Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Information flow in networks based on nonstationary multivariate neural recordings

Presented by: Natalie Klein

Accepted by: Department of Statistics, Machine Learning Department

Readers:

Robert E. Kass, Advisor

Valérie Ventura

Max G'Sell

Tobias Teichert

Approved by the Committee on Graduate Degrees:

Richard Scheines, Dean Date

Carnegie Mellon University

Information flow in networks based on nonstationary multivariate neural recordings

A Dissertation Submitted to the Graduate School
in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics and Machine Learning

by

Natalie Klein

Department of Statistics, Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

August 2019

c

For my parents.

Acknowledgements

I would like to thank those who inspired me to pursue a PhD and those who supported me, financially and otherwise, throughout my time at CMU. Special thanks to my parents and husband for encouraging me to pursue higher education and continue on to a PhD and to my favorite undergraduate math professor, Michael Nakamaye, for encouragement and advice when I was applying to graduate programs. This work would not have been possible without supportive peers at CMU, my advisor, my thesis committee, and my husband Tim (who endured a move to the faraway land of Pittsburgh, and without whom I'm not sure I could have survived five years!).

Takes time, you pick a place to go, and just keep truckin' on...
Sometimes the light's all shinin' on me,
Other times I can barely see.
Lately it occurs to me what a long, strange trip it's been.

The Grateful Dead

Abstract

Neural recordings, such as local field potentials (LFPs), reflect the activity of populations of neurons in time-varying voltage traces and, due to high temporal resolution, they are well-suited for identifying networks of interacting brain regions. Typical analyses are performed on the average across many repetitions of the same task, which eliminates the variation needed to quantify statistical associations between nonstationary signals. In this thesis, I extend statistical and machine learning methodology from graphical models, time series and spatiotemporal models, and Bayesian hierarchical models to develop new tools appropriate for identifying networks of interacting brain regions from multivariate neural recordings. I discuss three different methods, each designed to focus on different a characterization of association. First, we developed dynamic kernel canonical correlation analysis (DKCCA) to identify time-varying lagged correlations between multi-electrode LFPs from two brain regions (Rodu, Klein, *et al*, *J. Neurophys.*, 2018). Second, in work submitted for publication, we explored a novel undirected graphical model suitable for identifying lagged synchronization of neural oscillations via phase coupling and provided inference methods for graphical structure learning. Finally, I sought to infer neural circuitry during stimulus processing on a finer spatial scale using LFP recordings in one cortical area. In particular, I used a biophysically-motivated spatiotemporal Gaussian process model to solve an ill-posed inverse problem and recover the current source density (CSD) generating the observed LFPs. In addition, I implemented a Bayesian hierarchical model for variation in nonstationary stimulus responses, where correlated current source variation across cortical layers may indicate information flow. I demonstrate these methods in laminar LFP recordings from primate primary auditory cortex and in two-dimensional LFP recordings from a Neuropixel probe in mouse visual cortex.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Electrical signals recorded from electrodes placed in brain tissue reflect time-varying neural activity. The high-frequency content of the recorded signal indicates spikes of individual neurons in the vicinity of the electrode, while the lower-frequency content (consisting of timescales slower than about 500Hz) is termed the *local field potential* (LFP) and reflects synaptic processes across a population of nearby neurons (Buzsáki et al., 2012; Einevoll et al., 2013). Therefore, LFPs offer measures of time-varying neural activity that represent populations of neurons and are a suitable data source for investigating associations between different neural populations or brain regions. Understanding these associations is of great interest to neuroscientists, as complex behaviors arise not from the isolated activity of specific neural populations, but from communication and coordination across many neurons and brain regions. LFP data are typically multivariate (recorded from a number of electrodes) and repetitions (trials) are available to serve as repeated observations of the neural activity in response to the same stimulus or task. The goal of the work presented in this dissertation is to determine potentially time-lagged associations between neural populations based on LFPs, taking into account the dynamic nature of the data and likely nonstationary nature of the signals and associations. Chapter 2 consists of published work addressing identification of time-varying cross-correlations between two brain areas, each with multivariate LFP recordings. Chapter 3, which has been submitted for publication, develops undirected graphical models based on phase coupling between oscillatory components of the LFP signals. The final chapter focuses on analysis not of the LFPs directly, but of aggregate current flow in and out of specific populations of neurons, called the *current source density* (CSD); my developments include new methodology for inferring the CSD from the LFP and the application of my method to two different LFP data sets. Importantly, each of these methodological developments are demonstrated to reveal patterns of association based on LFPs that were not found by existing methods.

