high temporal resolution [3, 17, 18, 19]. Existing work in the context of auditory processing suggests that certain features of speech, such as the acoustic envelope, can be used as reliable linear predictors of the neural response manifested in M/EEG. The corresponding linear filters are referred to as temporal response functions (TRFs) [19, 20, 21].

The resulting encoding model considers the whole brain along with the neuroimaging measurement mechanism as a linear time invariant (LTI) system, collectively characterized by the TRFs, where the the system takes one or several representations of the auditory stimuli (e.g., the broadband acoustic envelope) as input and produces the M/EEG recordings as output. While the functional roles of specific components of the TRF are well-studied and linked to behavioral attributes such as attention, the cortical origins of the underlying neural processing are not as well understood due to the spatial mixing at the sensor level. Existing methods for demixing the TRFs at the cortical level work in a two-stage fashion: either the TRFs are first estimated at the sensor level, and then mapped to the cortex via source localization (See, for example, [4, 18]), or the neuroimaging data are first mapped to the cortex followed by estimating TRFs for each of the resulting cortical sources [22]. Given that each stage is biased towards specific goals, such as enforcing sparsity and smoothness, the end result typically suffers from destructive propagation of biases across stages, which in turn hinders a valid statistical interpretation of the results and requires significant post-hoc processing to summarize the results in a meaningful fashion.

In Chapter 4, we address this challenge by *directly* estimating linear filter representations of cortical sources from neuroimaging data in the context of speech processing [23, 24]. To this end, we introduce the Neuro-Current Response Functions (NCRFs), a set of linear filters, spatially distributed throughout the cortex, that predict the cortical currents giving rise to the observed ongoing MEG (or EEG) data in response to continuous speech. We cast NCRF estimation within a Bayesian framework, which allows unification of the TRF and source estimation problems, and also facilitates the incorporation of prior information on the structural properties of the NCRFs. To generalize this analysis to M/EEG recordings which lack individual structural magnetic resonance (MR) scans, NCRFs are extended to free-orientation dipoles and a novel regularizing scheme is introduced to mitigate dependence on fine-tuned coordinate co-registration. We present a fast estimation algorithm, which we refer to as the Champ-Lasso algorithm, by leveraging recent advances in optimization, and demonstrate its utility through application to simulated and experimentally recorded MEG data under auditory experiments. Our simulation studies reveal significant improvements over existing two-stage methods, in terms of spatial resolution, filter reconstruction, and recovering dipole orientations. The analysis of experimentally-recorded MEG data without MR scans corroborates existing findings, by delineating the distinct cortical distribution of the underlying neural processes at high spatiotemporal resolution and thus obviating the need for post-processing steps such as clustering and denoising. In summary, we provide a principled whole brain encoding model for sensory processing as well as an estimation paradigm for MEG source analysis tailored to extracting the cortical origin of neural responses to continuous stimuli.

Part III: Granger Causal Inference from Sparse Autoregressive Models

Reliable identification of causal influences is one the central challenges in neural data analysis [25, 26, 27]. Granger causal (GC) characterization of time-series is among the widely used data-driven methods in this regard [28, 29, 30]. The notion of GC influence pertains to assessing the improvements in predicting the future samples of one time-series by incorporating the past samples of another one.

Conventionally, the prediction task is cast within the multivariate autoregressive modeling framework, in which the optimal linear predictors are obtained by the ordinary least squares (OLS) method coupled with the AIC [31] or BIC [32] procedures to determine optimal model orders. Then, the GC measure is defined as the logarithmic ratio of the two prediction error variances, and its statistical significance is assessed based on the corresponding asymptotic distributions [33, 34]. Two of the main practical challenges of this methodology are (1) parameter estimation and model selection under limited data durations, which leads to over-fitting and hence errors in identifying the causal influences [25, 27, 35, 36], and (2) correlated process noise as a confounding factor that hinders accurate identification of the causal effects [37].

The theory of sparse estimation, and particularly the LASSO, has successfully addressed these challenges for parameter estimation. The LASSO and its variants have already been utilized in existing work to identify graphical GC influences based on the estimated model parameters, either directly [38] or by appropriate thresholding [39, 40]. Another strand of results uses de-biasing techniques in order to construct confidence intervals and thereby identify the significant causal interactions (See, for example, [41, 42]). There is, however, an evident disconnect between these LASSO-based approaches and the classical OLS-based GC inference: while the LASSO-based approaches aim at identifying the GC effects based on the estimates of the model parameters, the classical GC methodology relies on the comparison of the prediction errors between two models (i.e., an unconstrained model and a constrained model) by resorting to asymptotic distributions.

In Chapter 5, we close the gap by providing a LASSO-based Granger causal analysis for a canonical bivariate autoregressive model with correlated process noise. Under the hypothesis that the true model admits a sparse autoregressive representation, we study the non-asymptotic properties of a likelihood-based scaled F-statistic under the null (i.e., absence of a Granger causal effect) and alternative (i.e., presence of a Granger causal effect) conditions and establish that the well-known sufficient conditions of LASSO also suffice for robust identification of Granger causal influences. By slightly weakening these sufficient conditions, we also characterize the false positive error probability of a simple thresholding rule for identifying Granger causal effects. We present simulation studies to compare the performance of the conventional ordinary least squares method to that of the LASSO in detecting Granger causal influences to demonstrate the validity of our theoretical claims and to explore the key underlying trade-offs. We also present an application to experimentally-recorded neural data from general anesthesia to assess the causal influence of the LFP on spiking activity. In summary, our main contribution is to extend the non-asymptotic results of the LASSO to the classical asymptotic characterization of GC influences, and to identify the key trade-offs in terms of sampling requirements and strength of the causal effects that result in robust GC identification.

Finally we close this thesis by discussing some of the future directions of research along the same line, in Chapter 6. It is worth nothing that in addition to their utility in analyzing neuronal data, our techniques have potential application in other domains beyond neuroscience, thanks to the plug-and-play nature of the algorithms used in our inference frameworks: for example, the spectral analysis techniques may have applications in domains such as economics, forensics, oceanography, climatology, seismology; the LASSO-based Granger causal analysis can be used to extract causal influences in social networks or gene regulatory networks. In order to facilitate usage by the broader systems neuroscience community, MATLAB/Python implementations of the algorithms developed in this dissertation are archived as open source repositories on GitHub: https://github.com/proloyd.

Part I

Spectral Analysis of Neural Data

Chapter 2: Dynamic Bayesian Multitaper Spectral Analysis

Spectral analysis techniques are among the most important tools for extracting information from time series data recorded from naturally occurring processes. Examples include speech [43], images [44], electroencephalography (EEG) [45], oceanography [46], climatic time series [47] and seismic data [48]. Due to the exploratory nature of most of these applications, non-parametric techniques based on Fourier methods and wavelets are among the most widely used. In particular, the multitaper (MT) method excels among the available non-parametric techniques due to both its simplicity and control over the bias-variance trade-off via bandwidth adjustment [11, 12, 13].

Most existing spectral analysis techniques assume that the time series is stationary. In many applications of interest, however, the energy of the various oscillatory components in the data exhibits dynamic behavior. Extensions of stationary time series analysis to these non-stationary processes have led to 'time-varying' spectral descriptions such as the Wigner-Ville distribution [49, 50], the evolutionary spectra and its generalizations [51, 52], and the time-frequency operator symbol formulation [53] (See [54] for a detailed review). A popular approach to estimating such time-varying spectra is to subdivide the data into overlapping windows or segments and estimate the spectrum locally for each window using various Fourier or wavelet-based methods[55], assuming the underlying process is quasi-stationary, i.e., the spectrum changes slowly with time. Thereby, the so-called spectrogram analysis is obtained by using sliding windows with overlap in order to capture non-stationarity.

Although sliding window processing is widely used due to its fast implementation, it has several major drawbacks. First, the window length and extent of overlap are subjective choices and can drastically change the overall attribute of the spectrogram if chosen poorly. Second, given that the estimate associated to a given window is obtained by only the data within, it ignores the common dynamic trends shared across multiple windows, and thereby fails to fully capture the degree of smoothness inherent in the signal. Instead, the smoothness of the estimates is enforced by the amount of overlap between adjacent windows. Third, although techniques such as the MT analysis are able to mitigate the variabilities arising from finite data duration or the so-called 'sampling' noise by averaging over multiple tapers, their spectral resolution degrades when applied to data within small windows due to the increase in the Rayleigh resolution [56]. In addition, they do not have a mechanism in place to suppress the additive measurement noise that commonly contaminates empirical observations. Fourth, the overlap between adjacent windows hinders a precise statistical assessment of the estimates, such as constructing confidence intervals due to the high dependence of estimates across windows. To address this issue, statistical corrections for multiple comparisons need to be employed [57], which in turn limit the resulting test powers when multiple windows are involved.

In recent years, several alternative approaches to non-stationary spectral analysis have been proposed, such as the empirical mode decomposition (EMD) [58], synchrosqueezed wavelet transform [59, 60], time-frequency reassignment [61], time-frequency ARMA models [62], and spectrotemporal pursuit [63]. These techniques aim at decomposing the data into a small number of smooth oscillatory components in order to produce spectral representations that are smooth in time but sparse or structured in frequency. Although they produce spectral estimates that are highly localized in the time-frequency plane, they require certain assumptions on the data to hold. For example, EMD analysis assumes the signal to be deterministic and does not take into account the effect of observation noise [58]. Other methods assume that the underlying spectrotemporal components pertain to certain structures such as amplitude-modulated narrowband mixtures [59, 60], sparsity [63] or chirp-like dynamics [61]. In addition, they lack a statistical characterization of the estimates. Finally, although these sophisticated methods provide spectrotemporal resolution improvements, they do not yield implementations as simple as those of the sliding window-based spectral estimators.

In this chapter, we address the above-mentioned shortcomings of sliding window multitaper estimators by resorting to state-space modeling. State-space models provide a flexible and natural framework for analyzing systems that evolve with time [64, 65, 66, 67], and have been previously used for parametric [62, 68] and non-parametric [63] spectral estimation. The novelty of our approach is in the integration of techniques from MT analysis and state-space modeling in a Bayesian estimation framework. To this end, we construct state-space models in which the underlying states pertain to the eigen-spectral quantities, such as the empirical eigen-coefficients and eigen-spectra arising in MT analysis. We employ state dynamics that capture the evolution of these quantities, coupled with observation models that reflect the effect of measurement and sampling noise. We then utilize Expectation-Maximization (EM) to find the maximum *a posteriori* (MAP) estimate of the states given the observed data to construct our spectral estimators as well as statistical confidence intervals.

We provide theoretical analysis of the bias-variance trade-off, which reveals two major features of our proposed framework:

1. Our methodology inherits the control mechanism of the bias-variance trade-off from the MT framework by means of changing the design bandwidth parameters [13], and

2. Our algorithms enjoy the optimal data combining and denoising features of Bayesian filtering and smoothing.

In addition, due to the simplicity and wide usage of Bayesian filtering and smoothing algorithms, our algorithms are nearly as simple to implement as the sliding window-based spectrograms. To further demonstrate the performance of our algorithms, we apply them to synthetic as well as real data including human EEG recordings during sleep and electric network frequency data from audio recordings. Application of our proposed estimators to these data provides spectrotemporal features that are significantly denoised, are smooth in time, and enjoy high spectral resolution, thereby corroborating our theoretical results.

The rest of the chapter is organized as follows: In Section 2.1, we present the preliminaries and problem formulation. In Section 2.2, we develop our proposed estimators. Application of our estimators to synthetic and real data are given in Section 2.3, followed by our theoretical analysis in Section 2.4. Finally, our concluding remarks are presented in Section 2.5.

2.1 Preliminaries and Problem Formulation

2.1.1 Non-stationary Processes and Time-Varying Spectrum

Consider a finite realization of T samples from a discrete-time non-stationary process $y_t, t = 1, 2, \cdots T$, obtained via sampling a continuous-time signal above the Nyquist rate. We assume that the non-stationary process y_t is harmonizable so that it admits a Cramér representation [69, p. 150] of the form:

$$y_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi ft} dz(f), \tag{2.1}$$

where $dz(f)$ is the generalized Fourier transform of the process. This process has a covariance function of the form:

$$\Gamma_L(t_1, t_2) := \mathbb{E}[y_{t_1} y_{t_2}^*] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi(t_1 f_1 - t_2 f_2)} \gamma_L(f_1, f_2) df_1 df_2, \tag{2.2}$$

where $\gamma_L(f_1, f_2)$:=$\mathbb{E}[dz(f_1)dz^*(f_2)]$ is referred to as the generalized spectral density or the Loève spectrum [70]. Due to the difficulty in extracting physically-plausible spectrotemporal information from the two-dimensional function $\gamma_L(f_1, f_2)$, other forms of spectrotemporal characterization that are two-dimensional functions over time and frequency have gained popularity [55]. To this end, by defining the coordinate rotations $t := (t_1 + t_2)/2$, $\tau := t_1 - t_2$, $f := (f_1 + f_2)/2$, and $g := f_1 - f_2$ and by substituting in the definition of the covariance function in Eq. (2.2), we obtain:

$$\Gamma(\tau, t) := \Gamma_L(t_1, t_2) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi(tg+\tau f)} \gamma(g, f) df dg,$$

where f and g are referred to as the ordinary and non-stationary frequencies, respectively[70], and $\gamma(g, f)dfdg := \gamma_L(f_1, f_2)df_1df_2$ is the Loève spectrum in the rotated coordinates. To obtain one such two-dimensional spectral density representation over time and frequency, we define:

$$D(t, f) := \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi tg} \gamma(g, f) dg = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-i2\pi\tau f} \mathbb{E}[y_{t+\frac{\tau}{2}} y^*_{t-\frac{\tau}{2}}] d\tau, \tag{2.3}$$

which coincides with the expected value of the Wigner-Ville distribution [49]. The 'time-varying' spectral representation $D(t, f)$ captures the spectral information of the data as a function of time, and thus provides a useful framework for analyzing non-stationary time series. However, estimating $D(t, f)$ from finite samples of the process is challenging, considering that the expectation needs to be replaced by time averages which may smooth out the time-varying features of the signal [50].

In order to address this challenge, certain additional assumptions need to be imposed on the underlying process, which as a matter of fact restrict the extent of temporal or spectral variations the signal can exhibit. In this regard, two such popular assumptions are posed in terms of the *quasi-stationary* and *underspread* properties. In order to define these properties quantitatively, let us first define the Expected Ambiguity Function (EAF) [52, 54] as:

$$\Delta(\tau, g) := \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi\tau f} \gamma(g, f) df = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-i2\pi t g} \mathbb{E}[y_{t+\frac{\tau}{2}} y^*_{t-\frac{\tau}{2}}] dt \tag{2.4}$$

A signal whose EAF has a limited spread along g (i.e., negligible spectral correlation) is called *quasi-stationary.* If the signal is concentrated around the origin with respect to both τ and g (i.e., small spectral and temporal correlation), it is called *underspread.* There exists a large body of work on estimating time-varying spectra under the quasi-stationarity assumption, such as short-time periodograms, pseudo-Wigner estimators [50], and estimates of the evolutionary spectra [51]. More recent methods such as the Generalized Evolutionary Spectra (GES) estimators [52] and time-frequency auto-regressive moving-average (TFARMA) [62] estimators rely on the underspread property of the underlying signals. We refer the reader to [71, Chapter 10] for a detailed discussion of these properties.

It is noteworthy that both assumptions are fairly general, encompass a broad range of naturally-occurring processes, and have resulted in successful applications in real life problems. We will next discuss one of the widely used methods for estimating time-varying spectra under the quasi-stationary assumption, which extends MT spectral analysis beyond second-order stationary processes and has gained popularity in exploratory studies of naturally occurring processes [70].

2.1.2 The Sliding Window MT Spectral Analysis

One of the popular non-parametric techniques for estimating the ‘time-varying’ spectral representation $D(t, f)$ is achieved by subdividing the data into overlapping windows or segments and estimating the spectrum for each window independently using the MT method [70]. This method naturally and intuitively extends the popular non-parametric MT method to the non-stationary scenario under the assumption of quasi-stationarity, which enables one to treat the time series within segments (locally) as approximately second-order stationary [72]. The resulting spectrotemporal representations are smoothed version of the Wigner-Ville distribution [54] and are referred to as spectrogram. In what follows, we briefly describe the MT spectrogram method, since it provides the foundation of our treatment.

The MT method is an extension of single-taper spectral analysis, where the data is element-wise multiplied by a taper prior to forming the spectral representation to mitigate spectral leakage [11, 12]. In the MT method, spectral representation is computed as the average of several such single-taper PSDs, where the tapers are orthogonal to each other and exhibit good leakage properties. This can be achieved by using the *discrete prolate spheroidal sequences* (dpss) or *Slepian sequences* [73], due to their orthogonality and optimal leakage properties.

Another viewpoint of the MT method with this particular choice of data tapers is the decomposition of the spectral representation of the process over a set of orthogonal basis functions. Indeed, these basis functions originate from an approximate solution to the integral equation expressing the projection of $dz(f)$ onto the

Fourier transform of the data:

$$y(f) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\sin W\pi(f-\zeta)}{\sin \pi(f-\zeta)} e^{-i2\pi(f-\zeta)\frac{W-1}{2}} dz(\zeta).$$

where W is window length, i.e., number of samples, and $dz(\zeta)$ is an orthogonal increment process. This integral equation can be approximated using a local expansion of the increment process over an interval $[-B, B]$, for some small design band-width B, in the space spanned by the eigenfunctions of the Dirichlet kernel $\frac{\sin W\pi f}{\sin \pi f}$ [11, 12].

These eigenfunctions are known as the prolate spheroidal wave functions (PSWFs), which are a set of doubly-orthogonal functions over $[-B, B]$ and $[-\frac{1}{2}, \frac{1}{2}]$, with time-domain representations given by the dpss sequences. Let $u_l^{(k)}$ be the lth sample of the kth dpss sequence, for a given bandwidth B and window length W. The kth PSWF is then defined as:

$$U^{(k)}(f) := \sum_{l=0}^{W-1} u_l^{(k)} e^{-i2\pi fl}.$$

Choosing $K \leqslant \lfloor 2WB \rfloor - 1$ dpss having eigenvalues close to 1 as data tapers, the MT spectral estimate can be calculated as follows:

$$\widehat{S}^{(\mathsf{mt})}(f) := \frac{1}{K} \sum_{k=1}^{K} |\widehat{x}^{(k)}(f)|^2, \tag{2.5}$$

where $\widehat{x}^{(k)}(f) := \sum_{l=0}^{W-1} e^{-i2\pi fl} u_l^{(k)} y_l$ for $k = 1, 2, \cdots, K$ are called the 'eigen-coefficients'. The 'eigen-spectra', $\widehat{S}^{(k)}(f) := |\widehat{x}^{(k)}(f)|^2$ can be viewed as the expansion coefficients of the decomposition.

To estimate time-varying spectra under the MT framework, sliding windows with overlap are used to enforce temporal smoothness and increase robustness of the estimates [70, 74], resulting in MT spectrogram estimates. Although this 'overlapping' MT procedure overcomes frequency leakage issues and produces consistent estimates, subjective choices of the window length and the degree of overlap can change the overall appearance of the spectrogram drastically when poor choices of these parameters are used. In addition, these estimates lack precise statistical inference procedures, such as hypothesis testing, due to the statistical dependence induced by the overlaps. The objective of this chapter is to overcome these limitations by directly modeling and estimating the evolution of the process without committing to overlapping sliding windows, while achieving fast and efficient implementations. Before presenting our proposed solutions, we give a brief overview of other existing approaches in the literature in order to put our contributions in context.

2.1.3 Motivation and Connection to Existing Literature

Our goal is to overcome the foregoing challenges faced by the sliding window MT spectrogram analysis in estimating the expected value of the Wigner-Ville distribution (See Eq. (2.3)) under the quasi-stationarity assumption. In addition, our framework can be viewed in the context of the spectrogram approximation to the evolutionary spectra [51]. The evolutionary spectra is obtained by considering an expansion of y_t over the set of complex sinusoids as

$$y_t = \int_{-1/2}^{1/2} x_t(f) e^{j2\pi f t} df, \tag{2.6}$$

with uncorrelated time-varying expansion coefficients, i.e., $\mathbb{E}[x_t(f_1)x_t^*(f_2)] = D(t, f_1)$ $\delta(f_1 - f_2)$. In standard MT spectrogram analysis, the extent of overlap between consecutive segments dictates the amount of temporal smoothness in the estimates. Our approach is to avoid the usage of overlapping windows by modeling and estimating the dependence of the spectra across windows using state-space models, while retaining the favorable leakage properties of the MT analysis. As will be revealed in the subsequent sections, in the same vein as the sliding window multitaper analysis our methods pertain to the class of spectrogram estimates, which are viewed as smoothed versions of the Wigner-Ville spectrum [54].

Due to the underlying quasi-stationarity assumption, i.e., negligible spectral correlation, the domain of applicability of our methods might be narrower than the more general non-stationary spectral analysis methods such as GES and Weyl spectral estimation and TFARMA modeling; however, our methods admit simple and efficient implementations, which makes them attractive for exploratory applications in which sliding window processing is widely used with subjective and ad hoc choices of design parameters. In this context, the novelty of our contributions lies in:

1. Capturing the evolution of the spectra across windows by modeling the dynamics of certain eigen-spectral quantities arising in MT analysis (e.g., spectral eigen-coefficients and eigen-spectra);

2. Addressing the additive measurement noise and multiplicative sampling noise, which severely distort the spectrograms obtained by the multitaper framework; and

3. Constructing a framework for precise statistical assessment of the estimates, by addressing the dependency among windows using a Bayesian formulation.

As it will be evident in Section 2.4, the use of state-space models in the context of MT analysis results in adaptive weighting of the estimates of the eigen-coefficients or eigen-spectra across windows, thanks to the optimal data combining feature of Bayesian smoothing. These adaptive weights depend on the common dynamic trends shared across windows and hence result in capturing the degree of smoothness inherent in the signal, while producing estimates robust against uncertainties due to observation noise and limited data.

It is noteworthy that the use of state-space models here is significantly different from those used in *parametric* non-stationary spectral analysis methods such as TFARMA modeling [62]. In the TFARMA formulation, time delays and frequency shifts are used to model the non-stationary dynamics of the process in a physically intuitive way. These state-space models therefore determine the functional form of the resulting spectral estimates in closed form in terms of the finite set of ARMA coefficients. In contrary, the state-space models used here do not determine the functional form of the spectral estimates at each window, and rather control the temporal smoothness of the eigen-spectral quantities via forming a regularization mechanism in the underlying Bayesian estimation framework (See Section 2.1.5). In our approach, we indeed estimate the spectrogram at a given number of frequency bins in each window, which scales with the total number of samples, in the same vein as sliding window MT spectrogram.

In light of the above, our algorithms belong to the class of *semi-parametric* estimation methods, as the underlying model is a hybrid of *parametric* unobservable state evolution process and a *non-parametric* data generating process [75]. In Section 2.3.1, we will compare our proposed semi-parametric methodology with both non-parametric and parametric techniques, namely the MT spectrogram analysis and the Time-Frequency Autoregressive (TFAR) modeling technique [62].

2.1.4 Problem Formulation

Assume, without loss of generality, that an arbitrary window of length W is chosen so that for some integer N, $NW = T$ and let $\mathbf{y}_n = \big[y_{(n-1)W+1}, y_{(n-1)W+2}, \cdots, y_{nW}\big]^\top$ for $n = 1, 2 \cdots N$, denotes the data in the nth window. This way, the entire data is divided into N non-overlapping segments of length W each. To this end, we invoke quasistationarity assumption by modeling y_t to be stationary within each of these segments of length W. With this assumption, motivated by the major sources of uncertainty in spectral estimation, i.e., measurement noise and sampling noise, we formulate two state-space frameworks in the following subsections.

2.1.4.1 Mitigating the Measurement Noise

Suppose that $\widetilde{y}_t$ is the noise corrupted observation obtained from the true signal y_t, i.e., $\widetilde{y}_t = y_t + v_t$, where $(v_t)_{t=1}^T$ is an i.i.d. zero-mean Gaussian noise sequence with fixed variance σ^2. By discretizing the representation in Eq. (3.2) at a frequency spacing of $2\pi/J$ with J an integer, at any arbitrary window n, we have

$$\widetilde{\mathbf{y}}_n = \mathbf{F}_n \mathbf{x}_n + \mathbf{v}_n, \tag{2.7}$$

where $\mathbf{F}_n$ is a matrix with elements $(\mathbf{F}_n)_{l,j} := \exp\left(i2\pi(((n-1)W+l)\frac{j-1}{J})\right)$ for $l = 1, 2, \cdots, W$ and $j = 1, 2, , \cdots J$; $\widetilde{\mathbf{y}}_n := \left[\widetilde{y}_{(n-1)W+1}, \widetilde{y}_{(n-1)W+2}, \cdots, \widetilde{y}_{nW}\right]^\top$ is the noisy observation of the true signal $\mathbf{y}_n$; $x_n(f)$ and $\mathbf{x}_n := \left[x_n(0), x_n(2\pi\frac{1}{J}), \cdots, x_n(2\pi\frac{J-1}{J})\right]^\top$ denote the orthogonal increment process and its discretized version, respectively at window n and $\mathbf{v}_n = \left[v_{(n-1)W+1}, v_{(n-1)W+2}, \cdots, v_{nW}\right]^\top$ is zero-mean Gaussian noise with covariance $\text{Cov}\{\mathbf{v}_i, \mathbf{v}_j\} = \sigma^2 \mathbf{I}\delta_{i,j}$.

Let $\mathbf{u}^{(k)} := \left[u_1^{(k)}, u_2^{(k)}, \cdots, u_W^{(k)}\right]^\top$ denotes the kth dpss taper and $\widetilde{\mathbf{y}}_n^{(k)} := \mathbf{u}^{(k)} \odot \widetilde{\mathbf{y}}_n$, where $\odot$ denotes element-wise multiplication. Let $x_n^{(k)}(f)$ and $\mathbf{x}_n^{(k)} := [x_n^{(k)}(0), x_n^{(k)}(2\pi\frac{1}{J}), \cdots, x_n^{(k)}(2\pi\frac{J-1}{J})]^\top$ denote the kth spectral eigen-coefficient of $\mathbf{y}_n$ and its discretized version, respectively, for $k = 1, 2, \cdots, K$. Then, following Eq. (2.7) we consider the following spectrotemporal representation of the tapered data segments:

$$\widetilde{\mathbf{y}}_n^{(k)} = \mathbf{F}_n \mathbf{x}_n^{(k)} + \mathbf{v}_n^{(k)}, \tag{2.8}$$

where $\mathbf{v}_n^{(k)}$ is the contribution of $\mathbf{v}_n$ to the kth tapered data, assumed to be independent of $\mathbf{x}_{1:n-1}^{(k)}$, and identically distributed according to a zero-mean Gaussian distribution with covariance $\text{Cov}\{\mathbf{v}_i^{(k)}, \mathbf{v}_j^{(k)}\} = {\sigma^{(k)}}^2 \mathbf{I}\delta_{i,j}$. We view $\widetilde{\mathbf{y}}_n^{(k)}$ as a noisy observation corresponding to the true eigen-coefficient $\mathbf{x}_n^{(k)}$, which provides a linear Gaussian forward model for the observation process.

In order to capture the evolution of the spectrum and hence systematically enforce temporal smoothness, we impose a stochastic continuity constraint on the eigen-coefficients $(\mathbf{x}_n^{(k)})_{n=1}^N$ for $k = 1, 2, \cdots K$, using a first-order difference equation:

$$\mathbf{x}_n^{(k)} = \alpha^{(k)} \mathbf{x}_{n-1}^{(k)} + \mathbf{w}_n^{(k)}, \tag{2.9}$$

starting with an initial condition $\mathbf{x}_0^{(k)} = [0, 0, \cdots, 0]^\top \in \mathbb{R}^J$, where $0 \leqslant \alpha^{(k)} < 1$, and $\mathbf{w}_n^{(k)}$ is independent of $\mathbf{x}_{1:n-1}^{(k)}$ and assumed to be independently distributed according to a zero-mean Gaussian distribution with diagonal covariance $\text{Cov}\{\mathbf{w}_i^{(k)}, \mathbf{w}_j^{(k)}\} = \mathbf{Q}_i^{(k)} \delta_{i,j}$. Under this assumption, the discrete-time process, $(\mathbf{x}_n^{(k)})_{n=1}^N$ forms a jointly Gaussian random process with independent increments, while the process itself is statistically dependent. An estimate of the unobserved states (true eigen-coefficients) from the observations (tapered data) under this model suppresses the measurement noise and captures the state dynamics.

It is worth mentioning that, here, the terms smoothness or continuity do not adhere to their usual notions used for continuous valued functions. Instead, here we say the discrete states, $\{\mathbf{x}\}_{k=0}^K$:

- are temporally smooth, if the states satisfy:

$$\|\mathbf{x}_k - \alpha \mathbf{x}_{k-1}\|_2 \leq M \|\mathbf{x}_k\|_2 \ \forall \ k = 1, \cdots, K, \ \ M < \infty$$

- follow a stochastic continuity constraint if the following statements are true:

$$\max_{k=1,\cdots,K} \mathbb{E}\left[\mathbf{x}_k - \alpha\mathbf{x}_{k-1}\right] = 0, \quad \text{and} \quad \max_{k=1,\cdots,K} \mathbb{E}\left[\|\mathbf{x}_k - \alpha\mathbf{x}_{k-1}\|_2^2\right] < \infty,$$

  i.e., one possible to way to enforce temporal smoothness in probabilistic case, for some α satisfying $0 < \alpha \leq 1$.

2.1.4.2 Mitigating the Sampling Noise

Suppose the additive measurement noise is negligible, i.e. $v_t \approxeq 0$. For now, consider only a single window of length W. It is known that when the spectrum does not rapidly vary over the chosen design bandwidth B, the eigen-spectra are approximately uncorrelated and the following approximation holds for $k = 1, 2 \cdots, K$ [12, 70]:

$$\frac{\widehat{S}^{(k)}(f)}{S(f)} \sim \frac{\chi_2^2}{2}, \ 0 < f < 1/2, \tag{2.10}$$

where $\widehat{S}^{(k)}(f)$ and $S(f)$ are the tapered estimate and true PSD, respectively. In other words, the empirical eigen-spectra of the process can be thought of as the true spectra corrupted by a multiplicative noise, due to sampling and having access to only a single realization of the process. We refer to this uncertainty induced by sampling as *sampling noise*. By defining $\psi^{(k)}(f) := \log \widehat{S}^{(k)}(f) + \log 2$ and $s^{(k)}(f) := \log S(f)$, we can transform the multiplicative effect of the sampling noise in Eq.

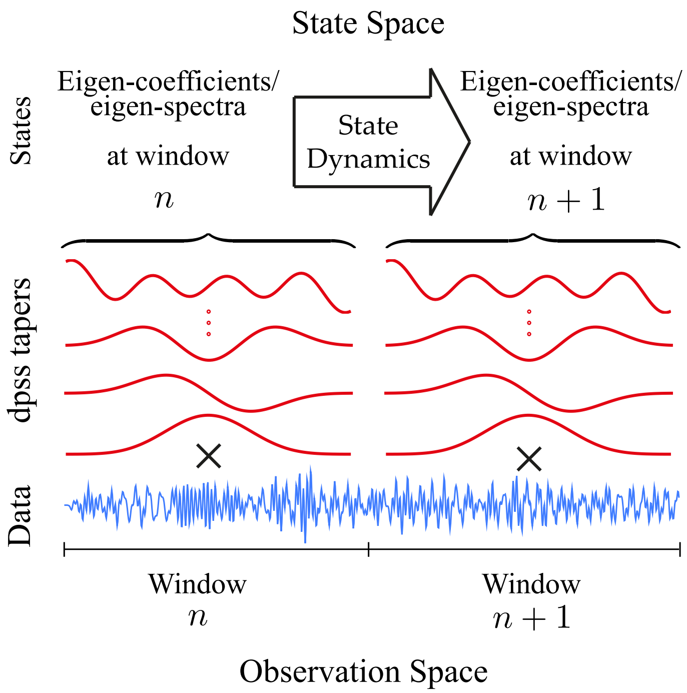

State Space
States
Eigen-coefficients/ eigen-spectra at window n
State Dynamics
Eigen-coefficients/ eigen-spectra at window $n+1$
dpss tapers
×
×
Data
Window n
Window $n+1$
Observation Space

Figure 2.1: Schematic depiction of the proposed models.

(2.10) to the following additive forward model[76]:

$$\psi^{(k)}(f) = s^{(k)}(f) + \phi^{(k)}(f), \tag{2.11}$$

where $\phi^{(k)}(f)$ is a log-chi-square distributed random variable, capturing the uncertainty due to sampling noise. It can be shown that $\phi^{(k)}(f)$ has a density given by:

$$p(\phi) = \frac{1}{2}\exp\left(\phi - \frac{1}{2}\exp(\phi)\right), \tag{2.12}$$

which belongs to the family of log-Gamma distributions, including the Gumbel and Bramwell-Holdsworth-Pinton distributions common in extreme value statistics [77].

In order to incorporate this observation model in our dynamic framework, we define the state vector, $\mathbf{s}_n^{(k)} := \left[s_n^{(k)}(0), s_n^{(k)}\left(2\pi\frac{1}{J}\right), \cdots, s_n^{(k)}\left(2\pi\frac{J-1}{J}\right)\right]^\top$; the observation vector, $\boldsymbol{\psi}_n^{(k)} := \left[\psi_n^{(k)}(0), \psi_n^{(k)}\left(2\pi\frac{1}{J}\right), \cdots, \psi_n^{(k)}\left(2\pi\frac{J-1}{J}\right)\right]^\top$ and the observation noise vector, $\boldsymbol{\phi}_n^{(k)} := \left[\phi_n^{(k)}(0), \phi_n^{(k)}\left(2\pi\frac{1}{J}\right), \cdots, \phi_n^{(k)}\left(2\pi\frac{J-1}{J}\right)\right]^\top$. Then, the forward model at window n can be stated as:

$$\boldsymbol{\psi}_n^{(k)} = \mathbf{s}_n^{(k)} + \boldsymbol{\phi}_n^{(k)}, \tag{2.13}$$

where each element of $\boldsymbol{\phi}_n^{(k)}$ is log-chi-square distributed. Similar to the preceding model, we impose a stochastic continuity constraint over the logarithm of the eigen-spectra as follows:

$$\mathbf{s}_n^{(k)} = \theta^{(k)}\mathbf{s}_{n-1}^{(k)} + \mathbf{e}_n^{(k)}, \tag{2.14}$$

starting with an initial condition $\mathbf{s}_0^{(k)} = [0, 0, \cdots, 0]^\top \in \mathbb{R}^J$, where $0 \leqslant \theta^{(k)} < 1$, and $\mathbf{e}_n^{(k)}$ is assumed to be a zero-mean Gaussian vector independent of $\mathbf{s}_{1:n-1}^{(k)}$ and with a diagonal covariance $\mathrm{Cov}\{\mathbf{e}_i^{(k)}, \mathbf{e}_j^{(k)}\} = \mathbf{R}_i^{(k)}\delta_{i,j}$. Note that the logarithm function maps the range of the eigen-spectra in $[0, \infty)$ to $(-\infty, \infty)$ which makes the Gaussian state evolution plausible. An estimate of the unobserved states (logarithm of the true spectra) from the observations (logarithm of the empirical eigen-spectra) under this model suppresses the sampling noise and captures the state dynamics.

In summary, through these models we project the data of each short window onto the functional space spanned by the PSWFs and impose stochastic continuity constraints (Eq. (2.9) and Eq. (2.14)) on these projections (eigen-coefficients or eigen-spectra) in order to recover spectral representations that are smooth in time and robust against measurement or sampling noise. Fig. 2.1 provides a visual illustration of the proposed modeling paradigm.

2.1.5 The Inverse Problem

We formulate the spectral estimation problem as one of Bayesian estimation, in which the Bayesian risk/loss function, fully determined by the posterior density of $(\mathbf{x}_n^{(k)})_{n=1,k=1}^{N,K}$ (resp. $(\mathbf{s}_n^{(k)})_{n=1,k=1}^{N,K}$) given the observations $(\widetilde{\mathbf{y}}_n^{(k)})_{n=1,k=1}^{N,K}$ (resp. $(\boldsymbol{\psi}_n^{(k)})_{n=1,k=1}^{N,K}$) is minimized. We first consider the forward model of Eq. (2.8), which provides the observed data likelihood given the states. Under the state-space model of Eq. (2.9), the kth eigen-coefficient can be estimated by solving the following maximum *a posteriori* (MAP) problem:

$$\min_{\mathbf{x}_1^{(k)},\mathbf{x}_2^{(k)},\cdots,\mathbf{x}_N^{(k)}} \sum_{n=1}^{N} \left[\frac{1}{\sigma^2} \left\| \widetilde{\mathbf{y}}_n^{(k)} - \mathbf{F}_n \mathbf{x}_n^{(k)} \right\|_2^2 + (\mathbf{x}_n^{(k)} - \alpha \mathbf{x}_{n-1}^{(k)})^H {\mathbf{Q}_n^{(k)}}^{-1} (\mathbf{x}_n^{(k)} - \alpha \mathbf{x}_{n-1}^{(k)}) \right], \tag{2.15}$$

for $k = 1, 2, \cdots, K$. Similarly, in the second state space framework Eq. (2.13), the eigen-spectra can be obtained by solving another MAP estimation problem:

$$\min_{\mathbf{s}_1^{(k)},\mathbf{s}_2^{(k)},\cdots,\mathbf{s}_N^{(k)}} \sum_{n=1}^{N} \Bigg[\mathbf{1}_J^\top [\mathbf{s}_n^{(k)} - \boldsymbol{\psi}_n^{(k)} + \frac{1}{2} \exp(\boldsymbol{\psi}_n^{(k)} - \mathbf{s}_n^{(k)})] \\ + (\mathbf{s}_n^{(k)} - \theta \mathbf{s}_{n-1}^{(k)})^H {\mathbf{R}_n^{(k)}}^{-1} (\mathbf{s}_n^{(k)} - \theta \mathbf{s}_{n-1}^{(k)}) \Bigg], \tag{2.16}$$

for $k = 1, 2, \cdots, K$, where $\mathbf{1}_J$ is the vector of all ones of length J. We call the MAP estimation problems in Eq. (2.15) and Eq. (2.16) the Dynamic Bayesian Multitaper (DBMT) and the log-DBMT estimation problems, respectively. Similarly, the respective spectrogram estimates will be denoted by the DBMT and log-DBMT estimates.

Eq. (2.15) is a strictly convex function of $\mathbf{x}_n^{(k)} \in \mathbb{C}^W$ and $\mathbf{Q}_n^{(k)} \in \mathbb{S}_{++}^W$ for $n = 1, 2 \cdots, N$, which can be solved using standard eqimization techniques. However, these techniques do not scale well with the data length N. A careful examination of the log-posterior reveals a block tri-diagonal structure of the Hessian, which can be used to develop efficient recursive solutions that exploit the temporal structure of the problem. A similar argument holds for the eqimization problem in Eq. (2.16). However, the parameters of these state-space models need to be estimated from the data. In the next section, we show how the EM algorithm can be used to both estimate the parameters and states efficiently from the eqimization problems Eq. (2.15) and Eq. (2.16).

2.2 Fast Recursive Solutions via the EM Algorithm

In order to solve the MAP problem in Eq. (2.15), we need to find the parameters $\mathbf{Q}_n^{(k)} \in \mathbb{S}_{++}^W$ and $\alpha^{(k)} \in (0, 1]$ for $n = 1, 2 \cdots, N$ and $k = 1, 2, \cdots, K$. Similarly $\mathbf{R}_n^{(k)} \in \mathbb{S}_{++}^W$ and $\theta^{(k)} \in (0, 1]$ need to be estimated for the problem in Eq. (2.16). If the underlying states were known, one could further maximize the log-posterior with respect to the parameters. This observation can be formalized in the EM framework [65, 66, 78]. To avoid notational complexity, we drop the dependence of the various variables on the taper index k in the rest of this subsection.

2.2.1 The DBMT Spectrum Estimation Algorithm

By treating $\mathbf{x}_n, n = 1, 2, \cdots, N$ as the hidden variables and $\alpha, \mathbf{Q}_n, n = 1, 2, \cdots, N$ as the unknown parameters to be estimated, we can write the complete log-likelihood as:

$$\log L(\alpha, \mathbf{Q}_{1:N}) := -\sum_{n=1}^{N} \left[\frac{1}{\sigma^2} \|\widetilde{\mathbf{y}}_n - \mathbf{F}_n \mathbf{x}_n\|_2^2 + \log |\det \mathbf{Q}_n| \right. \\ \left. + (\mathbf{x}_n - \alpha \mathbf{x}_{n-1})^H \mathbf{Q}_n^{-1} (\mathbf{x}_n - \alpha \mathbf{x}_{n-1}) \right] + c, \qquad (2.17)$$

where c represents the terms that do not depend on α, $(\mathbf{Q}_n)_{n=1}^N$ or $(\mathbf{x}_n)_{n=1}^N$. For simplicity of exposition, we assume that $\mathbf{Q}_n = \mathbf{Q}$ for $n = 1, 2, \cdots, N$. The forthcoming treatment can be extended to the general case with little modification. Also, note that σ^2 can be absorbed in $\mathbf{Q}$, and thus is assumed to be known. At the lth iteration, we have:

E-Step

Given $\alpha^{[l]}, \mathbf{Q}^{[l]}$, for $n = 1, 2, \cdots, N$, the expectations, $\mathbf{x}_{n|N} := \mathbb{E}[\mathbf{x}_n | \widetilde{\mathbf{y}}_{1:N}, \alpha^{[l]}, \mathbf{Q}^{[l]}]$, $\mathbf{\Sigma}_{n|N} := \mathbb{E}[(\mathbf{x}_n - \mathbf{x}_{n|N})(\mathbf{x}_n - \mathbf{x}_{n|N})^H | \widetilde{\mathbf{y}}_{1:N}, \alpha^{[l]}, \mathbf{Q}^{[l]}]$, and $\mathbf{\Sigma}_{n,n-1|N} := \mathbb{E}[(\mathbf{x}_n - \mathbf{x}_{n|N})(\mathbf{x}_{n-1} - \mathbf{x}_{n-1|N})^H | \widetilde{\mathbf{y}}_{1:N}, \alpha^{[l]}, \mathbf{Q}^{[l]}]$, can be calculated using the Fixed Interval Smoother (*FIS*) [79] (lines 4 and 5) and the state-space covariance smoothing algorithm [80] (line 6). These expectations can be used to compute the expectation of the complete data log-likelihood $\mathbb{E}\big[\log L(\alpha, \mathbf{Q}) | \widetilde{\mathbf{y}}_{1:N}, \alpha^{[l]}, \mathbf{Q}^{[l]}\big]$.

Algorithm 1: The DBMT Estimate of the kth Eigen-coefficient

```
Initialize: observations $\widetilde{\mathbf{y}}^{(k)}_{1:N}$; initial guess $\mathbf{x}^{(0)}_{1:N}$; initial guess $\mathbf{Q}^{[0]}$; initial conditions $\mathbf{\Sigma}_{0|0}$; tolerance $\mathsf{tol} \in (0, 10^{-3})$, Maximum Number of iteration $L_{\max} \in \mathbb{N}^{+}$.
repeat
    $l = 0$.
    Forward filter for $n = 1, 2, \cdots, N$:
        $\mathbf{x}_{n|n-1} = \alpha^{[l]}\mathbf{x}_{n-1|n-1}$
        $\mathbf{\Sigma}_{n|n-1} = {\alpha^{[l]}}^2\mathbf{\Sigma}_{n-1|n-1} + \mathbf{Q}^{[l]}$
        $\mathbf{K}_n = \mathbf{\Sigma}_{n|n-1}\mathbf{F}_n^H(\mathbf{F}_n\mathbf{\Sigma}_{n|n-1}\mathbf{F}_n^H + \sigma^2\mathbf{I})^{-1}$
        $\mathbf{x}_{n|n} = \mathbf{x}_{n|n-1} + \mathbf{K}_n(\widetilde{\mathbf{y}}_n - \mathbf{F}_n\mathbf{x}_{n|n-1})$
        $\mathbf{\Sigma}_{n|n} = \mathbf{\Sigma}_{n|n-1} - \mathbf{K}_n\mathbf{F}_n\mathbf{\Sigma}_{n|n-1}$
    Backward smoother for $n = N-1, N-2, \cdots, 1$:
        $\mathbf{B}_n = \alpha^{[l]}\mathbf{\Sigma}_{n|n}\mathbf{\Sigma}^{-1}_{n+1|n}$
        $\mathbf{x}_{n|N} = \mathbf{x}_{n|n} + \mathbf{B}_n(\mathbf{x}_{n+1|N} - \mathbf{x}_{n+1|n})$
        $\mathbf{\Sigma}_{n|N} = \mathbf{\Sigma}_{n|n} + \mathbf{B}_n(\mathbf{\Sigma}_{n+1|N} - \mathbf{\Sigma}_{n+1|n})\mathbf{B}_n^H$
    Covariance smoothing for $n = N-1, N-2, \cdots, 1$:
        $\mathbf{\Sigma}_{n,n-1|N} = \mathbf{B}_{n-1}\mathbf{\Sigma}_{n|N}$
    Let $\widehat{\mathbf{X}}^{[l]} := [\mathbf{x}^H_{1|N}, \mathbf{x}^H_{2|N}, \cdots, \mathbf{x}^H_{N|N}]^H$.
    Update $\alpha^{[l+1]}$ and $\mathbf{Q}^{[l+1]}$ as:
        $\alpha^{[l+1]} = \frac{\sum_{n=2}^{N} \mathrm{Tr}(\mathbf{\Sigma}_{n,n-1|N}{\mathbf{Q}^{[l]}}^{-1}) + \mathbf{x}^H_{n-1|N}{\mathbf{Q}^{[\mathbf{l}]}}^{-1}\mathbf{x}_{n|N}}{\sum_{n=2}^{N} \mathrm{Tr}(\mathbf{\Sigma}_{n-1|N}{\mathbf{Q}^{[l]}}^{-1}) + \mathbf{x}^H_{n-1|N}{\mathbf{Q}^{[l]}}^{-1}\mathbf{x}_{n-1|N}}$,
        $\mathbf{Q}^{[l+1]} = \frac{1}{N}\sum_{n=1}^{N}\Big[\mathbf{x}_{n|N}\mathbf{x}^H_{n|N} + \mathbf{\Sigma}_{n|N} + {\alpha^{[l+1]}}^2(\mathbf{x}_{n-1|N}\mathbf{x}^H_{n-1|N} + \mathbf{\Sigma}_{n-1|N})$
            $- \alpha^{[l+1]}(\mathbf{x}_{n-1|N}\mathbf{x}^H_{n|N} + \mathbf{x}_{n|N}\mathbf{x}^H_{n-1|N} + 2\mathbf{\Sigma}_{n,n-1|N})\Big]$.
    Set $l \leftarrow l + 1$.
until $\frac{\|\widehat{\mathbf{X}}^{[l]} - \widehat{\mathbf{X}}^{(l-1)}\|_2}{\|\widehat{\mathbf{X}}^{[l]}\|_2} < \mathsf{tol}$ or $l = L_{\max}$.
Output: Denoised eigen-coefficients $\widehat{\mathbf{X}}^{[L]}$ where $L$ is the index of the last iteration of the algorithm, and error covariance matrices $\mathbf{\Sigma}_{n|N}$ for $n = 1, 2, \cdots, N$ in from last iteration of the algorithm.
```

M-Step

The parameters for subsequent iterations, $\alpha^{[l+1]}$ and $\mathbf{Q}^{[l+1]}$ can be obtained by maximizing the expectation of Eq. (2.17). Although this expectation is convex in α and $\mathbf{Q}$ individually, it is not a convex function of both. Hence, we perform cyclic iterative updates for $\alpha^{[l+1]}$ and $\mathbf{Q}^{[l+1]}$ given by:

$$\alpha^{[l+1]} = \frac{\sum_{n=2}^{N} \mathrm{Tr}(\boldsymbol{\Sigma}_{n,n-1|N}\mathbf{Q}^{[l]^{-1}}) + \mathbf{x}_{n-1|N}^{H}\mathbf{Q}^{[\mathbf{l}]^{-1}}\mathbf{x}_{n|N}}{\sum_{n=2}^{N} \mathrm{Tr}(\boldsymbol{\Sigma}_{n-1|N}\mathbf{Q}^{[l]^{-1}}) + \mathbf{x}_{n-1|N}^{H}\mathbf{Q}^{[l]^{-1}}\mathbf{x}_{n-1|N}} \tag{2.18}$$

and

$$\begin{aligned}\mathbf{Q}^{[l+1]} = \frac{1}{N}\sum_{n=1}^{N}\Big[&\mathbf{x}_{n|N}\mathbf{x}_{n|N}^{H} + \boldsymbol{\Sigma}_{n|N} + {\alpha^{[l+1]}}^{2}(\mathbf{x}_{n-1|N}\mathbf{x}_{n-1|N}^{H}+\boldsymbol{\Sigma}_{n-1|N})\\ &-\alpha^{[l+1]}(\mathbf{x}_{n-1|N}\mathbf{x}_{n|N}^{H}+\mathbf{x}_{n|N}\mathbf{x}_{n-1|N}^{H}+2\boldsymbol{\Sigma}_{n,n-1|N})\Big].\end{aligned} \tag{2.19}$$

These iterations can be performed until convergence to a possibly local maximum. However, with even one such update, the overall algorithm forms a majorization-minimization (MM) procedure, generalizing the EM procedure and enjoying from similar convergence properties [81]. One possible implementation of this iterative procedure is described in Algorithm 1. Once the DBMT estimates of all the K eigen-coefficients $\widehat{\mathbf{x}}_{n}^{(k)}$ are obtained, for $n = 1, 2, \cdots, N$ and $k = 1, 2, \cdots, K$, the DBMT spectrum estimate is constructed similar to Eq. (2.5):

$$\widehat{D}_{n}(f_j) = \frac{1}{K}\sum_{k=1}^{K}\left|\left(\widehat{\mathbf{x}}_{n}^{(k)}\right)_{j}\right|^{2}, \tag{2.20}$$

where $f_j := \frac{2\pi(j-1)}{J}$ for $j = 1, 2, \cdots, J$ and $n = 1, 2, \cdots, N$. Confidence intervals can be computed by mapping the Gaussian confidence intervals for $\widehat{\mathbf{x}}_n^{(k)}$'s to the final DBMT estimate.