Chapter 2 investigates time-varying and time-lagged connectivity between two brain regions, each with multiple recorded LFPs, through the development of a dynamic extension of kernel canonical correlation analysis (Rodu et al., 2018). Canonical correlation analysis (CCA) involves linear combinations of each of two sets of variables in which the weights are chosen to maximize the correlation between the linear combinations of each set of variables; kernel CCA generalizes CCA by mapping the sets of variables to higher-dimensional feature spaces prior to optimizing over the weights (Hardoon et al., 2004a). Even if nonlinear feature maps are not used, kernel CCA provides a framework for computationally-efficient regularized CCA when the number of variables is large relative to the number of observations, as is often the case in LFP data. To understand cross-correlation between LFPs in two brain regions based on multi-electrode recordings within each region, we developed *dynamic kernel CCA* (DKCCA) as an extension of kernel CCA to multivariate time series. By estimating the kernel CCA weights in a sliding-window fashion, the weights vary smoothly over time and provide coherent time-varying linear combinations for each set of variables at each time point. The resulting one-dimensional projections for each region provide an informative cross-correlation matrix that may indicate lagged correlations between brain regions, and we developed permutation-based excursion tests to locate significant cross-correlations. We compared DKCCA to simpler methods based on assuming equal weights for all electrodes over time (equivalent to averaging across all signals within a brain region before computing cross-correlation) or on computing all-pairwise cross-correlation matrices and then averaging the matrices across all pairs. In simulations, we found that DKCCA recovered the correct lagged structure while the other methods did not. In real LFP data from hippocampus and prefrontal cortex in an associative learning task (Brincat and Miller, 2015a), DKCCA suggested task-relevant lagged correlations between the two regions during memory retrieval that the other methods did not.

Chapter 3 consists of work submitted to *Annals of Applied Statistics* in which we developed tools for identifying multivariate phase coupling, which, in contrast to most existing approaches for phase coupling, constructs undirected conditional independence graphs. The concept of phase coupling is frequently applied to neural recordings with substantial oscillatory activity, where phase coupling indicates across-trial phase-locked relationships between oscillations in two different brain regions that are a potential marker of long-range neural integration (Lachaux et al., 1999a; Marek et al., 2018). However, existing methods for assessing phase coupling are inherently bivariate in nature, meaning they construct graphs representing interactions between brain regions based on evaluating phase coupling for each pair of regions individually. In contrast to bivariate methods, undirected graphical models provide the very nice interpretation that an edge between a pair of nodes is absent if and only if the corresponding pair of nodal random variables are independent after conditioning on all the other nodal variables; in this sense, an edge represents a unique association of two nodal variables that can not be explained by the other variables. To estimate undirected graphical models based on phase relationships, we developed a model suitable for multivariate phase angles, the torus graph model, in which the parameters correspond to conditional independence relationships. We showed that this model generalizes previous work in multivariate circular statistics and explored its properties as a full regular exponential family model. Furthermore, we proposed a computationally efficient and statistically consistent estimation method based on score matching (Hyvärinen, 2005a), as an intractable normalization constant complicates the use of more standard techniques such as maximum likelihood. Finally, we detail asymptotic inference or regularization approaches to determining an undirected graph structure using the model. In the context of multivariate phase coupling of neural oscillations, we illustrated using simulations how torus graphs can overcome drawbacks of bivariate phase coupling measures; specifically, bivariate measures are sensitive to the marginal distributions of the variables and are influenced by both direct dependencies and indirect dependencies (mediated by other nodes). In neural oscillation phase angle data from 24-dimensional LFPs, previously studied in Brincat and Miller (2015a), we demonstrated that torus graphs were able to recover meaningful phase-based functional connectivity structures between prefrontal cortex and three hippocampal subregions that indicated that during the cue presentation period of the task, hippocampal activity appeared to be leading prefrontal cortex activity.

Chapter 4 consists of work in preparation for publication and makes use of biophysical models that relate the CSD to the measured LFPs through a *forward model.* Attempting to invert the forward model to infer the CSD from measured LFPs is the goal of CSD estimation methods (Pitts, 1952; Nicholson and Llinas, 1971). I show that the CSD provides a measure of neural activity that may better represent the activity of specific neural populations near the electrodes, and thereby help to understand information flow within a neural circuit. In particular, trial-to-trial variation in responses to a stimulus (Arieli et al., 1996) are of great interest, and previous work in the auditory system suggests that both phase coupling and trial-to-trial variation in stimulus-evoked responses are better understood using the CSD than the LFPs (Szymanski et al., 2011). I develop a Gaussian process spatiotemporal model for the latent CSD (GPCSD) that gives rise to the measured LFPs; in addition, I model each trial as the sum of a trial-specific stimulus-evoked response and ongoing activity to make inferences about trial-to-trial variation. I show that GPCSD performs better on simulated data than previous CSD estimation methods, likely because it is the only CSD method that models correlation over time and because I use a principled approach to tuning all hyperparameters (which are often fixed ahead of time in other methods). In one-dimensional auditory LFPs, I demonstrate that GPCSD can isolate oscillatory components to specific spatial locations, leading to more interesting and interpretable torus graphs than those constructed directly on the LFPs. I also propose a model for the CSD evoked response with time shift and amplitude scale variation on across trials, and show that many CSD components in the early evoked response are correlated in shift and amplitude not only within one probe, but also across two simultaneously recording probes. In two-dimensional Neuropixels LFPs, I applied GPCSD followed by PCA to extract timecourses associated with spatial patterns accounting for the most variance in the CSD across time and trials. The spatial components from the CSD contained much more localized detail than those from the LFPs, and in particular, the visual evoked response appeared to be extracted by the first component in both CSD and LFP. Overall, these results suggest that GPCSD is a promising new method for recovering the CSD from LFP recordings, and that downstream analyses based on the estimated CSDs can lead to different and possibly more interpretable results than those based directly on the LFPs.