2.2.2 The log-DBMT Spectrum Estimation Algorithm

We utilize a similar iterative procedure based on the EM algorithm to find the log-DBMT spectrum estimate. As before, we treat $\mathbf{s}_n, n = 1, 2, \cdots, N$ as hidden variables and $\theta, \mathbf{R}_n, n = 1, 2, \cdots, N$ as the unknown parameters to be estimated. In order to give more flexibility to the observation model, we consider the observation noise to be distributed as log-chi-square with degrees of freedom 2ν, for some positive integer ν to be estimated. The density of each element of $\boldsymbol{\phi}_n^{(k)}$ is then given by:

$$p(\phi) = \frac{1}{2^\nu \Gamma(\nu)} \exp\left(\nu\phi - \frac{1}{2}\exp(\phi)\right). \tag{2.21}$$

We can express the complete data log-likelihood as:

$$\log L(\nu, \theta, \mathbf{R}_{1:n}) := -\sum_{n=1}^{N} \Bigg[\mathbf{1}_J^\top \Big(\nu(\mathbf{s}_n - \boldsymbol{\psi}_n) + \tfrac{1}{2}\exp(\boldsymbol{\psi}_n - \mathbf{s}_n)\Big) + J\Big(\nu \log 2 + \log\Gamma(\nu)\Big) \\ + (\mathbf{s}_n - \theta\mathbf{s}_{n-1})^H \mathbf{R}_n^{-1} (\mathbf{s}_n - \theta\mathbf{s}_{n-1}) + \log|\det \mathbf{R}_n| \Bigg] + c, \tag{2.22}$$

where c represents the terms that do not depend on ν, θ, $(\mathbf{R}_n)_{n=1}^N$ or $(\mathbf{s}_n)_{n=1}^N$. Again assuming $\mathbf{R}_n = \mathbf{R}$ for all $n = 1, 2, \cdots, N$ for simplicity, the following EM algorithm can be constructed:

E-step

Computation of the conditional expectation of the log-likelihood in Eq. (2.22) requires evaluating $\mathbb{E}[\mathbf{s}_n|\boldsymbol{\psi}_{1:N}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}]$ and $\mathbb{E}[\exp(-\mathbf{s}_n)|\boldsymbol{\psi}_{1:N}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}]$ for $n = 1, 2, \cdots, N$. Unlike the DBMT estimation problem, the forward model in this case is non-Gaussian, and hence we cannot apply the Kalman filter and FIS to find the state expectations. To compute the conditional expectation, the distribution of $\mathbf{s}_n|\boldsymbol{\psi}_{1:n}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}$ or its samples are required [67]. Computation of the distribution $\mathbf{s}_n|\boldsymbol{\psi}_{1:n}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}$ involves intractable integrals and sampling from the distribution using numerical methods such as Metropolis-Hastings is not computationally efficient, especially for long data, given that it has to be carried out at every iteration. Since the posterior distribution is unimodal and a deviation from the Gaussian posterior, we approximate the distribution of $\mathbf{s}_n|\boldsymbol{\psi}_{1:n}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}$ as a Gaussian distribution by matching its mean and covariance matrix to the log-posterior in Eq. (2.22). To this end, the mean is approximated by the mode of $f_{\mathbf{s}_n|\boldsymbol{\psi}_{1:n}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}}$ and the covariance is set to the inverse of the negative Hessian of the log-likelihood in Eq. (2.22) [66, 82]. Under this approximation, computing $\mathbb{E}[\exp(-\mathbf{s}_n)|\boldsymbol{\psi}_{1:n}, \mathbf{R}^{[l]}, \theta^{[l]}, \nu^{[l]}]$ is also facilitated thanks to the closed-form moment generating function of $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$\mathbb{E}\left[\exp\left(\mathbf{a}^\top \mathbf{z}\right)\right] = \exp\left(\mathbf{a}^\top \boldsymbol{\mu} + \frac{1}{2}\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}\right). \qquad (2.23)$$

Algorithm 2: The log-DBMT Estimate of the kth log-Eigen-spectra

```
Initialize: observations $\boldsymbol{\psi}_{1:N}^{(k)}$; initial guess $\mathbf{s}_{1:N}^{[0]}$; initial guess $\mathbf{R}^{[0]}$; initial conditions $\boldsymbol{\Omega}_{0|0}$; tolerance $\mathsf{tol} \in (0, 10^{-3})$, Maximum Number of iteration $L_{\max} \in \mathbb{N}^{+}$.
repeat
    $l = 0$.
    Forward filter for $n = 1, 2, \cdots, N$:
        $\mathbf{s}_{n|n-1} = \theta^{[l]}\mathbf{s}_{n-1|n-1}$
        $\boldsymbol{\Omega}_{n|n-1} = {\theta^{[l]}}^2\boldsymbol{\Omega}_{n-1|n-1} + \mathbf{R}^{[l]}$
        $\mathbf{s}_{n|n} = \mathbf{s}_{n|n-1} + \boldsymbol{\Omega}_{n|n-1}\left[\frac{1}{2}\exp(\boldsymbol{\psi}_n - \mathbf{s}_{n|n}) - \nu^{[l]}\mathbf{1}_J\right]$
        $\boldsymbol{\Omega}_{n|n} = \boldsymbol{\Omega}_{n|n-1}^{-1} - \frac{1}{2}\mathrm{diag}\{\exp(\boldsymbol{\psi}_n - \mathbf{s}_{n|n})\}$
    Backward smoother for $n = N-1, N-2, \cdots, 1$:
        $\mathbf{A}_n = \theta^{[l]}\boldsymbol{\Omega}_{n|n}\boldsymbol{\Omega}_{n+1|n}^{-1}$
        $\mathbf{s}_{n|N} = \mathbf{s}_{n|n} + \mathbf{A}_n(\mathbf{s}_{n+1|N} - \mathbf{s}_{n+1|n})$
        $\boldsymbol{\Omega}_{n|N} = \boldsymbol{\Omega}_{n|n} + \mathbf{A}_n(\boldsymbol{\Omega}_{n+1|N} - \boldsymbol{\Omega}_{n+1|n})\mathbf{A}_n^H$
    Covariance smoothing for $n = N-1, N-2, \cdots, 1$:
        $\boldsymbol{\Omega}_{n,n-1|N} = \mathbf{A}_{n-1}\boldsymbol{\Omega}_{n|N}$
    Let $\widehat{\mathbf{S}}^{[l]} := [\mathbf{s}_{1|N}^H, \mathbf{s}_{2|N}^H, \cdots, \mathbf{s}_{N|N}^H]^H$.
    Update $\nu^{[l+1]}$, $\theta^{[l+1]}$ and $\mathbf{R}^{[l+1]}$ as:
        $\theta^{[l+1]} = \frac{\sum_{n=2}^N \mathrm{Tr}(\boldsymbol{\Omega}_{n,n-1|N}{\mathbf{R}^{[l]}}^{-1}) + \mathbf{s}_{n-1|N}^H{\mathbf{R}^{[l]}}^{-1}\mathbf{s}_{n|N}}{\sum_{n=2}^N \mathrm{Tr}(\boldsymbol{\Omega}_{n-1|N}{\mathbf{R}^{[l]}}^{-1}) + \mathbf{s}_{n-1|N}^H{\mathbf{R}^{[l]}}^{-1}\mathbf{s}_{n-1|N}}$,
        $\nu^{[l+1]} = \frac{1 - \log 2 + \frac{1}{JN}\sum_{n=1}^N \mathbf{1}_J^\top(\boldsymbol{\psi}_n^{[l]} - \mathbf{s}_n^{[l]}) - \digamma(\nu^{[l+1]})}{\frac{2}{JN}\sum_{n=1}^N \mathbf{1}_J^\top \exp(\boldsymbol{\psi}_n^{[l]} - \mathbf{s}_n^{[l]})}$,
        $\mathbf{R}^{[l+1]} = \frac{1}{N}\sum_{n=1}^N \big[\mathbf{s}_{n|N}\mathbf{s}_{n|N}^H + \boldsymbol{\Omega}_{n|N} + {\theta^{[l+1]}}^2(\mathbf{s}_{n-1|N}\mathbf{s}_{n-1|N}^H + \boldsymbol{\Omega}_{n-1|N})$
            $- \theta^{[l+1]}(\mathbf{s}_{n-1|N}\mathbf{s}_{n|N}^H + \mathbf{s}_{n|N}\mathbf{s}_{n-1|N}^H + 2\boldsymbol{\Omega}_{n,n-1|N})\big]$.
    Set $l \leftarrow l + 1$.
until $\frac{\|\widehat{\mathbf{S}}^{[l]} - \widehat{\mathbf{S}}^{(l-1)}\|_2}{\|\widehat{\mathbf{S}}^{[l]}\|_2} < \mathsf{tol}$ or $l = L_{\max}$.
Output: Denoised log-eigen-spectra $\widehat{\mathbf{S}}^{[L]}$ where $L$ is the index of the last iteration of the algorithm, and error covariance matrices $\boldsymbol{\Omega}_{n|N}$ for $n = 1, 2, \cdots, N$ in from last iteration of the algorithm.
```

Similar to the case of DBMT, we can exploit the block tri-diagonal structure of the Hessian in Eq. (2.16) to carry out the E-step efficiently using forward filtering and backward smoothing.

M-step

Once the conditional expectation of the log-likelihood in Eq. (2.22) given $\boldsymbol{\psi}_{1:n}$, $\mathbf{R}^{[l]}$, $\theta^{[l]}$, $\nu^{[l]}$ is available, we can update $\mathbf{R}^{[l+1]}$ and $\theta^{[l+1]}$ using similar closed form equations as in Eq. (2.19). But updating $\nu^{[l+1]}$ by maximizing the conditional expectation of the log likelihood in Eq. (2.22) wrt. $\nu^{[l+1]}$ requires solving following nonlinear equation:

$$\nu^{[l+1]} = \frac{1 - \log 2 + \frac{1}{JN}\sum_{n=1}^{N} \mathbf{1}_J^\top(\boldsymbol{\psi}_n^{[l]} - \mathbf{s}_n^{[l]}) - \digamma(\nu^{[l+1]})}{\frac{2}{JN}\sum_{n=1}^{N} \mathbf{1}_J^\top \exp(\boldsymbol{\psi}_n^{[l]} - \mathbf{s}_n^{[l]})}, \tag{2.24}$$

where $\digamma(\cdot)$ is the digamma function. We can use Newton's method to solve this equation up to a given precision. An implementation of the log-DBMT is given by Algorithm 2. Note that unlike the DBMT algorithm which pertains to a Gaussian observation model, the forward filtering step to compute $\mathbf{s}_{n|n}$ is nonlinear, and standard techniques such as Newton's method can be used to solve for $\mathbf{s}_{n|n}$. We use the log-DBMT algorithm to find all the K estimates of true log-spectra and construct the log-DBMT estimate as:

$$\widehat{D}_n(f_j) = \frac{1}{K}\sum_{k=1}^{K} \exp\left(\left(\widehat{\mathbf{s}}_n^{(k)}\right)_j\right), \tag{2.25}$$

where $f_j := \frac{2\pi(j-1)}{J}$ for $j = 1, 2, \cdots, J$ and $n = 1, 2, \cdots, N$. Again, confidence intervals can be computed by mapping the Gaussian confidence intervals for $\widehat{\mathbf{s}}_n^{(k)}$'s to the final log-DBMT estimate.

2.2.3 Parameter Selection

The window length W, design bandwidth B, and the number of tapers K need to be carefully chosen. Since both proposed algorithms are motivated by the standard overlapping MT method, we use the same guidelines for choosing these parameters [70, 74]. The window length W is determined based on the expected rate of change of the PSD (given domain-specific knowledge) in order to make sure that the quasi-stationarity assumption holds. The design bandwidth B is chosen small enough to be able to resolve the dominant frequency components in the data, while being large enough to keep the time-bandwidth product $\rho := WB \geqslant 1$. The number of tapers K is then chosen as $K \leqslant \lfloor 2\rho \rfloor - 1$ [70].

2.3 Application to Synthetic and Real Data

Before presenting our theoretical analysis, we examine the performance of DBMT and log-DBMT spectrogram estimators on synthetic data, and then demonstrate their utility in two real world data applications, namely spectral analysis of human EEG during sleep and Electric Network Frequency signal detection.

2.3.1 Application to Synthetic Data

The synthetic data consists of a linear combination of an amplitude-modulated and a frequency-modulated processes with high dynamic range (i.e., high-Q). The amplitude-modulated component $y_t^{(1)}$ is generated through modulating an AR(6) process tuned around 11 Hz by a cosine at a low frequency $f_0 = 0.02\,\text{Hz}$. The frequency-modulated component $y_t^{(2)}$ is a realization of an ARMA$(6,4)$ with varying pole loci. To this end, the process has a pair of 3rd order poles at $\omega_t := 2\pi f_t$ and $-\omega_t$, where f_t increases from 5 Hz, starting at $t = 0$, every $\sim 26\,\text{s}$ by increments of 0.48 Hz, to achieve frequency modulation. In summary, the noisy observations are given by:

$$y_t = y_t^{(1)} \cos(2\pi f_0 t) + y_t^{(2)} + \sigma v_t, \tag{2.26}$$

where v_t is a white Gaussian noise process and σ is chosen to achieve an SNR of 30 dB. The process is truncated at 600 s to be used for spectrogram analysis. Fig. 2.2 shows a 12 s sample window of the process.

In addition to the standard overlapping MT, we present comparison to the TFAR method as an example of parametric state-space modeling approaches to non-stationary spectral analysis. This method is known to be well suited to processes whose time-varying spectra exhibits sharp peaks, i.e., signals consisting of several narrow-

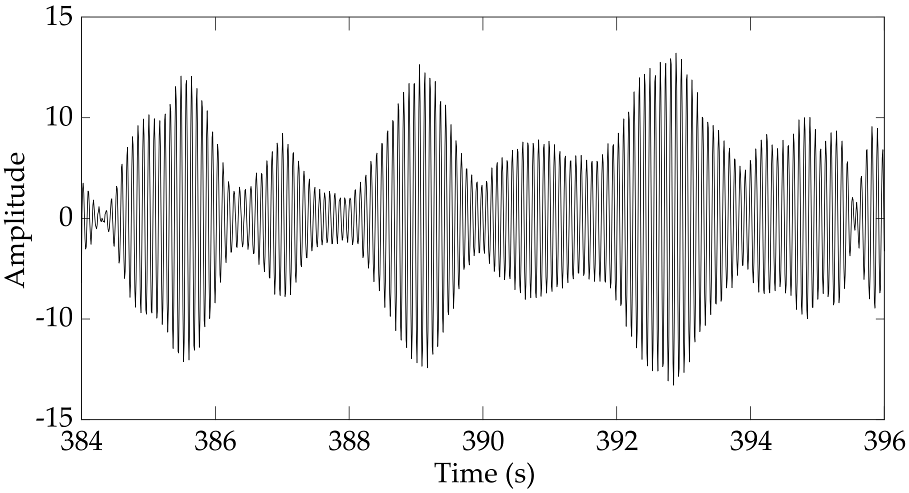


Figure 2.2: Sample from the synthetic data from t =384 s to 396 s.

band components [62]. The TFAR model is defined by the input-output relation

$$y_t := -\sum_{m=1}^{M_A} \sum_{l=-L_A}^{L_A} a_{m,l} e^{i\frac{2\pi}{N}lt} y_{t-m} + \sum_{l=-L_B}^{L_B} b_{0,l} e^{i\frac{2\pi}{N}lt} e_t,$$

where e_t is a stationary white noise process with unit variance, and $(a_{m,l})_{m=1,l=-L_A}^{M_A,L_A}$ and $(b_{0,l})_{l=-L_B}^{L_B}$ are the autoregressive (AR) and zero-delay moving average (MA) parameters, respectively. The integers M_A and L_A are respectively the delay and Doppler model orders of the AR component and L_B denotes the Doppler model order of the zero-delay MA component. The AR and MA parameters are estimated by solving the time-frequency Yule-Walker equations, from which the evolutionary spectra can be constructed (See the methods described in [62] for more details).

Fig. 2.3 shows the true as well as estimated spectrograms by the standard overlapping MT, DBMT, log-DBMT and the TFAR estimators. Each row consists of three panels: the left panel shows the entire spectrogram; the middle panel shows a zoomed-in spectrotemporal region marked by the dashed box in the left panel; and the right panel shows the PSD along with confidence interval (CI) in gray hull, at a selected time point marked by a dashed vertical line in the middle panel. Note that for the standard MT estimates, the CIs are constructed assuming a χ^2_{2K} distribution of the estimates around the true values [12, 70], whereas for DBMT and log-DBMT estimate by mapping the Gaussian confidence intervals for eigen-coefficients or eigen-spectra to the final estimates. We were not able to evaluate the CIs for the TFAR estimates, since to the best of our knowledge we are not aware of any method to do so. Fig. 2.3*A* shows true spectrogram of the synthetic process, in which the existence of both amplitude and frequency modulations makes the spectrogram estimation a challenging problem.

Fig. 2.3*B* shows the standard overlapping MT spectrogram estimate. We used windows of length 6 s and the first 3 tapers corresponding to a time-bandwidth product of 3 and 50% overlap to compute the estimates (note that the same window length, tapers and time-bandwidth product are used for the DBMT and log-DBMT estimators). Although the standard MT spectrogram captures the dynamic evolution of both components, it is blurred by the background noise and picks up spectral artifacts (i.e., vertical lines) due to window overlap, frequency mixing, and sampling noise. Fig. 2.3*C* demonstrates how the DBMT spectrogram estimate overcomes these deficiencies of the overlapping MT spectrogram: the spectrotemporal localization is sharper and smoother across time, artifacts due to overlapping between windows are vanished, and frequency mixing is further mitigated. By comparing the right panel of the second and third rows, two important observations can be made: first, the DBMT captures the true dynamic range of the original noiseless PSD,

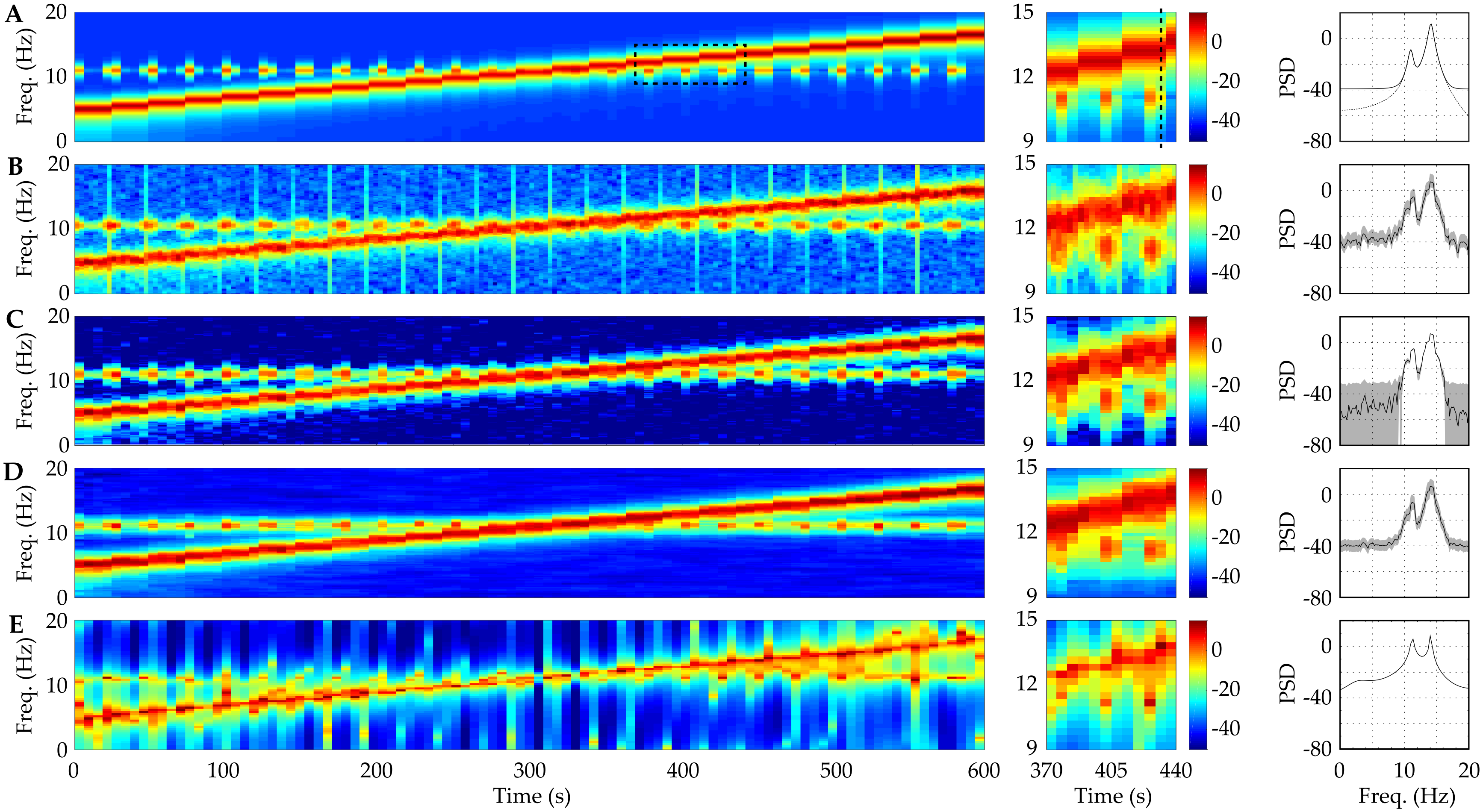


Figure 2.3: Spectrogram analysis of the synthetic data. (A) Ground truth, (B) overlapping MT estimates, (C) DBMT estimates, (D) log-DBMT estimates, and (E) TFAR estimates. Left: spectrograms. Middle: zoomed-in views from $t = 370\,\mathrm{s}$ to $t = 440\,\mathrm{s}$. The color scale is in decibels. Right: PSDs corresponding to a window of length 6 s starting at $t = 474\,\mathrm{s}$. Dashed and solid lines in row A show respectively the noiseless and noisy PSDs. Grey hulls show 95% confidence intervals.

the standard MT estimate fails to do so. Second, the CIs in Fig. 2.3C as compared to 2.3B are wider when the signal is weak (e.g., near 5 Hz) and tighter when the signal is strong (e.g., near 11 Hz). The latter observation highlights the importance of the model-based confidence intervals in interpreting the denoised estimates of DBMT: while the most likely estimate (i.e., the mean) captures the true dynamic range of the noiseless PSD, the estimator does not preclude cases in which the noise floor of -40 dB is part of the true signal, while showing high confidence in detecting the spectral content of the true signal that abides by the modeled dynamics.

Next, Fig. 2.3D shows the log-DBMT spectrogram estimate, which shares the artifact rejection feature of the DBMT spectrogram. However, the log-DBMT estimate is smoother than both the standard overlapping MT and DBMT spectrograms in time as well as in frequency (see the zoomed-in middle panels), due to its sampling noise mitigation feature (by design). However, similar to the standard MT estimate, the log-DBMT estimator treats the observation noise as part of the signal, and thus does not suppress it. Though, the confidence intervals of the log-DBMT PSD estimate are tighter than those of the standard overlapping MT estimate due to averaging across multiple windows via Bayesian filtering/smoothing. As we will show in Section 2.4, these qualitative observations can be established by our theoretical analysis.

Finally, Fig. 2.3E shows the TFAR spectral estimate. The model orders are chosen as $M_A = 20$, $L_A = 25$, $L_B = 25$, large enough to allow the parametric model to achieve high time-frequency resolution. As it is shown in the right panel, the TFAR method provides the smoothest estimate along the frequency axis. However, it is not successful in capturing the true dynamic range of the signal due to spectral leakage (See middle and right panels). In addition, it is contaminated by similar vertical frequency artifacts as in the case of the MT spectrogram (See Fig. 2.3A).

In the spirit of easing reproducibility, we have deposited a MATLAB implementation of these algorithms on the open source repository GitHub [83], which generates Fig. 2.3.

2.3.2 Application to EEG data

To illustrate the utility of our proposed spectrogram estimators, we apply them to human EEG data recorded during sleep. In the interest of space, in the remainder of this section, we only present comparisons with the MT spectrogram as a non-parametric benchmark. The EEG data set is available online as part of the SHHS Polysomnography Database (`https://www.physionet.org/pn3/shhpsgdb/`). The data is 900 s long during stage 2 sleep, and sampled at 250 Hz. During stage 2 sleep, the EEG is known to manifest delta waves (0 Hz to 4 Hz) and sleep spindles (transient wave packets with frequency 12 Hz to 14 Hz) [84, 85]. Accurate localization of these spectrotemporal features has significant applications in studying sleep disorders and cognitive function [84]. Since the transient spindles occur at a time scale of seconds, we choose a window length of 2.25 s for all algorithms (with 50% overlap for the standard overlapping MT estimate). We also chose a time-bandwidth product of 2.25 for all algorithms, in order to keep the frequency resolution at 2 Hz. Figs. 2.4A,

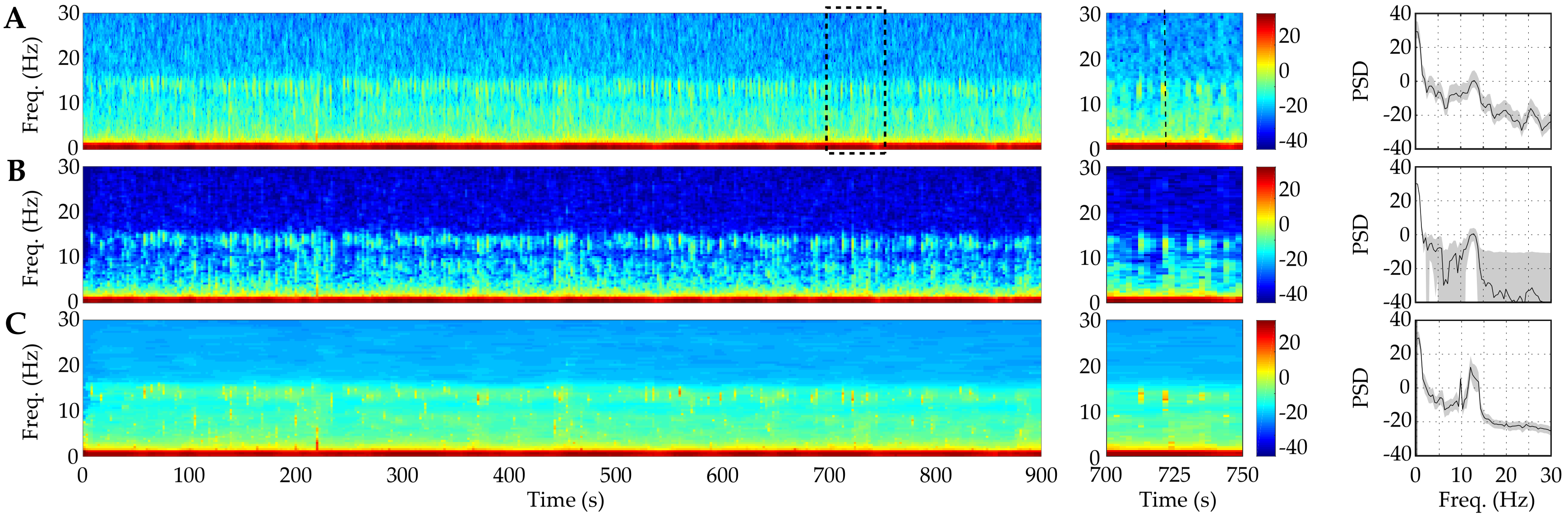


Figure 2.4: Spectrogram analysis of the EEG data. (A) overlapping MT estimates, (B) DBMT estimates, and (C) log-DBMT estimates. Left: spectrograms. Middle: zoomed-in views from t =700 s to 750 s. The color scale is in decibels. Right: PSD estimate corresponding to a window of length 2.25 s starting at t =722.25 s. Grey hulls show 95% confidence intervals.

B and *C* show the MT, DBMT and log-DBMT spectrogram estimates, respectively, with a similar presentational structure as in Fig. 2.3. As the middle panels reveal, the overlapping MT estimate is not able to clearly distinguish the delta waves and sleep spindles due to high background noise. The DBMT estimate shown in Fig. 2.4*B*, however, provides a significantly denoised spectrogram, in which the delta waves and sleep spindles are visually separable. The log-DBMT estimator shown in Fig. 2.4*C* provides significant spectrotemporal smoothing, and despite not fully reducing the background noise, provides a clear separation of the delta waves and spindles (see the PSD in the right panel). Similar to the analysis of synthetic data, the same observations regarding the confidence intervals of the estimators can be made.

2.3.3 Application to ENF data

Finally, we examine the performance of our proposed algorithms in tracking the Electrical Network Frequency (ENF) signals from audio recordings. The ENF signal corresponds to the supply frequency of the power distribution network which is embedded in audio recordings [86, 87]. The instantaneous values of this time-varying frequency and its harmonics form the ENF signal. The ability to detect and track the spectrotemporal dynamics of ENF signals embedded in audio recordings has shown to be crucial in data forensics applications [86].

Fig. 2.5*A* shows the spectrogram estimates around the sixth harmonic of the nominal 60 Hz ENF signal (data from [87]). We used 1000 s of audio recordings, and constructed spectrograms with windows of length 5 s and using the first 3 tapers

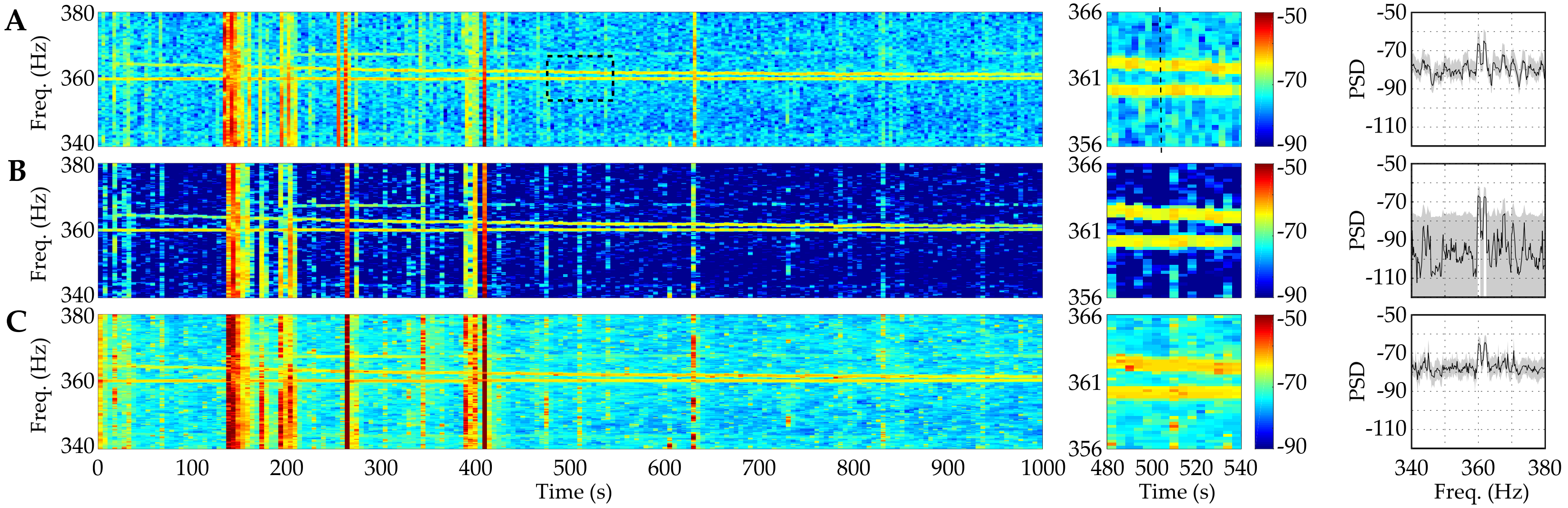


Figure 2.5: Spectrogram analysis of the ENF data. (A) overlapping MT estimates, (B) DBMT estimates, and (C) log-DBMT estimates. Left: spectrograms. Middle: zoomed-in views from $t = 480\,s to 540\,$s. The color scale is in decibels. Right: PSD estimate corresponding to a window of length 5 s starting at $t = 505\,$s. Grey hulls show 95% confidence intervals.

corresponding to a time-bandwidth product of 3 for all three methods (with 25% overlap for the overlapping MT estimate). The two dominant components around the sixth ENF harmonic exhibit temporal dynamics, but are hard to distinguish from the noisy background. Fig. 2.5B shows the DBMT spectrogram, in which the background noise is significantly suppressed, yielding a crisp and temporally smooth estimate of the ENF dynamics. The log-DBMT estimate is shown in Fig. 2.5C, which provides higher spectrotemporal smoothness than the standard MT estimate. Although the log-DBMT shows smaller variability in the estimates (middle and right panels), the gain is not as striking as in the cases of synthetic data and EEG analysis, due to the usage of longer windows which mitigates the sampling noise for all algorithms. Similar observations as in the previous two cases regarding the statistical confidence intervals can be made, which highlight the advantage of modeling the spectrotemporal dynamics in spectrogram estimation.

2.4 Theoretical Analysis

2.4.1 Filter Bank Interpretation

In order to characterize the spectral properties of any non-parametric spectrum estimator, the tapers applied to the data need to be carefully inspected. In the MT framework, the dpss sequences are used as tapers, which are known to produce negligible side-lobes in the frequency domain [11, 12]. The DBMT and log-DBMT algorithms also use the dpss tapers to alleviate the problem of frequency leakage. However, because of the stochastic continuity constraint we introduced, the estimate associated to any given window is now a function of the data in *all* the windows. Therefore, the theoretical properties of the MT method do not readily apply to our estimators.

To characterize the statistical properties of our estimates, we first need take a detour from the usual analysis of spectrum estimation techniques. In what follows, we mainly focus on the DBMT algorithm for the sake of presentation. By virtue of the FIS procedure under the assumptions that:

1. the window length W is an integer multiple of J, the number of discrete frequencies, so that $\mathbf{F}_n = \mathbf{F}_1, \forall n$, and

2. the state noise covariance matrices are time-invariant, i.e., $\mathbf{Q}_n = \mathbf{Q}, \forall n$,

one obtains the following expansion of $\mathbf{x}_{n|N}^{(k)}$ in terms of the observed data [63]:

$$\mathbf{x}_{n|N}^{(k)} = \sum_{s=1}^{n-1} \prod_{m=s}^{n-1} \big[\alpha(\mathbf{I}-\mathbf{K}_m\mathbf{F}_m)\big]\mathbf{K}_s\mathbf{U}^{(k)}\widetilde{\mathbf{y}}_s + \mathbf{K}_n\mathbf{U}^{(k)}\widetilde{\mathbf{y}}_n + \sum_{s=n+1}^{N} \prod_{m=n}^{s} \mathbf{B}_m\mathbf{K}_s\mathbf{U}^{(k)}\widetilde{\mathbf{y}}_s. \tag{2.27}$$

In other words, the DBMT algorithm maps the entire data $\widetilde{\mathbf{y}} := [\widetilde{y}_1, \widetilde{y}_2, \cdots, \widetilde{y}_T]^\top$ to the vector of coefficients $\widehat{\mathbf{X}}^{(k)}$ according to [63]:

$$\widehat{\mathbf{X}}^{(k)} = \mathbf{G}^{(k)}\mathbf{F}^H\mathbf{U}^{(k)}\widetilde{\mathbf{y}}, \tag{2.28}$$

where $\mathbf{F}$ and $\mathbf{U}^{(k)}$ are block-diagonal matrices with $\mathbf{F}_1$ and $\mathbf{U}_k := \text{diag}[\mathbf{u}^{(k)}]$ as the diagonal blocks, respectively, and $\mathbf{G}$ is a weighting matrix which depends only on $\mathbf{Q}_\infty = \lim_{l\to\infty}\mathbf{Q}^{[l]}$, $\alpha_\infty = \lim_{l\to\infty}\alpha^{[l]}$, and window length, W. The rows of

$\mathbf{G}^{(k)}\mathbf{F}^H\mathbf{U}^{(k)}$ form a filter bank whose output is equivalent to the time-frequency representation.

In order to continue our analysis, we make two assumptions common in the analysis of adaptive filters[88, 89]. First, we assume that the parameter estimates $\mathbf{Q}_\infty$ and α_∞ are close enough to the true values of $\mathbf{Q}$ and α, and therefore replace them by $\mathbf{Q}$ and α, i.e., as if the true parameters were known. Note that we have discarded the dependence of $\mathbf{Q}$ and α on k for lucidity of analysis. Second, noting that $\alpha(\mathbf{I} - \mathbf{K}_m\mathbf{F}_m) = \alpha\boldsymbol{\Sigma}_{m|m}\boldsymbol{\Sigma}^{-1}_{m|m-1}$ and $\mathbf{B}_m = \alpha\boldsymbol{\Sigma}_{m|m}\boldsymbol{\Sigma}^{-1}_{m+1|m}$ and that in steady state we have $\boldsymbol{\Sigma}_{m|m} := \boldsymbol{\Sigma}_\infty$ and $\boldsymbol{\Sigma}_{m|m-1} = \alpha^2\boldsymbol{\Sigma}_\infty + \mathbf{Q}$, Eq. (2.27) can be approximated by:

$$\mathbf{x}^{(k)}_{n|N} = \sum_{s=1}^{N} \boldsymbol{\Lambda}^{|s-n|}\boldsymbol{\Gamma}\mathbf{F}_s^H\mathbf{U}^{(k)}\widetilde{\mathbf{y}}_s, \tag{2.29}$$

for $1 \ll n \ll N$, where $\boldsymbol{\Lambda} = \alpha\boldsymbol{\Sigma}_\infty(\alpha^2\boldsymbol{\Sigma}_\infty + \mathbf{Q})^{-1}$, and $\boldsymbol{\Gamma} = (\alpha^2\boldsymbol{\Sigma}_\infty + \mathbf{Q})\big[\mathbf{I} - rW\big((\alpha^2\boldsymbol{\Sigma}_\infty + \mathbf{Q})^{-1} + rW\mathbf{I}\big)^{-1}\big]$. That is, for values of n far from the data boundaries, the weighting matrix is equivalent to a weighted set of dpss tapers in matrix form acting on all the data windows with an exponential decay with respect to the nth window.

As an example, the equivalents filters of the DBMT estimator corresponding to the first taper for the 11 Hz and 9 Hz frequencies around 300 s, from the synthetic data example are shown in Fig. 2.6. They are also compared to the equivalent filters corresponding to first taper of standard MT method in the frequency domain. As apparent from Fig. 2.6, the weighting matrix sets the gain of these filters in an adaptive fashion across *all* windows, unlike the standard MT method which only uses the data in window n. In addition, the filter corresponding to frequency of 9Hz, which is negligible in the data, is highly attenuated, resulting in significant noise suppression. In this sense, the proposed estimation method can be identified as a *data-driven denoising method* for constructing time-frequency representations given noisy time series data. Next, we will characterize the performance of the DBMT estimator in terms of bias-variance trade-off.

2.4.2 Bias and Variance Analysis

We first consider the implication of the stochastic continuity constraint of Eq. (2.9) on the evolution of the orthogonal increment processes governing the time series data. We first assume that the parameters $\alpha^{(k)} = \alpha$, for all $k = 1, 2, \cdots, K$. Suppose that the data in window n has a Cramér representation with an orthogonal increment process $dz_n(f)$, $n = 1, 2, \cdots, N$. Then, one way to achieve the stochastic

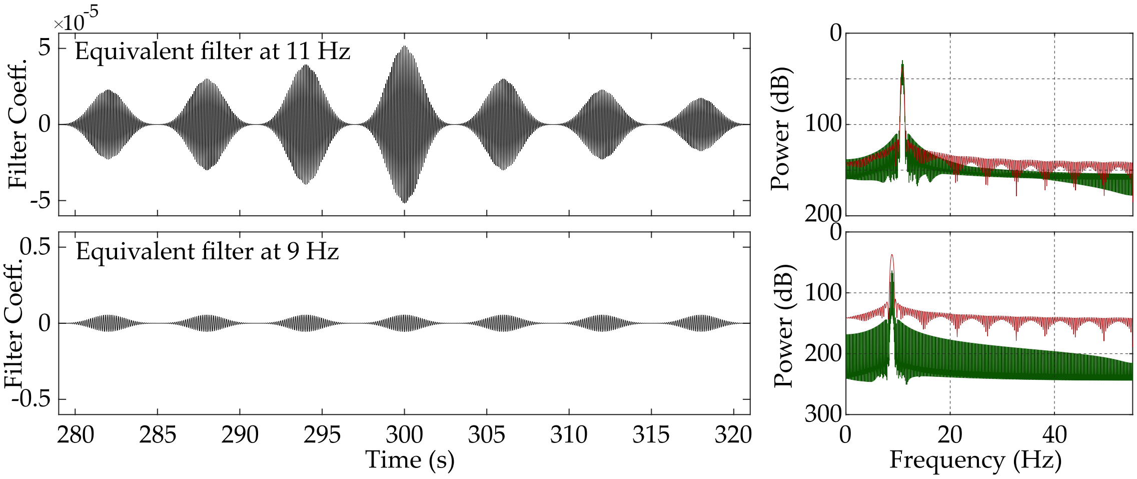


Figure 2.6: Equivalent filters corresponding to the first taper of the DBMT estimate of the synthetic data example. Left: equivalent filters in time around $t = 300$ s. Right: equivalent filters of MT (red) and DBMT (green) in frequency.

continuity of Eq. (2.9) is to assume:

$$dz_{n+1}(f) = \alpha dz_n(f) + d\epsilon_n(f), \tag{2.30}$$

where $d\epsilon_n(f)$ is a Gaussian orthogonal increment process, independent of $dz_n(f)$. In the forthcoming analysis we also assume the locally stationarity condition, i.e., the generalized Fourier transform of the process remains stationary within each window. This assumption is common in the analysis of non-parametric spectral estimators [70, 74]. Finally, we assume a scaling of $K, N, W \to \infty$, $B \to 0$, $BW \to \rho$, for some constant ρ [90]. The following theorems characterize the bias and variance of the DBMT estimator:

Theorem 2.1. *Suppose that the locally stationary process y_t is governed by orthogonal increment processes evolving according to the dynamics $dz_{n+1}(f) = \alpha dz_n(f) + d\epsilon_n(f)$, with $\alpha < 1$, where the noise process $d\epsilon_n(f), \forall f \in (-1/2, 1/2]$ is a zero-mean Gaussian increment process with variance $q(f) > 0$, independent of $dz_n(f)$. If the process is corrupted by additive zero-mean Gaussian noise with variance σ^2, then for $f \in \{f_1, f_2, \cdots, f_J\}$, the* DBMT *estimate satisfies:*

$$\left|\mathbb{E}[\widehat{D}_n(f)] - D(f)\right| \leqslant \left(1 - \frac{1}{K}\sum_{k=1}^{K} \lambda_k\right) \kappa_n(f) \sup_f \{D(f)\} + |1 - \kappa_n(f)| D(f) + \mu_n(f)\sigma^2 + \kappa_n(f) o(1),$$

where λ_k is the eigenvalue associated with the kth PSWF, $D(f) := q(f)/(1-\alpha)$, and $\kappa_n(f), \mu_n(f)$ are functions of α and $q(f)$ and explicitly given in the proof.

Theorem 2.2. *Under the assumptions of Theorem 2.1, the variance of the* DBMT *estimate $\widehat{D}_n(f)$ satisfies:*

$$\mathsf{Var}\left\{\widehat{D}_n(f)\right\} \leqslant \frac{2}{K}\left[\sup_f\{\kappa_n(f)D(f)+\mu_n(f)\sigma^2\}\right]^2.$$

The proofs of Theorems 2.1 and 2.2 integrate the treatment of [90] with the structure of the FIS estimates, and are presented in Appendix A. In order to illustrate the implications of these theorems, several remarks are in order:

Remark 2.1. The function $\kappa_n(f)$ controls the trade-off between bias and variance: for values of $\kappa_n(f) < 1$, the bound on the variance decreases while the bias bound increases, and for $\kappa_n(f) \approx 1$, all the terms in the bias bound become negligible, while the variance bound increases. The function $\mu_n(f)$, on the other hand, reflects observation noise suppression in both the bias and variance. Note that these upper bounds are tight and achieved for a signal with flat spectrum.

Remark 2.2. The bias and variance bounds of [90] for the standard MT method can be recovered by setting $\kappa_n(f) = 1$, $\mu_n(f) = 1$, and $\sigma^2 = 0$ in the results of Theorems 2.1 and 2.2, i.e., in the absence of signal dynamics and measurement noise. For the DBMT estimator, signal and measurement noise variances, respectively contribute to the bias/variance upper bounds in different fashions through $\kappa_n(f)$ and $\mu_n(f)$, due to the distinction of the signal and measurement noise in our state-space model. In contrast, in the standard MT method, possible measurement noise is treated in the same way as the true data, and hence both the signal and noise variances have equal contributions in the estimator bias/variance.

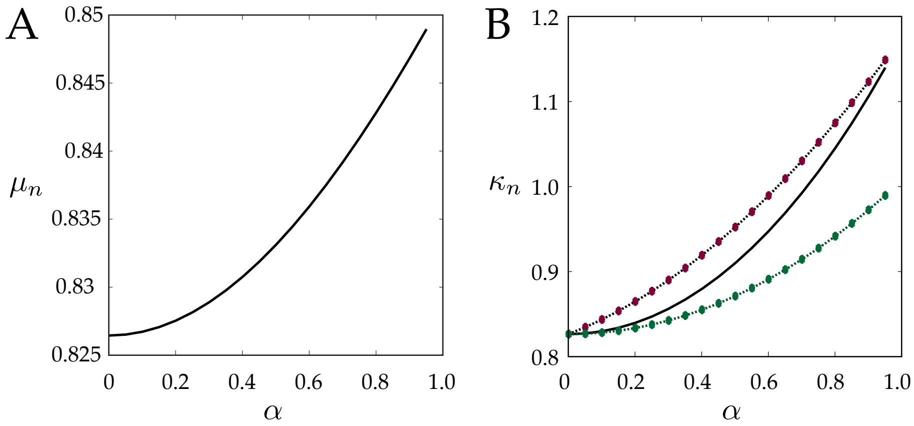


Figure 2.7: (A) μ_n versus α, (B) κ_n and its upper/lower bounds versus α for $N = 100$, $n = 50$, and $q/\sigma^2 = 10$.

The functions $\kappa_n(f)$ and $\mu_n(f)$ do not have closed-form expressions with respect to the state-space parameters α, σ^2 and $q(f)$. In order to illustrate the roles of $\kappa_n(f)$ and $\mu_n(f)$, we consider the scenario under which the upper bounds on the bias and variance are achieved, i.e., $q(f)$ being independent of f, and hence $\kappa_n(f) = \kappa_n$ and $\mu_n(f) = \mu_n$, $\forall f$. In this scenario, even though the dependence of μ_n and κ_n on the state-space parameters are quite involved, it is possible to obtain upper and lower bounds on κ_n and μ_n. As it is shown in Proposition A.1 in Appendix A, the main parameters determining the behavior of κ_n and μ_n are q_n/σ (i.e., the SNR) and α (i.e., temporal signal dependence). Here, we present a numerical example for clarification. Fig. 2.7*A* and *B* shows the plot of μ_n vs. α and κ_n vs. α for $n = 50$ and $q/\sigma^2 = 10$. It is apparent that μ_n increases with α and does not exceed 1. The fact that $\mu_n < 1$ implies that the DBMT estimator achieves a higher noise suppression compared to the standard MT method. This fact agrees with the noise suppression performances observed in Section 2.3.

Fig. 2.7*B* shows the plot of κ_n vs. α, which exhibits a similar increasing

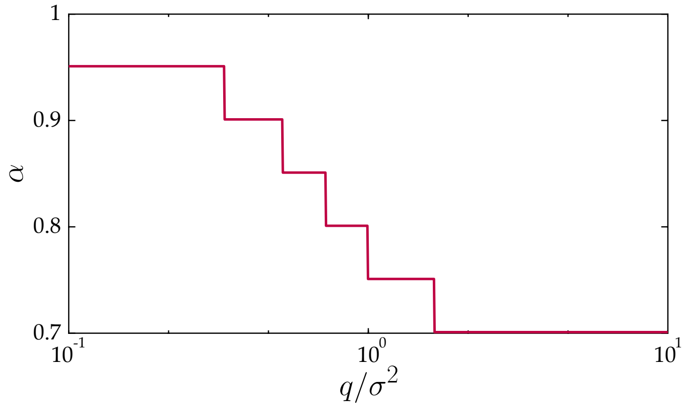


Figure 2.8: α corresponding to $\kappa_n \approx 1$ against q/σ^2

trend, but eventually exceeds 1. This result implies that with a careful choice α, it is possible to achieve $\kappa_n < 1$, and hence obtain lower variance than that of the standard MT estimate. Fig. 2.8 illustrates this statement by showing the value of α for which $\kappa_n \approx 1$ vs. q_n/σ^2. For models with high temporal dependence (i.e., α close to 1), it is possible to achieve $\kappa_n < 1$ and hence reduce the estimator variance, due to the increase in the weight of data pooled from adjacent windows, even for small values of q_n/σ^2 (i.e., low SNR). However, this reduction in variance comes with the cost of increasing the bias. On the contrary when the data across windows have low temporal dependence (i.e., $\alpha \ll 1$), it is only possible to achieve a reduction in variance for high values of q_n/σ^2 (i.e., high SNR). This is due to the fact that at low SNR with low temporal dependence, pooling data from adjacent windows is not beneficial in reducing the variance in a particular window, and indeed can result in higher bias.

Remark 2.3. Even though the parameter α is estimated in a data-driven fashion, it can be viewed as a tuning parameter controlling the bias-variance trade-off, given the foregoing discussion. For a given SNR, fixing α at a small value can help reduce the variance but with a cost of increasing bias, and vice versa. In light of this observation, Fig. 2.8 can be thought of as a guideline for choosing α to achieve $\kappa_n \approx 1$, so that the estimator is nearly unbiased, and achieves a lower variance than that of the standard overlapping MT estimator due to the noise suppression virtue of the state-space model. Although we focused on the case of flat spectrum in the foregoing discussion, it is possible to numerically compute these trade-off curves for more general cases, given the general expressions for $\kappa_n(f)$ and $\mu_n(f)$ given in Appendix A.

Remark 2.4. Extending Theorems 2.1 and 2.2 to the log-DBMT algorithm is not straightforward, due to the high nonlinearity of the underlying state-space models. However, under the common Gaussian approximation of the log-posterior density, the well-known variance reduction property of the fixed interval smoother [79, 89] carries over to the estimator of $\log S_n(f)$. That is, the variance of the log-DBMT estimate of $\log S_n(f)$ obtained using the state-space model is lower than that of the standard MT estimate which only uses the data within window n. This fact agrees with our earlier observation in Section 2.3.1 regarding the tightening of the confidence intervals for log-DBMT as compared to the overlapping MT estimates.

2.5 Concluding Remarks

Spectral analysis of non-stationary time series data poses serious challenges for classical non-parametric techniques, in which temporal smoothness of the spectral representations are implicitly captured using sliding windows with overlap. This widely-practiced approach tends to ignore the inherent smoothness of the data and is not robust against measurement/sampling noise. In this chapter, we address these issues and provide an alternative to the sliding window spectrogram analysis paradigm. We propose two semi-parametric spectrogram estimators, namely the DBMT and log-DBMT estimators, by integrating techniques from MT analysis and Bayesian estimation. To this end, we explicitly model the temporal dynamics of the spectrum using a state-space model over the spectral features obtained by multitapering. Therefore our algorithms inherit the optimality features of both Bayesian estimators and MT analysis.

Our algorithms admit efficient and simple implementations, thanks to the Expectation-Maximization algorithm and the well-known fixed interval state-space smoothing procedure. Unlike existing approaches, our algorithms require no *a priori* assumptions about the structure of the spectral representation and operate in a fully data-driven fashion. While both algorithms yield spectral estimates that are smooth in time, by design the DBMT algorithm significantly suppresses the measurement noise in forming the spectrogram and the log-DBMT algorithm mitigates sampling noise due to small observation length. We establish the performance gains provided by our algorithms through theoretical analysis of the bias-variance trade-off, as well as application to synthetic and real data from human EEG recording during sleep and ENF signals from audio recordings.

Chapter 3: Multitaper Spectral Analysis of Neuronal Spiking Activity Driven by Latent Stationary Processes

The advent of invasive recording technologies from the brains of animals and humans, such as multi-electrode arrays and electrocorticography (ECoG), has resulted in abundant pools of neuronal spiking data, which often exhibit oscillatory features [91]. Characterizing the properties of these oscillations with high spectral resolution is crucial to understanding their role in cognitive functions.

Most existing spectral analysis techniques, e.g. nonparametric techniques based on Fourier methods and Wavelets, the multitaper method [11, 74, 92], or their extensions such as DBMT analysis in Chapter 2, State-space multitaper time-frequency analysis [93], however, are designed for continuous-time data and cannot be readily applied to binary spiking data. There have been efforts aimed at addressing this challenge, which consider the periodogram of smoothed spike trains using kernel methods as the spectral representation [94, 95, 96]. Another strand of results are based on the theory of point processes, which has been widely used in recent years to model and analyze the statistical properties of binary spike trains [97, 98, 99, 100]. These techniques relate the Conditional Intensity Function (CIF) or the spiking rate of a point process governing the spiking statistics to intrinsic and external neural covariates using state-space models. Then, the spectrum of the estimated CIF is characterized using standard nonparametric or parametric techniques [66, 101, 102].

Despite their relative success in application, these methods have several shortcomings from a theoretical perspective. First, it is known that explicit smoothing of the signal using kernels or implicit smoothing using state-space models, results in the distortion of the spectrum [12]. Second, time-domain smoothing alleviates the variance of the estimates at the cost of increasing the bias. On the other hand, existing techniques which avoid time-domain smoothing (e.g., [103]) may exhibit high variability. Third, these modeling frameworks often require a priori information (e.g., sparsity) or may suffer from model mismatch (e.g., overly-smoothed state estimates).

To address these issues, in this chapter we introduce a novel multitaper spectral analysis method, which we call the Point Process Multitaper Method (PMTM), to be directly applied to binary data. To this end, we generate auxiliary spiking statistics which correspond to the tapered versions of the CIF, which are then used to independently estimate the eigen-spectra of the tapered CIFs via the Maximum Likelihood (ML) procedure. The multitaper spectral estimate is formed by averaging the corresponding eigen-spectral estimates. Our approach distinguishes itself from existing work by providing a direct spectral estimator from binary observations via a novel adaptation of multitaper analysis, with no recourse to intermediate time-domain smoothing or need for a priori information. We demonstrate the performance of PMTM using simulated spike trains driven by an autoregressive (AR) process and experimentally recorded neural data under general anesthesia. Our results reveal substantial gains achieved by PMTM as compared to existing nonparametric techniques, in terms of the bias-variance trade-off.

3.1 Problem Formulation

Let $N(t)$ and H_t denote the point process representing the number of spikes and spiking history in $[0, t)$, respectively, where $t \in [0, T]$ and T denotes the observation duration. The CIF of a point process $N(t)$ is defined as:

$$\lambda(t|H_t) := \lim_{\Delta \to 0} \frac{P[N(t+\Delta) - N(t) = 1|H_t]}{\Delta}. \tag{3.1}$$

To discretize the continuous process, we consider time bins of length Δ, small enough that the probability of having two or more spikes in an interval of length Δ is negligible. Thus, the discretized point process can be modeled by a Bernoulli process with success probability $\lambda_k := \lambda(k\Delta|H_k)\Delta$, for $1 \leq k \leq K$, where $K := T/\Delta$ and is assumed to be an integer with no loss of generality. Note that λ_k forms the CIF of the discretized process, which we refer to as CIF hereafter for brevity. Let $n_k \in \{0, 1\}$ be the number of spikes in bin k, for $0 \leq k \leq K$. Our objective is to estimate the Power Spectral Density (PSD) of the CIF from the observed spike train $\{n_k\}_{k=1}^{K}$, under the assumption that the CIF is a second-order stationary process.

More generally, we consider an ensemble of L neurons or L trials from a single neuron driven by the same CIF, and denote the observed spike trains by $\mathcal{D} := \left\{n_k^{(l)}\right\}_{k=1,l=1}^{K,L}$. When considering an ensemble of L neurons, this setting may only be valid for neuronal recordings from a small area of cortex using multi-electrode arrays (See, for example [104]), and when considering L trials from the same neuron, it is assumed that all trials pertain to the same stimulus (See, for example [105]). We model the CIF using a zero-mean second-order stationary random process, $\{x_k\}_{k=1}^K$, which by virtue of the Spectral Representation Theorem [12] admits a Cramér representation [69, p. 150] of the form:

$$x_k = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{i2\pi fk} dz(f), \tag{3.2}$$

where $dz(f)$ is a complex-valued orthogonal increment process (i.e., $z(f)$ and $z(f')$ are uncorrelated for $f \neq f'$) and the PSD, $S(f)$ of the process is defined as: $S(f)df = \mathbb{E}[|dz(f)|^2]$. Finally, we use a linear link for the CIF so that the model can be summarized as

$$\lambda_k = \mu + x_k, \quad n_k^{(l)} \sim \mathsf{Bernoulli}(\lambda_k), \tag{3.3}$$

for $1 \leq k \leq K$ and $1 \leq l \leq L$, where μ is the baseline spiking probability. The choice of the linear link, as opposed to more common links such as the logistic function [106], is for the sake of simplicity of the auxiliary data generation process that will be described in Section 3.2.1. Acknowledging the non-linearity of the model Eq. (3.3) and the availability of only a *finite* number of samples, we consider a piecewise continuous approximation to the PSD, i.e, $dz(f)$ is constant over the intervals $[\frac{m-1}{2N}, \frac{m}{2N})$, for large enough N, for $m = 1, 2, \cdots, N$. This enables us to express

$dz(f) = (a_m + ib_m)df$ for $f \in [\frac{m-1}{2N}, \frac{m}{2N})$, where a_m and b_m are random variables for $m = 1, 2, \cdots, N$ [103]. Invoking the conjugate symmetry of $dz(f)$ for real valued $\{x_k\}_{k=1}^K$, Eq. (3.2) can be written as

$$x_k = \sum_{m=1}^{N} \frac{2}{N}\left[a_m \cos\frac{\pi(m-1)}{N} - b_m \sin\frac{\pi(m-1)}{N}\right], \tag{3.4}$$

with a PSD of $S(f) = \frac{1}{N}\mathbb{E}[a_m^2 + b_m^2]$ for $f \in [\frac{m-1}{2N}, \frac{m}{2N})$.

Denoting $\mathbf{x} := [x_1, x_2, \cdots, x_K]^\top$ and $\mathbf{z} := [a_1, a_2, b_2, \cdots, a_N, b_N]^\top$ and defining $\mathbf{A}$ as

$$\mathbf{A} := \frac{2}{N}\begin{bmatrix} 1 & \cos\frac{\pi}{N} & -\sin\frac{\pi}{N} & \dots & \cos\frac{(N-1)\pi}{N} & -\sin\frac{(N-1)\pi}{N} \\ 1 & \cos\frac{2\pi}{N} & -\sin\frac{2\pi}{N} & \dots & \cos\frac{2(N-1)\pi}{N} & -\sin\frac{2(N-1)\pi}{N} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos\frac{K\pi}{N} & -\sin\frac{K\pi}{N} & \dots & \cos\frac{K(N-1)\pi}{N} & -\sin\frac{K(N-1)\pi}{N} \end{bmatrix},$$

one can write Eq. (3.4) in the vector form as $\mathbf{x} = \mathbf{A}\mathbf{z}$. By the orthogonality of the increment process $dz(f)$, z_i's are uncorrelated. We further assume that z_i's, for $i = 1, 2, \cdots, 2N-1$, are independent and each z_i follows a truncated zero-mean Gaussian distribution, with a density

$$\frac{f_i(z_i)\mathbb{1}[-\tilde{\mu} \leq z_i \leq \tilde{\mu}]}{\int f_i(\xi)\mathbb{1}[-\tilde{\mu} \leq \xi \leq \tilde{\mu}]d\xi}, \tag{3.5}$$

where $f_i(\cdot)$ is the Gaussian density $\mathcal{N}(0, {\sigma_i}^2)$, $\mathbb{1}$ is the indicator function, and $\tilde{\mu} := \frac{N}{2(2N-1)}\mu$. While a Gaussian assumption (instead of truncated Gaussian) is more common in application, our choice of truncation ensures that $\lambda_k = \mu + x_k$ is a feasible spiking probability: Given that each entry of $\mathbf{A}$ is bounded in absolute value by $\frac{2}{N}$, this ensures that each x_k is restricted to $[-\mu, \mu]$ for $\mu > 0$. In addition, by selecting $\mu \leq 1/2$, one can ensure that $0 \leq \lambda_k \leq 1$, for all $1 \leq k \leq K$.

3.2 The Point Process Multitaper Method

The MTM is an extension of tapered PSD estimation, where the spectral estimate is computed by averaging several tapered PSD estimates corresponding to orthogonal tapers with optimal spectral leakage properties [12]. The set of tapers from the *Discrete Prolate Spheroidal Sequences (dpss)* [73] provides excellent control over the bias-variance trade-off [11, 74, 92].

Let $v_k^{(j)}$ be the k^{th} sample of the j^{th} dpss sequence, for a given design bandwidth W, $k = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, J$ such that $J < \lfloor 2KW \rfloor - 1$. Given the time-series data $\{x_k\}_{k=1}^{K}$, the j^{th} eigen-spectrum is given by:

$$\widehat{S}^{(j)}(f) := \left| \sum_{k=1}^{K} e^{-i2\pi f k} v_k^{(j)} x_k \right|^2 \text{ for } j = 1, 2, \cdots J \tag{3.6}$$

from which the MTM PSD estimate can be computed as:

$$\widehat{S}^{(\mathsf{mtm})}(f) := \frac{1}{J} \sum_{j=1}^{J} \widehat{S}^{(j)}(f). \tag{3.7}$$

Due to the non-linear nature of the model Eq. (3.3), forming the MTM PSD estimate based on spiking data becomes non-trivial. In what follows, we indeed address this issue by devising a novel variant of MTM.

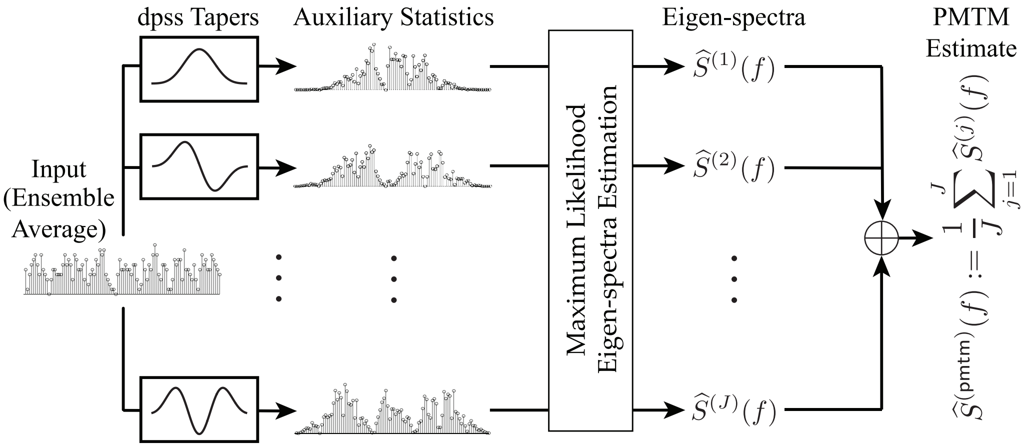


Figure 3.1: Schematic depiction of the proposed method. Stem plots show the ensemble average of the underlying spike trains.

3.2.1 Generating Auxiliary Spiking Statistics

Given that x_k is not directly observable, we instead modify the observed spike trains as if they were generated by a CIF comprising $v_k^{(j)} x_k$ (instead of x_k). For non-negative $v_k^{(j)}$ (e.g., for $j = 1$), generating such modified spike trains is usually carried out using the thinning method [107, 108]. The basic idea of the thinning method is to retain the original spikes with a probability determined by the ratio of the target and original CIFs, and thus to obtain a spike train corresponding to the tapered CIF, $v_k^{(j)}\lambda_k$. Given that the dpss tapers take negative values (for $j > 1$), the thinning method is not readily applicable and its naïve application results in negative-valued spiking probabilities.

To resolve this issue, we leverage the virtue that the spike train is generated according to a Bernoulli process and therefore its complement given by $\check{n}_k^{(l)} := 1 - n_k^{(l)}$ has a CIF given by $\check{\lambda}_k := 1 - \lambda_k = 1 - \mu - x_k$. Suppose that each taper $v_k^{(j)}$

is normalized by its maximum absolute value such that $|v_k^{(j)}| \leq 1$ for all k. Let $\zeta_k^{(j)} \in \{0,1\}$ be independently drawn Bernoulli variables with $p\big[\zeta_k^{(j)} = 1\big] = \big|v_k^{(j)}\big|$, for $j = 1, 2, \cdots, J$ and $k = 1, 2, \cdots, K$. Consider the following *thinned* spike train constructed from $n_k^{(l)}$ and its complement $\check{n}_k^{(l)}$:

$$n_k^{(l,j)} = \zeta_k^{(j)} \left(n_k^{(l)} \mathbb{1}[v_k^{(j)} \geq 0] + \check{n}_k^{(l)} \mathbb{1}[v_k^{(j)} < 0] \right). \tag{3.8}$$

In words, the sequence $n_k^{(l,j)}$ contains the thinned version of $n_k^{(l)}$, wherever $v_k^{(j)}$ is non-negative, and the thinned version of $\check{n}_k^{(l)}$, wherever $v_k^{(j)}$ is negative. As such, $n_k^{(l,j)} \in \{0,1\}$ for all k by construction. This makes $n_k^{(l,j)}$ a feasible spike train in the point process framework. Furthermore, noting that $\check{n}_k^{(l)}$ has a CIF of $\check{\lambda}_k = 1 - \lambda_k$, from Eq. (3.8) the CIF of $n_k^{(l,j)}$ can be expressed as:

$$\begin{aligned}
\lambda_k^{(j)} &:= (\mu + x_k)\big|v_k^{(j)}\big|\mathbb{1}[v_k^{(j)} \geq 0] + (1 - \mu - x_k)\big|v_k^{(j)}\big|\mathbb{1}[v_k^{(j)} < 0] \\
&\overset{(a)}{=} (\mu + x_k) v_k^{(j)} \mathbb{1}[v_k^{(j)} \geq 0] - (1 - \mu - x_k) v_k^{(j)} \mathbb{1}[v_k^{(j)} < 0] \\
&\overset{(b)}{=} \mu_k^{(j)} + v_k^{(j)} x_k,
\end{aligned} \tag{3.9}$$

where $\mu_k^{(j)} = \mu v_k^{(j)} \mathbb{1}[v_k^{(j)} \geq 0] - (1-\mu) v_k^{(j)} \mathbb{1}[v_k^{(j)} < 0]$, for $j = 1, 2, \cdots, J$, (a) follows from the definition of absolute value, and the equality (b) follows from rearranging the terms. The expression of $\lambda_k^{(j)}$ in Eq. (3.9) indeed contains the desired tapered version of x_k, which is required for multitaper spectral estimation.

By considering multiple independent realizations of $\{\zeta_k^{(j)}\}_{j=1,K=1}^{J,K}$, one can generate multiple realizations of $n_k^{(l,j)}$, and take their ensemble average as a smoothed sequence of auxiliary statistics to estimate the spectral representation of the tapered process $v_k^{(j)} x_k$. It is not difficult to see that the ensemble average converges in probability to [109]:

$$v_k^{(j)} n_k^{(l)} \mathbb{1}[v_k^{(j)} \geq 0] - v_k^{(j)} \check{n}_k^{(l)} \mathbb{1}[v_k^{(j)} < 0], \tag{3.10}$$

which can be directly computed from the original spiking activity $n_k^{(l)}$. Thus, for the sake of robustness we use this limit as the auxiliary spiking statistic hereafter, and refer to it also as $n_k^{(l,j)}$, for notational brevity. This choice is motivated by the direct-averaging method common in adaptive filtering [110]. Given that each taper $v_k^{(j)}$ was initially normalized by its maximum absolute value, the estimated eigen-spectra need to be accordingly rescaled. It is noteworthy that while in principle this procedure can be extended to more general link functions, the generation of the corresponding auxiliary statistics may be more intricate. Fig. 3.1 provides a visual summary of our proposed framework. The time-bandwidth product and the number of tapers are chosen following guidelines from the MTM literature [11, 74].

3.2.2 Maximum Likelihood Estimation of the Eigen-spectra

Once the auxiliary spiking statistics $\mathcal{D}^{(j)} = \{n_k^{(l,j)}\}_{k=1,l=1}^{K,L}$, $j = 1, 2, \cdots, J$ are available, the eigen-spectra need to be estimated to construct the PSD. Given the modeling framework of Section 3.1, estimation of the j^{th} eigen-spectrum reduces to estimating the parameters $\boldsymbol{\theta}^{(j)} := [\sigma_1^{(j)2}, \sigma_2^{(j)2}, \cdots, \sigma_{2N-1}^{(j)2}]^\top$, where $\sigma_m^{(j)2}$ is the variance of the random variable $z_i^{(j)}$ corresponding to the j^{th} eigen-spectra, $i =$

$1, 2, \cdots, 2N-1$. The ML estimate of the parameter $\boldsymbol{\theta}^{(j)}$ is given by:

$$\widehat{\boldsymbol{\theta}}^{(j)}_{\mathsf{ML}} = \underset{\boldsymbol{\theta}^{(j)}}{\arg\max}\, P(\mathcal{D}^{(j)}|\boldsymbol{\theta}^{(j)}) \tag{3.11}$$

Note that expressing $P(\mathcal{D}^{(j)}|\boldsymbol{\theta}^{(j)})$ solely in terms of $\mathcal{D}^{(j)}$, i.e., eliminating $\mathbf{z}^{(j)} := [z_1^{(j)}, z_2^{(j)}, \cdots, z_{2N-1}^{(j)}]^\top$, introduces computational intricacies, which we avoid by using the Expectation-Maximization (EM) algorithm [78] as our solution method. In what follows we drop the superscript j for the sake of clarity, as the same procedure will be used for estimating each eigen-spectrum. If $\mathbf{z}$ is known, the *complete* data log-likelihood of the observations $\mathcal{D}$ can be written as:

$$\begin{aligned}\log L(\boldsymbol{\theta}|\mathbf{z},\mathcal{D}) = &\sum_{l=1}^{L}\sum_{k=1}^{K}\Big[n_k^{(l)} \log \frac{\mu_k + (\mathbf{A}\mathbf{z})_k}{1-\big(\mu_k+(\mathbf{A}\mathbf{z})_k\big)} + \log\big(1-\big(\mu_k + (\mathbf{A}\mathbf{z})_k\big)\big)\Big] \\ &- \sum_{m=1}^{2N-1}\left(\log\int_{-\tilde{\mu}}^{\tilde{\mu}} f_m(\xi)d\xi + \frac{z_m^2}{2\sigma_m^2} + \frac{1}{2}\log\sigma_m^2\right) + C,\end{aligned} \tag{3.12}$$

which could be efficiently maximized to estimate $\boldsymbol{\theta}$ (terms independent of $\boldsymbol{\theta}$ are denoted by C). Note that the *linear* dependence of the complete data log-likelihood on $n_k^{(l)}$ makes the direct-averaging technique described in Section 3.2.1 plausible, as it corresponds to smoothing the data log-likelihood.

Before presenting the EM algorithm, we note that $\int_{-\tilde{\mu}}^{\tilde{\mu}} f_m(\xi)d\xi = 1-2\Phi(\gamma_m) \approx 1$ for moderate values of $\gamma_m := \frac{\tilde{\mu}}{\sigma_m}$ and small enough σ_m^2, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus, we drop this term henceforth to avoid unnecessary complexity. One may choose to work with this term included, at the expense of additional computational costs. At the i^{th} iteration, we have:

E-step

Given $\boldsymbol{\theta}^{[i]}$, the Q-function is given by

$$\mathbf{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]}) = -\sum_{m=1}^{2N-1}\left(\frac{1}{2}\log\sigma_m^2 + \frac{1}{2\sigma_m^2}\mathbb{E}[z_m^2|\mathcal{D},\boldsymbol{\theta}^{[i]}]\right) + C', \tag{3.13}$$

which requires $f_{\mathbf{z}|\mathcal{D},\boldsymbol{\theta}^{[i]}}$ or samples from it, and can thus be computationally demanding to compute (terms independent of $\boldsymbol{\theta}$ are denoted by C'). Instead, we use the unimodality of the density and approximate it by a multivariate Gaussian density $\mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}^{[i]}}, \boldsymbol{\Sigma}_{\mathbf{z}^{[i]}})$ [66, 97]. By invoking the fact that the mode and mean of a multivariate Gaussian density coincide and the Hessian of its natural logarithm is equal to $-\left(\boldsymbol{\Sigma}_{\mathbf{z}^{[i]}}\right)^{-1}$, we get:

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{z}^{[i]}} = \underset{\mathbf{z}\in D}{\arg\max}\ \sum_{l=1}^{L}\sum_{k=1}^{K}\left[n_k^{(l)}\log\frac{\mu_k + (\mathbf{Az})_k}{1-(\mu_k+(\mathbf{Az})_k)} + \log\left(1-(\mu_k+(\mathbf{Az})_k)\right)\right] \\ -\sum_{m=1}^{2N-1}\frac{z_m^2}{2\sigma_m^2}\ ,\end{aligned} \tag{3.14}$$

and $\boldsymbol{\Sigma}_{\mathbf{z}^{[i]}}$ is given by the Hessian of the log-likelihood in Eq. (3.12) evaluated at $\boldsymbol{\mu}_{\mathbf{z}^{[i]}}$. The maximization problem Eq. (3.14) is concave over

$$D = \{\mathbf{z}\in\mathbb{R}^{2N-1} : 0 \le \mu_k + (\mathbf{Az})_k \le 1, k = 1, 2, \cdots, K\}$$

and the Hessian is negative definite, so Newton-type methods for bound-constrained optimization can be used to compute $\boldsymbol{\mu}_{\mathbf{z}^{[i]}}$ efficiently. We use a line-search method

Algorithm 3: The Point Process Multitaper Method

```
Input: Ensemble of neuronal spiking data, $\{n_k^{(l)}\}_{k=1,l=1}^{K,L}$ for $k = 1, 2, \cdots, K$; Design bandwidth, $W$, such that $\alpha := KW \geq 1$; Number of tapers, $J$.
Generate $J < \lfloor 2\alpha \rfloor$ dpss corresponding to data length $K$ and half time-bandwidth product $\alpha$
for $j = 1$ to $J$ do
    Generate $\{n_k^{(l,j)}\}_{k=1}^{K}$, for $l = 1, 2, \cdots, L$
    Compute $\widehat{S}^{(j)}(f)$ using ML estimation
end for
Output: PMTM estimate, $\widehat{S}^{(\mathsf{pmtm})}(f) = \frac{1}{J}\sum_{j=1}^{J} \widehat{S}^{(j)}(f)$
```

[111], which generates a sequence of iterates by setting $\boldsymbol{\mu}_{\mathbf{z}^{[i]}}^{[r+1]} = \boldsymbol{\mu}_{\mathbf{z}^{[i]}}^{[r]} + \alpha^{[r]}\mathbf{d}^{[r]}$, where $\boldsymbol{\mu}_{\mathbf{z}^{[i]}}^{[r+1]}$ is a feasible approximation to the solution, $\alpha^{[r]}$ is the step-size and $\mathbf{d}^{[r]}$ is the Newton's step for that iteration. Then, $\boldsymbol{\Sigma}_{\mathbf{z}^{[i]}}$ can be computed by evaluating the Hessian of Eq. (3.12) at $\mathbf{z} = \boldsymbol{\mu}_{\mathbf{z}^{[i]}}$, which allows $\mathbb{E}[z_m^2|\mathcal{D}, \boldsymbol{\theta}^{[i]}]$ to be calculated as $\left((\boldsymbol{\mu}_{\mathbf{z}^{[i]}})_m\right)^2 + \left(\boldsymbol{\Sigma}_{\mathbf{z}}^{[i]}\right)_{m,m}$.

M-step

The parameter vector $\boldsymbol{\theta}^{[i+1]}$ is updated by maximizing the expectation in Eq. (3.13). Given that $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{[i]})$ is concave over the positive orthant, its unique maximizer is given by $\widehat{\boldsymbol{\theta}}_m^{[i+1]} = \mathbb{E}[z_m^2|\mathcal{D}, \boldsymbol{\theta}^{[i]}]$.

Note that we have assumed $\mu_k^{(j)}$'s to be known. Since it is not the case for most practical purposes, we first estimate μ as $\hat{\mu} = \frac{1}{LK}\sum_{l,k=1}^{L,K} n_k^{(l)}$ and compute $\mu_k^{(j)}$ in Eq. (3.9) using $\hat{\mu}$. We terminate the EM algorithm after a fixed large number of iterations or until some convergence criterion is met. A similar stopping rule for the maximization problem inside each EM step is used. We initialize $\boldsymbol{\theta}^{[0]}$ as an arbitrary vector in the positive orthant. Following the termination of the EM algorithm, the eigen-spectra are calculated as $\widehat{S}(0) = \widehat{\sigma}_1^2$ and $\widehat{S}(f_m) = \widehat{\sigma}_{2m}^2 + \widehat{\sigma}_{2m+1}^2$ for $f_m = \frac{m}{2N}$ and $m = 1, 2, \cdots N-1$. Finally, the PMTM estimate can be computed using Eq. (3.7). Algorithm 3 summarizes the proposed PMTM procedure.

3.3 Simulation Results

We simulate x_k as an AR(4) process given by

$$x_k = 0.4152x_{k-1} - 0.0922x_{k-2} + 0.4170x_{k-3} - 0.8852x_{k-4} + 0.025\epsilon_k,$$

where ϵ_k is zero-mean i.i.d. Gaussian noise with unit variance. We compute the CIF as $\lambda_k = \mu + x_k$ for $\mu = 0.12$ (truncated to $[0, 1]$, if necessary), to generate the binary spiking activity for $K = 512$ samples. A snapshot of one realization of this AR process and the raster plot of $L = 10$ spike trains are depicted in Fig. 3.2*A* and *B*, respectively.

We apply PMTM to this simulated data and benchmark it against two existing meth-

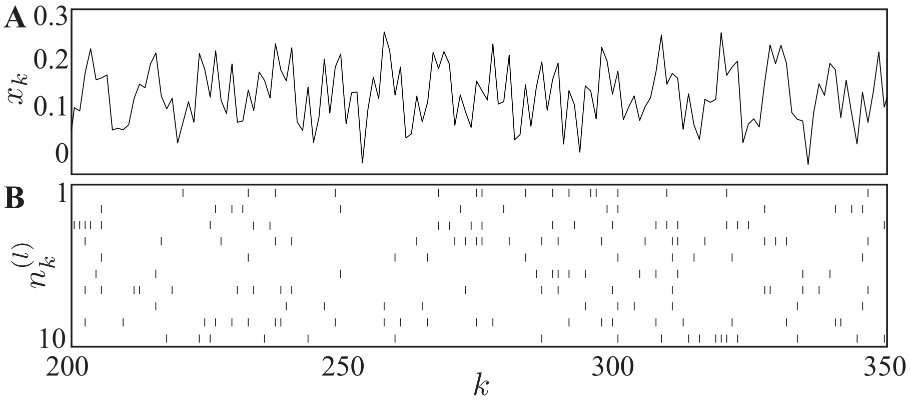


Figure 3.2: (*A*) A snapshot of the simulated AR process for $200 \leq k \leq 350$. (*B*) Raster plot of the corresponding neuronal ensemble activity.

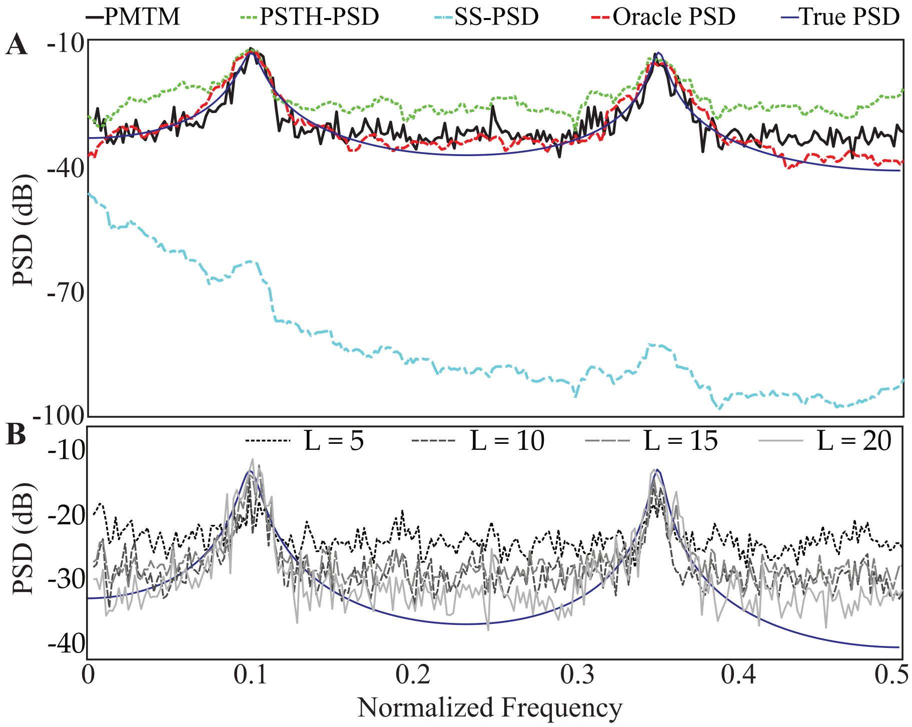


Figure 3.3: Comparison of the PSD estimates. (A) PMTM, PSTH-PSD, SS-PSD, Oracle PSD (all using $\alpha = 5, J = 8$), and the true PSD. (B) PMTM estimates for $L = 5,\ 10,\ 15$, and 20.

ods: (1) PSTH-PSD, where the PSD is computed by forming the MTM estimate of the ensemble peristimulus time histogram (PSTH), i.e., the average spike trains, and (2) SS-PSD, where x_k is first estimated using a state-space model $x_k = x_{k-1} + w_k$, followed by forming its MTM PSD estimate [66].

Fig. 3.3A shows the PMTM (black), PSTH-PSD (green), SS-PSD (aqua) and the true PSD (blue) for the realization shown in Fig. 3.2 in log-scale. For comparison purposes, we have also included the MTM PSD estimate of x_k, assuming that an oracle has access to it (Oracle PSD, in red). We have used the first 8 dpss tapers corresponding to $\alpha = 5$. As it can be observed from Fig. 3.3A, PSTH-PSD suffers from high bias, though exhibiting reduced variability, and the spectral peaks are difficult to distinguish from the background. On the other hand, SS-PSD suffers from model mismatch, as it over-smooths the CIF due to the usage of a state-space model, and as a result the spectral peaks are nearly absent in the estimate. The PMTM estimate, however, closely follows the true PSD by reducing spectral leakage and producing a nearly unbiased estimate on par with the Oracle PSD estimate, though it exhibits some variability. In order to quantify these comparisons, we computed a normalized measure of MSE by averaging the squared-error of the PSD normalized by the true PSD values in the log-scale. The normalized MSE values (± 2 STD) corresponding to 10 different AR process realizations and 5 different spike-train ensemble realizations are given in Table 3.1, which corroborates our foregoing qualitative comparison. It is noteworthy that the improved performance of the PMTM method comes with the cost of higher computational complexity as compared to the PSTH-PSD and SS-PSD methods. To ease reproducibility, we have deposited a MATLAB implementation of PMTM on the open source repository Github [112], which fully regenerates Fig. 3.3*A*.

PMTM	SS-PSD	PSTH-PSD
0.4733 ± 0.0072	7.7772 ± 2.0641	$0.8164 \pm 7.9592 \times 10^{-8}$

Table 3.1: Normalized MSE Comparisons

Finally, Fig. 3.3*B* examines the improvement of the PMTM estimates with respect to the ensemble size, for $L = 5, 10, 15$ and 20. As L increases, the PSD estimates improve, but with a seemingly saturating effect for $L \geq 10$.

3.4 Application to Experimentally Recorded Data

We next demonstrate the utility of PMTM in application to experimentally recorded data. The data consist of both spike trains and local field potential (LFP) recorded by a multi-electrode array from a human subject undergoing general anesthesia (See Lewis et al. [104] for details). It is known that the LFP signal–capturing the synchronous activity of a large population of neurons–is a salient covariate of rhythmic neuronal spiking under anesthesia [103, 104]. As such, it is expected that the PSD of the latent process that drives spiking activity resembles that of LFP. Fig. 3.4 shows the raster plot of 26 neurons considered for analysis, as well as the corresponding PSTH and LFP signals.

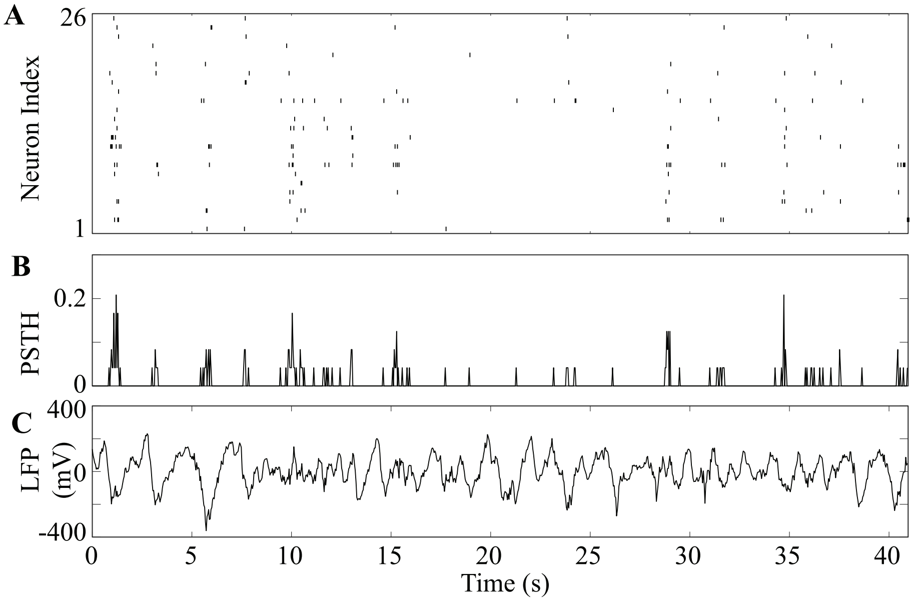


Figure 3.4: Experimentally recorded data under general anesthesia. (A) Raster plot of 26 neurons, (B) the corresponding PSTH, and (C) LFP, for a window of length 40.92 s.

To reduce computational complexity, the spike trains and LFP signal were down-sampled to 25 Hz, from their original sampling rates of 1 KHz and 250Hz, respectively. We considered a window of $K = 1024$ samples (40.92 s) for analysis. We used $N = 512$ frequency bins to construct the matrix $\mathbf{A}$. Given that the activity is known to be dominated by sub-hertz frequency components [104], we restricted PSD estimation to $[0, 3]$ Hz (i.e, the first 123 frequency bins). We used the first 4 dpss tapers corresponding to $\alpha = 3$. Fig. 3.5 shows the PSD estimates obtained by the PSTH-PSD, SS-PSD and PMTM methods, as well as the multitaper PSD estimate of the LFP signal, in linear scale and normalized by the magnitude of

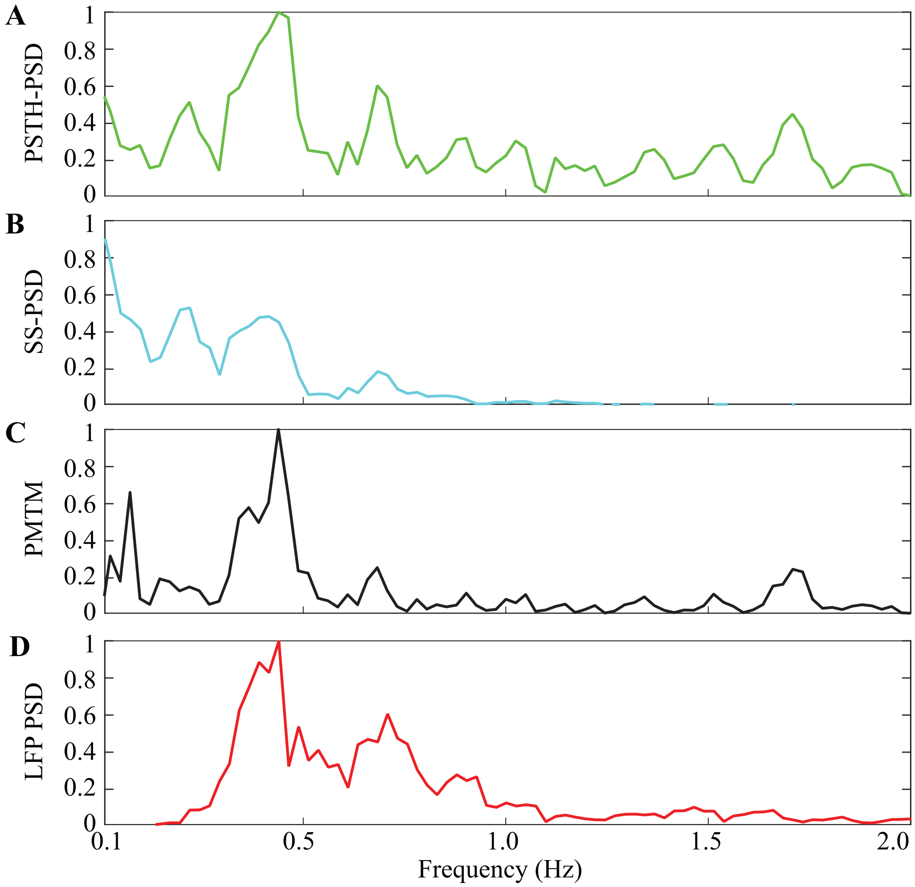


Figure 3.5: PSD estimates from experimentally recorded data under general anesthesia. (A) PSTH-PSD, (B) SS-PSD, (C), PMTM, and (D) multitaper PSD of the LFP signal. The PSDs are presented in linear scale and normalized to the magnitude of their respective largest peak. PSD estimates above 2 Hz are negligible and are cropped for visual convenience.

the largest peak, for ease of comparison. The PSTH-PSD estimate (Fig. 3.5A) is highly variable and exhibits multiple spurious peaks above 1 Hz. The SS-PSD estimate (Fig. 3.5B) overly concentrates the PSD within low frequencies due to the underlying temporal smoothing. The PMTM estimate (Fig. 3.5C), however, successfully suppresses the spectral energy above 1 Hz, and as expected, captures the main spectral features of the LFP signal (Fig. 3.5D).

3.5 Concluding Remarks

Spectral estimation of continuous time-series is a well-established domain, as hallmarked by the multitaper method known for its favorable control over the bias-variance trade-off. Computing the spectral representation of the neural covariates that underlie spiking activity, however, sets forth various challenges due to the intrinsic non-linearities involved. In this chapter, we addressed this problem by proposing a multitaper method specifically tailored for binary spiking data, which we refer to as PMTM. We compared the performance of PMTM to that of two existing techniques using simulated and experimentally recorded data, which revealed significant gains in terms of estimation accuracy. The PMTM can be extended to a wide variety of binary data, such as rainfall and earthquake data, to extract spectral representations of the underlying latent processes.

Part II

A Cortically-distributed Encoding Model of Speech Processing for M/EEG Analysis

Chapter 4: Neuro-Current Response Functions: an integrated approach to MEG source analysis for continuous stimuli paradigm

The human brain routinely processes complex information as it unfolds over time, for example, when processing natural speech, information from lower levels has to be continuously processed to build higher level representations, from the acoustic signal to phonemes to words to sentence meaning. Quantitative characterization of the neural dynamics underlying such sensory processing is not only important in understanding brain function, but it is also crucial in the design of neural prostheses and brain-machine interface technologies.

In modeling neural activity at the meso-scale using neuroimaging modalities such as electroencephalography (EEG) and magnetoencephalography (MEG), experimental evidence suggests that linear encoding models can be beneficial in predicting the key features of sensory processing; examples include encoding models of visual and auditory stimuli [3, 17, 18, 19].

Arguably the earliest and most widely used technique to construct neural encoding models is the ‘reverse correlation’ technique, in which neural responses time-locked to multiple repetitions of simple stimuli (such as acoustic tones and visual gratings) are averaged, weighted by the instantaneous value of the preceding stimulus, to form the so-called evoked response function. Originally devised to study the tuning properties of sensory neurons [113, 114, 115], it was later incorporated into MEG/EEG analysis. In probing the neural response to more sophisticated stimuli such as continuous speech and video, the goal is to understand the encoding of the continuous stimuli as a whole, which is composed of both low level (e.g., acoustics) and high level (e.g., semantics) features which are bound together and distributed across time [116, 117].

To address this issue, techniques from linear systems theory have been successfully utilized to capture neural encoding using MEG/EEG under the continuous stimuli paradigm. In this setting, the encoding model takes the form of a linear filter which predicts the MEG/EEG response from the features of the stimulus. For example, it has been shown that the acoustic envelope of speech is a suitable predictor of the EEG response [118]. These filters, or impulse response functions, play a crucial role in characterizing the temporal structure of auditory information processing in the brain, and are often referred to as Temporal Response Function (TRF) [19, 20, 21]. For instance, in a competing-speaker environment in the presence of two speech streams, it has been observed that the TRF extracted from MEG response to the acoustic power consists of an early component at around 50 ms representing the acoustic power of the speech mixture, while a later peak at around 100 ms preferentially encodes the acoustic power of the attended speech stream [4, 119]. Building up on evidence from ferret electrophysiology [120] and human electrocorticography (ECoG) [121], more recent studies have expanded the TRF framework beyond the acoustic level to account for phoneme-level processing [116], lexical processing [117] and semantic processing [122].

Thanks to the grounding of the TRF model in linear system theory, several techniques from the system identification literature have been utilized for TRF estimation, such as the *normalized reverse correlation* [123], *ridge regression* [124], *boosting* [6], and *SPARLS* [125], some of which are available as software packages [126, 127, 128, 129]. While these methods have facilitated the characterization of the functional roles of various TRF components in sensory and cognitive processing of auditory stimuli, they predominantly aim at estimating TRFs over the MEG/EEG sensor space. While recent studies, using electrophysiology in animal models and ECoG in humans, have provided new insights into the cortical origins of auditory processing [see, for example, 120, 121, 130, 131], they do not account for the whole-brain distribution of the underlying sources due to their limited spatial range. As such, the whole-brain cortical origins of the TRF components are not well studied.

To address this issue using neuroimaging, current dipole fitting methods have been utilized to map the sensor space distribution of the estimated TRF components onto cortical sources [4, 18]. Given that the processing of sophisticated stimuli such as speech is known to be facilitated by a widely distributed cortical network, single dipole sources are unlikely to capture the underlying cortical dynamics. More recent results have used the minimum norm estimate (MNE) source localization technique to first map the MEG activity onto the cortical mantle, followed by estimating a TRF for each of the resulting cortical sources [22]. While these methods have shed new light on the cortical origins of the TRF, they have several limitations that need to be addressed. First, the ill-posed nature of the MEG/EEG source localization problem under distributed source models results in cortical estimates with low spatial resolution [132, 133]. Given the recent and ongoing advances in MEG/EEG source localization towards improving the spatial resolution of the inverse solutions [see, for example, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149], it is tempting to simply use more advanced source localization techniques followed by fitting TRFs to the resulting cortical sources. However, these techniques are typically developed for the event-related potential (ERP) paradigm [150, 151, 152] and leverage specific prior knowledge on the spatiotemporal orgranization of the underlying sources. While these assumptions bias the solution towards *source* estimates with high spatiotemporal resolution under specific repetition-based experimental settings, they do not account for the key structural properties of the underlying *neural processes* that extract information from continuous sensory stimuli. These key properties include the smoothness and/or sparsity of the response functions in the lag domain and their spatial correlation over the cortex, which may not be captured by merely enforcing spatiotemporal priors over the source domain.

Second, the single-trial nature of experiments involving continuous auditory stimuli, does not allow to leverage the time-averaging across multiple trials common in source localization of evoked responses from MEG/EEG. Third, the two-stage procedures of first fitting TRFs over the sensor space followed by localizing the peaks using dipole fitting, or first finding source estimates over the cortex followed by fitting TRFs to cortical sources, results in so-called bias propagation: the inherent biases arising from the estimation procedure in the first stage propagate to the second stage, often destructively so, and limit, sometimes severely, the statistical interpretability of the resulting cortical TRFs (see Section 4.2.1, for example). Finally, high resolution inverse solutions require precise forward models that are constructed based on high resolution MR scans with accurate sensor registration, which may not be readily available.

In order to address these limitations, here we provide a methodology for *direct* estimation of cortical TRFs from MEG observations, taking into account their spatiotemporal structure. We refer to these cortical TRFs as neuro-current response functions (NCRFs). We construct a unified estimation framework by integrating the linear encoding and distributed forward source models, in which the NCRFs linearly process different features of the continuous stimulus and result in the observed neural responses at the sensor level. We cast the inverse problem of estimating the NCRFs as a Bayesian optimization problem where the likelihood of the recorded MEG response is directly maximized over the NCRFs, thus eliminating the need for the aforementioned two-stage procedures.

In addition, to address the lack of accurate cortical surface patch statistics in the head model due to unavailability of MR scans, the NCRFs are extended to free-orientation dipoles by tripling them at each dipole location to account for vector valued current moments in 3D space. To guard against over-fitting and ensure robust recovery of such 3D NCRFs, we design a regularizer that captures the spatial sparsity and temporal smoothness of the NCRFs (e.g., minimizing the number of peaks or troughs) while eliminating any dependency on the choice of coordinate system for representing the vector valued dipole currents.

While the resulting optimization problem turns out to be non-convex, we provide an efficient coordinate-descent algorithm that leverages recent advances in evidence maximization to obtain the solution in a fast and efficient manner.

We empirically evaluate the performance of the proposed NCRF estimation framework using a simulation study mimicking continuous auditory processing, which reveals that the proposed method is not only capable of identifying active sources with better spatial resolution compared to existing methods, but can also infer the orientation of the dipoles as well as the time course of the response functions accurately. Lastly, we demonstrate the utility of estimation framework by analyzing experimentally recorded MEG data from young adult individuals listening to speech for NCRFs at different hierarchical levels of speech processing. A data set, initially recorded by [153] and lacking individual MR scans, was analyzed previously by [22] for source response functions using two-stage MNE followed by *boosting*-based TRF estimation. Our estimated NCRFs not only corroborate existing findings, but they are also readily interpretable in a meaningful fashion without any recourse to post-hoc processing (i.e. hierarchal clustering, sparse principal component analysis etc.) necessary for the previous study, thanks to improved spatial localization. In summary, our method successfully delineates the distinct cortical distribution of the underlying neural processes at high spatiotemporal resolution, providing new insights into the cortical dynamics of speech processing.

4.1 Theory and Methods

We develop our theory and methods for a canonical MEG auditory experiment in which the subject is listening to a single speech stream. Our goal is to determine how the different features of the speech stream are processed at different cortical stages and evoke specific neural responses that give rise to the recorded MEG data. For clarity of description and algorithm development, we first consider a single-trial experiment, and take the momentary acoustic power of the speech stream, i.e., the speech envelope, as the feature of interest. We will discuss below the more general scenarios including multiple trials, multiple speech stimuli, and multiple, and possibly competing, features reflecting different levels of cognitive processing.

4.1.1 Preliminaries and Notation

Let e_t, $1 \leq t \leq T$ denote the speech envelope at discrete time index t for a duration of T samples taken at a sampling frequency of f_s. We consider a distributed cortical source model composed of M dipole sources $\mathbf{d}_m = (\mathbf{r}_m, \mathbf{j}_{m,t})$, $1 \leq m \leq M$, where $\mathbf{r}_m \in \mathbb{R}^3$ denotes the right-anterior-superior (RAS) coordinates of the m^{th} dipole and $\mathbf{j}_{m,t} := [j_{m,t,R}, j_{m,t,A}, j_{m,t,S}]^\top \in \mathbb{R}^3$ denotes the dipole current vector at time t in the same coordinate system. The dipole locations can be obtained by standard tessellation of the 3D structural MR images of the cortex and assigning dipoles to the corresponding vertices [133, 154]. Furthermore, the MR images can also be utilized to approximate the orientation of the current vector, assuming current flow is orthogonal to the cortical surface and replacing the dipole current vector by a scalar value [155, 156]. However, this approach requires precise knowledge of the cortical geometry [137], and still might not result in ideal approximation of the cortical current orientations [157]. So, it is often desirable to retain the vectorial nature of current dipoles, even though the resulting process is more complex.

Next, we assume that these current dipoles are in part stimulus-driven, i.e., each component of the current dipole relies on contributions from the preceding stimulus:

$$j_{m,t,i} = f_i(e_t, e_{t-1}, \cdots, e_1) + v_{m,t,i} \tag{4.1}$$

where the placeholder i takes the values of one of the coordinate axes, $\{R, A, S\}$, f_i is a generic function, and $\mathbf{v}_{m,t} := [v_{m,t,R}, v_{m,t,A}, v_{m,t,S}]^\top$ accounts for the stimulus-independent background activity. Following the common modeling approaches in this context [3, 18, 115, 118], we take f_i to represent a linear finite impulse response (FIR) filter of length L:

$$f_i(e_t, e_{t-1}, \cdots, e_1) = \sum_{l=0}^{L-1} \tau_{m,i,l} e_{t-l} = (\boldsymbol{\tau}_{m,i})^\top \mathbf{e}_t, \quad i \in \{R, A, S\}, \tag{4.2}$$

where $\boldsymbol{\tau}_{m,i} := [\tau_{m,i,0}, \tau_{m,i,1} \cdots, \tau_{m,i,L-1}]^\top$ and $\mathbf{e}_t := [e_t, e_{t-1}, \cdots, e_{t-L+1}]^\top$. Note that $\boldsymbol{\tau}_{m,i}$ can be thought of as a TRF corresponding to the activity of dipole source m along the coordinate axis determined by i. The length of the filter L is typically determined by a priori assumptions on the effective integration window of the underlying neural process. When stacked together, the 3D linear filters

$\boldsymbol{\tau}_m := [\boldsymbol{\tau}_{m,R}, \boldsymbol{\tau}_{m,A}, \boldsymbol{\tau}_{m,S}] \in \mathbb{R}^{L\times 3}$ are vector-valued TRFs at each source m, capturing the linear processing of the stimuli at the cortical level. As such, we refer to these vector-valued filters as Neuro-Current Response Functions (NCRFs) henceforth. Intuitively speaking, the 3D vector $(\tau_{m,R,l}, \tau_{m,A,l}, \tau_{m,S,l})^\top$ is the vector-valued dipole activity at a lag of l/f_s second arising from a putative stimulus impulse at time 0.

Let $\mathbf{j}_t := [\mathbf{j}_{1,t}^\top, \mathbf{j}_{2,t}^\top, \cdots, \mathbf{j}_{M,t}^\top]^\top \in \mathbb{R}^{3M}$ be a vector containing all the current dipoles at time t, and $\mathbf{J} := [\mathbf{j}_1, \cdots, \mathbf{j}_T] \in \mathbb{R}^{3M\times T}$ be the matrix of current dipoles obtained by concatenating the instantaneous current dipoles across time $t = 1, 2, \cdots, T$. Similarly, let $\boldsymbol{V} \in \mathbb{R}^{3M\times T}$ denote the matrix of stimulus-independent background activity, $\boldsymbol{\Phi} := [\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \cdots, \boldsymbol{\tau}_M]^\top \in \mathbb{R}^{3M\times L}$ denote the matrix of NCRFs, and $\mathbf{S} := [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_T] \in \mathbb{R}^{L\times T}$ denote the matrix of features. Eq. (4.1) and Eq. (4.2) can then be compactly expressed as:

$$\mathbf{J} = \boldsymbol{\Phi}\mathbf{S} + \mathbf{V}. \tag{4.3}$$

As for the sensor space, we assume a conventional MEG setting with N sensors placed at different positions over the scalp, recording magnetic fields/gradients as a multidimensional time series. The MEG observation at the i^{th} sensor at time t is denoted by $y_{i,t}$, $1 \leq i \leq N$ and $t \in [1, \cdots, T]$. Let $\mathbf{Y} \in \mathbb{R}^{N\times T}$ be the MEG measurement matrix with the $(i,t)^{\text{th}}$ element given by $y_{i,t}$. The MEG measurement matrix is related to the matrix of current dipoles $\mathbf{J}$ according to the following forward model [133, 158, 159]:

$$\mathbf{Y} = \mathbf{L}\mathbf{J} + \mathbf{W}, \tag{4.4}$$

where $\mathbf{L} \in \mathbb{R}^{N\times dM}$ maps the *source space* activity to the *sensor space* and is referred to as the *lead-field matrix*, and $\mathbf{W} \in \mathbb{R}^{N\times T}$ is the matrix of additive measurement noise. The lead-field matrix can be estimated based on structural MRI scans by solving Maxwell's equations under the quasi-static approximation [160].

4.1.2 Problem Formulation

Given the stimulus-driven and current-driven forward models of Eq. (4.3) and Eq. (4.4), our main goal is to estimate the matrix $\mathbf{\Phi}$, i.e., the NCRFs. To this end, we take a Bayesian approach, which demands distributional assumptions on the various uncertainties involved, i.e., the stimulus-independent background activity and the measurement noise. For the measurement noise, we adopt the common temporally uncorrelated multivariate Gaussian assumption, i.e.,

$$p\left(\mathbf{Y}|\mathbf{J}\right) = \left|(2\pi)\mathbf{\Sigma}_w\right|^{-T/2} \exp\left(-\frac{1}{2}\|\mathbf{Y}-\mathbf{L}\mathbf{J}\|^2_{\mathbf{\Sigma}_w^{-1}}\right), \tag{4.5}$$

where $\|\mathbf{A}\|^2_{\mathbf{B}} := \mathsf{tr}\left\{\mathbf{A}^\top\mathbf{B}\mathbf{A}\right\}$ and $\mathbf{\Sigma}_w \in \mathcal{S}^N_+$ denotes the unknown noise covariance matrix. The covariance matrix $\mathbf{\Sigma}_w$ can be estimated from either empty-room or pre-stimulus recordings [161]. Next, let $\mathbf{V}_m \in \mathbb{R}^{3\times T}$ denote the matrix of background activity at source m, for $m = 1, 2, \cdots, M$. We adopt the following distribution for the background activity $\mathbf{V}$:

$$p(\mathbf{V}|\mathbf{\Gamma}) = \left(\prod_{m=1}^{M} |(2\pi)\mathbf{\Gamma}_m|^{-T/2}\right) \exp\left(-\frac{1}{2}\sum_{m=1}^{M} \|\mathbf{V}_m\|^2_{\mathbf{\Gamma}_m^{-1}}\right), \tag{4.6}$$

i.e., the portion of the current dipoles reflecting the background activity are modeled as zero-mean independent Gaussian random vectors with unknown 3D covariance matrix $\mathbf{\Gamma}_m \in \mathcal{S}_+^3$. Under this assumption, Eq. (4.3) can be expressed as:

$$p\left(\mathbf{J}|\mathbf{\Phi}, \mathbf{\Gamma}\right) = |(2\pi)\mathbf{\Gamma}|^{-T/2} \exp\left(-\frac{1}{2}\|\mathbf{J} - \mathbf{\Phi}\mathbf{S}\|^2_{\mathbf{\Gamma}^{-1}}\right), \tag{4.7}$$

where $\mathbf{\Gamma} \in \mathcal{S}_+^{3M}$ is a block-diagonal covariance matrix with its m^{th} diagonal block given by $\mathbf{\Gamma}_m$, for $m = 1, 2, \cdots, M$.

Under these assumptions, the joint distribution of the MEG measurement and current dipole matrices is given by:

$$p\left(\mathbf{Y}, \mathbf{J}|\mathbf{\Phi}, \mathbf{\Gamma}\right) = |(2\pi)\mathbf{\Sigma}_w|^{-T/2}|(2\pi)\mathbf{\Gamma}|^{-T/2} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{L}\mathbf{J}\|^2_{\mathbf{\Sigma}_w^{-1}} - \frac{1}{2}\|\mathbf{J} - \mathbf{\Phi}\mathbf{S}\|^2_{\mathbf{\Gamma}^{-1}}\right), \tag{4.8}$$

By marginalizing over $\mathbf{J}$ (see Section B.1 for details), we obtain the distribution of the MEG measurement matrix parametrized by the NCRF matrix $\mathbf{\Phi}$ and the source covariance matrix $\mathbf{\Gamma}$:

$$p\left(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Gamma}\right) = \left|(2\pi)\left(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top\right)\right|^{-T/2} \exp\left(-\frac{1}{2}\|\mathbf{Y} - \mathbf{L}\mathbf{\Phi}\mathbf{S}\|^2_{\left(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top\right)^{-1}}\right). \tag{4.9}$$

It is now possible to cast the problem of finding $\mathbf{\Phi}$ as a Bayesian estimation problem, in which a loss function fully determined by the posterior distribution of NCRF matrix $\mathbf{\Phi}$ given the MEG measurement matrix $\mathbf{Y}$ is minimized. In other words, if $\mathbf{\Gamma}$ were known, the NCRF matrix estimation would amount to the following *maximum likelihood* problem:

$$\min_{\mathbf{\Phi}} \qquad \frac{1}{2}\|\mathbf{Y} - \mathbf{L\Phi S}\|^2_{(\mathbf{\Sigma}_w + \mathbf{L\Gamma L}^\top)^{-1}} \ . \tag{4.10}$$

Another advantage of this Bayesian framework is the possibility of introducing regularization schemes that can mitigate the ill-posed nature of this problem, and instead work with regularized maximum likelihood problems. Note that this optimization problem makes a direct connection between the MEG measurement matrix,$\mathbf{Y}$ and the NCRF matrix $\mathbf{\Phi}$ and allows us to avoid the aforementioned two-stage procedures in finding TRFs at the cortical level [18, 22].

4.1.2.1 Regularization

As is the case in other source imaging methods, there are many fewer constraints than the free parameters determining the NCRFs. This makes the problem severely ill-posed. As such, proceeding with the maximum likelihood problem in Eq. (4.10) is likely to result in overfitting. In order to ensure robust recovery of a meaningful solution to this ill-posed problem, we need to include prior knowledge on the structure of the NCRFs in the form of regularization.

To this end, we construct regularizers based on a convex norm of the NCRF matrix $\mathbf{\Phi}$, to both capture the structural properties of the NCRFs and facilitate algorithm development. The structural properties of interest in this case are spatial sparsity over the cortical source space, sparsity of the peaks/troughs, smoothness in the lag domain, and rotational invariance [19, 125].

In order to promote smoothness in the lag domain and sparsity of the peaks/troughs, we adopt a concept from Chen et al. [162], in which a temporally smooth time series is approximated by a small number of Gabor atoms over an over-complete dictionary $\mathbf{G} \in \mathbb{R}^{L\times\tilde{L}}$, for some $\tilde{L} \geq L$ [125, 163]. To this end, we first perform a change of variables $\boldsymbol{\tau}_m := \mathbf{G}\boldsymbol{\theta}_m$, $\mathbf{\Phi} = \mathbf{\Theta}\mathbf{G}^\top$, and $\widetilde{\mathbf{S}} := \mathbf{G}^\top\mathbf{S}$, where $\boldsymbol{\theta}_m \in \mathbb{R}^{\tilde{L}\times 3}$ are the coefficients of the m^{th} NCRF over the dictionary $\mathbf{G}$ and $\mathbf{\Theta} \in \mathbb{R}^{3M\times\tilde{L}}$ is a matrix containing $\boldsymbol{\theta}_m$s across its rows. Then, to enforce sparsity of the peaks/troughs, spatial sparsity, and rotational invariance, we use the following mixed-norm penalty

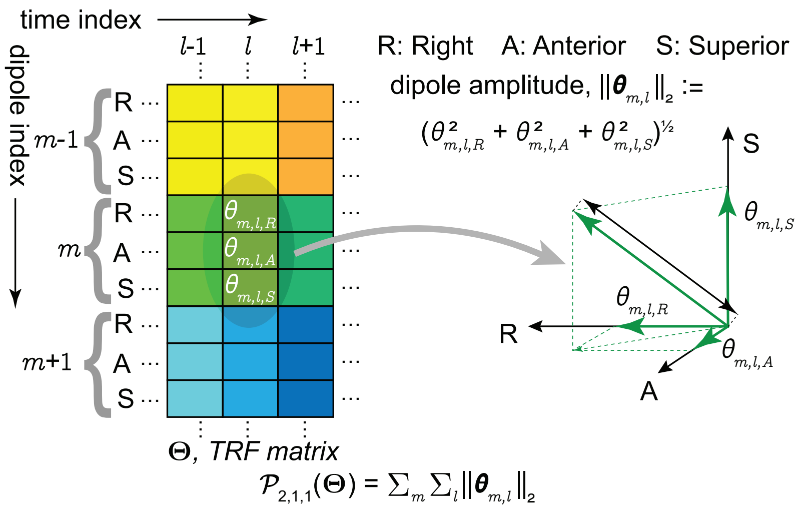

time index
dipole index
l-1
l
l+1
R: Right A: Anterior S: Superior
dipole amplitude, $\|\boldsymbol{\theta}_{m,l}\|_2 :=$
$(\theta^2_{m,l,R} + \theta^2_{m,l,A} + \theta^2_{m,l,S})^{½}$
m-1
m
m+1
R ...
A ...
S ...
$\theta_{m,l,R}$
$\theta_{m,l,A}$
$\theta_{m,l,S}$
S
R
A
$\mathbf{\Theta}$, TRF matrix
$\mathcal{P}_{2,1,1}(\mathbf{\Theta}) = \Sigma_m \Sigma_l \|\boldsymbol{\theta}_{m,l}\|_2$

Figure 4.1: Mixed-norm penalty term $\mathcal{P}_{2,1,1}(\mathbf{\Theta})$ for regularizing the loss function. The penalty term is constructed by first isolating all 3D Gabor coefficient vectors across the dictionary elements and space, and then aggregating their ℓ_2 norm. As a result, it promotes sparsity in space and Gabor coefficients, while being invariant to the orientation of the dipole currents.

over $\boldsymbol{\theta}_m$s, i.e., the Gabor coefficients:

$$\mathcal{P}_{2,1,1}(\boldsymbol{\Theta}) := \sum_{m=1}^{M} \sum_{l=1}^{\tilde{L}} \sqrt{\theta_{m,l,R}^2 + \theta_{m,l,A}^2 + \theta_{m,l,S}^2}. \tag{4.11}$$

Let $\boldsymbol{\theta}_{m,l} \in \mathbb{R}^3$ be the l^{th} Gabor coefficient vector for the m^{th} NCRF. Note that the summand is $\|\boldsymbol{\theta}_{m,l}\|_2$, which is a rotational invariant norm with respect to the choice of dipole RAS coordinate system. This structural feature allows the estimates to be robust to coordinate rotations (see Section B.2). The inner summation of $\|\boldsymbol{\theta}_{m,l}\|_2$ (as opposed to $\|\boldsymbol{\theta}_{m,l}\|_2^2$) over $l = 1, 2, \cdots, \tilde{L}$ enforces group sparsity of the Gabor coefficients (i.e., the number of peaks/troughs), akin to the effect of ℓ_1-norm. Finally, the outer summation over $m = 1, 2, \cdots, M$ promotes spatial sparsity of the NCRFs (see Fig. 4.1, and also Section B.2).

Using this change of variables and regularization scheme, we can reformulate Eq. (4.10) as the following regularized maximum likelihood problem:

$$\min_{\boldsymbol{\Theta}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{L}\boldsymbol{\Theta}\widetilde{\mathbf{S}}\|^2_{(\boldsymbol{\Sigma}_w + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top)^{-1}} + \eta\mathcal{P}_{2,1,1}(\boldsymbol{\Theta}). \tag{4.12}$$

The parameter $\eta > 0$ controls the trade-off between data fidelity and regularization, i.e., the complexity of the resulting model grows inversely with the magnitude of η. This parameter can be chosen in a data-driven fashion using cross-validation (see Section 4.2.2). Fig. 4.2 provides a visual illustration of the proposed modeling and estimation paradigm.

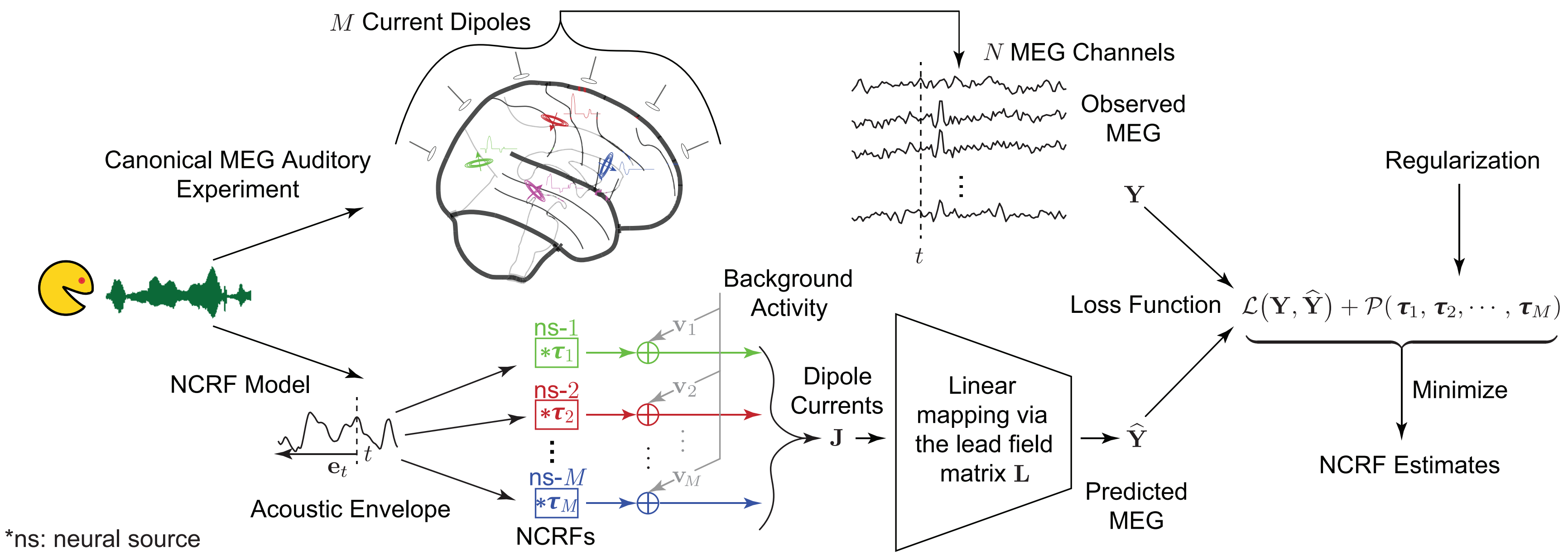


Figure 4.2: Schematic depiction of the proposed modeling and estimation framework. Upper branch: the experimental setting in which the subject is listening to speech while MEG neural responses are being recorded. Lower branch: the modeling framework in which the speech waveform is transformed into a feature variable representation, and is thereby processed via M linear filters (i.e., NCRFs) to generate time-varying current dipoles at each of the corresponding M source locations. Note that each NCRF in the lower branch corresponds to a 3D vector of dipole activity with a specific temporal profile, as shown in the upper branch with matching colors. These dipoles give rise to the predicted MEG response via a source-to-sensor mapping (i.e., the lead-field matrix). The two branches converge on the right hand side, where the NCRFs are estimated by minimizing a regularized loss function.

4.1.2.2 Source Covariance Matrix Adaptation

Note that the objective function in Eq. (4.12) is convex in $\mathbf{\Theta}$ and thus one can proceed to solve for $\mathbf{\Theta}$ by standard convex optimization techniques. However, this requires the knowledge of the source covariance matrix $\mathbf{\Gamma}$, which is unknown in general. From Eq. (4.7), it is evident that $\mathbf{\Gamma}$ implicitly offers adaptive penalization over the source space through spatial filtering. As such, the source covariance matrix serves as a surrogate for depth compensation [164], by reducing the penalization level at locations with low SNR. One data-independent approach for estimating $\mathbf{\Gamma}$ is based on the lead-field matrix [137]. Here, thanks to the Bayesian formulation of our problem, we take a data-driven approach to adapt the source covariance matrix to the background activity not captured by the stimulus [165]. One principled way to do so is to estimate both $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ from the observed MEG data by solving the following optimization problem:

$$\min_{\mathbf{\Theta},\mathbf{\Gamma}} \quad \frac{T}{2}\log\left|\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top\right| + \frac{1}{2}\|\mathbf{Y} - \mathbf{L}\mathbf{\Theta}\widetilde{\mathbf{S}}\|^2_{(\mathbf{\Sigma}_w+\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top)^{-1}} + \eta\mathcal{P}_{2,1,1}(\mathbf{\Theta}) \qquad (4.13)$$

Unfortunately, the loss function in Eq. (4.13) is not convex in $\mathbf{\Gamma}$. However, given an estimate of $\mathbf{\Theta}$, solving for the minimizer of Eq. (4.13) in $\mathbf{\Gamma}$ is a well-known problem in Bayesian estimation and is referred to as evidence maximization or empirical Bayes [166]. Although a general solution to this problem is not straightforward to obtain, there exist several Expectation-Maximization (EM)-type algorithms, such as ReML [136], sMAP-EM [167], and the conjugate function-based algorithm called

Champagne [138], which might be employed to estimate $\mathbf{\Gamma}$ given an estimate of $\mathbf{\Theta}$. In the next section, we present an efficient recursive coordinate descent-type algorithm that leverages recent advances in evidence maximization and proximal gradient methods to solve the problem of Eq. (4.13).

4.1.3 Inverse Solution: The Champ-Lasso Algorithm

Since simultaneous minimization of Eq. (4.13) with respect to both $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ is not straightforward, we instead aim to optimize the objective function by alternatingly updating $\mathbf{\Theta}$ and $\mathbf{\Gamma}$, keeping one fixed at a time. Suppose after the r^{th} iteration, the updated variable pair is given by $\left(\mathbf{\Theta}^{(r)}, \mathbf{\Gamma}^{(r)}\right)$, then the update rules for $(r+1)^{\mathsf{th}}$ iteration are as given as follows:

Updating $\mathbf{\Gamma}$

With $\mathbf{\Theta} = \mathbf{\Theta}^{(r)}$, Eq. (4.13) reduces to the following optimization problem:

$$\min_{\mathbf{\Gamma}} \quad \mathsf{tr}\left(\mathbf{\Sigma}_v^{-1}\mathbf{C}_v\right) + \log|\mathbf{\Sigma}_v|, \tag{4.14}$$

with $\mathbf{C}_v = (\mathbf{Y} - \mathbf{L}\mathbf{\Theta}^{(r)}\widetilde{\mathbf{S}})(\mathbf{Y} - \mathbf{L}\mathbf{\Theta}^{(r)}\widetilde{\mathbf{S}})^\top/T$ and $\mathbf{\Sigma}_v = \mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$. Although the problem is non-convex in $\mathbf{\Gamma}$, it can be solved via the Champagne algorithm [138], which solves for $\mathbf{\Gamma}$ by updating a set of auxiliary variables iteratively. Though the solution $\mathbf{\Gamma}^{(r+1)}$ is not guaranteed to be a global minimum, the convergence rate is fast (with computation cost per iteration being linear in N), and more importantly each iteration is guaranteed not to increase the loss function in Eq. (4.14).

Updating $\mathbf{\Theta}$

Fixing $\mathbf{\Gamma} = \mathbf{\Gamma}^{(r+1)}$, results in the following convex optimization problem:

$$\min_{\mathbf{\Theta}} \quad \frac{1}{2}\|\mathbf{L\Theta}\widehat{\mathbf{S}} - \mathbf{Y}\|^2_{\mathbf{\Sigma}_v^{(r+1)^{-1}}} + \eta\mathcal{P}_{2,1,1}(\mathbf{\Theta}), \tag{4.15}$$

over $\mathbf{\Theta}$, where $\mathbf{\Sigma}_v^{(r+1)} = \mathbf{\Sigma}_w + \mathbf{L\Gamma}^{(r+1)}\mathbf{L}^\top$. The first term in Eq. (4.15) is a smooth differentiable function whose gradient is straightforward to compute, and the proximal operator for the penalty term $\mathcal{P}_{2,1,1}(\mathbf{\Theta})$ has a closed-form expression and can be computed in an efficient manner [168]. Regularized optimization problems of this nature can be efficiently solved using an instance of the forward-backward splitting (FBS) method [169, 170]. We use an efficient implementation of FBS similar to FASTA (Fast Adaptive Shrinkage/Thresholding Algorithm) software package [171] to obtain $\mathbf{\Theta}^{(r+1)}$ from Eq. (4.15).

Although the loss function is not jointly-convex in $(\mathbf{\Theta}, \mathbf{\Gamma})$, the foregoing update steps ensure that the loss in Eq. (4.13) is not increased at any iteration and stops changing when a fixed-point or limit-cycle is reached [172]. Finally, $\mathbf{\Gamma}^0$ can be initialized according to MNE-python recommendations for choosing the source covariance matrix in computing linear inverse operators. Also note that due to the efficiency of the overall solver, it is possible to start the optimization with several randomized initializations, and choose the best solution among several potential alternatives.

4.1.4 Extension to Multiple Feature Variables

The preceding sections focused on the case of a single stimulus feature variable, i.e., the speech envelope. However, complex auditory stimuli such as natural speech, are processed at various levels of hierarchy. Upon entering the ears, the auditory signal is decomposed into an approximate spectrogram representation at the cochlear level prior to moving further into the auditory pathway [173]. Beyond these low-level acoustic features, higher-level phonemic, lexical, and semantic features of the natural speech are also processed in the brain. Thus, to obtain a complete picture of complex auditory cortical processing, it is desirable to consider response functions corresponding to more than one feature variable.

One can proceed to estimate response functions for each feature variable separately. But, since many of these features have significant temporal correlations, the resulting response functions do not readily provide unique information regarding the different levels of the processing hierarchy. To investigate simultaneous processing of these various feature variables and allow them to compete in providing independently informative encoding models, we consider a multivariate extension of the response functions [19, 116].

Suppose that there are $F \geq 1$ feature variables of interest. We modify Eq. (4.3) by replacing each column of the NCRF matrix $\mathbf{\Phi}$ by F columns (one for each temporal response function) and each row of the stimulus matrix by F rows (one for each feature variable). As we will demonstrate below in Section 4.2, this will enable us to distinguish between different cortical regions in terms of their response latency across a hierarchy of features.

4.1.5 Extension to Multiple Trials with Different Stimuli

Next, we consider extension to K different trials corresponding to possibly different auditory stimuli. Let the stimuli, MEG observation, and background activity covariance matrices for the k^{th} trial be denoted by $\widetilde{\mathbf{S}}^k$, $\mathbf{Y}^k$, and $\mathbf{\Gamma}^k$, respectively, for $k = 1, \cdots, K$. We can extend the optimization problem of Eq. (4.13) as follows:

$$\min_{\mathbf{\Theta}, \mathbf{\Gamma}} \quad \sum_{k=1}^{K} \left[\frac{T}{2} \log \left| \mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}^k\mathbf{L}^\top \right| + \frac{1}{2} \| \mathbf{Y}^k - \mathbf{L}\mathbf{\Theta}\widehat{\mathbf{S}}^k \|^2_{(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}^k\mathbf{L}^\top)^{-1}} \right] + \eta \mathcal{P}_{2,1,1}(\mathbf{\Theta}). \tag{4.16}$$

In doing so, we have assumed that the background activity is a stationary Gaussian process within a trial (with covariance $\mathbf{\Gamma}^k$ at trial k), and that the NCRFs remain unchanged across trials, which promotes integration of complementary information from different trials (without direct averaging). Note that this assumption intentionally suppresses the trial-to-trial variability of the NCRFs by adaptively weighting the contribution of each trial according to its noise level (i.e., $\mathbf{\Gamma}^k$), in favor of recovering NCRFs that can explain common cortical patterns of auditory processing. In contrast, if all the trials were to be concatenated or directly averaged to form a unified trial (with a single covariance matrix $\mathbf{\Gamma}$), the trial-to-trial variability would not necessarily be suppressed, especially when there are few trials available. Furthermore, this formulation allows to incorporate trials with different lengths into the same framework.

Algorithm 4: The Champ-Lasso Algorithm over Multiple Trials

```
Input: MEG observations $\mathbf{Y}^k$, modified stimuli matrix $\widehat{\mathbf{S}}^k$, for $k = 1, 2, \cdots, K$; Lead-field matrix $\mathbf{L}$; Regularization parameter $\eta$; initial values of $\boldsymbol{\Theta}^0$; Tolerance parameter $\mathsf{tol} \in (0, 10^{-3})$, Maximum number of outer iterations $R_{\max} \in \mathbb{N}^+$.
$r = 0$ .
repeat
    for $k = 1, \cdots, K$ do
        $\mathbf{C}_v^{k(r)} = \frac{1}{T}(\mathbf{Y}^k - \mathbf{L}\boldsymbol{\Theta}^{(r)}\widehat{\mathbf{S}}^k)(\mathbf{Y}^k - \mathbf{L}\boldsymbol{\Theta}^{(r)}\widehat{\mathbf{S}}^k)^\top$
        $\boldsymbol{\Gamma}^{k(r+1)} = \arg\min_{\boldsymbol{\Gamma}} \ \mathsf{tr}\left(\boldsymbol{\Sigma}_v^{-1}\mathbf{C}_v^{k(r)}\right) + \log|\boldsymbol{\Sigma}_v| \text{ s.t. } \boldsymbol{\Sigma}_v = \boldsymbol{\Sigma}_w + \mathbf{L}^\top\boldsymbol{\Gamma}\mathbf{L}$
                                                            ▷ Champagne iterations
        $\boldsymbol{\Sigma}_v^{k(r+1)} = \left(\boldsymbol{\Sigma}_w + \mathbf{L}\boldsymbol{\Gamma}^{k(r+1)}\mathbf{L}^\top\right)^{-1}$
    end for
    $\boldsymbol{\Theta}^{(r+1)} = \arg\min_{\boldsymbol{\Theta}} \sum_{k=1}^{K} \frac{1}{2}\|\mathbf{L}\boldsymbol{\Theta}\widehat{\mathbf{S}}^k - \mathbf{Y}^k\|^2_{{\boldsymbol{\Sigma}_v^{k(r+1)}}^{-1}} + \eta\mathcal{P}_{2,1,1}(\boldsymbol{\Theta})$
                                                            ▷ FASTA iterations
until $\frac{\|\boldsymbol{\Theta}^{(r+1)} - \boldsymbol{\Theta}^{(r)}\|_2}{\|\boldsymbol{\Theta}^{(r)}\|_2} < \mathsf{tol}$ or $r = R_{\max}$.
Set $r \leftarrow r + 1$.
Output: $\boldsymbol{\Theta}^{(R)}$ where $R$ is the index of the last outer iteration of the algorithm.
```

The optimization problem of Eq. (4.16) can be solved via a slightly modified version of the solution presented in Section 4.1.3. The resulting algorithm is summarized in Algorithm 4, which we refer to as the Champ-Lasso algorithm. A python implementation of the Champ-Lasso algorithm is archived on the open source repository Github [174] to ease reproducibility and facilitate usage by the broader systems neuroscience community.

4.1.6 Subjects, Stimuli, and Procedures

The data used in this chapter are a subset of recordings presented in Presacco et al. [153], and is publicly available in the Digital Repository at the University of Maryland [175]. The auditory experiments were conducted under the participation of 17 young adult subjects (aged 18-27 years), recruited from the Maryland, Washington D.C. and Virginia area. The participants listened to narrated segments from the audio-book, *The Legend of Sleepy Hollow by Washington Irving* (https://librivox.org/the-legend-of-sleepy-hollow-by-washington-irving/), while undergoing MEG recording. Although the dataset contains recordings under different background noise levels, for the current analysis we considered recordings of two 1 min long segments of the audio-book with no background noise presented as single-speaker audio. Each of these segments was repeated three times to every individual, yielding a total 6 min of data per subject. To ensure that the participants actively engage in the listening task, they were tasked to also silently count the number of specific words that they would hear in the story.

4.1.7 Recording and Preprocessing

The data were acquired using a whole head MEG system (KIT, Nonoichi, Ishikawa, Japan) consisting of 157 axial gradiometers, at the University of Maryland Neuroimaging Center, with online low-pass filtering (200 Hz) and notch filtering (60 Hz) at a sampling rate of 1 kHz. Data were pre-processed with MNE-python 0.18.1 [143, 154]. After excluding flat and noisy channels, temporal signal space separation was applied to remove extraneous artifacts [176]. Data were then filtered between 1 Hz to 80 Hz using a zero-phase FIR filter with the default filter parameter options of the software. Independent component analysis [extended infomax, 177] was applied to remove ocular and cardiac artifacts. Finally, 60 s long data epochs corresponding to the stimuli were extracted and downsampled to 200 Hz.

4.1.8 Source Space Construction

At the beginning of each recording session, each participant's head shape was digitized with a 3-Space Fastrak system (Polhemus), including 3 fiducial points and 5 marker positions. Five marker coils attached to these five marker positions were used to localize the head position of the participant relative to the MEG sensors. The head position measurement was recorded twice: at the beginning and end of the recording session and the average measured head positions were used. Since MR scans of the participants were not performed, the 'fsaverage' brain model [178] was co-registered (via uniform scaling, translation and rotation) to each participant's head, using the digitized head shapes.

A volumetric source space for the 'fsaverage' brain was defined on a regular grid with spacing of 7 mm between two neighboring points, and then morphed to individual participants. These morphed source spaces were then used to compute lead-field matrices by placing 3 orthogonal virtual current dipoles on each of the grid points. The computed lead-field matrices contained contribution from 3222 virtual current dipoles, after removing those within subcortical structures along the midline. No cortical patch statistics were available due to the lack of MR scans, so the current dipoles were allowed to have arbitrary orientations in 3D.

4.1.9 Stimulus Feature Variables

We included predictor variables reflecting three different hierarchical levels of speech processing, including acoustic, lexical, and semantic features. These feature variables are described in detail in Brodbeck et al. [22]:

- *Envelope:* The speech envelope was found by averaging the auditory spectrogram representation generated using a model of the auditory periphery [173] across frequency bands. This continuous univariate feature variable reflects the momentary acoustic power of the speech signal.

- *Word Frequency:* First, logarithmic word frequency measures, $\log_{10} \mathsf{wf}$, were extracted from the SUBTLEX database [179] for each word. Then, a piecewise-continuous feature variable was constructed by representing each word in the speech segment by a rectangular pulse with height given by $6.33 - \log_{10} \mathsf{wf}$. Note that in this coding scheme, infrequent words are assigned higher values, while common words get lower values. Windows of silence were assigned 0.

- *Semantic Composition:* Lastly, to probe semantic processing, the semantic composition patterns identified by Westerlund et al. [180], including adjective-noun, adverb-verb, adverb-adjective, verb-noun, preposition-noun and determiner-noun pairs, were used. To generate the feature variable, the second word in each pair was represented by a rectangular window of height 1, and 0 elsewhere. This binary-valued feature variable identifies the semantic binding of word pairs within the speech stream.

All three variables were constructed from the speech segments at the same sampling frequency as the preprocessed MEG data (i.e. 200 Hz). All feature variables were centered and scaled by their mean absolute value, to facilitate comparison of NCRF components pertaining to different feature variables.

4.1.10 Estimation Setup, Initialization and Statistical Tests

We estimated 1000 ms-long NCRFs ($L = 200$) corresponding to each of these three stimulus variables ($F = 3$). This choice leads to a high-dimensional NCRF matrix $\mathbf{\Phi} \in \mathbb{R}^{9666 \times 600}$. The noise covariance matrix, $\mathbf{\Sigma}_w$, was estimated from empty-room data using MNE-python 0.18.1 [143, 154] following an automatic model selection procedure. The regularization parameter was tuned on the basis of generalization error via a 3-fold cross-validation procedure: from a predefined set of regularization parameters (equally spaced in logarithmic scale), the one resulting in the least generalization error was chosen to estimate the NCRFs for each subject. To maintain low running time, instead of utilizing a randomized initialization scheme for $\mathbf{\Gamma}^0$, we initialized it according to the MNE-python recommendation for source covariances. The NCRF matrix $\mathbf{\Theta}^0$ was initialized as an all zero matrix. In the consecutive iterations of the Champagne and FASTA algorithms, a warm starting strategy was followed, i.e., initializing each iteration by the solution of the previous one.

To check whether inclusion of each of the feature variables improves the overall NCRF model significantly, the original model fit being tested for significance (i.e., its cost function evaluated at the estimated NCRF parameters) was compared against the average of three other model fits constructed by deliberately misaligning one feature variable via 4-fold cyclic permutations, using one-tailed t-tests.

To evaluate the group-level significance of the estimated NCRF components, the NCRF estimates were first smoothed with a Gaussian kernel (with standard deviation of 10 mm) over the source locations to compensate for possible head misalignments and anatomical differences across subjects. Then, at each dipole location and time index, the magnitudes of the vector-valued NCRFs were tested for significance using a permutation test via the threshold-free cluster-enhancement (TFCE) algorithm [181] (see Section B.3 for details).

4.1.11 Simulation Setup

Before applying the Champ-Lasso algorithm to localize NCRFs from experimentally recorded data, we assessed its performance using realistic simulation studies with known ground truth. In accordance with our experimental settings, we synthesized six 1 min long MEG data segments according to the forward model of Eq. (4.3) and Eq. (4.4), mimicking the neural processing of the speech envelope. To this end, we simulated temporal response functions of length 500 ms (with significant M50 and M100 components) associated with dipole current sources within the auditory and motor cortices. Fig. 4.3A shows the simulated activity over the cortical surface at specific time lags (top and bottom panels) as well as the temporal profile of the NCRFs (middle panel). The cortical activity was simulated using patches defined over a finely-discretized source space (namely, ico-5, with average dipole spacing of 3.1 mm) with the dipole directions constrained to be normal to the ‘fsaverage’ surface patches. To make the simulation as realistic as possible, we used real MEG recordings corresponding to a different speech stream as background noise (i.e., stimulus-independent background activity), maintaining a -5 dB signal-to-noise ratio.

In order to avoid any favorable bias in the inverse solution, we used a different source space for NCRF estimation, i.e., the aforementioned volumetric source space with unconstrained dipole orientations (Section 4.1.8), than the one used for simulating the data, i.e., ico-5. As a comparison benchmark, we also applied the two-stage method of Brodbeck et al. [22], MNE-boosting, and one of its variants, Champagne-boosting, to first localize the cortical sources using MNE and Champagne, respectively, followed by boosting with 10-fold cross-validation and ℓ_1-norm error of the standardized source estimates, for independently estimating TRFs for all sources. The boosting was initialized with a zero response function, and iteratively modified it in small increments (typically 0.001) at a single time-lag in which a change led to the largest ℓ_1-norm error reduction in the training set. The process stopped when the training error no longer decreased, or testing error increased in two successive steps.

In order to compare the spatial spread across different methods, we computed the *dispersion* metric as the ratio of total NCRF power outside and inside of spheres of radius r (for $r = 10, 15, 20$ mm) around the center of mass of the simulated cortical patches (i.e., lower is better). To quantify the response function reconstruction performance, the 3-dimensional NCRFs within radius of $r = 15$ mm around the center of mass of the simulated cortical patches were averaged and then separated into *principal orientation* and *principal time course*, using singular value decomposition. The principal orientations and time courses were compared to the ground truths using the Pearson correlation (i.e., higher is better). Finally, we quantified the *selectivity* of the principal orientation and time course in the recovered NCRFs, by the ratio of the principal singular value to the sum of all three (i.e., higher is better).

4.2 Results

4.2.1 Simulation Studies

The two two-stage localized TRFs and estimated NCRFs are shown in Fig. 4.3B, C and D, respectively. Since boosting tends to result in temporally sparse response functions [6], response functions were smoothed with a 50 ms Hamming window. The anatomical plots show the spatial response function profile at the same temporal peaks selected in Fig. 4.3A, with direction of the vectors projected onto the lateral plane. The Champ-Lasso algorithm successfully recovers both the smooth temporal profile of the NCRFs and the spatial extent and location of the active sources, and provides estimates that closely resemble the ground truth.

The two-stage localized TRFs, however, fail to recover the true extent of the sources due to the destructive propagation of biases: MNE-boosting estimates are spatially dispersed while Champagne-boosting estimates are overly sparse. Also, the poor signal-to-noise ratio caused the estimates to exhibit spurious peaks in the anterior temporal and inferior frontal lobes: the prominence of these spurious peaks

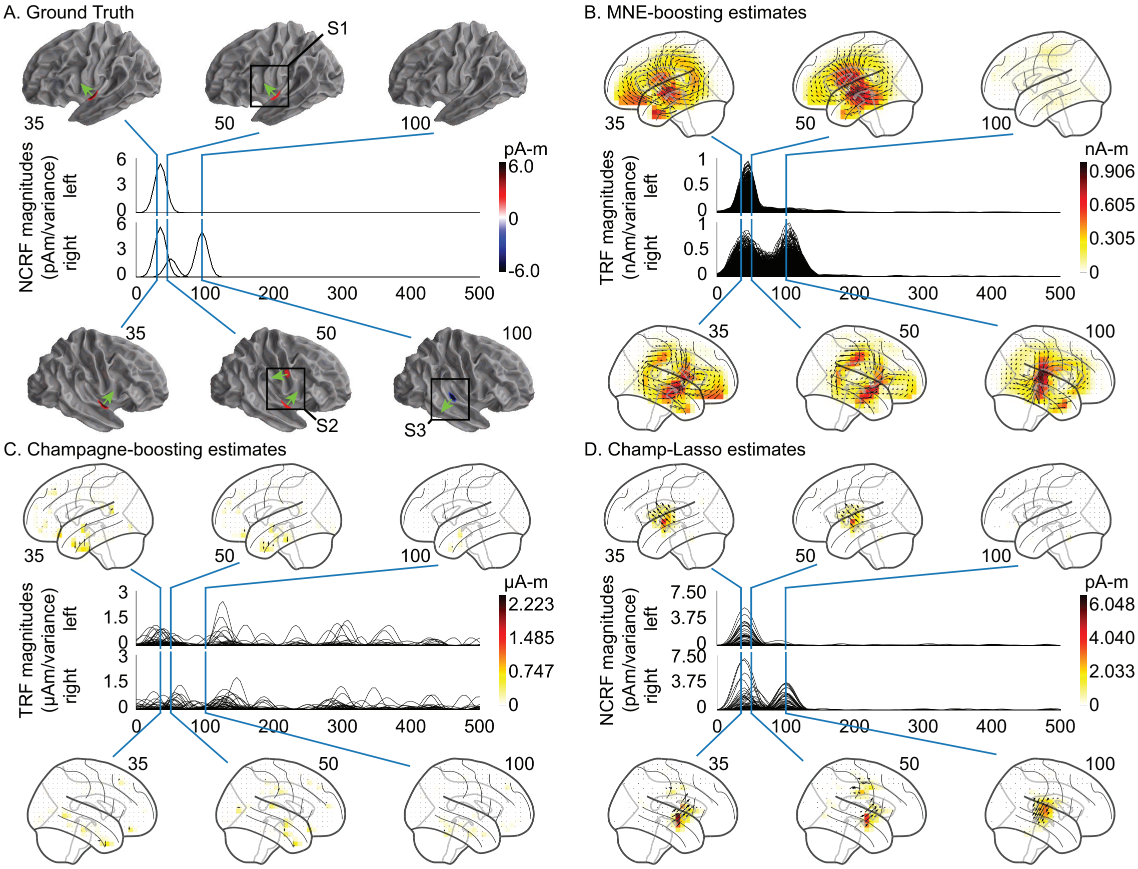


Figure 4.3: Results for a simulated auditory experiment. The top and bottom portions of each subplot pertain to the left and right hemispheres, respectively. A. The anatomical plots show the simulated neural sources normal to cortical surface, and the traces show the overlaid temporal profiles. The colorbar encodes directional intensity normal to the cortical surface (shown by the green arrows). B & C. The two-stage localized (i.e. MNE-boosting, and one of its variants, Champagne-boosting, respectively) TRFs (free-orientation) are shown on the anatomical plots, where the 3D dipoles are projected onto the lateral plane. D. NCRF estimates from the Champ-Lasso algorithm. The colorbar encodes dipole magnitudes. The spatial extent, dipole moment scale, temporal profile, and orientations of the neural sources are faithfully recovered by the Champ-Lasso algorithm, whereas the two-stage localized TRFs are either spatially dispersed (MNE-boosting) or overly sparse (Champagne-boosting) and exhibit spurious peaks in the anterior temporal and inferior frontal lobes.

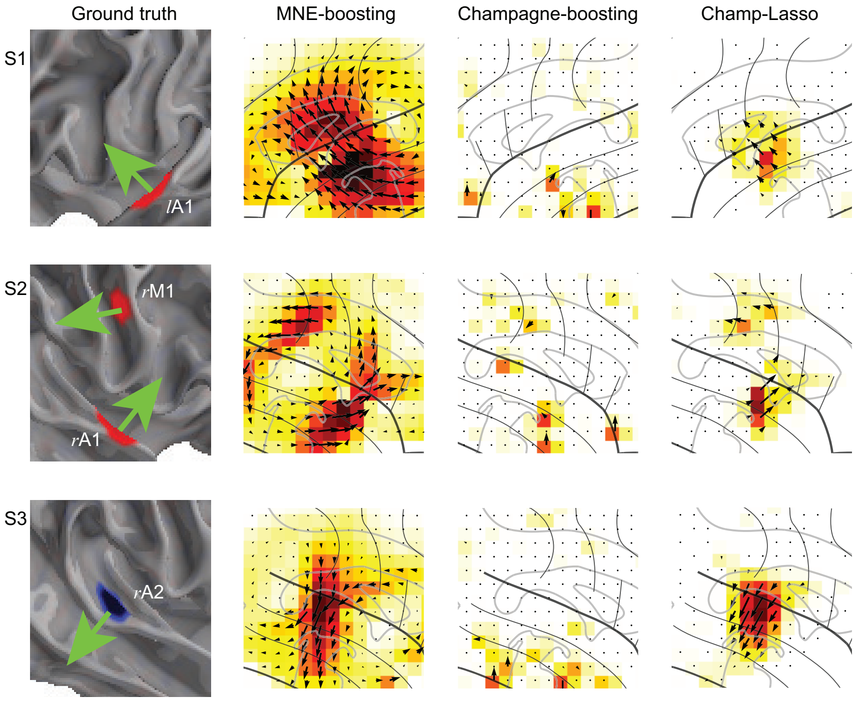


Figure 4.4: Results for the simulated auditory experiment (continued): Zoomed in views of the active cortical patches (marked as S1, S2 and S2 in Fig. 4.3A) emphasizing the orientations of the simulated current dipoles (green arrows) alongside the estimated current dipole directions.

in the Champagne-boosting estimates results in fully overshadowing the true sources. In addition, the two-stage localized TRFs are rescaled using the standard deviation of the sources before plotting. This rescaling, combined with the poor signal-to-noise ratio, leads to the large scaling discrepancy between the estimates and the ground truth. It is worth noting that despite the fact that the Champ-Lasso algorithm is unaware of the true dipole orientations, the resulting NCRF orientations closely match the normal directions of the patches (see Fig. 4.4). The Champ-Lasso also successfully suppresses spurious peaks in the anterior temporal and inferior frontal lobes, demonstrating its robustness to background activity.

Benchmarking metrics described in Section 4.1.11 are listed in Table 4.1 and Table 4.2. Table 4.1 lists the *dispersion* metric, demonstrating how the different

r (mm)	Champ-Lasso	MNE-boosting	Champagne-boosting
0.010	**1.139**	4.005	6.446
0.015	**0.630**	2.229	3.172
0.020	**0.229**	1.491	2.963

Table 4.1: Comparison with respect to the dispersion metric, defined as the ratio of the total NCRF power outside and inside of spheres of radius r (for $r = 10, 15, 20$ mm) around the center of mass of the simulated cortical patches (lower is better). The bold numerical values indicate the best performance among the different estimation methodologies for each sphere radius.

algorithms perform in localizing the neural sources correctly. Table 4.2 contains the *correlation* measures for the principal orientation and time course along with the selectivity of these principal orientations and time courses across the different simulated regions. While the principal orientations of Champ-Lasso and MNE-boosting are similarly correlated with the true orientation, the selectivity of this orientation in MNE-boosting is inferior to Champ-Lasso. Champagne-boosting exhibits the poorest performance overall. Unlike the other methods, the Champ-Lasso principal time

		lA1	rA1	rM	rA2
Principal Orientation Correlation	Champ-Lasso	**0.991**	**0.992**	**0.923**	**0.996**
	MNE-boosting	0.978	0.989	**0.923**	0.995
	Champagne-boosting	0.124	0.207	-0.395	0.039
Principal Time Course Correlation	Champ-Lasso	0.968	**0.953**	**0.972**	**0.958**
	MNE-boosting	0.959	0.856	0.741	0.918
	Champagne-boosting	**0.994**	0.131	0.114	0.015
Selectivity	Champ-Lasso	**0.999**	**0.977**	**0.932**	0.984
	MNE-boosting	0.977	0.794	0.469	0.845
	Champagne-boosting	0.993	0.891	0.904	**1.000**

Table 4.2: Comparison with respect to the reconstruction metrics: Pearson correlation coefficients between the estimated *principal orientation* and *principal time course* and the ground truth, as well as their selectivity (higher is better) for different cortical patches (lA1, rA1, rM and rA2 as in Fig. 4.4). The bold numerical values indicate the best performance among the different estimation methodologies in each category.

courses consistently show high correlation with the ground truth.

4.2.2 Application to Experimentally Recorded MEG Data

4.2.2.1 Analysis of the Acoustic Envelope NCRFs

Fig. 4.5A depicts the group average of estimated NCRFs, corresponding to the acoustic envelope, masked by a significance level of $p = 0.05$. The time traces show the magnitude of the average NCRFs (gray segments are statistically insignificant) and the anatomical plots show the spatial NCRF profile at selected temporal peaks, with direction of the vectors projected onto the lateral plane. Fig. 4.5B shows the temporal profiles (masked at a significance level of $p = 0.05$) of 6 selected NCRFs exhibiting peak spatial activity (collapsed across time). The colored dots on the anatomical plots show the locations of these NCRFs, with matching colors to those of the traces. The traces are grouped by hemisphere and dorsoventrally ordered. The left and right NCRFs in the motor areas are referred to as $l\mathsf{M}_{\mathsf{env}}$ and $r\mathsf{M}_{\mathsf{env}}$, respectively. The left and right auditory NCRF pairs are labeled as $l\mathsf{A1}_{\mathsf{env}}$, $l\mathsf{A2}_{\mathsf{env}}$ and $r\mathsf{A1}_{\mathsf{env}}$, $r\mathsf{A2}_{\mathsf{env}}$, respectively. The NCRFs corresponding to the acoustic envelope in Fig. 4.5A exhibit two prominent temporal peaks: an early peak at around 30–35 ms, bilaterally centered over the auditory cortex (AC), and a later peak at around 100 ms, dorsal to the first peak and stronger in the right hemisphere. The latter is evident from comparing the left temporal profiles $l\mathsf{A1}_{\mathsf{env}}$ and $l\mathsf{A2}_{\mathsf{env}}$, with their right hemisphere counterparts $r\mathsf{A1}_{\mathsf{env}}$ and $r\mathsf{A2}_{\mathsf{env}}$. Note that the orientations of the NCRFs at the second peak (blue arrow, bottom panel of Fig. 4.5A) are nearly the

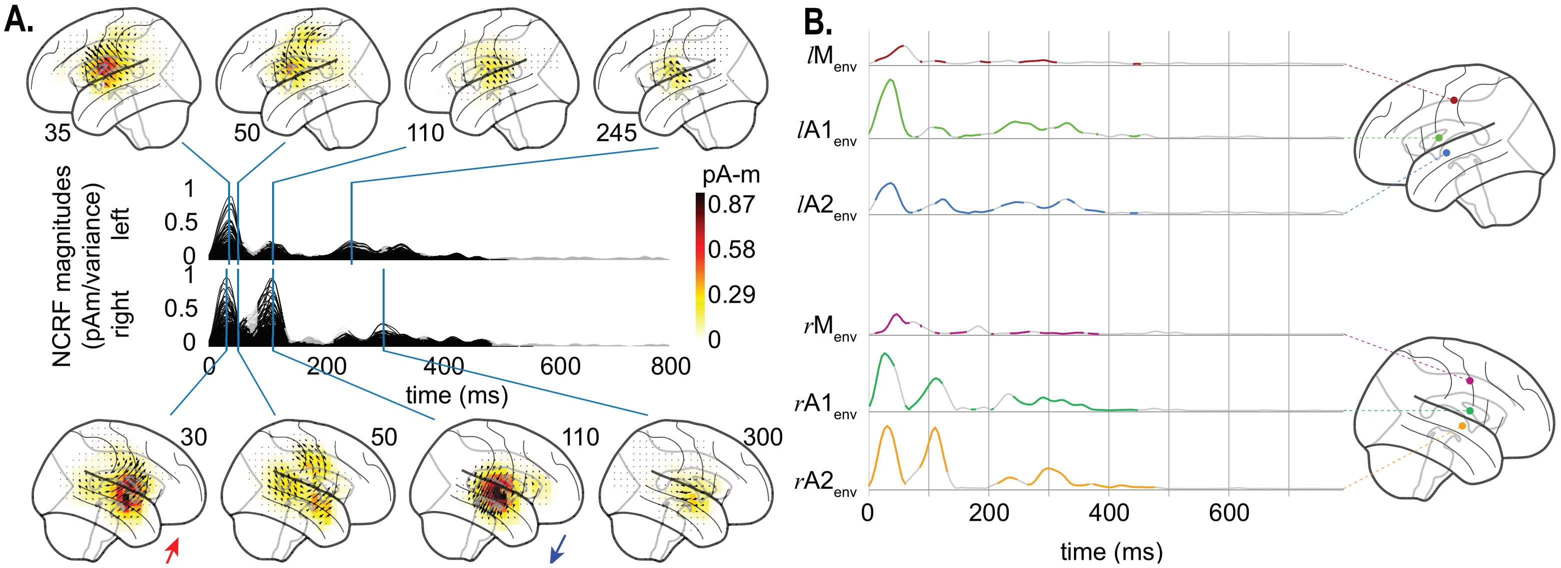


Figure 4.5: Estimated NCRFs for the acoustic envelope feature variable. A. The anatomical plots show the group-level average NCRFs projected onto the lateral plane (top and bottom panels) corresponding to selected visually salient peaks in the temporal profiles (middle panels). The top and bottom portions of the subplot pertain to left and right hemisphere, respectively. Numerical labels of each anatomical subplot indicates the corresponding time lag in ms. B. The time traces show the temporal profile of 6 selected NCRFs exhibiting peak spatial activity (collapsed across time), grouped by hemisphere and dorsoventrally ordered. The locations of the selected NCRFs are shown on the anatomical plots, with colors matching the time traces and linked by dashed lines. The gray portions of the traces in both subplots indicate statistically insignificant NCRFs at the group level (significance level of 5%). Note that the last 200 ms segments of the temporal profiles are cropped, as they did not correspond to any significant components at the group level. The prominent NCRFs consist of a bilateral auditory component at $\sim 30 - 35$ ms, a bilateral motor component at ~ 50 ms, and an auditory component at ~ 110 ms (stronger in right hemisphere and with nearly opposing polarity with respect to the earlier auditory component, indicated by the two colored arrows pointing to the average direction of the NCRFs in subplot A). See Supplementary Movie 01 for a detailed animation showing how the acoustic envelope NCRF components change as function of time lags.

opposite of those at the first peak (red arrow, bottom panel of Fig. 4.5A), which accounts for the negative polarity of the $\mathsf{M100}$ peak with respect to $\mathsf{M50}$ in standard TRF analysis. Furthermore, after the appearance of the first peak ($\sim$ 35 ms, auditory traces $l\mathsf{A1}_{\mathsf{env}}$, $l\mathsf{A2}_{\mathsf{env}}$, $r\mathsf{A1}_{\mathsf{env}}$, and $r\mathsf{A2}_{\mathsf{env}}$) at the AC, the activity appears to gradually shift towards the primary motor cortex (PMC) in both hemispheres ($\sim$ 50 ms, motor traces $l\mathsf{M}_{\mathsf{env}}$ and $r\mathsf{M}_{\mathsf{env}}$). Additionally, the NCRFs show small bilateral late auditory components at around $\sim$ 250–350 ms.

4.2.2.2 Analysis of the Word Frequency NCRFs

Fig. 4.6A shows the NCRFs for the word frequency feature variable, in the same format as in Fig. 4.5. Fig. 4.6B shows the temporal profiles (masked at a significance level of $p = 0.05$) of 4 selected NCRFs exhibiting peak spatial activity (collapsed across time) also in the same format as in Fig. 4.5. These include a left auditory ($l\mathsf{A}_{\mathsf{wf}}$), a left frontal ($l\mathsf{F}_{\mathsf{wf}}$), a left inferior temporal (IT) ($l\mathsf{IT}_{\mathsf{wf}}$), and a right auditory ($r\mathsf{A}_{\mathsf{wf}}$) NCRF. The significant NCRF components manifest predominantly in the left hemisphere (Fig. 4.6A). The earliest peak in the left AC occurs at around 50 ms, followed by a much stronger peak at around 150 ms, slightly posterior to the former (see $l\mathsf{A}_{\mathsf{wf}}$ in Fig. 4.6B). The earlier peak also has contributions from the inferior temporal gyrus, as indicated by $l\mathsf{IT}_{\mathsf{wf}}$. In addition, the left frontal cortex exhibits weak activity at around 150 ms (see $l\mathsf{F}_{\mathsf{wf}}$ in Fig. 4.6B). A weak but localized peak centered over the left superior temporal sulcus (STS) is visible at around 240 ms. The only significant component in the right hemisphere occurs around the same

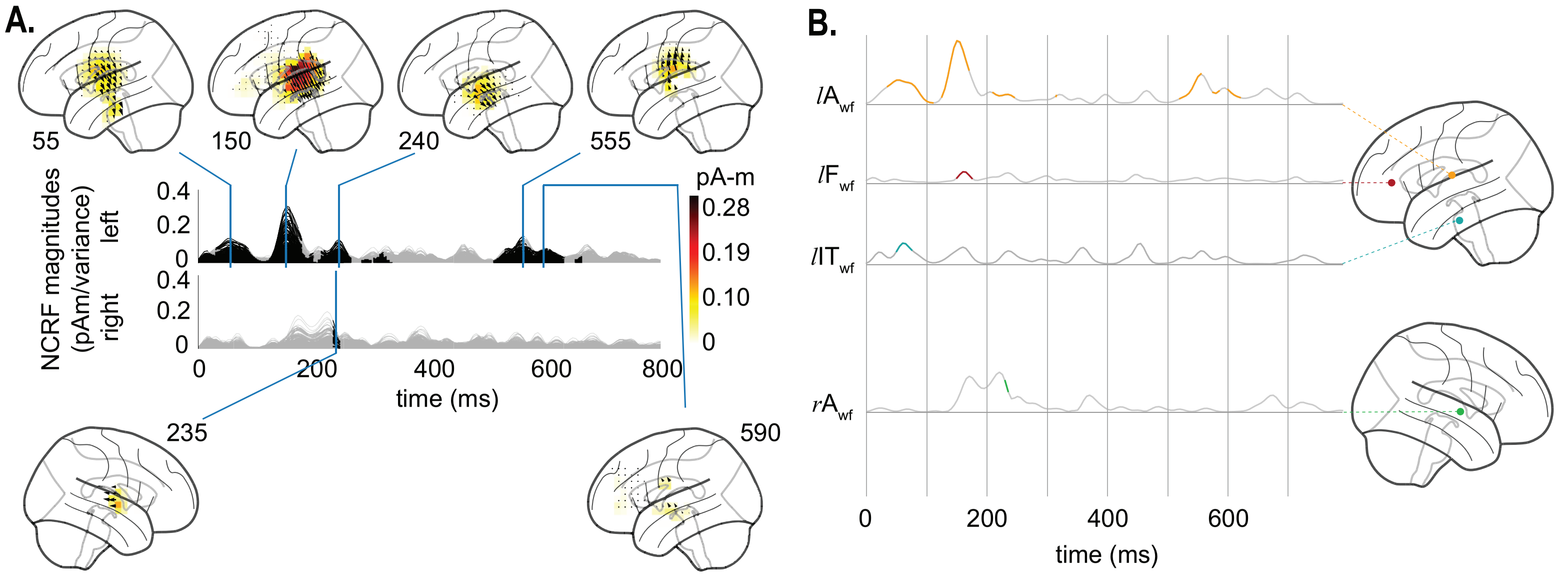


Figure 4.6: Estimated NCRFs for the word frequency feature variable. A. The anatomical plots show the group-level average NCRFs projected onto the lateral plane (top and bottom panels) corresponding to selected visually salient peaks in the temporal profiles (middle panels). The top and bottom portions of the subplot pertain to left and right hemisphere, respectively. Numerical labels of each anatomical subplot indicates the corresponding time lag in ms. B. The time traces show the temporal profile of 4 selected NCRFs exhibiting peak spatial activity (collapsed across time), grouped by hemisphere and and dorsoventrally ordered. The locations of the selected NCRFs are shown on the anatomical plots, with colors matching the time traces and linked by dashed lines. The anatomical plots show the locations of the selected NCRFs with colors matching the time traces. The gray portions of the traces in both subplots indicate statistically insignificant NCRFs at the group level (significance level of 5%). The prominent NCRFs manifest in the left hemisphere, dominated by an auditory component at ∼ 150 ms. See Supplementary Movie 02 for a detailed animation showing how the word frequency NCRF components change as function of time lags.

time. Finally, the late NCRF components (at around 500–600 ms) mostly originate from the left AC and STS, with weak contributions from the right frontal cortex.

4.2.2.3 Analysis of the Semantic Composition NCRFs

The estimated NCRFs corresponding to the semantic composition feature variable are shown in Fig. 4.7A, along with 5 representative NCRFs in Fig. 4.7B. These include two left auditory ($l\mathsf{A1}_{\mathsf{sc}}$ and $l\mathsf{A2}_{\mathsf{sc}}$), two right frontal ($r\mathsf{F1}_{\mathsf{sc}}$ and $r\mathsf{F2}_{\mathsf{sc}}$), and one right middle temporal ($r\mathsf{MT}_{\mathsf{sc}}$) NCRF. The main NCRF components in the left AC peak at around 155 ms and 475 ms, with the earlier peak being ventral to the later one (see $l\mathsf{A1}_{\mathsf{sc}}$ and $l\mathsf{A2}_{\mathsf{sc}}$ in Fig. 4.7B). The significant right hemispheric NCRFs are temporally concentrated between 155 to 210 ms, and appear superior to those in the left hemisphere, involving inferior frontal gyrus (IFG). Strikingly, these NCRFs in the right hemisphere seem to move in the anterosuperior direction until around 185 ms, at which point the right hemisphere exhibits strong frontal activity (Fig. 4.7A). The NCRFs return to their initial location afterwards at around 210 ms. This sequence of spatiotemporal changes is also evident in the sequence of temporal peaks in Fig. 4.7B, given by $r\mathsf{MT}_{\mathsf{sc}} \rightarrow r\mathsf{F2}_{\mathsf{sc}} \rightarrow r\mathsf{F1}_{\mathsf{sc}} \rightarrow r\mathsf{F2}_{\mathsf{sc}}$.

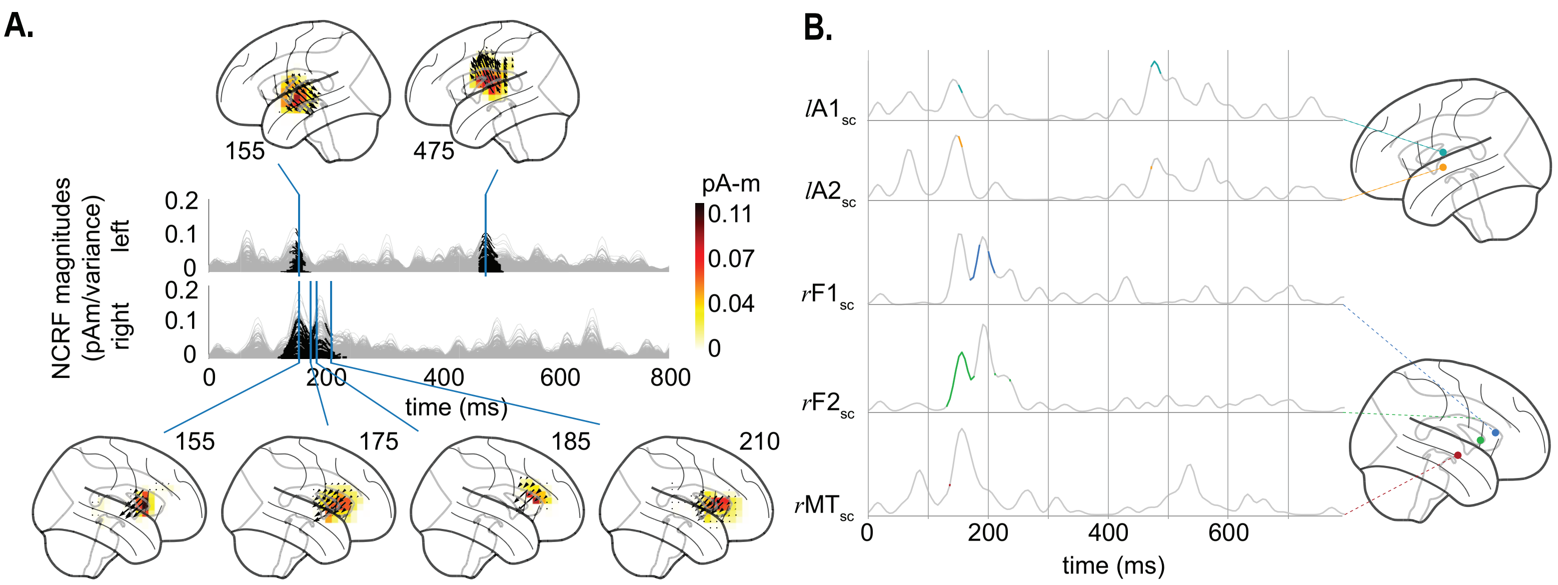


Figure 4.7: Estimated NCRFs for the semantic composition feature variable. A. The anatomical plots show the group-level average NCRFs projected onto the lateral plane (top and bottom panels) corresponding to selected visually salient peaks in the temporal profiles (middle panels). The top and bottom portions of the subplot pertain to left and right hemisphere, respectively. Numerical labels of each anatomical subplot indicates the corresponding time lag in ms. B. The time traces show the temporal profile of 5 selected NCRFs exhibiting peak spatial activity (collapsed across time), grouped by hemisphere and dorsoventrally ordered. The locations of the selected NCRFs are shown on the anatomical plots, with colors matching the time traces and linked by dashed lines. The gray portions of the traces in both subplots indicate statistically insignificant NCRFs at the group level (significance level of 5%). The prominent NCRF components consist of a bilateral auditory component at $\sim$ 155 ms, a right auditory-frontal component from $\sim$ 180 ms to $\sim$ 210 ms, and a left auditory late component at $\sim$ 475 ms. The sequence of peaks given by $r\mathsf{MT}_{\mathsf{sc}} \rightarrow r\mathsf{F2}_{\mathsf{sc}} \rightarrow r\mathsf{F1}_{\mathsf{sc}} \rightarrow r\mathsf{F2}_{\mathsf{sc}}$ show the back and forth movement of the NCRFs from the auditory to frontal cortices. See Supplementary Movie 03 for a detailed animation showing how the semantic composition NCRF components change as function of time lags.

4.3 Discussion and Concluding Remarks

Characterizing the dynamics of cortical activity from noninvasive neuroimaging data allows us to probe the underlying mechanisms of sensory processing at high spatiotemporal resolutions. In this chapter, we demonstrated a framework for direct estimation of such cortical dynamics in response to various features of continuous auditory stimuli from the MEG response. To this end, we developed a fast inverse solution under a Bayesian estimation setting, the Champ-Lasso algorithm, for inferring the Neuro-Current Response Functions (as spatiotemporal models of cortical processing) in a robust and scalable fashion.

One of the key features of the Champ-Lasso algorithm is the ability to simultaneously estimate cortical source covariances in a data-driven fashion (as opposed to relying on data-agnostic depth-weighting procedures) and finding the NCRF model parameters. The interplay between the two as well as incorporating the structural properties of the NCRFs into the model, taking advantage of the Bayesian nature of the estimation framework, ultimately leads to spatially focal NCRFs, with smooth temporal profiles. In other words, the NCRF and source covariance estimation procedures function in tandem to best explain the observed MEG data while minimizing the spatial leakage and capturing the smoothness of the temporal responses. In contrast, previously existing methodologies result in estimates that are spatially broad, which then require post-hoc clustering procedures to meaningfully summarize the underlying spatiotemporal cortical dynamics. These serialized procedures in turn introduce biases to the estimates, and hinder meaningful statistical interpretation of the results.

To demonstrate the utility of our proposed framework, we estimated NCRFs corresponding to several feature variables of speech, reflecting different levels of cognitive processing and comprehension from MEG. The data analyzed here were analyzed by an earlier method in Brodbeck et al. [22], where a two-stage procedure was utilized to probe the cortical processing of speech: the MEG data was first cortically localized using an MNE inverse solver, followed by estimating individual temporal response functions for each source. In order to summarize the resulting estimates in a meaningful fashion, yet another processing step was necessary to disentangle the different spatially dispersed and highly overlapping cortical sources. Our results corroborate those obtained in Brodbeck et al. [22], while obviating the need for any such post-processing, by providing a one-step estimation procedure with the substantial benefit of greatly improved spatial resolution. In addition, the three-dimensional nature of the NCRFs in our framework allows the segregation of different spatial activation patterns that are temporally overlapping. For example, the bilateral activity components in the primary motor cortex in response to the acoustic envelope are automatically clearly distinguishable from the early activation in the auditory cortex, without the need for any post-hoc processing. To ease the visual comparison, Fig. 4.8 compares the estimated NCRF distributions (transparent cortex) to those of Brodbeck et al. [22] (inflated cortical surface), at several time points for each of the three stimulus features.

Our results also support other neuroimaging evidence for the hierarchical model of speech processing, involving not only the temporal lobe, but also the motor

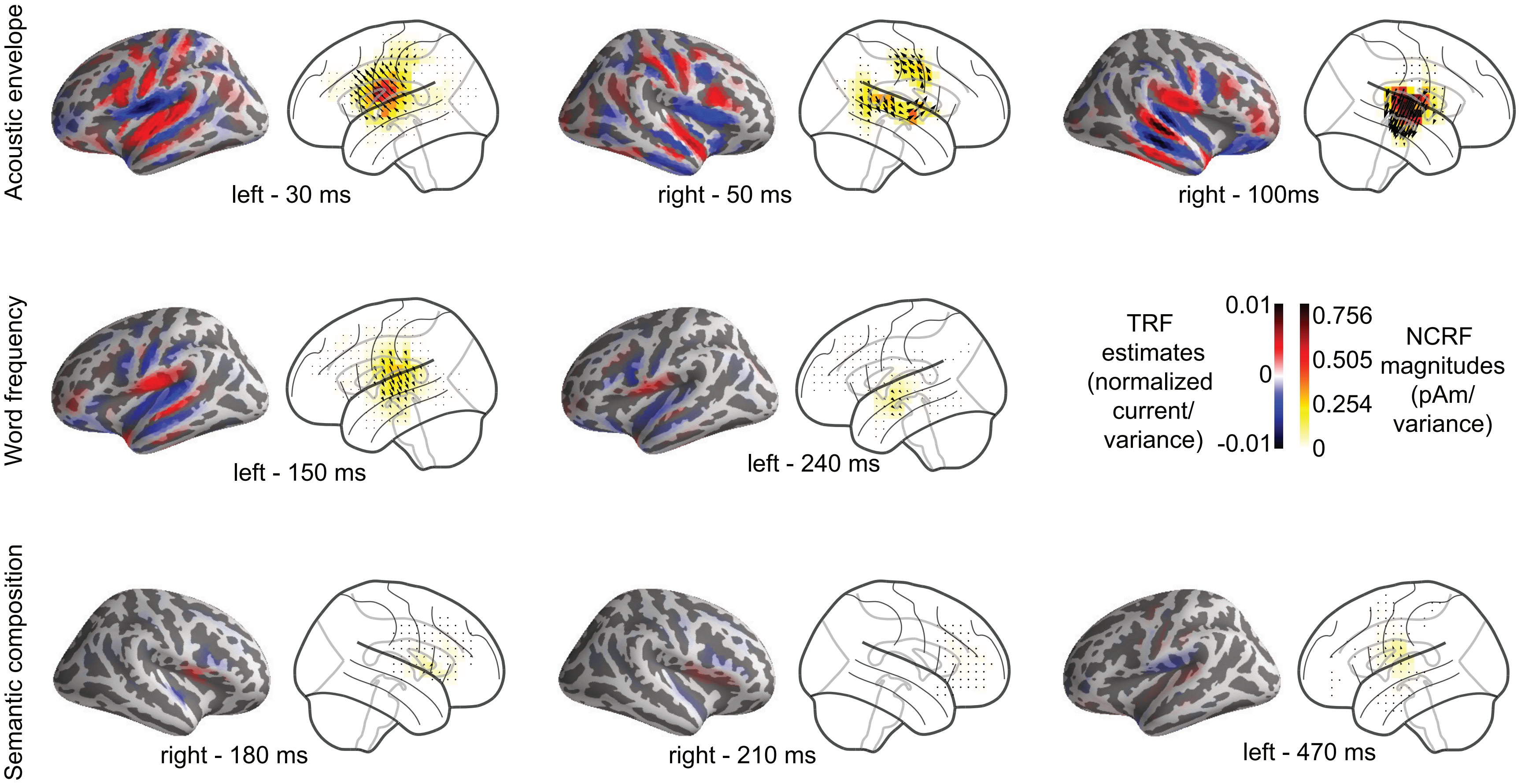


Figure 4.8: Comparison of cortical spread of estimated NCRFs against MNE-boosting TRFs. The pairwise anatomical plots show the group-level average MNE-boosting TRFs from Brodbeck et al. [22] (left) alongside the group-level average NCRFs projected onto the lateral plane (right) for a few selected visually salient peaks in the temporal profiles, corresponding to acoustic envelope (top), word frequency (middle) and semantic composition (bottom) feature variables.

and frontal cortices [182, 183, 184, 185, 186, 187]. To probe the functional organization of this hierarchy, we estimated NCRFs corresponding to features extracted from speech at the acoustic, lexical and semantic levels and found distinct patterns of cortical processing at high spatiotemporal resolutions. Our results indeed imply that while the acoustic and lexical features are processed primarily within the temporal and motor cortical regions [121, 188, 189, 190, 191, 192, 193], phrase-level processing, assessed here using the semantic composition variable, is carried out through the involvement of the frontal cortex [194, 195, 196].

Another advantage of our proposed methodology is mitigating the dependence of the solution on the precise geometry of the underlying cortical source models. In conventional neuromagnetic source imaging, individual structural MR images are utilized in the construction of source space models, particularly for retrieving the cortical surface segmentation. The normal direction to the so-called cortical patches in these models is key in determining the lead-field matrix, which are often referred to as orientation-constrained source models. However, in many available neuroimaging datasets (including the one analyzed in here), MR images are not available, relying only on an average head model, instead of one informed by the subject-specific cortical geometry. In order to mitigate the need for such information, we utilized a free-orientation volumetric source space in our estimation framework. While this makes the underlying optimization problem more involved and computationally intensive, it adds more than a compensatory amount of flexibility to the underlying models and allows them to recover missing information regarding the cortical source space geometry. To this end, we used rotationally invariant sparsity-inducing priors to regularize the spatiotemporal distribution of the NCRFs. Together with the aforementioned data-driven source covariance adaptation, this regularization scheme results in consistent source orientation estimates and provides a degree of immunity to unwanted side-effects of error-prone coordinate-frame rotations. To confirm these theoretical expectations, we validated this feature of our framework using simulation studies with known ground truth. In light of the above, posing NCRF estimation over an orientation-free volumetric source space can also be thought of as unifying the virtues of distributed source imaging and single dipole fitting: we aim at estimating both the orientations and magnitudes of spatially sparse dipole currents within the head volume that can best linearly predict the MEG responses to continuous stimuli.

This flexibility encourages applications of Champ-Lasso algorithm beyond MEG, for example, to EEG or simultaneous M/EEG recordings. In theory, any source localization method is equally applicable to all such scenarios (albeit with varying performance, due to the intrinsic differences between MEG and EEG), once the lead-field matrix is computed precisely. The main challenge is thus the placement of the current dipoles over the cortical mantle and correctly inferring the orientation of the dipoles from the structural MR scans. Unfortunately, a large majority of EEG experiments do not contain structural MR scans, eliminating the possibility of precise source-space analysis. Our analysis pipeline could be particularly useful for these scenarios, as the particular formulation aims to eliminate this strict requirement on dipole placements by making the solution robust against the unavailability of the precise geometry of the cortical mantle. The favorable performance of the

Champ-Lasso algorithm in application to MEG data gives promise of its utility in application to EEG or simultaneous M/EEG recordings, which would still need to be verified in future studies.

To facilitate such verification as well as usage by the broader systems neuroscience community a python implementation of the Champ-Lasso algorithm is archived on the open source repository Github [174]. The current implementation of our algorithm uses the aforementioned regularization scheme to recover temporally smooth and spatially sparse NCRFs. Due to the plug-and-play nature of the proposed Bayesian estimation framework, one can easily utilize other relevant regularization schemes to promote spatial smoothness or incorporate spectro-temporal prior information, by just modifying the penalty term.

Part III

Granger Causal Inference from Sparse Autoregressive Models

Chapter 5: Non-Asymptotic Guarantees for Robust Identification of Granger Causality via the LASSO

Reliable identification of causal influences is one the central challenges in time-series analysis, with implications for various domains such as economics [197], neuroscience [25, 26, 27] and computational biology [198, 199]. Granger causal (GC) characterization of time-series is among the widely used methods in this regard. This framework was pioneered by Granger [28], with subsequent key generalizations provided by Geweke [29, 30]. The notion of GC influence pertains to assessing the improvements in predicting the future samples of one time-series by incorporating the past samples of another one.

While causality, as the relationship between cause and effect, is a philosophically well-defined concept, it eludes a universal definition in empirical sciences and engineering. Granger causality is one of many definitions used in time series models [see 200, 201, for other notions], with an explicit data-driven form that admits statistical testing. The stochastic nature of the time series model, i.e., the uncertainty and the direction of time flow are the central features of GC definition. In principle, given two time-series x_t and y_t, one asserts that y_t has a GC influence on x_t when the posterior densities $p(x_t|x_{t-1}, x_{t-2}, \cdots, y_{t-1}, y_{t-2}, \cdots)$ and $p(x_t|x_{t-1}, x_{t-2}, \cdots)$ differ significantly. However, estimating these posterior densities from the observed data is a difficult task in general, and requires additional modeling assumptions. A popular set of such assumptions pertains to parametric multi-variate auto-regressive (MVAR) models along with certain distributional specifications (e.g., zero-mean Gaussian process noise). In these models, the aforementioned posterior densities can be fully characterized by the estimates of parameters and prediction error variances. As a consequence, one first aims at predicting x_{t+1} by a linear combination of the joint past observations $\{x_t, \cdots x_0\}$, $\{y_t, \cdots y_0\}$ (i.e., the *full* model), followed by repeating this task by excluding the past observations of y_t (i.e., the *reduced* model). If the prediction error variance in the former case is significantly smaller than the latter, we say that y_t has a GC influence on x_t.

Conventionally, the optimal linear predictors are obtained by the ordinary least squares (OLS), and the model orders are determined by the AIC [31] or BIC [32] procedures. Then, the GC measure is defined as the logarithmic ratio of the two prediction error variances, and its statistical significance is assessed based on the corresponding asymptotic distributions [33, 34, 202]. While the aforementioned procedure is relatively simple to carry out, it faces two key challenges. First, in order to obtain reliable MVAR parameter estimates via OLS, a relatively long observation horizon is required. In datasets with small sample size (e.g., gene expression data [203]), the regression models typically over-fit the observed data, causing both parameter estimation and model order selection to break down [27, 35]. In addition, AIC/BIC may restrict the order of the MVAR in a way that the resulting model fails to capture the complex and long-range dynamics of the underlying couplings

[25, 36]. Secondly, correlated process noise arising from latent processes, may lead to misidentification of GC influences, which is often referred to as the confounding effect [37].

These challenges have been successfully addressed in the context of regularized MVAR estimation [204, 205, 206, 207, 208, 209, 210, 211, 212]. In particular, the theory of sparse estimation via the LASSO [213, 214, 215, 216] provides a principled methodology for simultaneous parameter estimation and model selection in high dimensional MVAR models [205, 206, 207, 208, 209, 210, 217, 218, 219]. In addition, the Oracle property of the LASSO in presence of correlated noise ensures robust recovery of the set of MVAR parameters arising from the direct causal influences while discarding any spurious couplings due to correlated process noise, thus alleviating the confounding effect. The LASSO and its variants have already been utilized in existing work to identify graphical GC influences based on the recovered sparsity patterns [38, 39, 40, 220, 221, 222]. These methods construct the GC graph based on the estimated model parameters, either directly [38] or by appropriate thresholding [39, 40] to control false positive errors. This idea has even been extended to time series models that account for nonlinear dynamics using structured multilayer perceptrons or recurrent neural networks [223]. Another related class of results uses de-biasing techniques in order to construct confidence intervals and thereby identify the significant causal interactions [41, 42, 224, 225, 226, 227, 228]. There is, however, an evident disconnect between these LASSO-based approaches and the classical OLS-based GC inference: while the LASSO-based approaches aim at identifying the GC effects based on consistent estimates of the parameters in the non-asymptotic regime, the classical GC methodology relies on the comparison of the prediction errors across the *full* and *reduced* models by resorting to asymptotic distributions. There exist a slew of partial correlation based nonparametric methods that employ conditional independence tests for causality detection [229, 230, 231], thus avoiding time-series modeling assumptions altogether.

In this chapter, we close the gap between currently available LASSO-based approaches and the classical OLS-based GC inference by unifying these two approaches via introducing a new LASSO-based GC statistic that resembles the classical GC measure, and by leveraging the consistency properties of the LASSO to characterize the non-asymptotic properties of said GC statistic. In particular, we consider a canonical bivariate autoregressive (BVAR) process with correlated process noise. We then propose a likelihood-based scaled F-statistic as the relevant GC statistic, which we call the LGC statistic, and study its non-asymptotic properties under both the presence and absence of a GC influence. Our analysis reveals that the well-known sufficient conditions of the LASSO for stable BVAR estimation are also sufficient for accurate detection of the GC influences, if the strength of the causal effect satisfies additional mild conditions. Furthermore, by slightly weakening these sufficient conditions, we characterize the false positive error probability of a simple thresholding scheme for identification of GC influences.

We present simulation studies to compare the performance of the classical OLS-based and the proposed LGC-based approaches in detecting GC influences, in order to demonstrate the validity of our theoretical claims and to explore the key underlying trade-offs. We also present an application to experimentally-recorded neural data from general anesthesia to assess the causal role of the local field potential (LFP) on spiking activity. Our results based on LGC analysis corroborate existing hypotheses on the causal role of LFP in mediating local spiking activity, whereas these effects are concealed by the classical GC analysis due to significant over-fitting.

In summary, our main contribution is to extend the theoretical results of the LASSO to the classical characterization of GC influences, and to identify the key trade-offs in terms of sampling requirements and strength of the causal effects that result in robust GC identification. The rest of this chapter is organized as follows: Section 5.1 provides background and our problem formulation. Our main theoretical contributions are given in Section 5.2. Section 5.3 presents application to simulated and experimentally-recorded neural data, followed by our concluding remarks in Section 5.4.

5.1 Background and Problem Formulation

5.1.1 Granger Causality in a Canonical BVAR Regression Model

Consider finite-duration observations from two time-series x_t and y_t, given by $\{x_t, y_t\}_{t=-p+1}^{n}$, where n is the sample size and p is the order. The BVAR(p) model can be expressed as:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \mathbf{A}_1 \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \mathbf{A}_2 \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \cdots + \mathbf{A}_p \begin{bmatrix} x_{t-p} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \epsilon'_t \end{bmatrix}, \tag{5.1}$$

with $\mathbf{A}_i \in \mathbb{R}^{2\times 2}$ for $i \in \{1, 2, \cdots, p\}$ denoting the BVAR parameters and $[\epsilon_t, \epsilon'_t]^\top$ denoting the process noise with known distribution. It is commonly assumed that $[\epsilon_t, \epsilon'_t]^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$. Using this BVAR($p$) model and considering $\{x_t, y_t\}_{t=-p+1}^{0}$ as the initial condition, one can form a prediction model of x_t as follows:

$$\mathbf{x} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{5.2}$$

where the response $\mathbf{x}$, regressors $\mathbf{X}$, and residuals $\boldsymbol{\epsilon}$ are defined as:

$$\mathbf{x} = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_1 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{n-1} & \cdots & x_{n-p} & y_{n-1} & \cdots & y_{n-p} \\ x_{n-2} & \cdots & x_{n-p-1} & y_{n-2} & \cdots & y_{n-p-1} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ x_0 & \cdots & x_{-p+1} & y_0 & \cdots & y_{-p+1} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_n \\ \epsilon_{n-1} \\ \vdots \\ \epsilon_1 \end{bmatrix} \tag{5.3}$$

The regression coefficients $\boldsymbol{\theta}$ consist of $2p$ parameters: $\{\theta_i\}_{i=1}^{p}$, representing the autoregression coefficients obtained from $(\mathbf{A}_i)_{1,1}$, $i = 1, 2, \cdots, p$ and $\{\theta_i\}_{i=p+1}^{2p}$ representing the cross-regression coefficients obtained from $(\mathbf{A}_i)_{1,2}$, $i = 1, 2, \cdots, p$. Hereafter, we denote the true coefficients by $\boldsymbol{\theta}^* \in \mathbb{R}^{2p}$. Also, for a generic coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^{2p}$, the corresponding auto- and cross-regression components are denoted by $\boldsymbol{\theta}_{(1)} \in \mathbb{R}^p$ and $\boldsymbol{\theta}_{(2)} \in \mathbb{R}^p$, respectively, i.e., $\boldsymbol{\theta} =: [\boldsymbol{\theta}_{(1)}; \boldsymbol{\theta}_{(2)}]$.

The GC influence of y_t on x_t can then be assessed via hypothesis testing, with the null hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$. For testing, one considers the following

BVAR(p) models:

$$\text{Full Model: } \mathbf{x} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \text{Reduced Model: } \mathbf{x} = \mathbf{X}\widetilde{\boldsymbol{\theta}} + \widetilde{\boldsymbol{\epsilon}}, \text{ with } \widetilde{\boldsymbol{\theta}}_{(2)} = \mathbf{0}. \tag{5.4}$$

In other words, in the *full* model, all columns of $\mathbf{X}$ are used to estimate $\mathbf{x}$, but in the *reduced* model, only the first p columns are used. The conventional GC measure [30] is then defined as the logarithmic ratio of the residual variances: $\mathcal{F}_{y \mapsto x} := \log\left(\mathrm{var}(\widetilde{\boldsymbol{\epsilon}})/\mathrm{var}(\boldsymbol{\epsilon})\right)$. Note that when the residuals are Gaussian, $\mathcal{F}_{y \mapsto x}$ is the log-likelihood ratio statistic. Given that the *reduced* model is nested within the *full* model, we have $\mathcal{F}_{y \mapsto x} \geq 0$.

In order to compute $\mathcal{F}_{y \mapsto x}$ from the time-series data, empirical residual variances are used based on OLS parameter estimates under both models [197]:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{OLS}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \frac{1}{n} \left\| \mathbf{x} - \mathbf{X}\boldsymbol{\theta} \right\|^2, \quad \widehat{\widehat{\boldsymbol{\theta}}}_{\mathrm{OLS}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}_{(2)}=\mathbf{0}}{\operatorname{argmin}} \ \frac{1}{n} \left\| \mathbf{x} - \mathbf{X}\boldsymbol{\theta} \right\|^2. \tag{5.5}$$

The estimated $\mathcal{F}_{y \mapsto x}$ is a random variable over $\mathbb{R}_{\geq 0}$, and typically has a non-degenerate distribution. Thus, a non-zero $\mathcal{F}_{y \mapsto x}$ does not necessarily imply a GC influence. To control for false discoveries, the well-established results on the asymptotic normality of maximum likelihood estimators can be utilized: under mild assumptions, $n\mathcal{F}_{y \mapsto x}$ converges in distribution to a chi-square χ^2_p with degree p. In addition, under a sequence of local alternatives $H^n_{y \mapsto x,1} : \boldsymbol{\theta}^*_{(2)} = \boldsymbol{\delta}/\sqrt{n}$, for some constant vector $\boldsymbol{\delta}$, $n\mathcal{F}_{y \mapsto x}$ converges in distribution to a non-central chi-square $\chi^2_p(\nu)$ with degree p and non-centrality $\nu > 0$ [33, 34]. These asymptotic results lead to a simple thresholding strategy: rejecting the null hypothesis if $\mathcal{F}_{y \mapsto x}$ exceeds a fixed threshold. A key consideration in this framework is choosing the model order p. To this end, criteria such as the AIC [31] and BIC [32] are widely used to strike a balance between the variance accounted for and the number of coefficients to be estimated.

While the foregoing procedure works well in practice for large sample sizes, its performance sharply degrades as the sample size decreases. This performance degradation has two main reasons:

1. The regression models become under-determined and result in poor estimates of the parameters, and

2. The conventional model selection criteria fail to capture possible long-range temporal coupling of the underlying processes.

As a result, the classical GC measure is highly susceptible to over-fitting. In addition, when the process noise elements ϵ_t and ϵ'_t are highly correlated, the OLS estimates incur additional error in capturing the true BVAR parameters, and hence result in mis-detection of the GC influences. While some existing non-parametric methods aim at entirely bypassing MVAR estimation by utilizing spectral matrix factorization [232] or multivariate embedding [233] for system identification, they are similarly prone to the adverse effects of small sample size.

5.1.2 LASSO-based Causal Inference in the High-Dimensional Setting

In the so-called high-dimensional setting, where the model dimension becomes comparable to or even exceeds the sample size, regularization schemes are employed to guard against over-fitting. These schemes include Tikonov regularization [234, 235], ℓ_1-regularization or the LASSO [213, 214, 215, 216], smoothly clipped absolute deviation [236, 237], Elastic-Net [238], and their variants, and have particularly proven useful in MVAR estimation [204, 205, 206, 207, 211, 212]. Among these techniques, the LASSO has been widely used and studied in the high-dimensional sparse MVAR setting, under fairly general assumptions [205, 206, 207]. By augmenting the least squares error loss with the ℓ_1-norm of the parameters, the LASSO simultaneously guards against over-fitting and provides automatic model selection [216, 217, 218, 219, 239, 240], under the hypothesis that the true parameters are sparse. In the context of MVAR estimation, assuming that the time-series data admit a sparse MVAR representation, the LASSO estimates enjoy tight bounds on the estimation and prediction errors under suitable sample size requirements, even for models with correlated noise [209, 210].

By leveraging the foregoing properties, the LASSO and its variants have been utilized in existing work to identify GC influences in a graphical fashion [38, 39, 40, 220, 221, 222]. These approaches construct the GC graph either directly from the estimated coefficients (e.g., [38]) or by appropriate thresholding (e.g., [39, 40]) to control false positive errors. Alternatively, de-biasing techniques have been introduced for constructing confidence intervals over the estimated parameters and thereby identifying the significant causal effects [41, 42, 224, 225, 226, 227, 228].

5.1.3 Unifying the Classical OLS-based and LASSO-based Approaches

Comparing the classical OLS-based and the recent LASSO-based approaches to causal inference reveals an evident disconnect: the latter approach directly utilizes the estimated parameters in a single model to identify the GC influence with non-asymptotic performance guarantees, while the former is based on comparing the prediction performance of two different models by resorting to asymptotic distributions for statistical testing.

Our main goal here is to close this gap by unifying these two approaches. To this end, we first replace OLS estimation in Eq. (5.5) by its LASSO counterpart:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \frac{1}{n}\left\|\mathbf{x}-\mathbf{X}\boldsymbol{\theta}\right\|^2 + \lambda_n\|\boldsymbol{\theta}\|_1, \quad \widehat{\widehat{\boldsymbol{\theta}}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}_{(2)}=\mathbf{0}}{\operatorname{argmin}} \ \frac{1}{n}\left\|\mathbf{x}-\mathbf{X}\boldsymbol{\theta}\right\|^2 + \lambda_n\|\boldsymbol{\theta}\|_1, \tag{5.6}$$

where λ_n denotes the regularization parameter. Let $\ell(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}) := \frac{1}{n}\left\|\mathbf{x}-\mathbf{X}\boldsymbol{\theta}\right\|^2$ with $\boldsymbol{\theta} = [\boldsymbol{\theta}_{(1)}; \boldsymbol{\theta}_{(2)}]$. By similarly grouping the solutions of Eq. (5.6) as $\widehat{\boldsymbol{\theta}} = [\widehat{\boldsymbol{\theta}}_{(1)}; \widehat{\boldsymbol{\theta}}_{(2)}]$ and $\widehat{\widehat{\boldsymbol{\theta}}} = [\widehat{\widehat{\boldsymbol{\theta}}}_{(1)}; \mathbf{0}]$, we then propose to use the following statistic:

$$\mathcal{T}_{y\mapsto x} := \frac{\ell\left(\widehat{\widehat{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right)}{\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right)} - 1 = \frac{\ell\left(\widehat{\widehat{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right) - \ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right)}{\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right)}, \tag{5.7}$$

akin to a scaled likelihood-based version of the F-statistic, which we call the LASSO-based GC (LGC) statistic. Note that the LGC statistic can be related to the conventional GC statistic as $\mathcal{T}_{y \mapsto x} = \exp(\mathcal{F}_{y \mapsto x}) - 1$, when $\lambda_n = 0$. Therefore, it is expected for $\mathcal{T}_{y \mapsto x}$ to be near 0 under the null hypothesis. One advantage of using this statistic is that a simple thresholding strategy, similar to that used for the classical GC statistic, can be used to reject the null hypothesis $H_{y \mapsto x,0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$. In the next section, we will characterize the non-asymptotic properties of the LGC statistic and seek conditions that allow us to distinguish between the null (i.e., absence of a GC effect) and a suitably defined alternative (i.e., presence of a GC effect) hypothesis.

5.2 Main Theoretical Results

Our main theoretical contribution in this section is to characterize $\mathcal{T}_{y \mapsto x}$ under both the null and a suitably chosen alternative hypothesis, and establish sufficient conditions that guarantee distinguishing these hypotheses with high probability. We will also analyze the false positive error probability corresponding to the aforementioned thresholding strategy, under slightly weakened sufficient conditions. The latter result can be used to obtain suitable thresholds in practice, as we will demonstrate in Section 5.3.

Before presenting the main results, we state our key assumptions:

Assumption 5.I. We assume $\{x_t, y_t\}_{t=-p+1}^{n}$ to be a part of a realization of zero-mean bivariate stationary process[†] that admits a stable and invertible BVAR(p) representation, with a zero-mean i.i.d. Gaussian process noise with positive definite

[†] i.e., the initial condition is such that the samples under consideration attained the stationary distribution.

covariance $\mathbf{\Sigma}_{\epsilon}$. We further assume that $\|\boldsymbol{\theta}^*\|_0 = k$, i.e., $\boldsymbol{\theta}^*$ is k-sparse (See Section C.1 for more details and discussion).

Our main theorem can be stated as follows:

Theorem 5.1 (Main Theorem)**.** *Suppose that the key assumptions (5.I) hold. Then, for the proposed LGC statistic $\mathcal{T}_{y\mapsto x}$ evaluated at the BVAR(p) parameter estimates from the solutions of Eq. (5.6) with a regularization parameter $\lambda_n = 4\mathscr{A}\sqrt{\log(2p)/n}$, there exists a threshold that correctly distinguishes the null and the local alternative hypothesis $H^n_{y\mapsto x,1} : \|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathscr{B}k\log(2p)/n$ with probability at least $1 - K/p^d$, if $n \geq \max\{\mathscr{C}, \mathscr{D}k\}\log(2p)$, with $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}$, $\mathscr{D}$, K and $d > 0$ denoting constants that are explicitly given in the Appendix C.*

Proof Sketch. We present the proof sketch here, and defer the detailed proof to Section C.2. Under assumptions (5.I), it can be shown that the following conditions, adapted from [209], hold with high probability as long as $n = \mathcal{O}(k\log(2p))$ (See Section C.3):

Condition C1 (Restricted eigenvalue (RE) condition)**.** The symmetric matrix $\widehat{\mathbf{\Sigma}} = \mathbf{X}^\top\mathbf{X}/n \in \mathbb{R}^{2p\times 2p}$ satisfies restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau \geq 0$, i.e., $\widehat{\mathbf{\Sigma}} \sim \mathrm{RE}(\alpha, \tau)$:

$$\boldsymbol{\phi}^\top\widehat{\mathbf{\Sigma}}\boldsymbol{\phi} \geq \alpha\|\boldsymbol{\phi}\|_2^2 - \tau\|\boldsymbol{\phi}\|_1^2, \ \forall \ \boldsymbol{\phi} \in \mathbb{R}^{2p},$$

with $\tau := \frac{m-1}{m}\frac{\alpha}{32k}$ for some constant $m > 1$.

Condition C2 (Deviation condition)**.** There exist deterministic functions $\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$ and $\mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$ such that

$$\left\| \frac{1}{n} \mathbf{X}^\top (\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*) \right\|_\infty \leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon) \sqrt{\frac{\log(2p)}{n}}, \text{ and}$$
$$\left\| \frac{1}{n} \mathbf{X}_{(1)}^\top \left(\mathbf{x} - \mathbf{X}_{(1)} \widetilde{\boldsymbol{\theta}}^*_{(1)} \right) \right\|_\infty \leq \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon) \sqrt{\frac{\log(2p)}{n}},$$

where $\mathbf{X}_{(1)}$ denotes the first p columns of $\mathbf{X}$, and $\widetilde{\boldsymbol{\theta}}^*_{(1)}$ is a suitably defined surrogate of the true parameters under the *reduced* model (See Section C.1 for details).

Conditions (C1) and (C2) guarantee the consistency of the LASSO estimates with the given choice of λ_n, which allows us to obtain the following deviation inequalities for $\ell(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)})$ and $\ell(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0})$, under the *full* and *reduced* models, respectively:

$$\left| \ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) \right| \leq \Delta_F \quad \text{and} \quad \left| \ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right) - \ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) \right| \leq \Delta_R, \tag{5.8}$$

where Δ_R and Δ_F are quantities explicitly given in Section C.2. Also, it can be shown that:

$$\frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_R - \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) + \Delta_F} \leq \mathcal{T}_{y \mapsto x} \leq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) + \Delta_R + \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_F}. \tag{5.9}$$

Under $H_{y \mapsto x, 0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$, we can simplify the upper bound on $\mathcal{T}_{y \mapsto x}$ by the facts that

$\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) = \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right)$ and $\Delta_R = \Delta_F$, to get:

$$\mathcal{T}_{y\mapsto x} \leq \frac{\Delta_R + \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_F} \leq \frac{2\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} - \Delta_N - \Delta_F}, \tag{5.10}$$

with probability at least $1 - 2\exp\left(-n\Delta_N^2/8(\boldsymbol{\Sigma}_\epsilon)^2_{1,1}\right)$, for some constant $\Delta_N > 0$ to be specified. On the other hand, under a general alternative $H_{y\mapsto x,1} : \boldsymbol{\theta}^*_{(2)} \neq \mathbf{0}$, we can show that:

$$\mathcal{T}_{y\mapsto x} \geq \frac{D - (\Delta_D + \Delta_R + \Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + \Delta_N + \Delta_F}, \tag{5.11}$$

with probability at least $1 - 2/(2p)^{c_{11}} - 2\exp\left(-n\Delta_N^2/8(\boldsymbol{\Sigma}_\epsilon)^2_{1,1}\right)$, where D and Δ_D are deterministic quantities that are given in Section C.4, and c_{11} is a constant explicitly given in Lemma C.8. In order to be able to distinguish the null and alternative hypotheses, we show that it is sufficient to restrict the alternative hypothesis to take a local form $H^n_{y\mapsto x,1} : \|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathcal{B}k\log(2p)/n$. Finally, by incorporating the probability that conditions (C1) and (C2) hold, the claim of the theorem can be shown to hold with probability at least $1 - K/p^d$ for some constants K and $d > 0$. □

By slightly weakening the sufficient condition $n \geq \mathcal{D}k\log(2p)$ in Theorem 5.1, we arrive at the following corollary that upper bounds the false positive error probability:

Corollary 5.1.1 (False Positive Error Probability)**.** *Suppose that assumptions (5.I) and conditions (C1) and (C2) in the proof of Theorem 5.1 hold. Then, for any* $t_0 > 0$, *thresholding the proposed LGC measure* $\mathcal{T}_{y\mapsto x}$ *at a level* $\mathfrak{t} > 0$ *for re-*

jecting the null hypothesis results in a false positive error probability of at most $2\exp\left(-n\Big/8\left(1+\gamma t_0\sqrt{\log(2p)/n}\right)^2\right)$ *with* $\gamma := (\mathfrak{t}+2)/\mathfrak{t}$, *if*

$$n \geq \left(\widetilde{\mathfrak{D}}/t_0\right)^2 k^2 \log(2p) + 2\widetilde{\mathfrak{D}}\gamma k \log(2p),$$

for some constant $\widetilde{\mathfrak{D}}$ *that is explicitly given in the Appendix C.*

Proof Sketch. Note that under conditions (C1) and (C2) there exists some real numbers $s, t > 0$ such that,

$$\left|\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - (\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/s \quad \text{and} \quad \Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}t/s. \tag{5.12}$$

Then the upper bound on $\mathcal{T}_{y\mapsto x}$ under the null hypothesis given in Eq. (C.8) simplifies to:

$$\mathcal{T}_{y\mapsto x} \leq \frac{2t/s}{1-(1+t)/s}, \tag{5.13}$$

which allows us to set a problem independent threshold. Now, for any threshold $\mathfrak{t} > 0$, we can solve for s in terms of t and $\mathfrak{t}$ as:

$$s = 1 + \frac{2+\mathfrak{t}}{\mathfrak{t}}t. \tag{5.14}$$

With a choice of $t = t_0\sqrt{\log(2p)/n}$, the inequalities in Eq. (5.12) provide the false positive error probability expression and sampling requirement, respectively, to give the statement of the corollary. The detailed proof is given in Section C.2. □

To discuss the implications of these results, some remarks are in order:

Remark 5.1. Intuitively, detecting a GC effect arising from a small cross-regression coefficient $\boldsymbol{\theta}^*_{(2)}$ is challenging, and often requires a long observation horizon to be identified. Theorem 5.1 quantifies this intuition via a lower bound on the norm of the cross-regression coefficients in terms of the spectral properties of the process (via $\mathscr{B}$), sparsity k, sample size n and model order p. In particular, as $\|\boldsymbol{\theta}^*_{(2)}\|_2 \to 0$, a scaling of $n = \mathcal{O}(k \log(2p)/\|\boldsymbol{\theta}^*_{(2)}\|_2^2)$ maintains the sensitivity/specificity of $\mathcal{T}_{y \mapsto x}$ with high probability. The lower bound on $\|\boldsymbol{\theta}^*_{(2)}\|_2$ exhibits the same scaling as that in the thresholding procedure of [39] (i.e., the scaling of the LASSO estimation error), as well as the classical scaling of [33, 34] (up to logarithmic factors), and we thus believe is not significantly improvable.

Remark 5.2. Unlike the conventional estimation error results of the LASSO that specify a lower bound on the regularization parameter λ_n, Theorem 5.1 prescribes a fixed choice of λ_n for both the *full* and *reduced* estimation problems in Eq. (5.6). This is due to an interesting phenomenon revealed by our analysis: while conventional analyses of LASSO focus on the estimation performance of a single model and thus provide a lower bound on λ_n, in our framework we have two competing models (i.e., *full* and *reduced*) which need to be distinguishable under the null and alternative hypotheses in order to reliably detect the GC influences. The latter imposes an *upper* bound on λ_n. As such, there is a suitable interval for choosing λ_n that results in both consistent estimation and discrimination of the two models. We have presented our results using a single λ_n in this interval for both models, which is appealing from the user's perspective in practice, where cross-validation is often used for tuning λ_n. The user can select λ_n via cross-validation in solving the LASSO problem for the full model, and then use the same value of λ_n for the reduced model, thus avoiding extra computational costs of cross-validation.

Remark 5.3. Corollary 5.1.1 bounds the false positive error probability, i.e., Type I error rate, for a simple thresholding scheme for detecting GC influences from $\mathcal{T}_{y \mapsto x}$, under a slightly weakened sufficient condition on n, i.e., $n = \mathcal{O}(k^2 \log(2p))$ instead of $n = \mathcal{O}(k \log(2p))$. This non-asymptotic result provides a principled guideline for choosing a threshold that controls the false positive error rate. As such, this result extends the conventional statistical testing framework based on the asymptotics of log-likelihood ratio statistic using OLS to the non-asymptotic setting using the LASSO.

Remark 5.4. We have presented our results for a BVAR model in order to parallel the classical GC analysis. Our results can be extended to the general MVAR setting by using the *conditional* notion of Geweke [30] in a natural fashion, given that conditions (C1) and (C2) readily generalize to this setting.

Remark 5.5. The constants in the proof of Theorem 5.1 and Corollary 5.1.1, including the ones written in calligraphic letters solely depend on the joint spectrum of processes x_t, y_t as well as some absolute constants. As an illustrative example, by assuming $\mathbf{\Sigma}_\epsilon = 0.01\mathbf{I}$, $\mu_{\max}(\mathbf{A}) = 0.9$, $\mu_{\min}(\mathbf{A}) = \mu_{\min}(\breve{\mathbf{A}}) = 0.01$, $\widetilde{\Lambda}_{\min} = 0.7$, $\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2 = 0.5$, $\|[\mathbf{C}_{11}^{-1}\mathbf{C}_{12};\mathbf{I}]\boldsymbol{\theta}^*_{(2)}\|_2 = 1.5$, $d_0 = 4.2 \times 10^{-4}$, $C_0 = 10^{-6}$, and $c_2 = 10^5$, the key constants in Theorem 5.1 take the following numerical values:

$\mathscr{A} = 10^{-3}$, $\mathscr{B} = 9.84$, $\mathscr{C} = 236$, $\mathscr{D} = 43$, $K = 6$, and $d = 1$. These translate to $\lambda_n = 10^{-3}\sqrt{\log{(2p)}/n}$, a requirement of $n > \max\{236, 43k\}\log(2p)$, local alternative hypotheses satisfying $\|\boldsymbol{\theta}^*_{(2)}\|^2 > 9.84k\log{(2p)}/n$, and failure probability $< 6/p$. Similarly, for Corollary 5.1.1, we get $\widetilde{\mathscr{D}} = 10.67$, translating to a sample size requirement of $n > 0.046k^2\log(2p) + 393k\log(2p)$ (with $\mathscr{t} = 0.114$ and $t_0 = 100$). The potentially large numerical values of some of these constants suggest that the non-asymptotic advantage may come with large values of n and p.

5.3 Application to Simulated and Experimentally-Recorded Data

In this section, we evaluate our theoretical results through application to simulated and real data, and by comparing the performance of classical OLS-based method and the LGC statistics in detecting GC influences. We use the fast implementation in [241] to solve the LASSO problems. Unless otherwise stated, the regularization parameter λ_n is chosen via five-fold cross-validation performed over the *full* model, with the same λ_n used for the *reduced* model.

5.3.1 Simulation Studies

We simulated three time-series x_t, y_t, z_t according to the sparse MVAR(11) model:

$$x_t = -0.67x_{t-1} + 0.2x_{t-5} - 0.1x_{t-11} + 0.05z_{t-3} + \nu_{1,t}$$

$$y_t = -0.62y_{t-1} + 0.1y_{t-5} - 0.2y_{t-11} - 0.1x_{t-2} - 0.1x_{t-3} + 0.5x_{t-11}$$

$$-0.001z_{t-4} - 0.004z_{t-5} + \sqrt{0.6}\nu_{2,t}$$

$$z_t = -0.9025z_{t-2} + \nu_{3,t}$$

where $\nu_{i,t} \sim \mathcal{N}(0,1)$, i.i.d. for $i = 1, 2, 3$. In this model, x_t has a direct causal influence on y_t, but there is no causal influence from y_t to x_t. The *latent* process z_t, however, influences both x_t and y_t (Fig. 5.1(a)). As such, the correlated process noise components ϵ_t and ϵ'_t in Eq. (5.1) are modeled as $0.05z_{t-3} + \nu_{1,t}$ and $-0.001z_{t-4} - 0.004z_{t-5} + \sqrt{0.6}\nu_{2,t}$, respectively. As shown in Fig. 5.1(b), removing z_t from the analysis indeed induces a false (i.e., indirect) causal influence from y_t to x_t.

We performed two sets of comprehensive experiments to evaluate the effects of λ_n, n, and p on the identification of GC influences between x_t and y_t based on $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$:

Evaluating the Effect of λ_n. Fig. 5.1(c) shows the LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) for $n = 250$ and $p = 100$, obtained by varying λ_n in the interval

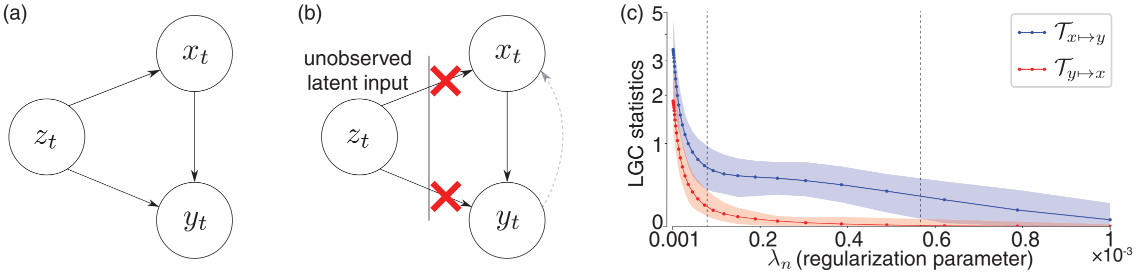


Figure 5.1: Simulation Results. (a) Ground truth causality pattern. (b) Estimation setup, in which z_t is latent and thus introduces a spurious causal link (dashed gray arrow) from y_t to x_t. (c) Effect of λ_n on the LGC statistics for $n = 250, p = 100$. The LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) are separable for a suitable range of λ_n, marked by the dashed vertical lines (colored hulls show the range of LGC over 30 realizations).

$[10^{-6}, 10^{-3}]$ uniformly in the log-scale. The dotted lines and colored hulls represent the average and range of the values, respectively, over 30 realizations. As discussed in *Remark 2*, there is an evident range of λ_n that provides a meaningful separation between $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$, which is marked by the dashed vertical lines in Fig. 5.1(c).

Evaluating the Effect of the Sample Size n. We fixed a model order of $p = 100$ and varied n uniformly in the interval $[100, 1000]$. Fig. 5.2(a) and (b) show the resulting LGC statistics $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$ corresponding to the LASSO and OLS ($\lambda_n = 0$), respectively. In Fig. 5.2(a), we also plotted the threshold t corresponding to a false positive probability of 0.01, according to Corollary 5.1.1 (dashed line). As n grows larger, the ranges of the two LGC statistics are saliently separated. The proposed thresholding rule of Corollary 5.1.1 is also able to correctly identify the true GC

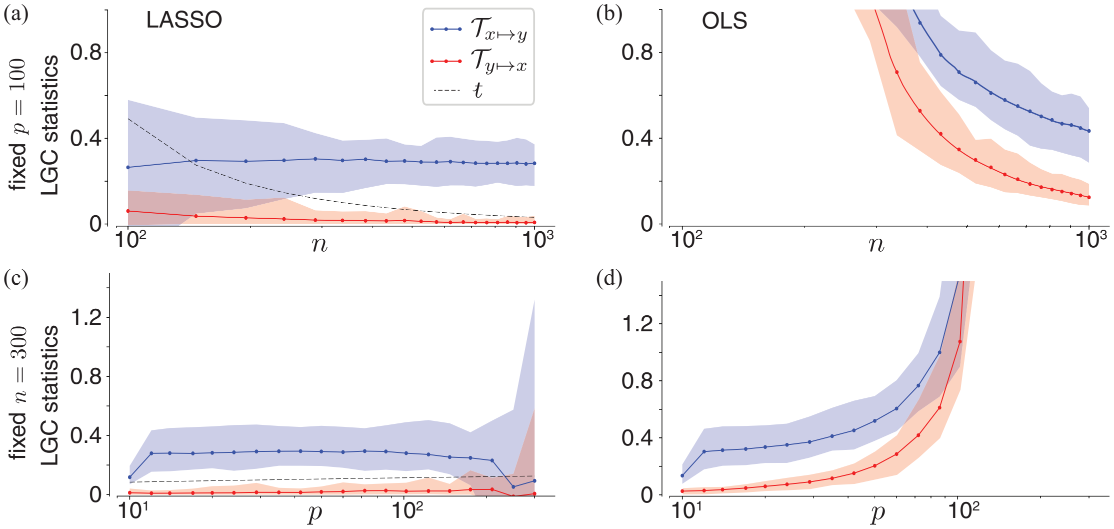


Figure 5.2: Simulation Results (continued). LASSO-based (left) and OLS-based ($\lambda_n = 0$, right) LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) obtained by varying n for fixed $p = 100$ (top panels (a) and (b)) and varying the model order p for fixed $n = 300$ (bottom panels (c) and (d)). The dashed lines in panels (a) and (c) show the threshold t at a false positive error level of 0.01 (colored hulls show the range LGC over 30 realizations).

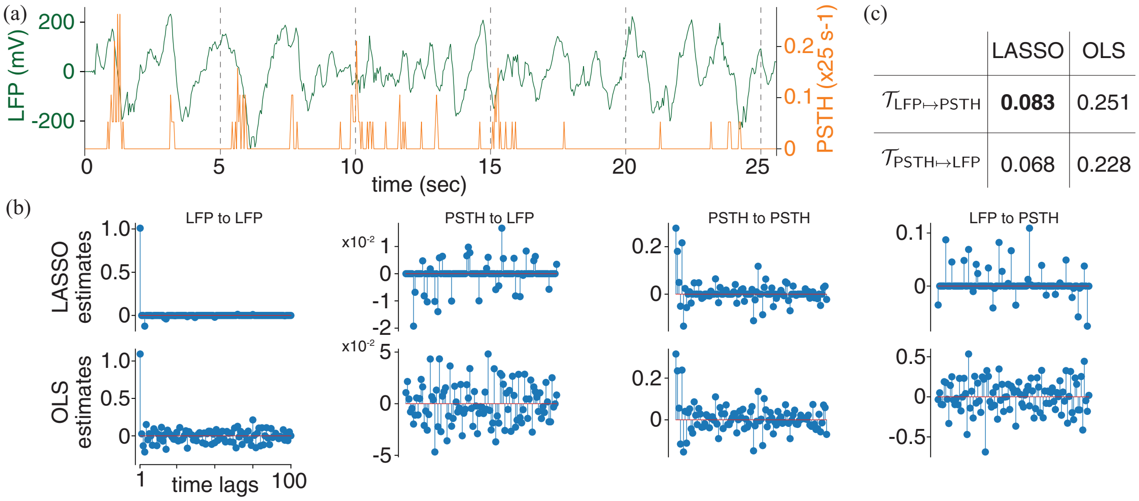
effects for $n \geq 250$.

The OLS results shown in Fig. 5.2(b), however, require much larger values of n to be stable, whereas the LGC measures provided by the LASSO (Fig. 5.2(a)) are stable even for $n < 2p$. In addition, OLS requires $n \geq 400$ for the ranges of the GC measures to be distinguishable.

Evaluating the Effect of the Model Order p. Finally, we fixed $n = 300$ and varied p in the interval $[10, 300]$ uniformly in the log-scale. Fig. 5.2(c) and (d) show the corresponding GC measures for the LASSO and OLS, respectively, along with the threshold t corresponding to a false positive probability of 0.01. For $p \ll n$, the LASSO and OLS exhibit similar performance. But, the OLS-based GC measures become unstable for $p \approx n$, whereas those of the LASSO remain stable throughout. The LGC statistics also remain saliently separable over a wider range of p for the LASSO, as compared to their OLS counterparts.

5.3.2 Application to Experimentally-Recoded Neural Data from General Anesthesia

Finally, we present an application to simultaneously recorded local field potential (LFP) and ensembles of single-unit spike recordings from a human subject under Propofol-induced general anesthesia (Data from [242]). The LFP signal, electrical field potential measured at the cortical surface, represents mesoscale dynamics of cortical activity with both cortical and subcortical (e.g., thalamic) origins. Single-unit spike recordings, on the other hand, represent the neuronal scale cortical dynam-



	LASSO	OLS
$\mathcal{T}_{\text{LFP}\mapsto\text{PSTH}}$	**0.083**	0.251
$\mathcal{T}_{\text{PSTH}\mapsto\text{LFP}}$	0.068	0.228

Figure 5.3: Analysis of neural data from general anesthesia. (a) LFP (green) and PSTH (orange) traces for a time window of duration 25.6 s. (b) BVAR parameter estimates corresponding to the *full* models for LASSO (top) and OLS (bottom). (c) Table summarizing the obtained LGC statistics. The LASSO-based measure detects a GC influence from LFP to the spiking activity (boldface number).

ics. It is believed that the spiking activity of cortical neurons is mediated by slow brain-wide oscillations under states such as anesthesia and sleep [104, 242, 243, 244]. Specifically, Lewis et al. [104] associate loss of consciousness under general anesthesia with emergence of periodic and profound suppression of neuronal spiking activity that is strongly phase-locked to the peaks of the LFP slow oscillations. Here, we examine the role of LFP slow oscillations in mediating the spiking activity by assessing the GC influences between them.

We use a time window of duration 25.6 s during anesthesia, corresponding to $n = 640$ samples (sampling frequency of 25 Hz). The multi-unit spiking activity is represented by its peristimulus time histogram (PSTH) (i.e., ensemble average over 19 units). Fig. 5.3(a) shows the LFP (green) and PSTH (orange) signals used in the analysis. We use a model order of $p = 100$, corresponding to a history length of

4 second, to ensure that slow oscillations ($\sim$0.25 Hz to 0.5 Hz) can be captured by the BVAR model.

Fig. 5.3(b) shows the estimated BVAR coefficients by the LASSO (top) and OLS (bottom). A visual comparison of the two sets of coefficients suggests that OLS has likely over-fitted the data. The corresponding LGC statistics $\mathcal{T}_{\mathsf{LFP} \mapsto \mathsf{PSTH}}$ and $\mathcal{T}_{\mathsf{PSTH} \mapsto \mathsf{LFP}}$ for both methods are reported in the table of panel Fig. 5.3(c). For a false positive error probability of 0.01, Corollary 5.1.1 prescribes a threshold $t =$ 0.0802, which results in detecting the GC effect $\mathsf{LFP} \mapsto \mathsf{PSTH}$ (boldface number) as significant, and discarding the GC effect $\mathsf{PSTH} \mapsto \mathsf{LFP}$. The conventional chi-square test applied to the classical GC statistics $\mathcal{F}_{\mathsf{LFP} \mapsto \mathsf{PSTH}}$ and $\mathcal{F}_{\mathsf{PSTH} \mapsto \mathsf{LFP}}$, however, fails to detect any GC influence, even at a significance level as high as 0.05. The outcome of the LASSO-based LGC analysis is therefore consistent with the aforementioned hypothesis on the role of LFP in mediating spiking activity.

5.4 Concluding Remarks

In this chapter, we proposed a GC statistic based on the LASSO parameter estimates, namely the LGC statistic, in order to identify GC influences in a canonical sparse BVAR model with correlated process noise. By analyzing the non-asymptotic properties of LGC statistic, we established that the well-known sufficient conditions for the consistency of LASSO also suffice for accurate identification of GC influences. By slightly weakening these conditions, we also analyzed the false positive error performance of a simple thresholding rule for detecting GC influences. We validated our theoretical claims through application to simulated and experimentally-recorded neural data from general anesthesia. In particular, we showed that the proposed LGC statistic is able to identify a GC effect from LFP to spiking activity under anesthesia, whereas the conventional OLS-based GC analysis does not detect this effect. Our contribution compared to existing literature is to provide a simple statistic inspired by the classical log-likelihood ratio statistic used for GC analysis, which can be directly computed from the LASSO estimates without the need to resort to de-biasing procedures or asymptotic results for testing. Future work includes extending our results to autoregressive generalized linear models with time-varying parameters.

Chapter 6: Concluding Remarks and Future Directions

6.1 Summary of Our Contributions

In this thesis, we devised novel statistical modeling frameworks and developed fast and efficient algorithms to analyze neuroimaging data from different modalities, ranging from the mesocale (i.e., M/EEG) to the neuronal level (i.e., singe/multi unit recordings), in order to investigate the mechanisms underlying brain function.

The first part of the dissertation concerned spectral analysis of spontaneous neural recordings. We developed a semi-parametric Bayesian framework, namely the DBMT method, for inferring the evolution of the spectral content in neuroimaging data, by integrating the multitaper method with state-space models. We also developed the PMTM algorithm, a multitaper method specifically tailored for neural spiking activity that takes into account the binary nature of the data, in order to analyze the spectra of underlying *latent* stationary process that govern spiking activity. We evaluated the performance of our proposed algorithms using synthetic and real data applications, and complemented the results with rigorous theoretical analyses.

In the second part, we focused on identifying the cortical sources responsible for the TRF components that give rise to the neural dynamics of speech processing manifested in M/EEG recordings. We introduced NCRFs for simultaneously determining the TRFs and their cortical distribution by unifying the TRF and distributed source localization models. We cast the joint estimation task as a Bayesian estimation problem and provided an efficient inference algorithm, namely the ChampLasso algorithm. Our simulation studies revealed significant improvements over existing two-stage methods, in terms of spatial resolution, filter reconstruction, and recovering dipole orientations. Application of our algorithm to experimentally recorded MEG responses provided new insight into the cortical encoding of various aspects of speech.

Finally, the third part of the dissertation considered the inference of GC influences in high-dimensional time series models with sparse coupling. We defined a LASSO-based statistic for inferring GC influences, which we referred to as the LGC statistic, and established non-asymptotic guarantees for robust identification of GC influences via the LGC statistic. Our theoretical and empirical analyses identified the key trade-offs in terms of sampling requirements and strength of the causal effects that result in the robust detection of GC influences, thus bridging the gap between the existing LASSO-based approaches and the classical least-squares based GC inference procedure.

All of the above methodologies take advantage of the Bayesian modeling framework in order to both employ arbitrarily complex models and learn them from observed data through solving optimization problems. Furthermore, the complexity of the models are tuned via incorporating domain knowledge in the form of prior distributions or regularization mechanisms. As demonstrated in this dissertation, the resulting learned models can be quite useful in providing interpretable and biologically-plausible descriptions of the underlying principles that govern neural activity under various tasks and conditions.

6.2 Future Directions

In closing, it is worth pointing out some of the potential future directions of research enabled by the results on this dissertation:

1. Integrating state-space models with more general notions of non-stationary spectra, such as the Generalized Evolutionary or Weyl spectra, in order to further improve the achievable spectrotemporal resolutions.

2. Developing a Bayesian framework to extracting a single set of group level NCRFs, with individual variations in component lags and amplitudes captured via appropriate distributions.

3. Exploiting the low-dimensional nature of the NCRFs to directly estimate the dominant factors and corresponding spatial distribution patterns, which in tandem describe cortical encoding of speech. One can potentially recover such factors at different time scales by utilizing Gaussian process priors.

4. Extending the LGC guarantees beyond stationary linear models, by considering generalized linear models or AR models with time-varying parameters.

Finally, it is noteworthy that our methodologies have potential application in other domains beyond neural data analysis: for example the spectral analysis techniques can prove to be fruitful in domains such as econometrics, data forensics, oceanography, climatology, and seismology; the LASSO-based Granger causal analysis can be used to extract functional connectivity in social networks or gene regulatory networks, thanks to the plug-and-play nature of the algorithms used in our inference framework. To ease the adoption of these methods in the aforementioned applications, we have archived implementations of our algorithms as open source repositories on GitHub: https://github.com/proloyd.

Appendix A: Supplementary Material on Chapter 2

This appendix contains the proof of the theorems charcterizeing the bias and variance for DBMT estimates as well as discusses the factors appearing in these theorems in detail.

Recall from Eq. (2.29) that the kth eigen-coefficient estimate at window n can be written in terms of the observed data as:

$$\widehat{\mathbf{x}}_{n|N}^{(k)} = \sum_{s=1}^{N} \mathbf{\Lambda}^{|s-n|} \mathbf{\Gamma} \mathbf{F}_s^H \widetilde{\mathbf{y}}_s^{(k)}, \tag{A.1}$$

where $\widetilde{\mathbf{y}}_s^{(k)} = \mathbf{u}^{(k)} \odot \widetilde{\mathbf{y}}_s = \mathbf{U}^{(k)} \widetilde{\mathbf{y}}_s$. Given that $\mathbf{Q}$ is a diagonal matrix with elements $q(f_j)$, $j = 1, 2, \cdots, J$, it can be shown that $\mathbf{\Sigma}_\infty$, $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are also diagonal matrices. Denoting the elements of $\mathbf{\Sigma}_\infty$, $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$, respectively by $\tau(f)$, $\eta(f)$ and $\gamma(f)$, for $f \in \{f_1, f_2, \cdots, f_J\}$, we have $\eta(f) = \dfrac{\alpha^2 \tau(f) + q(f)}{1 + rW(\alpha^2 \tau(f) + q(f))}$.

A.1 Proof of Theorem 2.1

First note that Eq. (2.30) implies that

$$\mathbb{E}[dz_n(f) dz_{n+t}^*(f')] = \alpha^t D(f) \delta(f - f') df df'. \tag{A.2}$$

By invoking the Cramér representation, the covariance of the data tapered by the kth and lth dpss sequences can be expressed as:

$$\begin{aligned}\mathbb{E}\left[(\widetilde{\mathbf{y}}_s^{(k)})_j(\widetilde{\mathbf{y}}_{s'}^{(l)})_{j'}^*\right] =&\alpha^{|s-s'|}\int_{-1/2}^{1/2} U_k(f_j-\beta)D(\beta)U_l^*(f_{j'}-\beta)d\beta\\ &+\int_{-1/2}^{1/2} U_k(f_j-\beta)\sigma^2\delta(s-s')U_l^*(f_{j'}-\beta)d\beta. \qquad \text{(A.3)}\end{aligned}$$

From Eq. (2.20) and Eq. (A.1), we get:

$$\widehat{D}_{n|N}(f_j) = \frac{\eta^2(f_j)}{K}\sum_{k=1}^{K}\sum_{s=1}^{N}\sum_{s'=1}^{N}\gamma(f_j)^{|s-n|}\gamma(f_j)^{|s'-n|}(\widetilde{\mathbf{y}}_s^{(k)})_j(\widetilde{\mathbf{y}}_{s'}^{(k)})_j^*. \qquad \text{(A.4)}$$

Taking the expectation of both sides and after some simplification, one arrives at:

$$\begin{aligned}\mathbb{E}[\widehat{D}_{n|N}(f_j)] &= \frac{\eta^2(f_j)}{K}\sum_{k=1}^{K}\sum_{s=1}^{N}\sum_{s'=1}^{N}\gamma(f_j)^{|s-n|}\gamma(f_j)^{|s'-n|}\mathbb{E}\left[(\widetilde{\mathbf{y}}_s^{(k)})_j(\widetilde{\mathbf{y}}_{s'}^{(k)})_j^*\right]\\ &= \left[\eta(f_j)^2\sum_{s=1}^{N}\sum_{s'=1}^{N}\gamma(f_j)^{|s-n|}\gamma(f_j)^{|s'-n|}\alpha^{|s-s'|}\right]\times\\ &\qquad \frac{1}{K}\sum_{k=1}^{K}\int_{-1/2}^{1/2}U_k(f_j-\beta)D(\beta)U_k^*(f_j-\beta)d\beta\\ &\qquad\qquad +\left[\eta(f_j)^2\sum_{s=1}^{N}\gamma(f_j)^{2|s-n|}\right]\sigma^2. \qquad \text{(A.5)}\end{aligned}$$

Using the orthogonality of the PSWFs as in [90], and using the fact that

$$D(\beta) \leqslant \sup_f D(f), \forall\beta,$$

we get:

$$\left|\mathbb{E}[\widehat{D}_{n|N}(f)] - \kappa_n(f)D(f)\right| \leqslant \kappa_n(f)(\sup_f\{D(f)\} - D(f))\left(1 - \frac{1}{K}\sum_{k=1}^{K}\lambda_k\right) + \mu_n(f)\sigma^2 + \kappa_n(f)o(1), \tag{A.6}$$

where

$$\kappa_n(f) := \eta(f)^2\sum_{s=1}^{N}\sum_{s'=1}^{N}\gamma(f)^{|s-n|}\gamma(f)^{|s'-n|}\alpha^{|s-s'|},$$
$$\mu_n(f) := \eta(f)^2\sum_{s=1}^{N}\gamma(f)^{2|s-n|},$$

for $f \in \{f_1, f_2, \cdots, f_J\}$. Using the triangle inequality, the bound of Theorem 2.1 on $\left|\mathbb{E}[\widehat{D}_{n|N}(f)] - D(f)\right|$ follows. □

A.2 Proof of Theorem 2.2

Using the notation of (A.1), we have:

$$\mathsf{Cov}\left\{\widehat{D}_n^{(k)}(f), \widehat{D}_m^{(l)}(f')\right\} = \eta(f)^4\sum_{s,s',t,t'=1}^{N}\gamma(f)^{|s-n|}\gamma(f)^{|s'-n|}\gamma(f')^{|t-m|}\gamma(f')^{|t'-m|}\times$$
$$\Bigg[\alpha^{|s-t|}\alpha^{|s'-t'|}\int\int U_k(f-\beta)U_l(f'+\beta)U_k(-\beta'-f)U_l(\beta'-f')\times$$
$$(D(\beta)+\sigma^2\delta(s-t))(D(\beta')+\sigma^2\delta(s'-t'))d\beta d\beta'+$$
$$\alpha^{|s-t'|}\alpha^{|s'-t|}\int\int U_k(f-\beta)U_l(\beta-f')U_k(-\beta'-f)U_l(\beta'+f')\times$$
$$(D(\beta)+\sigma^2\delta(s-t'))(D(\beta')+\sigma^2\delta(s'-t))d\beta d\beta'\Bigg] \tag{A.7}$$

Note that we have omitted the integral limits, as they are understood to be same as in (3.2) henceforth. After summing over all tapers and rearranging the summations, the first expression within the brackets in (A.7) becomes:

$$\begin{aligned}\int\int &\left[\eta(f)^2 \sum_{s,t=1}^{N} \gamma(f)^{|s-n|}\gamma(f')^{|t-m|}\alpha^{|s-t|}\right](D(\beta)+\sigma^2\delta(s-t))\times \\ &\sum_{k=1}^{K} U_k(f-\beta)U_k(-\beta'-f)\times \\ &\left[\eta(f)^2 \sum_{s',t'=1}^{N} \gamma(f)^{|s'-n|}\gamma(f')^{|t'-m|}\alpha^{|s'-t'|}\right](D(\beta')+\sigma^2\delta(s'-t'))\times \\ &\sum_{l=1}^{K} U_l(f'+\beta)U_l(\beta'-f')d\beta d\beta'. \end{aligned} \tag{A.8}$$

Letting

$$\mathbb{A}(n,m,f) := \left[\eta^2(f)\sum_{s,t=1}^{N}\gamma(f)^{|s-n|}\gamma(f')^{|t-m|}\alpha^{|s-t|}\right]D(f) + \left[\eta(f)^2\sum_{s=1}^{N}\gamma(f)^{2|s-n|}\right]\sigma^2$$

and using the Schwarz inequality, the integral in Eq. (A.8) can be bounded by:

$$\begin{aligned}&\left[\int\int\left|\sum_{k=1}^{K}U_k(f-\beta)U_k(-\beta'-f)\right|^2 \mathbb{A}(n,m,\beta)\mathbb{A}(n,m,\beta')d\beta d\beta' \right.\\ &\left.\times\int\int\left|\sum_{l=1}^{K}U_l(f'+\beta)U_l(\beta'-f')\right|^2 \mathbb{A}(n,m,\beta)\mathbb{A}(n,m,\beta')d\beta d\beta'\right]^{1/2}, \end{aligned} \tag{A.9}$$

Using bounds on the convolutions of PSWFs from [90], and upper bounding $\mathbb{A}(n,m,\beta)$ by $\sup_f\{\kappa_n(f)D(f)+\mu_n(f)\sigma^2\}$, the statement of the theorem on the variance of the DBMT estimate $\widehat{D}_n(f)$ follows. □

A.3 Characterization of $\kappa_n(f)$: Bounds and Parameter Dependence

A.3.1 Lower and Upper Bounds on $\kappa_n(f)$

Consider the scenario where $q(f) = q$, i.e., flat spectrum. Then, the dependent of $\gamma(f)$ and $\kappa_n(f)$ on f is suppressed. We have the following bound on κ_n:

Proposition A.1. *For $0 < \gamma, \alpha < 1$, the quantity κ_n can be bounded as:*

$$\left| \kappa_n - \left(1 - \frac{\gamma}{\alpha}\right)^2 \left[\frac{1 + \alpha\gamma - 2(\alpha\gamma)^N}{1 - \alpha\gamma} T_0 + \frac{\gamma}{(1-\gamma)^2} \right] \right| \leqslant \left(1 - \frac{\gamma}{\alpha}\right)^2 \frac{\gamma}{(1-\gamma)^2},$$

where $T_0 := \dfrac{1 + \gamma^2 - \gamma^{2n} - \gamma^{2(N-n+1)}}{1 - \gamma^2}$.

Proof. To get an upper bound on κ_n, we rewrite the expression defining κ_n as:

$$\begin{aligned} \kappa_n &= \eta^2 \sum_{s=1}^{N} \sum_{s'=1}^{N} \gamma^{|s-n|} \gamma^{|s'-n|} \alpha^{|s-s'|} \\ &= \eta^2 \sum_{t=-N+1}^{N-1} \alpha^{|t|} \sum_{s=1}^{N} \gamma^{|s-n|} \gamma^{|s-t-n|}. \end{aligned}$$

Now, let us define $T_t := \sum_{s=1}^{N} \gamma^{|s-n|} \gamma^{|s-t-n|}$. Then it can be verified that:

$$T_{t+1} \begin{cases} = \gamma T_t, \text{ when } t \geqslant N - n \\ \leqslant \gamma T_t + \gamma^{t+1}, \text{ when } 0 \leqslant t < N - n \end{cases} \tag{A.10}$$

and

$$T_{t-1}\begin{cases} = \gamma T_t, \text{ when } t \leqslant -n+1 \\ \\ \leqslant \gamma T_t + \gamma^{|t-1|}, \text{ when } -n+1 < t < 0. \end{cases} \tag{A.11}$$

Also, we have:

$$\begin{aligned} \sum_{t=0}^{N-1} \alpha^t T_t &\leqslant \sum_{t=0}^{N-1} \alpha^t \gamma^t T_0 + \sum_{t=1}^{N-n-1} t\gamma^t + \sum_{t=N-n}^{N-1} (N-n)\gamma^t \\ &\leqslant \frac{1-(\alpha\gamma)^N}{1-\alpha\gamma} T_0 + \frac{\gamma}{(1-\gamma)^2}. \end{aligned} \tag{A.12}$$

Similarly, we have:

$$\sum_{t=-N+1}^{0} \alpha^{|t|} T_t \leqslant \frac{1-(\alpha\gamma)^N}{1-\alpha\gamma} T_0 + \frac{\gamma}{(1-\gamma)^2}, \tag{A.13}$$

which along with (A.12) leads to the claimed upper bound. For the lower bound, we use the fact that

$$\gamma T_t = \begin{cases} \leqslant T_{t+1} \text{ when } t \geqslant 0 \\ \\ \leqslant T_{t-1} \text{ when } t \leqslant 0 \end{cases}, \tag{A.14}$$

which implies $\sum_{t=0}^{N-1} \alpha^t T_t \geqslant \frac{1-(\alpha\gamma)^N}{1-\alpha\gamma} T_0$ and $\sum_{t=-N+1}^{0} \alpha^{|t|} T_t \geqslant \frac{1-(\alpha\gamma)^N}{1-\alpha\gamma} T_0$. Using the latter lower bounds for κ_n yield the claimed lower bound. □

A.3.2 Relation between κ_n and $\mathbf{Q}$

The expressions for $\kappa_n(f)$ and $\mu_n(f)$ introduced at the beginning of current chapter depend on $\gamma(f)$ and α. But $\gamma(f)$ itself depends on $q(f)$ and α, and it is not straightforward to give a closed-form expression of $\gamma(f)$ merely in terms of $q(f)$ and α, since it requires computation of the filtered error covariance matrix $\mathbf{\Sigma}_{n|n}$ given $\mathbf{Q}$. Again, by invoking the stead-state approximation, and defining $\mathbf{\Sigma}_{n|n-1} =: \mathbf{\Sigma}, \forall n = 1, 2, \cdots, N$, the matrix $\mathbf{\Sigma}$ can be obtained by solving the following algebraic Riccati equation:

$$\mathbf{\Sigma} = \alpha^2\mathbf{\Sigma} - \alpha^2\mathbf{\Sigma}\mathbf{F}_n^H\left(\sigma^2\mathbf{I} + \mathbf{F}_n\mathbf{\Sigma}\mathbf{F}_n^H\right)^{-1}\mathbf{F}_n\mathbf{\Sigma} + \mathbf{Q} \tag{A.15}$$

and thereby the steady-state error covariance matrix $\mathbf{\Sigma}_{n|n} =: \mathbf{\Sigma}_\infty$ is given by $\mathbf{\Sigma}_\infty = \frac{1}{\alpha^2}(\mathbf{\Sigma} - \mathbf{Q})$.

Although this procedure can be carried out numerically, in general it is not possible to solve the Riccati equation to get a closed-form expression for arbitrary $\mathbf{Q}$. In order to illustrate the explicit dependent of $\kappa_n(f)$ on the state-space parameters, we consider the case of flat spectrum where $\mathbf{Q} = q\mathbf{I}$. In this case, it can be shown that $\mathbf{\Sigma} = \zeta\mathbf{I}$ for some $\zeta > 0$ and the matrix equation (A.15) reduces to a simpler scalar equation for ζ, given by:

$$\frac{\zeta}{\sigma^2} = \alpha^2\frac{\zeta}{\sigma^2}\left[1 - \frac{\zeta}{\sigma^2}\left(1 - \frac{rW(\zeta/\sigma^2)}{rW(\zeta/\sigma^2)+1}\right)rW\right] + \frac{q}{\sigma^2}. \tag{A.16}$$

Following simplification, a quadratic equation for ζ results, and since $\zeta \geqslant 0$, the positive solution for ζ is given by:

$$\frac{\zeta}{\sigma^2} = \frac{1}{2rW}\left[-\left(1-\alpha^2-rW\frac{q}{\sigma^2}\right)+\sqrt{\left(1-\alpha^2-rW\frac{q}{\sigma^2}\right)^2+4rW\frac{q}{\sigma^2}}\right]. \tag{A.17}$$

Then, γ can be computed as $\gamma = \frac{1}{\alpha}\big(1-\frac{q/\sigma^2}{\zeta/\sigma^2}\big)$. Using these expressions for γ and α, the functions κ_n and μ_n can be computed. Fig. 2.7B shows the upper and lower bounds on κ_n evaluated for $n = 50$ and different values of α for $q/\sigma^2 = 10$. The upper and lower bounds are simple functions of α and q/σ^2 and can be used to further inspect the performance trade-offs of the DBMT algorithm with respect to the state-space model parameters.

Appendix B: Supplementary Material on Chapter 4

B.1 Marginalization

To obtain the marginal distribution of Eq. (4.9) from the joint distribution of Eq. (4.8), one needs to integrate out $\mathbf{J}$ from the latter. Alternatively, thanks to the Gaussian assumption in Eq. (4.5) and Eq. (4.7), the marginalization can be carried out as follows. We start from the probabilistic generative model:

$$\mathbf{Y} = \mathbf{L}\mathbf{J} + \mathbf{W} \qquad \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{w}}), \tag{B.1}$$

$$\mathbf{J} = \mathbf{\Phi}\mathbf{S} + \mathbf{V} \qquad \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}). \tag{B.2}$$

Substituting the expression for $\mathbf{J}$ from Eq. (B.2) in Eq. (B.1), we arrive at:

$$\mathbf{Y} = \mathbf{L}(\mathbf{\Phi}\mathbf{S} + \mathbf{V}) + \mathbf{W} = \mathbf{L}\mathbf{\Phi}\mathbf{S} + \mathbf{L}\mathbf{V} + \mathbf{W} \tag{B.3}$$

Using the independence of $\mathbf{V}$ and $\mathbf{W}$, the distribution of the stimulus independent part can be derived as $\mathbf{L}\mathbf{V} + \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{w}} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^{\top})$. From here, the marginal distribution of $\mathbf{Y}$ can be written as given by Eq. (4.9).

B.2 Details of the Regularization Scheme

In this appendix, we provide more details on the regularization scheme used for NCRF estimation. Recall that the NCRF matrix estimation amounts to the following maximum likelihood problem:

$$\min_{\mathbf{\Theta}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{L}\mathbf{\Theta}\widehat{\mathbf{S}}\|^2_{(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top)^{-1}} \ . \tag{B.4}$$

given a particular choice of $\mathbf{\Gamma}$. With this choice, one can find the gradient of the objective as:

$$\mathbf{L}^\top(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top)^{-1}\left(\mathbf{Y} - \mathbf{L}\mathbf{\Theta}\widehat{\mathbf{S}}\right)\widehat{\mathbf{S}}^\top \tag{B.5}$$

and thus can attempt to solve the maximum likelihood problem using gradient descent techniques. The following observations on the gradient, however, show that the problem is ill-conditioned:

1. The left multiplier of $\mathbf{\Theta}$, i.e., $\mathbf{L}^\top(\mathbf{\Sigma}_w + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top)^{-1}\mathbf{L}$ is singular.

2. The right multiplier of $\mathbf{\Theta}$, i.e, $\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top$, which is the empirical stimulus correlation matrix is likely to be rank-deficient for naturalistic stimuli [128].

Therefore, a direct attempt at solving the problem via the gradient descent results in estimates of $\mathbf{\Theta}$ with high variability. In estimation theory, such ill-conditioning is handled by introducing a bias to the estimator, which contains a priori information about the problem, in order to reduce the estimation variance. In addition, the

NCRF model typically has many more free parameters than the observed data points, and without introducing prior information, the estimation problem is prone to over-fitting.

The prior information is often incorporated in the form of regularization. A commonly used regularization scheme in this context is the Tikhonov regularization and its variants for promoting smoothness [3]. Other estimation schemes such as boosting and ℓ_1-regularization promote sparse solutions [6, 125]. Here, we introduce a structured regularization by penalizing a specific mixed-norm of the NCRF matrix to recover spatio-temporally sparse solutions over the Gabor coefficients:

$$\mathcal{P}_{2,1,1}(\mathbf{\Theta}) = \sum_{m=1}^{M}\sum_{l=1}^{L} \|\boldsymbol{\theta}_{m,l}\|_2 = \sum_{m=1}^{M}\sum_{l=1}^{L} \sqrt{\theta_{m,l,R}^2 + \theta_{m,l,A}^2 + \theta_{m,l,S}^2}. \tag{B.6}$$

In words, for each current dipole location, we penalize the vector-valued response function by sum of the magnitude of its corresponding Gabor coefficients.

Note that the ℓ_1-regularization in this case, i.e., $\sum_{m=1}^{M}\sum_{l=1}^{L}\|\boldsymbol{\theta}_{m,l}\|_1$, is not compatible with the expected cortical distribution of the NCRFs. Since the ℓ_1-norm is separable with respect to the three 3 coordinates of $\boldsymbol{\theta}_{m,l}$, it tends to select a sparse subset of the 3D coordinates, rendering the recovered NCRF components parallel to the coordinate axes. In contrast, the proposed penalty aims to select the NCRF components as a single entity by penalizing the vector magnitudes at each lag. Indeed, if the current dipoles are constrained to be normal to the cortical patches in the NCRF formulation, the proposed penalty coincides with ℓ_1-regularization.

Another advantage of this mixed-norm penalty is its rotational invariance when working with 3D vector-valued response functions. Suppose the coordinate system is rotated by an orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{3\times 3}$. Then, the lead-field and NCRF matrices are transformed by: $\mathbf{L} \to \mathbf{L}\widetilde{\boldsymbol{U}}^\top =: \mathbf{L}'$, $\boldsymbol{\Theta} \to \widetilde{\boldsymbol{U}}\boldsymbol{\Theta} =: \boldsymbol{\Theta}'$ where $\widetilde{\boldsymbol{U}} = \boldsymbol{I}_M \otimes \boldsymbol{U}$. Then, $\mathbf{L}'\boldsymbol{\Theta}' = \mathbf{L}\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}}\boldsymbol{\Theta} = \mathbf{L}\boldsymbol{\Theta}$ and

$$\mathcal{P}_{2,1,1}(\boldsymbol{\Theta}') = \eta\sum_{m=1}^{M}\sum_{l=1}^{L}\|\boldsymbol{\theta}'_{m,l}\|_2 = \eta\sum_{m=1}^{M}\sum_{l=1}^{L}\|\boldsymbol{U}\boldsymbol{\theta}_{m,l}\|_2 = \eta\sum_{m=1}^{M}\sum_{l=1}^{L}\|\boldsymbol{\theta}_{m,l}\|_2 = \mathcal{P}_{2,1,1}(\boldsymbol{\Theta}),$$

which implies the aforementioned rotational invariance. As a result, the solutions are not dependent on any particular choice of coordinate system. Also, since the penalty does not prefer specific source orientations, it makes the solution more resilient to co-registration error than other approaches that do not consider the vector-valued nature of the current dipoles or constrain the solutions to be normal to the cortical surface.

B.3 Statistical Testing Procedures

To asses the statistical significance of the estimated NCRF components at the group level and across the source space, they need to be compared against suitable null hypotheses. The fact that the NCRF components are 3D vectors requires technical care in choosing the null hypotheses. Here, we provide two possible null hypotheses and testing methodologies: the Length Test, that only considers the length or magnitude of the NCRFs, and the Vector Test that takes into account both the magnitude and direction of the NCRFs. The corresponding source codes that implement these tests can be found at [129].

The Length Test

This test aims to assess the statistical significance of the NCRF components by comparing their magnitudes against a baseline 'null' NCRF model at the group level. To control for false positives arising from over-fitting, instead of using an all-zero null model of the NCRFs, we aim to learn the null model from the dataset itself. The time-series of the feature variables are split into four equal segments, and these segments are permuted cyclically to yield three 'misaligned' feature time-series. Then, for each feature variable, three 'misaligned' time-series are constructed by swapping its original time-series with the 'misaligned' ones, while keeping the other two feature variables intact. Then, the average NCRF magnitudes estimated from these three 'misaligned' time-series are considered as the null model for that feature variable. The NCRF magnitude pairs from the original data and the null model are tested for significance using mass-univariate tests based on related measures t-tests.

To control for multiple comparisons, nonparametric permutation tests [245, 246] based on the threshold-free cluster-enhancement (TFCE) algorithm [181] are used. First, at each dipole location and time point, the t-statistic is computed from the difference between the NCRF magnitude pairs. The resulting statistic-map is then processed by the TFCE algorithm, which boosts contiguous regions with high test statistic as compared to isolated ones, based on the assumption that spatial extent of the true sources is typically broader than those generated by noise. To find the distribution of these TFCE values under the null hypothesis, TFCE values are calculated following the same procedure, on 10000 different random permutations of the data. In each permutation, the sign of the NCRF magnitude differences is flipped for a randomly selected set of subjects, without resampling the same set of subjects. Then, at every permutation, the maximum value of the obtained TFCE values is recorded, thereby constructing a non-parametric distribution of the maximum TFCE values under the null hypothesis. The original TFCE values that exceed the $(1 - \alpha)$ percentile of the null distribution are considered significant at a level of α corrected for multiple comparisons across the sources.

The Vector Test

This test aims at quantifying the significance of the estimated 3D NCRF components at the group level, based on the one-sample Hotelling's T^2 test. In the one-sample Hotelling's T^2 test, the population mean of the sample vectors is tested against the null hypothesis of mean zero, i.e. $\boldsymbol{\mu}_0 = 0$. To control for multiple comparisons, a similar strategy based on nonparametric permutations as in the case of the Length Test is used. At every time lag, the T^2 statistic for each dipole is computed as:

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \bar{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \tag{B.7}$$

where $\bar{\mathbf{x}}$, $\bar{\Sigma}$ are the population mean and covariance matrix of the vector-valued NCRF components, respectively. The T^2 statistic quantifies the variability of vector-valued samples, akin to the role of the t-statistic for 1D samples [247]. The resulting T^2-maps are then processed by the TFCE algorithm. As before, to construct

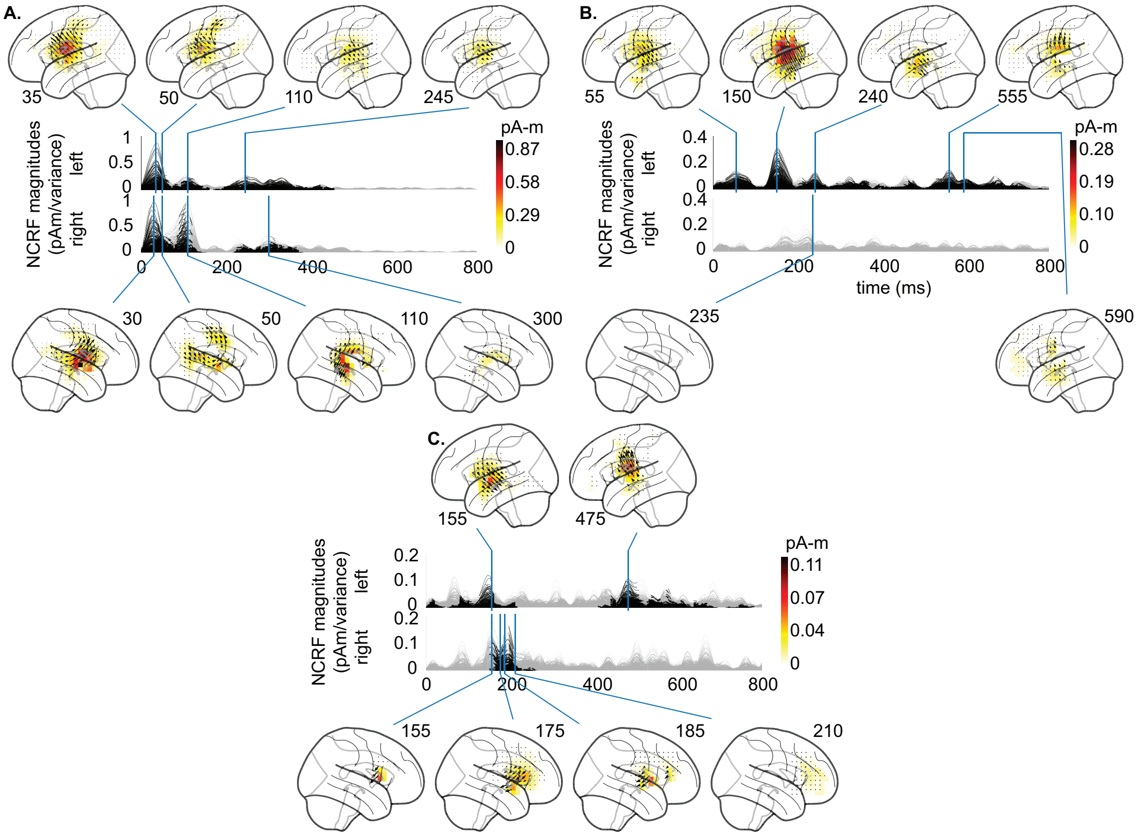


Figure B.1: Estimated NCRFs for acoustic envelope (A), word frequency (B), semantic composition (C): The anatomical plots show the group-level average NCRFs projected onto the lateral plane (top and bottom panels) corresponding to selected visually salient peaks in the temporal profiles (middle panels). The top and bottom portions of the subplot pertain to left and right hemisphere, respectively. Numerical labels of each anatomical subplot indicates the corresponding time lag in ms. The gray portions of the traces indicate statistically insignificant NCRFs at the group level (significance level of 5%). The significance levels are computed using the Vector Test, as opposed to the main manuscript where the significance levels are based on the Length Test. The main features of the NCRFs discussed in the Results section are similarly recovered by the Vector Test.

a non-parametric distribution of maximum TFCE values under the null hypothesis, maximum values of the TFCE-processed T^2 maps on 10000 different random permutations of the data are recorded. In each permutation, the vector-valued NCRF components of each subject undergo uniform random rotations in 3D [248]. The original TFCE values that exceed the $(1-\alpha)$ percentile of the null distribution are considered significant at a level of α, corrected for multiple comparisons across the sources.

Traditionally, response functions are estimated as scalar functions of the data, either over the sensor space or over the source space by orientation-constrained inverse solvers. Considering the directional variability of the NCRF estimates at the group level, however, takes into account the group level anatomical variability that may effect the current dipole orientations. In addition, the Vector Test is less computationally demanding than the Length Test, because it does not require refitting NCRFs for permuted models. In the Results section of the manuscript, we presented the NCRFs masked at a significance level of 5%, based on the Length Test. To demonstrate the difference between these two tests, here we also present the sames results using the Vector Test (Fig B.1).

Appendix C: Supplementary Material on Chapter 5

Here, we will first give an overview of the implications of assumptions (A) in Section C.1. We will then give the proofs of Theorem 5.1 and Corollary 5.1.1 in Section C.2, which utilize a number of auxiliary theorems and technical results. The auxiliary theorems, mainly on the prediction error analysis of the *full* and *reduced* models, ensure that the standard conditions required for the consistency of the LASSO hold, and are given in Section C.3. Finally, several technical results used in the proofs of Sections C.2 and C.3 are given in Section C.4.

C.1 Implications of the key assumptions

In this section, we give an overview of the implications of the key assumptions (A) on the BVAR model, adopted from Basu et al. [209]. Hereafter, we denote the maximum and minimum eigen-values of any matrix $\mathbf{M}$ by $\Lambda_{\max}(\mathbf{M})$ and $\Lambda_{\min}(\mathbf{M})$, respectively. We start with the following general assumption:

Assumption C.I. Let $\mathbf{\Gamma}(l)$ be the auto-covariance matrix of the BVAR process at lag l. The spectral density matrix $\mathbf{F}(\omega) =: \dfrac{1}{2\pi} \displaystyle\sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) \exp\left(-il\omega\right)$ exists, and its maximum eigen-value is bounded almost everywhere on $[-\pi, \pi]$, i.e., $\mathcal{M}(\mathbf{F}) := \underset{\omega\in[-\pi,\pi]}{\text{ess sup}}\Lambda_{\max}(\mathbf{F}(\omega)) < \infty$.

It can be shown that the spectral density exists, if $\sum_{l=0}^{\infty} \|\mathbf{\Gamma}(l)\|_2^2 < \infty$. Furthermore, if $\sum_{l=0}^{\infty} \|\mathbf{\Gamma}(l)\|_2 < \infty$, the spectral density is bounded and continuous, so that the essential supremum is indeed achieved.

For the BVAR(p) process in Eq. (5.1), let the matrix valued characteristic polynomial be defined as $\mathbf{A}(z) := \mathbf{I} - \sum_{j=1}^{p} \mathbf{A}_j z^j$. Then, the following conditions from Assumption 5.I provide a simple characterization of the spectral density matrix:

1. The process noise covariance matrix $\mathbf{\Sigma}_\epsilon$ is positive definite with bounded eigenvalues, i.e., $0 < \Lambda_{\min}(\mathbf{\Sigma}_\epsilon) \leq \Lambda_{\max}(\mathbf{\Sigma}_\epsilon) < \infty$.

2. The BVAR process is stable and invertible, i.e., $\det(\mathbf{A}(z)) \neq 0$ on or inside the unit circle $\{z \in \mathbb{C} : |z| \leq 1\}$.

Under these two conditions, the spectral density matrix satisfies Assumption C.I, is bounded and continuous, and admits the representation:

$$\mathbf{F}(\omega) = \frac{1}{2\pi} \mathbf{A}^{-1}(\exp(-i\omega)) \mathbf{\Sigma}_\epsilon \mathbf{A}^{-H}(\exp(-i\omega)).$$

Additionally, consider the infimum of the spectral density over unit circle:

$$\mathcal{m}(\mathbf{F}) := \operatorname*{ess\,inf}_{\omega \in [-\pi,\pi]} \Lambda_{\max}(\mathbf{F}(\omega)).$$

Then, the following useful bounds hold for the BVAR(p) process in Eq. (5.1):

$$\mathcal{M}(\mathbf{F}) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\mathbf{\Sigma}_\epsilon)}{\mu_{\min}(\mathbf{A})}, \qquad \mathcal{m}(\mathbf{F}) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\mathbf{\Sigma}_\epsilon)}{\mu_{\max}(\mathbf{A})}, \tag{C.1}$$

where

$$\mu_{\max}(\mathbf{A}) := \min_{|z|=1} \Lambda_{\max}\left(\mathbf{A}^H(z)\mathbf{A}(z)\right) \text{ and } \mu_{\min}(\mathbf{A}) := \max_{|z|=1} \Lambda_{\min}\left(\mathbf{A}^H(z)\mathbf{A}(z)\right).$$

The bounds in Eq. (C.1) are particularly useful in the prediction analysis of Section C.3, where we replace the quantities $\mathcal{M}(\mathbf{F})$ and $\mathcal{m}(\mathbf{F})$ arising from the application of [209, Proposition 2.4] by these bounds. The characteristic polynomial $\mathbf{A}(z)$ encodes the temporal dependencies of the process, whereas $\mathbf{\Sigma}_\epsilon$ captures the correlation between the process noise components, possibly due to latent processes. Expressing the error bounds in terms of $\mu_{\max}(\mathbf{A}), \mu_{\min}(\mathbf{A}), \Lambda_{\max}(\mathbf{\Sigma}_\epsilon), \Lambda_{\min}(\mathbf{\Sigma}_\epsilon)$, instead of $\mathcal{M}(\mathbf{F})$ and $\mathcal{m}(\mathbf{F})$, helps to separate the contributions of these two sources of BVAR dependencies. We also consider the $2p$-dimensional alternative BVAR(1) representation of the 2-dimensional BVAR(p) process: $\mathbf{X}_t = \breve{\mathbf{A}}_1\mathbf{X}_{t-1} + \breve{\boldsymbol{\epsilon}}$, where $\mathbf{X}_t$ is the first the row of $\mathbf{X}$ in Eq. (5.3) organized as a column vector, and $\breve{\mathbf{A}}_1$ and $\breve{\boldsymbol{\epsilon}}$ are constructed by the corresponding augmentation of $\mathbf{A}_i$'s and ϵ_t's, respectively. The process $\mathbf{X}_t$ has a characteristic polynomial, $\breve{\mathbf{A}}(z) := \mathbf{I} - \breve{\mathbf{A}}_1 z$ and is stable if and only if the original process is stable [249]. However, $\mu_{\min}(\breve{\mathbf{A}})$ and $\mu_{\max}(\breve{\mathbf{A}})$ are generally different than $\mu_{\min}(\mathbf{A})$ and $\mu_{\max}(\mathbf{A})$, respectively.

Let the columns of $\mathbf{X}$ corresponding to $\boldsymbol{\theta}_{(i)}$ be denoted by $\mathbf{X}_{(i)}$, for $i = 1, 2$. The remaining component of our key assumption is that the BVAR parameters are k-sparse. The implication of this assumption for the *full* model is fairly standard, under both the null and alternative hypotheses. However, for the *reduced* model under the alternative hypothesis, where only the auto-regressive parameters are unspecified and the cross-regression parameters are enforced to be $\mathbf{0}$, we need to define a suitable surrogate "true" model. To this end, we use the orthogonality principle to define the surrogate "true" auto-regression coefficients in the *reduced* model as:

$$\widetilde{\boldsymbol{\theta}}^*_{(1)} := \boldsymbol{\theta}^*_{(1)} + \mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)},$$

where

$$\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right] = \begin{bmatrix} \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\right] \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(1)}^\top\mathbf{X}_{(2)}\right] \\ \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(2)}^\top\mathbf{X}_{(1)}\right] \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(2)}^\top\mathbf{X}_{(2)}\right] \end{bmatrix} =: \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} =: \mathbf{C}.$$

Note that even though the MVAR coefficients under the alternative hypothesis are k-sparse, the surrogate "true" auto-regression coefficients under the *reduced* model may not be. To deal with this issue, we follow the treatment of Negahban et al. [215] in analyzing the LASSO under weakly sparse or compressible parameters, and further impose the norm condition on $\boldsymbol{\theta}^*_{(2)}$ (given in the statement of Theorem 5.1) to restrict the alternative hypothesis. The latter ensures that the *full* and *reduced* models are distinguishable under the alternative hypothesis (See Section C.3 for details).

C.2 Proofs of Theorem 5.1 and Corollary 5.1.1

The key idea in the proofs of Theorem 5.1 and Corollary 5.1.1 is to simultaneously analyze the *full* and *reduced* models, and to balance their consistency requirements in a unified fashion. These requirements for the *full* model are the same as those in standard sparse MVAR estimation. However, analysis of the *reduced* model, where the cross-regression coefficients are forced to be zero, requires further technical care.

C.2.1 Proof of Theorem 5.1

The proof has two main steps. First, we bound the deviation of the empirical quantities $\ell(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)})$ (*full* model) and $\ell(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0})$ (*reduced* model) with respect to their counterparts evaluated at the true parameters. In doing so, we first assume that the following conditions hold:

(C1) Restricted eigenvalue (RE) condition: the symmetric matrix $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^\top \mathbf{X}/n \in \mathbb{R}^{2p \times 2p}$ satisfies restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau \geq 0$, i.e., $\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}(\alpha, \tau)$:

$$\boldsymbol{\phi}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\phi} \geq \alpha \|\boldsymbol{\phi}\|_2^2 - \tau \|\boldsymbol{\phi}\|_1^2, \ \forall \ \boldsymbol{\phi} \in \mathbb{R}^{2p},$$

with $\tau := \frac{m-1}{m} \frac{\alpha}{32k}$ for some constant $m > 1$.

(C2) Deviation condition: there exist deterministic functions $\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$, $\mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$ such that

$$\left\| \frac{1}{n} \mathbf{X}^\top (\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*) \right\|_\infty \leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon) \sqrt{\frac{\log 2p}{n}},$$
$$\left\| \frac{1}{n} \mathbf{X}_{(1)}^\top \left(\mathbf{x} - \mathbf{X}_{(1)} \widetilde{\boldsymbol{\theta}}_{(1)}^* \right) \right\|_\infty \leq \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon) \sqrt{\frac{\log 2p}{n}}.$$

As a result, we can lower bound $\mathcal{T}_{y\mapsto x}$ under the alternative hypothesis and upper bound it under the null hypothesis, and invoke suitable concentration results to control these bounds. We then seek conditions under which the bounds do not coincide, which further restricts the alternative hypothesis. The second step of the proof establishes that (C1) and (C2) indeed hold with high probability.

Step 1. For the *full* model, we have:

$$\begin{aligned}\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) =& \frac{1}{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &+ \frac{2}{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{X}^\top(\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*). \end{aligned} \tag{C.2}$$

Using the consistency results of the LASSO (Theorem C.1), we can upper bound Eq. (C.2) as:

$$\begin{aligned}\Big|\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big)\Big| &\leq \frac{1}{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &\quad + \Big\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\Big\|_1 \Big\|\frac{2}{n}\mathbf{X}^\top(\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*)\Big\|_\infty \qquad \text{(C.3)} \\ &\leq \frac{18m}{m+1}\frac{k\lambda_n^2}{\alpha} + 2\frac{12m}{m+1}\frac{k\lambda_n^2}{\alpha} \\ &= \frac{42}{m+1}\frac{k\lambda_n^2}{\alpha/m} =: \Delta_F. \qquad \text{(C.4)}\end{aligned}$$

Let J denote the index set of the support of $\boldsymbol{\theta}^*_{(1)}$, with its complement denoted by J^c. Note that $|J| \leq k$. We use the shorthand notation $\boldsymbol{\theta}_J$ to denote the restriction of a vector $\boldsymbol{\theta}$ to its indices given by J. In a similar fashion to Eq. (C.4), by invoking the consistency of the *reduced* model (Theorem C.2) we have:

$$\begin{aligned}\left|\ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right) - \ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right)\right| \leq\, & 21\frac{k\lambda_n^2}{\alpha/m} + 28\sqrt{\frac{k\lambda_n^3}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1} \\ & + \left(14\left(\sqrt{2m}+1\right)+2\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 \\ \leq & 35\frac{k\lambda_n^2}{\alpha/m} + \left(14\left(\sqrt{2m}+1\right)+16\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 =: \Delta_R, \end{aligned} \tag{C.5}$$

where we have used the inequality of arithmetic and geometric means to further upper bound the middle term in the right hand side of the first line in Eq. (C.5).

Using the bounds in Eqs. (C.4) and (C.5), we get:

$$\frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \Delta_R}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) + \Delta_F} \leq \frac{\ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right)}{\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right)} \leq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) + \Delta_R}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_F}, \tag{C.6}$$

which gives the following lower and upper bounds on $\mathcal{T}_{y\mapsto x}$:

$$\begin{aligned}\frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_R - \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) + \Delta_F} & \leq \mathcal{T}_{y\mapsto x} \\ & \leq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) + \Delta_R + \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_F} \end{aligned} \tag{C.7}$$

Now, under the null hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}_{(2)} = \mathbf{0}$, we have $\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) = \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right)$ and $\Delta_R = \Delta_F$, which implies:

$$\mathcal{T}_{y\mapsto x} \leq \frac{\Delta_R + \Delta_F}{\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - \Delta_F} \leq \frac{2\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} - \Delta_N - \Delta_F}, \tag{C.8}$$

with probability at least $1-2\exp\left(-\frac{n\Delta_N^2}{8(\boldsymbol{\Sigma}_\epsilon)_{1,1}^2}\right)$, for some constant Δ_N (to be specified). This concentration result for $\ell\Big(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\Big)$ is established in Lemma C.7.

On the other hand, under a general alternative hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} \neq \mathbf{0}$, Lemma C.8 can be used to show that:

$$\mathcal{T}_{y\mapsto x} \geq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)},\mathbf{0}\right)-\ell\Big(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\Big)-\Delta_R-\Delta_F}{\ell\Big(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\Big)+\Delta_F} \geq \frac{D-(\Delta_D+\Delta_R+\Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}+\Delta_N+\Delta_F}, \tag{C.9}$$

with probability at least $1-2\exp\left(-\frac{n\Delta_N^2}{8(\boldsymbol{\Sigma}_\epsilon)_{1,1}^2}\right)-\frac{2}{(2p)^{c_{11}}}$, whenever $n \geq \frac{c_{11}}{c}\log(2p)$, for specific constants D, Δ_D, c, and c_{11} defined in Lemma C.8).

From the upper and lower bounds in Eq. (C.8) and Eq. (C.9), it is possible to choose a threshold to distinguish between the two hypothesis without ambiguity, if:

$$\frac{D-(\Delta_D+\Delta_R+\Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}+\Delta_N+\Delta_F} > \frac{2\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}-\Delta_N-\Delta_F}, \tag{C.10}$$

which after rearrangement translates to:

$$D > \Delta_D+\Delta_R+\Delta_F\left(1+2\frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}+\Delta_N+\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}-(\Delta_N+\Delta_F)}\right). \tag{C.11}$$

Next, we choose $\Delta_N = \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{4}$. Then, assuming $\Delta_F \leq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{4}$, the bound of Eq. (C.11) further simplifies to $D > \Delta_D+\Delta_R+7\Delta_F$. Note that this assumption on Δ_F requires:

$$\frac{42}{m+1}\frac{k\lambda_n^2}{\alpha/m} \leq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{4}. \tag{C.12}$$

and imposes an upper bound on λ_n, as discussed in *Remark 2* in Section 3.

Using the following inequality:

$$\left\|\widetilde{\boldsymbol{\theta}}_{(1)J^c}\right\|_1 \leq \left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*\right\|_1 \leq \left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_1 \left\|\boldsymbol{\theta}_{(2)}^*\right\|_1,$$

and with the choice of $\lambda_n = 4\mathscr{A}\sqrt{\frac{\log 2p}{n}}$, for $\mathscr{A}$ satisfying (See Proposition C.4 and Proposition C.5):

$$\mathscr{A} \geq \max\left\{\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon), \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\right\},$$

we obtain:

$$\Delta_D + \Delta_R + 7\Delta_F \leq \mathscr{a}\sqrt{\frac{\log 2p}{n}}\left\|\boldsymbol{\theta}_{(2)}^*\right\|_2^2 + \mathscr{b}\sqrt{\frac{k\log 2p}{n}}\left\|\boldsymbol{\theta}_{(2)}^*\right\|_2 + \mathscr{c}\frac{k\log 2p}{n}$$

where

$$\begin{aligned}
\mathscr{a} &= c_{10}, \\
\mathscr{b} &= \left[4\mathscr{A}\left(\left(14\left(\sqrt{2m}+1\right)+16\right)+\left(\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_1+1\right)\right)\right], \\
\mathscr{c} &= \frac{16\mathscr{A}^2}{\alpha/m}\left(\frac{294}{m+1}+35\right),
\end{aligned}$$

with c_{10} given in Lemma C.8. Also D in Lemma C.8 can be lower bounded as $D \geq \widetilde{\Lambda}_{\min}\|\boldsymbol{\theta}_{(2)}^*\|_2^2$, with $\widetilde{\Lambda}_{\min} := \Lambda_{\min}\left(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)$, which gives the following sufficient condition for inequality (C.11) to hold:

$$\widetilde{\Lambda}_{\min} \left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 \geq \mathscr{a}\sqrt{\frac{\log 2p}{n}} \left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 + \mathscr{b}\sqrt{\frac{k\log 2p}{n}} \left\|\boldsymbol{\theta}^*_{(2)}\right\|_2 + \mathscr{c}\frac{k\log 2p}{n}. \tag{C.13}$$

By further requiring $n \geq (2c_{10}/\widetilde{\Lambda}_{\min})^2 \log(2p)$, we have $\widetilde{\Lambda}_{\min} - \mathscr{a}\sqrt{\log 2p/n} \geq \widetilde{\Lambda}_{\min}/2$. The latter combined with Eq. (C.13), and an application of Lemma C.9 gives the sufficient condition: $\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 \geq \mathscr{B}k\log 2p/n$ for unambiguous discrimination between the null and the *local* alternative hypothesis $H^n_{y\mapsto x,0} : \left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 \geq \mathscr{B}k\log 2p/n$, as long as $n \geq \max\{\mathscr{C}', \mathscr{D}'k\}\log(2p)$, with probability at least

$$1 - 2\exp\left(-\frac{n}{128}\right) - \frac{2}{(2p)^{c_{11}}}, \tag{C.14}$$

where

$$\mathscr{B} = \left(\frac{4\mathscr{b}^2}{\widetilde{\Lambda}^2_{\min}} + \frac{4\mathscr{c}}{\widetilde{\Lambda}_{\min}}\right), \quad \mathscr{C}' = \frac{c_{11}}{c}\max\left\{1, \frac{2\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)(\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2^2 + 1)}{\widetilde{\Lambda}_{\min}\mu_{\min}(\breve{\mathbf{A}})}\right\}^2, \tag{C.15}$$

$$\text{and} \quad \mathscr{D}' := \frac{2688m}{m+1}\frac{\mathscr{A}^2}{\alpha(\boldsymbol{\Sigma}_\epsilon)_{1,1}}. \tag{C.16}$$

Note that the condition $n \geq \mathscr{D}'k\log(2p)$ ensures the upper bound (C.12) on λ_n.

Step 2. Proposition C.3 establishes that (C1) holds with probability at least

$$1 - c_1\exp(-c_2 n\min\{\zeta^{-2}, 1\}), \tag{C.17}$$

if $n \geq C_0 \max\{\zeta^2, 1\} k \log 2p$, for some constants C_0, c_1, c_2, and ζ (> 0). Also, Proposition C.4 and Proposition C.5 establish that condition (C2) holds with probability at least

$$1 - \frac{d_1}{(2p)^{d_2}} - \frac{d_1'}{(2p)^{d_2'}}, \tag{C.18}$$

if $n \geq \max\{D_0, D_0'\} \log(2p)$, for some constants d_1, d_1', d_2, d_2', D_0, and D_0' (> 0).

Combining the two steps, the claim of the theorem holds with probability at least

$$1 - 2\exp\left(-\frac{n}{128}\right) - \frac{2}{(2p)^{c_{11}}} - \frac{d_1}{(2p)^{d_2}} - \frac{d_1'}{(2p)^{d_2'}} - c_1 \exp(-c_2 n \min\{\zeta^{-2}, 1\}), \tag{C.19}$$

if $n \geq \max\{\mathscr{C}'', \mathscr{D}'' k\} \log(2p)$, where $\mathscr{D}'' := \max\{\mathscr{D}', C_0 \max\{\zeta^2, 1\}\}$ and $\mathscr{C}'' = \max\{\mathscr{C}', D_0, D_0'\}$. Finally, the probability in Eq. (C.19) can be further lower bounded by

$$1 - \frac{K}{p^d} \tag{C.20}$$

where

$$d := \min\{1, c_{11}, d_2, d_2'\}$$
$$K := \max\left\{\frac{4 + c_1}{2^{c_{11}}}, \frac{d_1}{2^{d_2}}, \frac{d_1'}{2^{d_2'}}\right\},$$

if $n \geq \max\{\mathscr{C}, \mathscr{D} k\} \log(2p)$, with

$$\mathscr{C} := \max\left\{\frac{c_{11}}{c} \max\left\{1, \frac{2\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)(\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2^2 + 1)}{\widetilde{\Lambda}_{\min}\mu_{\min}(\breve{\mathbf{A}})}\right\}^2, \frac{1}{c_2 \min\{\zeta^{-2}, 1\}},\right.$$

$$128, D_0, D_0' \Big\}, \tag{C.21}$$

$$\mathscr{D} := \max \left\{ \frac{2688m}{m+1} \frac{\mathscr{A}^2}{\alpha(\boldsymbol{\Sigma}_\epsilon)_{1,1}}, C_0 \max\{\zeta^2, 1\} \right\}. \tag{C.22}$$

This concludes the proof of Theorem 5.1. □

C.2.2 Proof of Corollary 5.1.1

First we note that under conditions (C1) and (C2) there exists some real numbers $s, t > 0$ such that,

$$\left| \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - (\boldsymbol{\Sigma}_\epsilon)_{1,1} \right| \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/s \quad \text{and} \quad \Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1} t/s. \tag{C.23}$$

This allows us to set a problem indenpendent threshold, since the upper bound on $\mathcal{T}_{y \mapsto x}$ under the null hypothesis given in Eq. (C.8) simplifies to:

$$\mathcal{T}_{y \mapsto x} \leq \frac{2t/s}{1 - (1+t)/s}. \tag{C.24}$$

Now, given any threshold $\mathscr{t} > 0$, we can solve for s in terms of t and $\mathscr{t}$ as:

$$s = 1 + \frac{2 + \mathscr{t}}{\mathscr{t}} t. \tag{C.25}$$

To ensure $\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1} t/s$, we need the following to hold:

$$\frac{(m+1)\alpha}{42m} \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{16\mathscr{A}^2} \frac{n}{k \log(2p)} - \frac{1}{t} - \frac{2+\mathscr{t}}{\mathscr{t}} \geq 0 \tag{C.26}$$

On the other hand, using the expression for s from Eq. (C.25) and invoking Lemma C.7 yield the following statement that can be used to bound the false positive error probability:

$$\mathbb{P}\left[\left|\ell\left(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\right)-(\boldsymbol{\Sigma}_\epsilon)_{1,1}\right|\geq\frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{s}\right]\leq 2\exp\left(-\frac{n}{8\left(1+\gamma t\right)^2}\right). \tag{C.27}$$

With a choice of $t=t_0\sqrt{\log(2p)/n}$ for any $t_0>0$, applying Lemma C.9 on Eq. (C.26) then gives the sampling requirement, $n\geq\frac{\widetilde{\mathscr{D}}^2}{t_0^2}k^2\log(2p)+2\widetilde{\mathscr{D}}\gamma k\log(2p)$, where

$$\widetilde{\mathscr{D}}=\frac{42m}{(m+1)\alpha}\frac{16A^2}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}\quad\text{and}\quad\gamma:=(2+\mathscr{t})/\mathscr{t}$$

and the false positive error probability given in the corollary. □

Remark C.1. Note that in this case the lower bound on $\mathcal{T}_{y\mapsto x}$ under the alternative hypothesis given by Eq. (C.9) simplifies to:

$$\mathcal{T}_{y\mapsto x}\geq\frac{\left(D-\Delta_D-\left(14\left(\sqrt{2m}+1\right)+16\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right)/(\boldsymbol{\Sigma}_\epsilon)_{1,1}-2t/s}{1+(1+t)/s}. \tag{C.28}$$

To ensure that the right hand side of Eq. (C.24) is less than the right hand side of Eq. (C.28), we need:

$$\frac{D-\Delta_D-\left(14\left(\sqrt{2m}+1\right)+16\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}\geq\frac{4t/s}{1-(1+t)/s}\geq 2\mathscr{t}. \tag{C.29}$$

Assuming that Δ_D and $\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1$ are small fractions of D, this condition implies that thresholding at a level $\mathscr{t}$ can also detect GC influences with "effect size"

$D/(\mathbf{\Sigma}_\epsilon)_{1,1}$ as small as $2t$.

C.3 Prediction error analysis of the *full* and *reduced* models

In this section, we establish that conditions (C1) and (C2) hold with high probability, under both the *full* and *reduced* models. Note that both the *full* and *reduced* models share the same RE condition (C1), since the *reduced* model is nested within the *full* model. However, the deviation conditions required in (C2) are different for the two models. For the *reduced* model, we require:

$$\left\| \frac{1}{n} \mathbf{X}^\top (\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*) \right\|_\infty \leq \mathbb{Q}(\boldsymbol{\theta}^*, \mathbf{\Sigma}_\epsilon) \sqrt{\frac{\log 2p}{n}}, \tag{C.30}$$

for some deterministic function $\mathbb{Q}(\boldsymbol{\theta}^*, \mathbf{\Sigma}_\epsilon)$. In the *reduced* model, however, we require

$$\left\| \frac{1}{n} \mathbf{X}_{(1)}^\top \left(\mathbf{x} - \mathbf{X}_{(1)} \widetilde{\boldsymbol{\theta}}_{(1)}^* \right) \right\|_\infty \leq \mathbb{Q}'(\boldsymbol{\theta}^*, \mathbf{\Sigma}_\epsilon) \sqrt{\frac{\log 2p}{n}} \tag{C.31}$$

for another deterministic function $\mathbb{Q}'(\boldsymbol{\theta}^*, \mathbf{\Sigma}_\epsilon)$.

First, we state a result adapted from Basu et al. [209] on the prediction error of LASSO under the *full* model:

Theorem C.1 (Prediction Error for the *full* Model)**.** *Suppose $\widehat{\mathbf{\Sigma}} \sim RE(\alpha, \tau)$, with τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ and $(\mathbf{X}, \mathbf{x})$ satisfying the deviation bound (C.30). Then for any $\lambda_n \geq 4\mathbb{Q}(\boldsymbol{\theta}^*, \mathbf{\Sigma}_\epsilon)\sqrt{\log(2p)/n}$, the solution to the* full *model in Eq.* (5.6) *satisfies:*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{3m}{m+1}\frac{\sqrt{k}\lambda_n}{\alpha}, \tag{C.32}$$

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{12m}{m+1}\frac{k\lambda_n}{\alpha}, \tag{C.33}$$

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \leq \frac{18m}{m+1}\frac{k\lambda_n^2}{\alpha}. \tag{C.34}$$

Proof. The proof closely follows that of Basu et al. [209, Proposition 4.1], and is thus omitted for brevity. □

In what follows, we show that the particular choice of τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ will simplify the prediction error analysis of the *reduced* model. As for the *reduced* model, the main technical difficulty in establishing prediction error bounds stems from the fact that $\widetilde{\boldsymbol{\theta}}^*_{(1)}$ is no longer k–sparse. We will address this issue in the following theorem:

Theorem C.2 (Prediction Error for the *reduced* Model)**.** *Suppose* $\widehat{\boldsymbol{\Sigma}} \sim RE(\alpha, \tau)$*, with* τ *satisfying the relation in Theorem C.1 and* $(\mathbf{X}_{(1)}, \mathbf{x})$ *satisfying the deviation bound* (C.31)*. Let* J *denote the support of* $\boldsymbol{\theta}^*_{(1)}$*, with its complement denoted by* J^c*. Then, for any* $\lambda_n \geq 4\mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\log 2p/n}$*, the solution to* reduced *model in Eq.* (5.6) *satisfies:*

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)} - \widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_2 \leq \frac{3}{2}\frac{\lambda_n\sqrt{k}}{\alpha/m} + \sqrt{\frac{2m}{k}}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 + \sqrt{\frac{4\lambda_n}{\alpha/m}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1},$$

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)} - \widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_1 \leq 6\frac{\lambda_n k}{\alpha/m} + 4(\sqrt{2m}+1)\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 + 8\sqrt{\frac{\lambda_n k}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1},$$

$$\frac{1}{n}\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)^{\top}\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\leq 9\frac{\lambda_n^2 k}{\alpha/m}+\left(6\left(\sqrt{2m}+1\right)+2\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1$$
$$+12\sqrt{\frac{\lambda_n^3 k}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1}.$$

Proof. Recall that $\widetilde{\boldsymbol{\theta}}^*_{(1)}:=\boldsymbol{\theta}^*_{(1)}+\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)}$. Let us define

$$F(\boldsymbol{\theta}_{(1)}):=\frac{1}{n}\left\|\mathbf{x}-\mathbf{X}_{(1)}\boldsymbol{\theta}_{(1)}\right\|_2^2+\lambda_n\|\boldsymbol{\theta}_{(1)}\|_1.$$

and consider the quantity, $\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right):=F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right)-F\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)$. Also, let $\mathbf{v}:=\boldsymbol{\theta}^*_{(1)}-\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}$ and $\widetilde{\mathbf{v}}:=\mathbf{v}+\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}$. Then, $\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right)$ can be simplified as:

$$\begin{aligned}
\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right)&=F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right)-F\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\\
&=\frac{1}{n}\mathbf{v}^{\top}\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\mathbf{v}+\frac{2}{n}\boldsymbol{\epsilon}^{\top}\mathbf{X}_{(1)}\mathbf{v}+\frac{2}{n}\boldsymbol{\theta}^{*\top}_{(2)}\mathbf{X}_{(2)}^{\top}\mathbf{X}_{(1)}\mathbf{v}\\
&\quad+\frac{2}{n}\boldsymbol{\theta}^{*\top}_{(2)}\mathbf{X}_{(2)}^{\top}\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)}-\frac{1}{n}(\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)})^{\top}\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)}\\
&\quad+\frac{2}{n}\boldsymbol{\epsilon}^{\top}\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)}+\lambda_n\left(\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right\|_1-\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_1\right)\\
&=\frac{1}{n}\widetilde{\mathbf{v}}^{\top}\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\widetilde{\mathbf{v}}+\frac{2}{n}\left(\boldsymbol{\epsilon}^{\top}+\boldsymbol{\theta}^{*\top}_{(2)}\mathbf{X}_{(2)}^{\top}-(\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)})^{\top}\mathbf{X}_{(1)}^{\top}\right)\mathbf{X}_{(1)}\widetilde{\mathbf{v}}\\
&\quad+\lambda_n\left(\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right\|_1-\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_1\right).
\end{aligned}$$

Using the facts that

$$\left\|\frac{2}{n}\left(\boldsymbol{\epsilon}^{\top}+\boldsymbol{\theta}^{*\top}_{(2)}\mathbf{X}_{(2)}^{\top}-(\mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)})^{\top}\mathbf{X}_{(1)}^{\top}\right)\mathbf{X}_{(1)}\right\|_{\infty}=\left\|\frac{2}{n}\mathbf{X}_{(1)}^{\top}\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\right\|_{\infty},$$

which is also greater than $\lambda_n/2$, we get:

$$\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) \geq \frac{1}{n}\widetilde{\mathbf{v}}^\top \mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\widetilde{\mathbf{v}} - \frac{\lambda_n}{2}(\|\widetilde{\mathbf{v}}_J\|_1 + \|\widetilde{\mathbf{v}}_{J^c}\|_1) \\ + \lambda_n \left(\|\widetilde{\mathbf{v}}_{J^c}\|_1 - \|\widetilde{\mathbf{v}}_J\|_1 - 2\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right).$$

Since $\widetilde{\mathbf{v}}^\top \mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\widetilde{\mathbf{v}}$ is non-negative, we get:

$$\begin{aligned} 0 &\geq -\frac{\lambda_n}{2}\left(\|\widetilde{\mathbf{v}}_J\|_1 + \|\widetilde{\mathbf{v}}_{J^c}\|_1\right) + \lambda_n \left(\|\widetilde{\mathbf{v}}_{J^c}\|_1 - \|\widetilde{\mathbf{v}}_J\|_1 - 2\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right) \\ &= -\frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1 - \|\widetilde{\mathbf{v}}_{J^c}\|_1 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right). \end{aligned} \tag{C.35}$$

The rest of the treatment follows the derivation of weakly sparse or compressible models (See, for example, the derivation of the main theorem in Negahban et al. [215]). To this end, we split the error $\widetilde{\mathbf{v}}$ into components within J and components within J^c. Using the inequality (C.35), we get:

$$\|\widetilde{\mathbf{v}}\|_1 = \|\widetilde{\mathbf{v}}_J\|_1 + \|\widetilde{\mathbf{v}}_{J^c}\|_1 \leq 4\|\widetilde{\mathbf{v}}_J\|_1 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 \leq 4\sqrt{k}\|\widetilde{\mathbf{v}}_J\|_2 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1, \tag{C.36}$$

where the last inequality follows from the fact $|J| \leq k$. Using the inequality (C.36) together with the RE condition and τ satisfying the relation in Theorem C.1, we can write:

$$\begin{aligned} \frac{1}{n}\widetilde{\mathbf{v}}^\top \mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\widetilde{\mathbf{v}} &\geq \alpha\|\widetilde{\mathbf{v}}\|_2^2 - \alpha\frac{m-1}{m}\|\widetilde{\mathbf{v}}\|_2^2 - \frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2 \\ &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2 - \frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2, \end{aligned}$$

were we have used the fact $\sqrt{2(a^2+b^2)} \geq (a+b)$. We finally arrive at:

$$\begin{aligned}\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2 - \frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2 - \frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1 - \|\widetilde{\mathbf{v}}_{J^c}\|_1 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right) \\ &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2 - \frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2 - \frac{\lambda_n}{2}\left(3\sqrt{k}\|\widetilde{\mathbf{v}}\|_2 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right) \qquad \text{(C.37)}\end{aligned}$$

where the last step follows from the inequalities $\|\widetilde{\mathbf{v}}_J\|_1 \leq \sqrt{k}\|\widetilde{\mathbf{v}}_J\|_2 \leq \sqrt{k}\|\widetilde{\mathbf{v}}\|_2$ and $\|\widetilde{\mathbf{v}}_{J^c}\|_1 \geq 0$. An application of Lemma C.9 establishes that the right hand side of the inequality (C.37) will be positive if:

$$\|\widetilde{\mathbf{v}}\|^2 \geq \frac{9}{4}\frac{\lambda_n^2}{\alpha^2/m^2}k + \frac{\lambda_n}{\alpha/m}\left(\frac{2}{\lambda_n}\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right). \qquad \text{(C.38)}$$

From the latter inequality, the first claim of the theorem follows using the fact that $\|\mathbf{a}\|_1 \geq \|\mathbf{a}\|_2$; the second claim follows form the first claim together with Eq. (C.36); the last claim follows from the fact that:

$$\frac{1}{n}\widetilde{\mathbf{v}}^\top \mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\widetilde{\mathbf{v}} \leq \frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1 - \|\widetilde{\mathbf{v}}_{J^c}\|_1 + 4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right), \qquad \text{(C.39)}$$

which concludes the proof of the theorem. □

Having established Theorem C.1 and Theorem C.2, the following proposition establishes that the RE condition (C1) holds with high probability:

Proposition C.3 (Verifying RE for $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^\top\mathbf{X}/n$)**.** *Let*

$$\zeta := 54\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)/\mu_{\min}(\breve{\mathbf{A}})}{\Lambda_{\min}(\boldsymbol{\Sigma}_\epsilon)/\mu_{\max}(\mathbf{A})}, \quad \alpha := \frac{\Lambda_{\min}(\boldsymbol{\Sigma}_\epsilon)}{2\mu_{\max}(\mathbf{A})}, \quad \tau := \frac{4\alpha\max\{\zeta^2,1\}}{c}\frac{\log 2p}{n}.$$

Then, for $n \geq C_0 \max\{\zeta^2, 1\} k \log 2p$, there exist constants c_1, c_2 such that

$$\mathbb{P}\left[\widehat{\boldsymbol{\Sigma}} \sim RE(\alpha, \tau)\right] \geq 1 - c_1 \exp(-c_2 n \min\{\zeta^{-2}, 1\}). \tag{C.40}$$

Proof. The proof closely follows that of [209, proof of Proposition 4.2], and is thus omitted for brevity. □

Remark C.2. In order for τ to satisfy $32k\tau/\alpha = (m-1)/m$ for some $m > 1$, we precisely need $n \geq (128/c)(m/(m-1)) \max\{\zeta^2, 1\} k \log(2p)$.

Finally, the following two propositions establish that the deviation conditions (C2) hold with high probability.

Proposition C.4 (Deviation Condition for the *full* Model)**.** *For $n \geq D_0 \log(2p)$, there exist constants d_0, d_1 and $d_2 > 0$ such that*

$$\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \geq \mathbb{Q}(\boldsymbol{\theta}^*, \Sigma_\epsilon)\sqrt{\frac{\log 2p}{n}}\right] \leq \frac{d_1}{(2p)^{d_2}}, \tag{C.41}$$

where

$$\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon) := d_0 \Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1 + \frac{1 + \mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})}\right).$$

Proof. The proof follows that of [209, Proposition 4.3], and is thus omitted for brevity. □

Proposition C.5 (Deviation Condition for the *reduced* Model)**.** *For $n \geq D_0' \log(2p)$,* *there exist constants* d_0', d_1', *and* $d_2' > 0$ *such that*

$$\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}_{(1)}^*\right)\right\|_\infty \geq \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log 2p}{n}}\right] \leq \frac{d_1'}{(2p)^{d_2'}}, \tag{C.42}$$

where

$$\mathbb{Q}'(\boldsymbol{\theta}^*, \Sigma_\epsilon) := d_0'\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1+\frac{1+\mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})}+\frac{3\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*;\boldsymbol{\theta}_{(2)}^*\right]\right\|_2}{\mu_{\min}(\breve{\mathbf{A}})}\right).$$

Proof. In the *reduced* model, the deviation can be expressed as:

$$\begin{aligned}\frac{1}{n}\mathbf{X}_{(1)}\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}_{(1)}^*-\mathbf{X}_{(2)}\mathbf{0}\right) &= \frac{1}{n}\mathbf{X}_{(1)}^\top\left(-\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*+\mathbf{X}_{(2)}\boldsymbol{\theta}_{(2)}^*+\boldsymbol{\epsilon}\right)\\ &= -\frac{1}{n}\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*+\frac{1}{n}\mathbf{X}_{(1)}^\top\mathbf{X}_{(2)}\boldsymbol{\theta}_{(2)}^*+\frac{1}{n}\mathbf{X}_{(1)}^\top\boldsymbol{\epsilon}\end{aligned}$$

The last term can be bounded in a similar fashion as done for the deviation in Proposition C.4. The i^{th} component of the first two terms can be expressed as $\mathbf{e}_i^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*;\boldsymbol{\theta}_{(2)}^*\right]$, for $i = 1, 2, \cdots, p$ where $\mathbf{e}_i$'s are the standard unit bases in $\mathbb{R}^{2p}$. Invoking Basu et al. [209, Proposition 2.4(a)] and noting that $\mathcal{M}(\mathbf{F}) \leq \Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\big/\mu_{\min}(\breve{\mathbf{A}})$, we get:

$$\mathbb{P}\left[\left|\mathbf{e}_i^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*;\boldsymbol{\theta}_{(2)}^*\right]\right| \geq 3\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\breve{\mathbf{A}})}\eta\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}_{(2)}^*;\boldsymbol{\theta}_{(2)}^*\right]\right\|_2\right] \leq$$

$$6\exp[-cn\min\{\eta,\eta^2\}].$$

Using the union bound, we can then get:

$$\mathbb{P}\left[\left|e_i^\top \frac{1}{n}\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)};\boldsymbol{\theta}^*_{(2)}\right] + \mathbf{e}_i^\top \frac{1}{n}\mathbf{X}_{(1)}\boldsymbol{\epsilon}\right| \geq \Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1+\frac{1+\mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})}+ \right.\right.$$
$$\left.\left. + \frac{3\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)};\boldsymbol{\theta}^*_{(2)}\right]\right\|_2}{\mu_{\min}(\breve{\mathbf{A}})}\right)\eta\right] \leq 12\exp[-cn\min\{\eta,\eta^2\}].$$

Using the latter inequality, the statement of the proposition follows from the same arguments used in the proof of Basu et al. [209, Proposition 4.3]. □

C.4 Concentration inequalities and technical lemmas

Lemma C.6. *Given i.i.d. samples from a normal distribution, i.e.,* $w_t \sim \mathcal{N}(0,\sigma^2)$*, the following holds:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n}\frac{w_t^2}{\sigma^2}-1\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{8}\right) \tag{C.43}$$

Proof. define $z_t = \frac{w_t}{\sigma} \sim \mathcal{N}(0,1)$. $\sum_{t=1}^n z_t^2 \sim \chi^2(n)$. z_t^2 is sub-exponential with parameter $(2,4)$, so is the sum $\sum_{t=1}^n z_t^2$ with parameter $(2\sqrt{n},4)$. The claim of the lemma then follows from standard sub-exponential tail bounds. □

Lemma C.7 (Concentration of the *full* Model Deviation)**.** *Under the* full *model, we have:*

$$\mathbb{P}\left[\left|\ell\Big(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\Big)-(\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \geq \Delta_N\right] \leq 2\exp\left(-\frac{n\Delta_N^2}{8(\boldsymbol{\Sigma}_\epsilon)_{1,1}^2}\right).$$

Proof. Note that $\ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) = \sum_{t=1}^n \epsilon_t^2/n$. Since $\epsilon_t \sim \mathcal{N}(0, (\boldsymbol{\Sigma}_\epsilon)_{1,1})$, using Lemma C.6 we get:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^n \frac{\epsilon_t^2}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}} - 1\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{8}\right). \tag{C.44}$$

By letting $\Delta_N := t(\boldsymbol{\Sigma}_\epsilon)_{1,1}$, the claim of the lemma follows. □

Lemma C.8. *Suppose that the deviation conditions (C2) hold. Then, there exist constants $c > 0$, c_{10} and $c_{11} > 0$ such that*

$$\left|\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) - D\right| \leq \Delta_D, \tag{C.45}$$

with probability at least $1 - 2/(2p)^{c_{11}}$, if $n \geq (c_{11}/c)\log(2p)$, where

$$\Delta_D := \left(\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)(\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_1 + 1)\left\|\boldsymbol{\theta}^*_{(2)}\right\|_1 + c_{10}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2\right)\sqrt{\frac{\log 2p}{n}},$$

and

$$D := {\boldsymbol{\theta}^*_{(2)}}^\top(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}.$$

Proof. Since $\widetilde{\boldsymbol{\theta}}^*_{(1)} = \boldsymbol{\theta}^*_{(1)} + \mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}$, we have:

$$\begin{aligned}
\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\left(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\right) &= \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)} - \mathbf{X}_{(2)}\mathbf{0}\right\|_2^2 - \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|-\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)} + \mathbf{X}_{(2)}\boldsymbol{\theta}^*_{(2)} + \boldsymbol{\epsilon}\right\|_2^2 - \frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2 \\
&= \frac{2}{n}\boldsymbol{\epsilon}^\top\mathbf{X}\boldsymbol{\vartheta} + \boldsymbol{\vartheta}^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\boldsymbol{\vartheta},
\end{aligned}$$

where $\boldsymbol{\vartheta} := \left[(-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)})^\top, \boldsymbol{\theta}^{*\top}_{(2)}\right]^\top$. Using the deviation conditions (C2), we get:

$$\begin{aligned}\left|\frac{1}{n}\boldsymbol{\epsilon}^\top \mathbf{X}\boldsymbol{\vartheta}\right| &\leq \left\|\frac{1}{n}\mathbf{X}^\top\epsilon\right\|_\infty \|\boldsymbol{\vartheta}\|_1\\ &\leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log 2p}{n}}\|\boldsymbol{\vartheta}\|_1\\ &\leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log 2p}{n}}\left(\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_1 + 1\right)\left\|\boldsymbol{\theta}^*_{(2)}\right\|_1.\end{aligned}$$

Furthermore, from Basu et al. [209, Proposition 2.4], for any $\boldsymbol{\vartheta} \in \mathbb{R}^{2p}$ and $\eta \geq 0$, there exists a constant $c > 0$ such that:

$$\mathbb{P}\left[\left|\boldsymbol{\vartheta}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \mathbf{C}\right)\boldsymbol{\vartheta}\right| \geq \eta\|\boldsymbol{\vartheta}\|^2\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\breve{\mathbf{A}})}\right] \leq 2\exp[-cn\min\{\eta, \eta^2\}]$$

By further bounding $\boldsymbol{\vartheta}$, we get:

$$\begin{aligned}\mathbb{P}\left[\left|\boldsymbol{\vartheta}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \mathbf{C}\right)\boldsymbol{\vartheta}\right| \geq \eta(\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2^2 + 1)\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\breve{\mathbf{A}})}\right]&\\ \leq 2\exp[-cn\min\{\eta, \eta^2\}],&\end{aligned}$$

where we have used $\boldsymbol{\vartheta}^\top\mathbf{C}\boldsymbol{\vartheta} = \boldsymbol{\theta}^{*\top}_{(2)}(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}$. Next, with the choice of

$$\eta = c_{10}\sqrt{\frac{\log 2p}{n}}\frac{\mu_{\min}(\breve{\mathbf{A}})}{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}\frac{1}{\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2^2 + 1},$$

for some constant c_{10}, the latter concentration inequality establishes that:

$$\left|\boldsymbol{\vartheta}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\boldsymbol{\vartheta} - \boldsymbol{\theta}^{*\top}_{(2)}(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}\right| \leq c_{10}\sqrt{\frac{\log 2p}{n}}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2,$$

with probability at least $1 - 2/(2p)^{c_{11}}$, where

$$c_{11} := c_{10}^2 c \left(\frac{\mu_{\min}(\breve{\mathbf{A}})}{\Lambda_{\max}(\mathbf{\Sigma}_\epsilon)} \frac{1}{\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2^2 + 1} \right)^2,$$

if $n \geq (c_{11}/c)\log(2p)$. This concludes the proof of the lemma. □

Finally, the following elementary lemma provides a useful technical tool for simplifying some of the algebraic inequalities:

Lemma C.9. *The quadratic function $f(x) = ax^2 - bx - c$ with $a, b, c > 0$ is positive for all real x satisfying $x^2 \geq \left(\frac{b}{a}\right)^2 + 2\frac{c}{a}$.*

Proof. Noting that,

$$\sqrt{1+z} \leq \frac{1}{2}z + 1 \quad \forall \quad z > 0, \tag{C.46}$$

the positive root of $f(x)$, denoted by x_+ can be upper bounded:

$$\begin{aligned} x_+^2 &= \left(\frac{b}{2a} + \sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}} \right)^2 && \text{(C.47)} \\ &= 2\left(\frac{b}{2a}\right)^2 + \frac{c}{a} + 2\frac{b}{2a}\sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}} && \text{(C.48)} \\ &\leq 2\left(\frac{b}{2a}\right)^2 + \frac{c}{a} + 2\left(\frac{b}{2a}\right)^2 \left(\frac{1}{2}\left(\frac{2a}{b}\right)^2 \frac{c}{a} + 1 \right) && \text{(C.49)} \\ &= \left(\frac{b}{a}\right)^2 + 2\frac{c}{a}. && \text{(C.50)} \end{aligned}$$

Then, $x^2 \geq x_+^2$ implies $ax^2 - bx - c > 0$. □

Bibliography

[1] D. J. Heeger and D. Ress. What does fMRI tell us about neuronal activity? *Nat. Rev. Neurosci.*, 3(2):142, 2002. doi: `10.1038/nrn730`.

[2] J. F. Hipp, D. J. Hawellek, M. Corbetta, M. Siegel, and A. K. Engel. Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nat. Neurosci.*, 15(6):884, 2012. doi: `10.1038/nn.3101`.

[3] E. C. Lalor, B. A. Pearlmutter, R. B. Reilly, G. McDarby, and J. J. Foxe. The VESPA: a method for the rapid estimation of a visual evoked potential. *Neuroimage*, 32(4):1549–1561, 2006. doi: `10.1016/j.neuroimage.2006.05.054`.

[4] N. Ding and J. Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.*, 109(29):11854–11859, 2012. doi: `10.1073/pnas.1205381109`.

[5] D. S. Bassett and E. Bullmore. Small-world brain networks. *Neuroscientist*, 12(6):512–523, 2006. doi: `10.1177/1073858406293182`.

[6] S. V. David, N. Mesgarani, and S. A. Shamma. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Netw. Comput. Neural Syst.*, 18(3):191–212, 2007. doi: `10.1080/09548980701609235`.

[7] N. A. Francis, D. E. Winkowski, A. Sheikhattar, K. Armengol, B. Babadi, and P. O. Kanold. Small networks encode decision-making in primary auditory cortex. *Neuron*, 97(4):885–897, 2018. doi: `10.1016/j.neuron.2018.01.019`.

[8] W. J. Freeman. *Mass action in the nervous system.* Academic Press, 1975.

[9] W. J. Freeman. Nonlinear gain mediating cortical stimulus-response relations. *Biological Cybernetics*, 33:237–247, 1979.

[10] J. Fritz, S. Shamma, M. Elhilali, and D. Klein. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.*, 6(11):1216, 2003. doi: `10.1038/nn1141`.

[11] D. J. Thomson. Spectrum Estimation and Harmonic Analysis. *Proc. IEEE*, 70(9):1055–1096, Sep 1982. doi: `10.1109/PROC.1982.12433`.

[12] D. B. Percival and A. T. Walden. *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993. ISBN 0521435412.

[13] T. P. Bronez. On the performance advantage of multitaper spectral analysis. *IEEE Trans. Signal Process*, 40(12):2941–2946, 1992. doi: `10.1109/78.175738`.

[14] P. Das and B. Babadi. A Bayesian multitaper method for nonstationary data with application to EEG analysis. In *2017 IEEE Signal Process. Med. Biol. Symp.*, pages 1–5, 2017. doi: `10.1109/SPMB.2017.8257055`.

[15] P. Das and B. Babadi. Dynamic Bayesian multitaper spectral analysis. *IEEE Trans. Signal Process.*, 66(6), 2018. doi: `10.1109/TSP.2017.2787146`.

[16] P. Das and B. Babadi. Multitaper spectral analysis of neuronal spiking activity driven by latent stationary processes. *Signal Process.*, 170:107429, 2020. doi: `10.1016/j.sigpro.2019.107429`.

[17] G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.*, 16(13):4207–4221, 1996. doi: `10.1523/JNEUROSCI.16-13-04207.1996`.

[18] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J. Neurophysiol.*, 102(1):349–359, 2009. doi: `10.1152/jn.90896.2008`.

[19] N. Ding and J. Z. Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.*, 107, 2012. doi: `10.1152/jn.00297.2011`.

[20] N. Ding and J. Z. Simon. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J. Neurosci.*, 33(13):5728–5735, 2013. doi: `10.1523/JNEUROSCI.5297-12.2013`.

[21] N. Ding and J. Z. Simon. Robust cortical encoding of slow temporal modulations of speech. In *Basic Asp. Hear.*, pages 373–381. Springer, 2013. doi: `10.1007/978-1-4614-1590-9_41`.

[22] C. Brodbeck, A. Presacco, and J. Z. Simon. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *Neuroimage*, 172:162–174, May 2018. doi: `10.1016/j.neuroimage.2018.01.042`.

[23] P. Das, C. Brodbeck, J. Z. Simon, and B. Babadi. Cortical Localization of the Auditory Temporal Response Function from MEG via Non-Convex Optimization. *Asilomar Conf. Signals, Syst. Comput. Pacific Grove, CA*, Oct 2018. doi: `10.1109/ACSSC.2018.8645204`.

[24] P. Das, C. Brodbeck, J. Z. Simon, and B. Babadi. Neuro-current response functions: A unified approach to MEG source analysis under the continuous stimuli paradigm. *Neuroimage*, 211:116528, May 2020. doi: 10.1016/j.neuroimage.2020.116528.

[25] S. L. Bressler and A. K. Seth. Wiener-Granger causality: A well established methodology. *Neuroimage*, 58(2):323–329, 2011. doi: 10.1016/j.neuroimage.2010.02.059.

[26] K. Friston, R. Moran, and A. K. Seth. Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.*, 23(2):172–178, 2013. doi: 10.1016/j.conb.2012.11.010.

[27] A. K. Seth, A. B. Barrett, and L. Barnett. Granger Causality Analysis in Neuroscience and Neuroimaging. *J. Neurosci.*, 35(8):3293–3297, 2015. doi: 10.1523/JNEUROSCI.4399-14.2015.

[28] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. URL http://www.jstor.org/stable/1912791.

[29] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.*, 77(378):304–313, 1982. doi: 10.1080/01621459.1982.10477803.

[30] J. F. Geweke. Measures of Conditional Linear Dependence and Feedback Between Time Series. *J. Am. Stat. Assoc.*, 79(388):907–915, 1984. doi: 10.1080/01621459.1984.10477110.

[31] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, 1974. doi: 10.1007/978-1-4612-1694-0_16.

[32] G. Schwarz. Estimating the Dimension of a Model. *Ann. Stat.*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

[33] A. Wald. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Trans. Am. Math. Soc.*, 54(3):426–482, 1943. URL http://www.jstor.org/stable/1990256.

[34] R. R. Davidson and W. E. Lever. The Limiting Distribution of the Likelihood Ratio Statistic under a Class of Local Alternatives. *Sankhyā Indian J. Stat. Ser. A*, 32(2):209–224, 1970. URL http://www.jstor.org/stable/25049656.

[35] A. K. Seth. A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Methods*, 186(2):262–273, 2010. doi: 10.1016/j.jneumeth.2009.11.020.

[36] P. P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophys. J.*, 76(2):691–708, 1999. doi: `10.1016/S0006-3495(99)77236-X`.

[37] M. T. Bahadori and Y. Liu. An examination of practical Granger causality inference. In *Proc. 2013 SIAM Int. Conf. data Min.*, pages 467–475. SIAM, 2013. doi: `10.1137/1.9781611972832.52`.

[38] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pages 66–75, 2007. doi: `10.1145/1281192.1281203`.

[39] S. Basu, A. Shojaie, and G. Michailidis. Network Granger Causality with Inherent Grouping Structure. *J. Mach. Learn. Res.*, 16(1):417–453, 2015. doi: `10.5555/2789272.2789285`.

[40] G. Michailidis and F. D'Alché-Buc. Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Math. Biosci.*, 246 (2):326–334, 2013. doi: `10.1016/j.mbs.2013.10.003`.

[41] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, Berlin, 2011. ISBN 9783642201929.

[42] A. Javanmard, H. Javadi, and Others. False discovery rate control via debiased lasso. *Electron. J. Stat.*, 13(1):1212–1253, 2019. doi: `10.1214/19-EJS1554`.

[43] T. F. Quatieri. *Discrete-time Speech Signal Processing: Principles and Practice.* Prentice Hall, 2008.

[44] J. S. Lim. *Two-dimensional signal and image processing.* 1990.

[45] G. Buzsáki and G. Buzsaki. *Rhythms of the Brain.* Oxford University Press, 2009. ISBN 9780199863716. doi: `10.1093/acprof:oso/9780195301069.001.0001`.

[46] W. J. Emery and R. E. Thomson. *Data analysis methods in physical oceanography.* Elsevier Science, 2001.

[47] M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and Others. Advanced spectral methods for climatic time series. *Rev. Geophys.*, 40(1):1003, 2002. doi: `10.1029/2000RG000092`.

[48] Ö. Yilmaz. *Seismic data analysis: processing, inversion, and interpretation of seismic data.* Number 10. SEG Books, 2001. doi: `10.1190/1.9781560801580`.

[49] L. Cohen. *Time-Frequency Analysis.* Prentice-Hall, Englewood Cliffs, NJ, 1995.

[50] W. Martin and P. Flandrin. Wigner-Ville spectral analysis of nonstationary processes. *IEEE Trans. Acoust.*, 33(6):1461–1470, 1985. doi: `10.1109/TASSP.1985.1164760`.

[51] M. B. Priestley. Evolutionary spectra and non-stationary processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 204–237, 1965. doi: `10.1111/j.2517-6161.1965.tb01488.x`.

[52] G. Matz, F. Hlawatsch, and W. Kozek. Generalized evolutionary spectral analysis and the Weyl spectrum of nonstationary random processes. *IEEE Trans. Signal Process.*, 45(6):1520–1534, 1997. doi: `10.1109/78.599994`.

[53] G. Matz and F. Hlawatsch. Nonstationary spectral analysis based on time-frequency operator symbols and underspread approximations. *IEEE Trans. Inf. Theory*, 52(3):1067–1086, 2006. doi: `10.1109/TIT.2005.864419`.

[54] W. Kozek. *Matched Weyl-Heisenberg expansions of nonstationary environments*. PhD thesis, University of Technology Vienna, 1997.

[55] J. K. Hammond and P. R. White. The analysis of non-stationary signals using time-frequency methods. *J. Sound Vib.*, 190(3):419–447, 1996. doi: `10.1006/jsvi.1996.0072`.

[56] B. J. W. S. Rayleigh. *The collected optics papers of Lord Rayleigh*. Optical Society of America, 1994.

[57] B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982. doi: `10.1137/1.9781611970319`.

[58] N. E. Huang and Others. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. Lond. A.*, 454(1971):903–995, 1998. doi: `10.1098/rspa.1998.0193`.

[59] I. Daubechies, J. Lu, and H.-T. Wu. Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal.*, 30 (2):243–261, 2011. doi: `10.1016/j.acha.2010.08.002`.

[60] I. Daubechies, Y. G. Wang, and H.-t. Wu. ConceFT: concentration of frequency and time via a multitapered synchrosqueezed transform. *Phil. Trans. R. Soc. A*, 374(2065):20150193, 2016. doi: `10.1098/rsta.2015.0193`.

[61] J. Xiao and P. Flandrin. Multitaper time-frequency reassignment for nonstationary spectrum estimation and chirp enhancement. *IEEE Trans. Signal Process.*, 55(6):2851–2860, 2007. doi: `10.1109/TSP.2007.893961`.

[62] M. Jachan, G. Matz, and F. Hlawatsch. Time-Frequency ARMA Models and Parameter Estimators for Underspread Nonstationary Random Processes. *IEEE Trans. Signal Process.*, 55(9):4366–4381, Sep 2007. doi: `10.1109/TSP.2007.896265`.

[63] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown. Robust spectrotemporal decomposition by iteratively reweighted least squares. *Proc. Natl. Acad. Sci.*, 111(50):E5336–E5345, 2014. doi: 10.1073/pnas.1320637111.

[64] L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013. doi: 10.1007/978-1-4757-3454-6.

[65] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J TIME SER ANAL*, 3(4):264–653, Jul 1982. doi: 10.1111/j.1467-9892.1982.tb00349.x.

[66] A. C. Smith and E. N. Brown. Estimating a State-Space Model from Point Process Observations. *Neural Comput.*, 15(5):965–991, 2003. doi: 10.1162/089976603765202622.

[67] G. Kitagawa. Non-Gaussian State-Space Modeling of Nonstationary Time Series. *J. Amer. Stat. Assoc.*, 82(400):1032–1041, 1987. doi: 10.1080/01621459.1987.10478534.

[68] T. Bohlin. Analysis of EEG signals with changing spectra using a short-word Kalman estimator. *Math. Biosci.*, 35(3-4):221–259, 1977. doi: 10.1016/0025-5564(77)90026-8.

[69] M. Loève. *Probability Theory II*. D. Van Nostrand Co., London, 1963. ISBN 978-1-4684-9466-2.

[70] D. J. Thomson. Multitaper Analysis of Nonstationary and Nonlinear Time Series Data. In *Nonlinear nonstationary signal Process.*, pages 317–394. London, UK: Cambridge Univ. Press, 2000.

[71] F. Hlawatsch and F. Auger. *Time-frequency analysis*. John Wiley & Sons, 2013.

[72] M. B. Priestley. *Spectral analysis and time series*. Academic press, 1982.

[73] D. Slepian. Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V: the discrete case. *Bell Syst. Tech. J.*, 57(5):1371–1430, May 1978. doi: 10.1002/j.1538-7305.1978.tb02104.x.

[74] B. Babadi and E. N. Brown. A review of multitaper spectral analysis. *IEEE Trans. Biomed. Eng.*, 61(5):1555–1564, May 2014. doi: 10.1109/TBME.2014.2311996.

[75] J. L. Powell. Estimation of semiparametric models. *Handb. Econom.*, 4:2443–2521, 1994.

[76] O. Rosen, D. S. Stoffer, and S. Wood. Local spectral analysis via a Bayesian mixture of smoothing splines. *J. Amer. Stat. Assoc.*, 104(485):249–262, 2009. doi: 10.1198/jasa.2009.0118.

[77] R. L. Prentice. A Log Gamma Model and Its Maximum Likelihood Estimation. *Biometrika*, 61(3):539–544, 1974. doi: 10.1093/biomet/61.3.539.

[78] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Ser. B*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

[79] H. E. Rauch, C. T. Striebel, and T. F. Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, 3:1445–1450, Aug 1965. doi: 10.2514/3.3166.

[80] P. D. E. Jong and M. J. Mackinnon. Covariances for smoothed estimates in state space models. *Biometrika*, 75(3):601, 1988. doi: 10.1093/biomet/75.3.601.

[81] K. Lange. *Optimization*. Springer, 2004. ISBN 978-1-4614-5838-8. doi: 10.1007/978-1-4757-4182-7.

[82] L. Fahrmeir. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *J. Amer. Stat. Assoc.*, 87 (418):501–509, 1992. doi: 10.1080/01621459.1992.10475232.

[83] P. Das. Dynamic Bayesian Multitaper Spectral Estimators, 2017. URL https://github.com/proloyd/DBMT.

[84] L. De Gennaro, M. Ferrara, L. D. Gennaro, and M. Ferrara. Sleep spindles: an overview. *Sleep Med. Rev.*, 7(5):423–440, 2003. doi: 10.1053/smrv.2002.0252.

[85] M. Steriade, D. A. McCormick, and T. J. Sejnowski. Thalamocortical oscillations in the sleeping and aroused brain. *Science*, 262(5134):679–685, Oct 1993. doi: 10.1126/science.8235588.

[86] R. Garg, A. L. Varna, A. Hajj-Ahmad, and M. Wu. Seeing ENF: Power-Signature-Based Timestamp for Digital Multimedia via Optical Sensing and Signal Processing. *IEEE Trans. Inf. Forensics Secur.*, 8(9):1417–1432, 2013. doi: 10.1109/TIFS.2013.2272217.

[87] A. Hajj-Ahmad, R. Garg, and M. Wu. Spectrum Combining for ENF Signal Estimation. *IEEE Signal Process. Lett.*, 20(9):885–888, Sep 2013. doi: 10.1109/LSP.2013.2272523.

[88] S. S. Haykin. *Adaptive filter theory*. Prentice-Hall information and system sciences series. Prentice Hall, 1991. ISBN 9780130132369.

[89] B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Englewood Cliffs, 1979.

[90] K. S. Lii and M. Rosenblatt. Prolate spheroidal spectral estimates. *Stat. Probab. Lett.*, 78(11):1339–1348, 2008. doi: 10.1016/j.spl.2008.05.022.

[91] L. M. Ward. Synchronous neural oscillations and cognitive processes. *Trends Cogn. Sci.*, 7(12):553–559, 2003. doi: 10.1016/j.tics.2003.10.012.

[92] A. T. Walden. A unified view of multitaper multivariate spectral estimation. *Biometrika*, 87(4):767–788, 2000. doi: 10.1093/biomet/87.4.767.

[93] S. E. Kim, M. K. Behr, D. Ba, and E. N. Brown. State-space multitaper time-frequency analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 115(1):E5–E14, 2018. doi: 10.1073/pnas.1702877115.

[94] M. Chalk, J. L. Herrero, M. A. Gieselmann, L. S. Delicato, S. Gotthardt, and A. Thiele. Attention Reduces Stimulus-Driven Gamma Frequency Oscillations and Spike Field Coherence in V1. *Neuron*, 66(1):114–125, 2010. doi: 10.1016/j.neuron.2010.03.013.

[95] B. C. Lewandowski and M. Schmidt. Short bouts of vocalization induce long-lasting fast gamma oscillations in a sensorimotor nucleus. *J. Neurosci.*, 31 (39):13936–13948, 2011. doi: 10.1523/JNEUROSCI.6809-10.2011.

[96] I. M. Park, S. Seth, A. R. Paiva, L. Li, and J. C. Principe. Kernel methods on spike train space for neuroscience: A tutorial. *IEEE Signal Process. Mag.*, 30(4):149–160, 2013. doi: 10.1109/MSP.2013.2251072.

[97] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004.

[98] L. Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Netw. Comput. Neural Syst.*, 15(4):243–262, 2004. doi: 10.1088/0954-898X_15_4_002.

[99] A. Wu, N. A. Roy, S. Keeley, and J. W. Pillow. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In *Adv. Neural Inf. Process. Syst. 30*, pages 3497–3506, 2017. URL http://papers.nips.cc/paper/6941-gaussian-process-based-nonlinear-latent-structure-discovery-in-multivariate-spike-train-data.

[100] S. Xu, Y. Li, T. Huang, and R. H. Chan. A sparse multiwavelet-based generalized Laguerre-Volterra model for identifying time-varying neural dynamics from spiking activities. *Entropy*, 19(8):425, 2017. doi: 10.3390/e19080425.

[101] P. M. Djuric, M. Vemula, and M. F. Bugallo. Target tracking by particle filtering in binary sensor networks. *IEEE Trans. Signal Process.*, 56(6):2229–2238, 2008. doi: 10.1109/TSP.2007.916140.

[102] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nat. Neurosci.*, 7(5):456–461, 2004. doi: 10.1038/nn1228.

[103] S. Miran, P. L. Purdon, E. N. Brown, and B. Babadi. Robust Estimation of Sparse Narrowband Spectra from Neuronal Spiking Data. *IEEE Trans. Biomed. Eng.*, 64(10):2468–2474, 2017. doi: 10.1109/TBME.2016.2642783.

[104] L. D. Lewis, V. S. Weiner, E. A. Mukamel, J. A. Donoghue, E. N. Eskandar, J. R. Madsen, W. S. Anderson, L. R. Hochberg, S. S. Cash, E. N. Brown, and P. L. Purdon. Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness. *Proc. Natl. Acad. Sci.*, 109(49):E3377–E3386, 2012. doi: 10.1073/pnas.1210907109.

[105] A. Sheikhattar, J. B. Fritz, S. A. Shamma, and B. Babadi. Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis. *IEEE Trans. Signal Process.*, 64(8):2026–2039, 2016. doi: 10.1109/TSP.2015.2512560.

[106] P. McCullagh and J. A. Nelder. Generalized linear models, 1989.

[107] P. A. Lewis and G. S. Shedler. Simulation of Nonhomogeneous Poisson Processes By Thinning. *Nav. Res. Logist. Q.*, 26(3):403–413, 1979. doi: 10.1002/nav.3800260304.

[108] Y. Ogaata. On Lewis' Simulation Method for Point Processes. *IEEE Trans. Inf. Theory*, 27(1):23–31, 1981. doi: 10.1109/TIT.1981.1056305.

[109] P. Billingsley. *Probability and measure.* John Wiley & Sons, New York, 1986.

[110] S. Haykin. *Adaptive Filter Theory.* Prentice-Hall information and system sciences series. Pearson, 4 edition, 2001. ISBN 9780130132369.

[111] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge university press, 2004. ISBN 0521833787.

[112] P. Das. The Point Process Multitaper Method, 2018. URL https://github.com/proloyd/PMTM.

[113] A. M. H. J. Aertsen and P. I. M. Johannesma. Reverse-correlation methods in auditory research. *Q. Rev. Biophys.*, 16(3):341–414, 1983. doi: 10.1017/S0033583500005126.

[114] A. M. H. J. Aertsen, J. H. J. Olders, and P. I. M. Johannesma. Biological Cybernetics Spectro-Temporal Receptive Fields of Auditory Neurons in the Grassfrog. *Biol. Cybern*, 39:209, 1981. doi: 10.1007/bf00342772.

[115] D. Ringach and R. Shapley. Reverse correlation in neurophysiology. *Cogn. Sci.*, 28(2):147–166, 2004. doi: 10.1207/s15516709cog2802_2.

[116] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.*, 25(19), 2015. doi: 10.1016/j.cub.2015.08.030.

[117] C. Brodbeck, L. E. Hong, and J. Z. Simon. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Curr. Biol.*, 28(24): 3976–3983, 2018. doi: 10.1016/j.cub.2018.10.042.

[118] E. C. Lalor and J. J. Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *Eur. J. Neurosci.*, 31(1): 189–193, 2010. doi: 10.1111/j.1460-9568.2009.07055.x.

[119] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi. Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *Neuroimage*, 124:906–917, 2016. doi: 10.1016/j.neuroimage.2015.09.048.

[120] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma. Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.*, 123, 2008. doi: 10.1121/1.2816572.

[121] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 2014. doi: 10.1126/science.1245994.

[122] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Curr. Biol.*, 28(5):803–809, Mar 2018. doi: 10.1016/j.cub.2018.01.080.

[123] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, 12(3):289–316, Aug 2001. doi: 10.1080/net.12.3.289.316.

[124] C. K. Machens, M. S. Wehr, and A. M. Zador. Linearity of cortical receptive fields measured with natural sounds. *J. Neurosci.*, 24(5):1089–1100, 2004. doi: 10.1523/JNEUROSCI.4445-03.2004.

[125] S. Akram, J. Z. Simon, and B. Babadi. Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments. *IEEE Trans. Biomed. Eng.*, 64(8):1896–1905, 2017. doi: 10.1109/TBME.2016.2628884.

[126] F. E. Theunissen. STRFPak: 5.3, 2007. URL http://strfpak.berkeley.edu/.

[127] F. E. Theunissen. STRFlab: 1.45, 2010. URL http://strflab.berkeley.edu/.

[128] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating

neural signals to continuous stimuli. *Front. Hum. Neurosci.*, 10:604, 2016. doi: 10.3389/fnhum.2016.00604.

[129] C. Brodbeck. Eelbrain: 0.27, 2017.

[130] N. Mesgarani and E. F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, pages 233–236, 2012. doi: 10.1038/nature11020.

[131] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. Reconstructing speech from human auditory cortex. *PLoS Biol.*, 10(1), Jan 2012. doi: 10.1371/journal.pbio.1001251.

[132] J.-M. Schoffelen and J. Gross. Source connectivity analysis with MEG and EEG. *Hum. Brain Mapp.*, 30(6):1857–1865, 2009. doi: 10.1002/hbm.20745.

[133] S. Baillet, J. C. Mosher, and R. M. Leahy. Electromagnetic brain mapping. *IEEE Signal Process. Mag.*, 18(6):14–30, 2001. doi: 10.1109/79.962275.

[134] I. F. Gorodnitsky, J. S. George, and B. D. Rao. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. *Electroencephalogr. Clin. Neurophysiol.*, 95(4):231–251, 1995. doi: 10.1016/0013-4694(95)00107-A.

[135] M. A. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian estimation for MEG inverse problem. *Neuroimage*, 23(3):806–826, 2004. doi: 10.1016/j.neuroimage.2004.06.037.

[136] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout. Multiple sparse priors for the M/EEG inverse problem. *Neuroimage*, 39(3):1104–1120, 2008. doi: 10.1016/j.neuroimage.2007.09.048.

[137] S. Haufe, V. V. Nikulin, A. Ziehe, K. R. Müller, and G. Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *Neuroimage*, 2008. doi: 10.1016/j.neuroimage.2008.04.246.

[138] D. P. Wipf, J. P. Owen, H. T. Attias, K. Sekihara, and S. S. Nagarajan. Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG. *Neuroimage*, 49(1):641–655, 2010. doi: 10.1016/j.neuroimage.2009.06.083.

[139] D. Wipf and S. Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *Neuroimage*, 44(3):947–966, 2009. doi: 10.1016/j.neuroimage.2008.02.059.

[140] M. Fukushima, O. Yamashita, A. Kanemura, S. Ishii, M. Kawato, and M. A. Sato. A state-space modeling approach for localization of focal current sources from MEG. *IEEE Trans. Biomed. Eng.*, 59(6):1561–1571, 2012. doi: `10.1109/TBME.2012.2189713`.

[141] M. Fukushima, O. Yamashita, T. R. Knösche, and M. aki Sato. MEG source reconstruction based on identification of directed source interactions on whole-brain anatomical networks. *Neuroimage*, 105:408–427, 2015. doi: `10.1016/j.neuroimage.2014.09.066`.

[142] W. Wu, S. Nagarajan, and Z. Chen. Bayesian Machine Learning: EEG/MEG signal processing measurements. *IEEE Signal Process. Mag.*, 33(1):14–36, 2016. doi: `10.1109/MSP.2015.2481559`.

[143] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and Others. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.*, 7:267, 2013. doi: `10.3389/fnins.2013.00267`.

[144] M. Dannhauer, E. Lämmel, C. H. Wolters, and T. R. Knösche. Spatio-temporal Regularization in Linear Distributed Source Reconstruction from EEG/MEG: A Critical Evaluation. *Brain Topogr.*, 26(2):229–246, Apr 2013. doi: `10.1007/s10548-012-0263-9`.

[145] T. R. Knösche, M. Gräser, and A. Anwander. Prior knowledge on cortex organization in the reconstruction of source current densities from EEG. *Neuroimage*, 67:7–24, 2013. doi: `10.1016/j.neuroimage.2012.11.013`.

[146] B. Babadi, G. Obregon-Henao, C. Lamus, M. S. Hämäläinen, E. N. Brown, and P. L. Purdon. A subspace pursuit-based iterative greedy hierarchical solution to the neuromagnetic inverse problem. *Neuroimage*, 87:427–443, 2014. doi: `10.1016/j.neuroimage.2013.09.008`.

[147] P. Krishnaswamy, G. Obregon-Henao, J. Ahveninen, S. Khan, B. Babadi, J. E. Iglesias, M. S. Hämäläinen, and P. L. Purdon. Sparsity enables estimation of both subcortical and cortical activity from MEG and EEG. *Proc. Natl. Acad. Sci.*, 114(48):E10465—-E10474, Nov 2017. doi: `10.1073/pnas.1705414114`.

[148] E. Pirondini, B. Babadi, G. Obregon-Henao, C. Lamus, W. Q. Malik, M. S. Hämäläinen, and P. L. Purdon. Computationally Efficient Algorithms for Sparse, Dynamic Solutions to the EEG Source Localization Problem. *IEEE Trans. Biomed. Eng.*, 65(6):1359–1372, 2018. doi: `10.1109/TBME.2017.2739824`.

[149] K. Liu, Z. Yu, W. Wu, Z. Gu, J. Zhang, L. Cen, S. Nagarajan, and Y. Li. Bayesian Electromagnetic Spatio-Temporal Imaging of Extended Sources based on Matrix Factorization. *IEEE Trans. Biomed. Eng.*, 66(9):2457–2469, 2019. doi: `10.1109/TBME.2018.2890291`.

[150] T. C. Handy. *Event-related potentials: A methods handbook*. MIT press, 2005.

[151] I. Gazzaniga and R. B. Ivry. *Cognitive Neuroscience: The Biology of the Mind*. New York: WW Norton and Company Press, 2009.

[152] S. J. Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[153] A. Presacco, J. Z. Simon, and S. Anderson. Effect of informational content of noise on speech representation in the aging midbrain and cortex. *J. Neurophysiol.*, 116(5):2356–2367, 2016. doi: `10.1152/jn.00373.2016`.

[154] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. MNE software for processing MEG and EEG data. *Neuroimage*, 86:446–460, 2014. doi: `10.1016/j.neuroimage.2013.10.027`.

[155] W. E. Kincses, C. Braun, S. Kaiser, W. Grodd, H. Ackermann, and K. Mathiak. Reconstruction of extended cortical sources for EEG and MEG based on a Monte-Carlo-Markov-Chain estimator. *Hum. Brain Mapp.*, 18(2):100–110, Feb 2003. doi: `10.1002/hbm.10079`.

[156] A. M. Dale and M. I. Sereno. Improved localization of cortical activity by combining EEG and MEG with mri cortical surface reconstruction: A linear approach. *J. Cogn. Neurosci.*, 5(2):162–176, 1993. doi: `10.1162/jocn.1993.5.2.162`.

[157] J. J. Bonaiuto, F. Afdideh, M. Ferez, K. Wagstyl, J. Mattout, M. Bonnefond, G. R. Barnes, and S. Bestmann. Estimates of cortical column orientation improve MEG source inversion. *bioRxiv*, art. 810267, Oct 2019. doi: `10.1101/810267`.

[158] J. Sarvas. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Phys. Med. Biol.*, 32(1):11, 1987. doi: `10.1088/0031-9155/32/1/004`.

[159] J. C. Mosher, R. M. Leahy, and P. S. Lewis. EEG and MEG: forward solutions for inverse methods. *IEEE Trans. Biomed. Eng.*, 46(3):245–259, 1999. doi: `10.1109/10.748978`.

[160] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65(2):413–497, Apr 1993. doi: `10.1103/RevModPhys.65.413`.

[161] D. A. Engemann and A. Gramfort. Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *Neuroimage*, pages 328–342, 2015. doi: `10.1016/j.neuroimage.2014.12.040`.

[162] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001. doi: 10.1137/S1064827596304010.

[163] H. G. Feichtinger and T. Strohmer. *Gabor analysis and algorithms: Theory and applications*. Springer Science & Business Media, 2012.

[164] F.-H. Lin, T. Witzel, S. P. Ahlfors, S. M. Stufflebeam, J. W. Belliveau, and M. S. Hämäläinen. Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage*, 31(1):160–171, 2006. doi: 10.1016/j.neuroimage.2005.11.054.

[165] C. Stahlhut, H. T. Attias, K. Sekihara, D. Wipf, L. K. Hansen, and S. S. Nagarajan. A hierarchical Bayesian M/EEG imagingmethod correcting for incomplete spatio-temporal priors. In *IEEE 10th Int. Symp. Biomed. Imaging*, pages 560–563. IEEE, 2013. ISBN 9781467364546. doi: 10.1109/ISBI.2013.6556536.

[166] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985. doi: 10.1007/978-1-4757-4286-2.

[167] C. Lamus, M. S. Hämäläinen, S. Temereanca, E. N. Brown, and P. L. Purdon. A spatiotemporal dynamic distributed solution to the MEG inverse problem. *Neuroimage*, 63(2):894–909, 2012. doi: 10.1016/j.neuroimage.2011.11.020.

[168] A. Gramfort, M. Kowalski, and M. Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Phys. Med. Biol.*, 57(7):1937, 2012. doi: 10.1088/0031-9155/57/7/1937.

[169] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009. doi: 10.1137/080716542.

[170] Y. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005. doi: 10.1007/s10107-004-0552-5.

[171] T. Goldstein, C. Studer, and R. Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. *arXiv Prepr. arXiv1411.3406*, 2014.

[172] S. J. Wright. Coordinate descent algorithms. *Math. Program.*, 151(1):3–34, Jun 2015. doi: 10.1007/s10107-015-0892-3.

[173] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory*, 38(2):824–839, Mar 1992. doi: 10.1109/18.119739.

[174] P. Das. *neuro-currentRF: A Unified Approach to MEG Source Analysis under the Continuous Stimuli Paradigm*. GitHub, 2019. URL https://github.com/proloyd/neuro-currentRF.

[175] A. Presacco, J. Simon, and S. Anderson. EEG-MEG. *available Digit. Repos. Univ. Maryl.*, 2018. URL http://hdl.handle.net/1903/21184.

[176] S. Taulu and J. Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys. Med. Biol.*, 51(7): 1759–1768, Mar 2006. doi: 10.1088/0031-9155/51/7/008.

[177] A. J. Bell and T. J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.*, 7(6):1129–1159, 1995. doi: 10.1162/neco.1995.7.6.1129.

[178] B. Fischl. FreeSurfer. *Neuroimage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

[179] M. Brysbaert and B. New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods*, 41(4):977–990, 2009. doi: 10.3758/BRM.41.4.977.

[180] M. Westerlund, I. Kastner, M. Al Kaabi, and L. Pylkkänen. The LATL as locus of composition: MEG evidence from English and Arabic. *Brain Lang.*, 141:124–134, 2015. doi: 10.1016/j.bandl.2014.12.003.

[181] S. M. Smith and T. E. Nichols. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98, 2009. doi: 10.1016/j.neuroimage.2008.03.061.

[182] S. K. Scott and I. S. Johnsrude. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.*, 26(2):100–107, Feb 2003. doi: 10.1016/S0166-2236(02)00037-1.

[183] G. Hickok and D. Poeppel. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92 (1-2):67–99, 2004. doi: 10.1016/j.cognition.2003.10.011.

[184] M. H. Davis and I. S. Johnsrude. Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hear. Res.*, 2007. doi: 10.1016/j.heares.2007.01.014.

[185] K. Okada, F. Rong, J. Venezia, W. Matchin, I. H. Hsieh, K. Saberi, J. T. Serences, and G. Hickok. Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cereb. Cortex*, 20(10):2486–2495, 2010. doi: 10.1093/cercor/bhp318.

[186] J. E. Peelle, I. S. Johnsrude, and M. H. Davis. Hierarchical processing for speech in human auditory cortex and beyond. *Front. Hum. Neurosci.*, 4, 2010. doi: 10.3389/fnhum.2010.00051.

[187] W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, X. Frédéric, and E. Theunissen. The Hierarchical Cortical Organization of Human Speech Processing. *J. Neurosci.*, 37(27):6539–6557, 2017. doi: `10.1523/JNEUROSCI.3267-16.2017`.

[188] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur. J. Neurosci.*, 15(2):399–402, 2002. doi: `10.1046/j.0953-816x.2001.01874.x`.

[189] S. M. Wilson, A. P. Saygin, M. I. Sereno, and M. Iacoboni. Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.*, 7 (7):701–702, Jul 2004. doi: `10.1038/nn1263`.

[190] F. Pulvermüller, M. Huss, F. Kherif, F. M. Del Prado Martin, O. Hauk, and Y. Shtyrov. Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U. S. A.*, 103(20):7865–7870, May 2006. doi: `10.1073/pnas.0509989103`.

[191] J. T. Crinion, M. A. Lambon-Ralph, E. A. Warburton, D. Howard, and R. J. S. Wise. Temporal lobe regions engaged during normal speech comprehension. *Brain*, 126(5):1193–1201, May 2003. doi: `10.1093/brain/awg104`.

[192] I. Dewitt and J. P. Rauschecker. Phoneme and word recognition in the auditory ventral stream. 109(8):2709, 2012. doi: `10.1073/pnas.1113427109`.

[193] P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.*, 36(6):2014–2026, Feb 2016. doi: `10.1523/JNEUROSCI.1779-15.2016`.

[194] J. H. Kaas and T. A. Hackett. 'What' and 'where' processing in auditory cortex. *Nat. Neurosci.*, 2(12):1045–1047, Dec 1999. doi: `10.1038/15967`.

[195] G. Hickok and D. Poeppel. The cortical organization of speech processing. *Nat. Rev. Neurosci.*, 8(5):393–402, May 2007. doi: `10.1038/nrn2113`.

[196] J. P. Rauschecker and S. K. Scott. Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat. Neurosci.*, 12 (6):718–724, Jun 2009. doi: `10.1038/nn.2331`.

[197] W. H. Greene. *Econometric Analysis.* Pearson Education, Inc., Upper Saddle River, New Jersey, 07458, The address, 5th edition, 2003. ISBN 0130661899.

[198] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application.* John Wiley & Sons, 2005.

[199] J. Feng, J. Jost, and M. Qian. *Networks: from biology to theory*, volume 80. Springer, 2007.

[200] I. E. Marinescu, P. N. Lawlor, and K. P. Kording. Quasi-experimental causality in neuroscience and behavioural research. *Nature human behaviour*, 2(12): 891–898, 2018. doi: 10.1038/s41562-018-0466-5.

[201] J. Pearl. *Causality.* Cambridge university press, 2009.

[202] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS Comput. Biol.*, 7(3), 2011. doi: 10.1371/journal.pcbi.1001110.

[203] C. Zou and J. Feng. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics*, 10(1):122, 2009. doi: 10.1186/1471-2105-10-122.

[204] A. Goldenshluger and A. Zeevi. Nonasymptotic Bounds for Autoregressive Time Series Modeling. *Ann. Stat.*, 29(2):417–444, Mar 2001. URL http://www.jstor.org/stable/2674109.

[205] H. Wang, G. Li, and C.-L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, 69(1):63–78, 2007. doi: 10.1111/j.1467-9868.2007.00577.x.

[206] Y. Nardi and A. Rinaldo. Autoregressive process modeling via the Lasso procedure. *J. Multivar. Anal.*, 102(3):528–549, 2011. doi: 10.1016/j.jmva.2010.10.012.

[207] F. Han and H. Liu. Transition matrix estimation in high dimensional time series. In *Int. Conf. Mach. Learn.*, pages 172–180, 2013. URL http://proceedings.mlr.press/v28/han13a.pdf.

[208] A. Kazemipour, S. Miran, P. Pal, B. Babadi, and M. Wu. Sampling requirements for stable autoregressive estimation. *IEEE Trans. Signal Process.*, 65 (9):2333–2347, 2017. doi: 10.1109/TSP.2017.2656848.

[209] S. Basu, G. Michailidis, and Others. Regularized estimation in sparse high-dimensional time series models. *Ann. Stat.*, 43(4):1535–1567, 2015. doi: 10.1214/15-AOS1315.

[210] K. C. Wong, Z. Li, and A. Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *Ann. Stat.*, 48(2):1124–1142, 2020. doi: 10.1214/19-AOS1840.

[211] A. Skripnikov and G. Michailidis. Regularized joint estimation of related vector autoregressive models. *Comput. Stat. Data Anal.*, 139:164–177, 2019. doi: 10.1016/j.csda.2019.05.007.

[212] S. Basu, X. Li, and G. Michailidis. Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions. *IEEE Trans. Signal Process.*, 67 (5):1207–1222, 2019. doi: 10.1109/TSP.2018.2887401.

[213] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[214] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Adv. Neural Inf. Process. Syst. 24*, pages 2726–2734, 2011. URL http://papers.nips.cc/paper/4454-high-dimensional-regression-with-noisy-and-missing-data-provable-guarantees-with-non-convexity.pdf.

[215] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Stat. Sci.*, 27(4):538–557, 2012. doi: 10.1214/12-STS400.

[216] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[217] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang. Subset selection for vector autoregressive processes using Lasso. *Comput. Stat. Data Anal.*, 52(7):3645–3657, 2008. doi: 10.1016/j.csda.2007.12.004.

[218] Y. Ren and X. Zhang. Subset selection for vector autoregressive processes via adaptive Lasso. *Stat. Probab. Lett.*, 80(23):1705–1712, 2010. doi: 10.1016/j.spl.2010.07.013.

[219] Y. Ren and X. Zhang. Model selection for vector autoregressive processes via adaptive Lasso. *Commun. Stat. Methods*, 42(13):2423–2436, 2013. doi: 10.1080/03610926.2011.611317.

[220] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, 25(12):i110–i118, 2009. doi: 10.1093/bioinformatics/btp199.

[221] A. Shojaie and G. Michailidis. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517—-i523, 2010. doi: 10.1093/bioinformatics/btq377.

[222] Y. Liu, M. T. Bahadori, and H. Li. Sparse-gev: Sparse latent space model for multivariate extreme value time serie modeling. In *Proc. 29 th Int. Conf. Mach. Learn. Edinburgh, Scotland, UK,*, 2012. URL http://icml.cc/2012/papers/404.pdf.

[223] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox. Neural granger causality for nonlinear time series. *arXiv preprint arXiv:1802.05842*, 2018.

[224] W. Tang, S. L. Bressler, C. M. Sylvester, G. L. Shulman, and M. Corbetta. Measuring Granger causality between cortical regions from voxelwise fMRI

BOLD signals with LASSO. *PLoS Comput. Biol.*, 8(5), 2012. doi: 10.1371/journal.pcbi.1002513.

[225] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.*, 42(3):1166–1202, 2014. doi: 10.1214/14-AOS1221.

[226] S. van de Geer. On the asymptotic variance of the debiased Lasso. *Electron. J. Stat.*, 13(2):2970–3008, 2019. doi: 10.1214/19-EJS1599.

[227] A. Javanmard and A. Montanari. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, 2014. URL http://www.jmlr.org/papers/volume15/javanmard14a/javanmard14a.pdf.

[228] A. Javanmard, A. Montanari, and Others. Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Stat.*, 46(6A):2593–2622, 2018. doi: 10.1214/17-AOS1630.

[229] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019. doi: 10.1038/s41467-019-10105-3.

[230] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019. doi: 10.1126/sciadv.aau4996.

[231] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, 6(1):1–10, 2015. doi: 10.1038/ncomms9502.

[232] M. Dhamala, G. Rangarajan, and M. Ding. Analyzing information flow in brain networks with nonparametric Granger causality. *Neuroimage*, 41(2):354–362, 2008. doi: 10.1016/j.neuroimage.2008.02.020.

[233] I. Vlachos and D. Kugiumtzis. Nonuniform state-space reconstruction and coupling detection. *Phys. Rev. E*, 82(1):16207, Jul 2010. doi: 10.1103/PhysRevE.82.016207.

[234] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov Regularization and Total Least Squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194, 1999. doi: 10.1137/S0895479897326432.

[235] F. Natterer. Error bounds for Tikhonov regularization in Hilbert scales. *Appl. Anal.*, 18(1-2):29–37, 1984. doi: 10.1080/00036818408839508.

[236] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273.

[237] H. Xie and J. Huang. SCAD-penalized regression in high-dimensional partially linear models. *Ann. Stat.*, 37(2):673–696, 2009. doi: 10.1214/07-AOS580.

[238] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (statistical Methodol.)*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[239] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. URL http://www.jmlr.org/papers/volume7/zhao06a/zhao06a.pdf.

[240] J. Ding, V. Tarokh, and Y. Yang. Model Selection Techniques: An Overview. *IEEE Signal Process. Mag.*, 35(6):16–34, Nov 2018. doi: 10.1109/MSP.2018.2867638.

[241] T. Goldstein and S. Osher. The Split Bregman Method for L1-Regularized Problems. *SIAM J. Imaging Sci.*, 2(2):323–343, 2009. doi: 10.1137/080725891.

[242] L. D. Lewis, S. Ching, V. S. Weiner, R. A. Peterfreund, E. N. Eskandar, S. S. Cash, E. N. Brown, and P. L. Purdon. Local cortical dynamics of burst suppression in the anaesthetized brain. *Brain*, 136(9):2727–2737, 2013. doi: 10.1093/brain/awt174.

[243] S. Chauvette, S. Crochet, M. Volgushev, and I. Timofeev. Properties of slow oscillation during slow-wave sleep and anesthesia in cats. *J. Neurosci.*, 31(42): 14998–15008, 2011. doi: 10.1523/JNEUROSCI.2339-11.2011.

[244] B. O. Watson, D. Levenstein, J. P. Greene, J. N. Gelinas, and G. Buzsáki. Network Homeostasis and State Dynamics of Neocortical Sleep. *Neuron*, 90 (4):839–852, 2016. doi: 10.1016/j.neuron.2016.03.036.

[245] T. E. Nichols and A. P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.*, 15(1):1–25, 2002. doi: 10.1002/hbm.1058.

[246] E. Maris and R. Oostenveld. Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods*, 164(1):177–190, 2007. doi: 10.1016/j.jneumeth.2007.03.024.

[247] K. V. Mardia. Assessment of Multinormality and the Robustness of Hotelling's T^2. Test. *J. R. Stat. Soc. Ser. C (Applied Stat.)*, 24(2):163–171, 1975.

[248] R. E. Miles. On random rotations in $\mathbb{R}^3$. *Biometrika*, 52(3/4):636–639, 1965. doi: 10.1093/biomet/52.3-4.636.

[249] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer-Verlag, Berlin, 2005. ISBN 3-540-40172-5. doi: `10.1007/978-3-540-27752-1`.