In lieu of a final discussion chapter summarizing the results pertaining to all three of the developed methods, I am including a discussion section within each chapter to give an overview of the results, ideas about future work, and discussion of potential drawbacks of each method. In all, this dissertation includes three methodological developments for assessing markers of information flow in networks of neural populations based on nonstationary multivariate neural recordings and demonstrates than the new methods can help uncover associations that existing methods cannot.

Chapter 2

Detecting multivariate cross-correlation between brain regions

This chapter is taken from work published in *Journal of Neurophysiology* in 2018 (Rodu et al., 2018). In this work, I collaborated with first author, Jordan Rodu, and we both worked with data collaborators Scott Brincat and Earl K. Miller and advisor Robert E. Kass.

2.1 Introduction

When recordings from multiple electrode arrays are used to establish functional connectivity across two or more brain regions, multiple signals within each brain region must be considered. If, for example, local field potential (LFP) signals in each of two regions are examined, the problem is to describe the multivariate relationship between all the signals from the first region and all the signals from the second region, as it evolves across time, during a task. One possibility is to take averages across signals in each region and then apply cross-correlation or Granger causality (Brovelli et al., 2004; Granger, 1969). Alternatively, one might apply these techniques across all pairs involving one signal from each of the two regions, and then average the results. Such averaging, however, may lose important information, as when only a subset of the time series from one region correlates well with a subset of time series from the other region. Furthermore, Granger causality is delicate in the sense that it can be misleading in some common situations (Wang et al., 2008; Ding and Wang, 2014; Barnett and Seth, 2011). We have developed and investigated a new method, which is descriptive (as opposed to involving generative models such as auto-regressive processes used in

Granger causality), and is capable of finding subtle multivariate interactions among signals that are highly non-stationary due to stimulus or behavioral effects.

Our approach begins with the familiar cross-correlogram, which is used to understand the correlation of two univariate signals, including their lead-lag relationship, and generalizes this in two ways: first, we extend it to a pair of multivariate signals using the standard multivariate technique known as canonical correlation analysis (CCA); second, we allow the correlation structure to evolve dynamically across time. In addition, we found that a comparatively recent variation on CCA, known as kernel-CCA (KCCA), provides a more flexible and computationally efficient framework. We call the initial CCA-based method dynamic CCA (DCCA) and the kernel-based version dynamic KCCA (DKCCA).

We assume the signals of interest are recorded across multiple experimental trials, and the correlations we examine measure the tendency of signals to vary together across trials: a positive correlation between two of the signals would indicate that trials on which the first signal is larger than average, tend also to be trials on which the second signal is larger than average. At a single point in time we could measure the correlation (across trials) between any two signals. At a single point in time we could also take any linear combination of signals in one region and correlate it with a linear combination of signals in the other region; the canonical correlation is the maximum such correlation among all possible linear combinations. A technical challenge is to find a way to compute canonical correlation while taking into account multiple time points at which the signals are collected.

In a different context, Lu (2013) proposed time-invariant weights over both multivariate signals and time points, such that the resulting vectors are maximally correlated. This does not solve satisfactorily the problem we face because it ignores the natural ordering of time and can therefore produce non-physiological combinations. Another proposal (Bießmann et al., 2010) applies to individual trials, and therefore examines correlation across time (as opposed to correlation across trials), which requires signals that are stationary (time-invariant), whereas we wish to describe the dynamic evolution of their correlation across time. Our DKCCA solution applies KCCA in sliding windows across time, similarly to the way a spectrogram computes frequency decompositions in sliding windows across time. After describing the method, we evaluate DKCCA on simulated data, where there is ground truth, and show that DKCCA can recover correlation structure where simpler averaging methods fail. We then apply DKCCA to data collected simultaneously from the prefrontal cortex and hippocampus of a rhesus macaque, and we uncover a novel connectivity pattern that is not detected by traditional averaging methods.

2.2 Materials and methods

In this section, we first review CCA and kernel CCA, and then describe our algorithms, DCCA and DKCCA, our artificial-data simulations, and our experimental methods.

2.2.1 CCA

CCA (Hotelling, 1936) provides a natural way to examine correlation between two sets of N multivariate observations. The algorithm finds maximally correlated linear combinations of the variables in each set, reducing a multivariate correlation analysis into several orthogonal univariate analyses. Specifically, given two multivariate, zero mean data sets $X \in \mathbb{R}^{N \times q_x}$ and $Y \in \mathbb{R}^{N \times q_y}$, where q_x and q_y are the number of variables in X and Y respectively, CCA seeks to find canonical weights w_X and w_Y, the coefficients for linear combinations of the columns of X and Y, respectively, such that

$$\rho_1 = \max_{w_X^1, w_Y^1} \frac{w_X^{1\top} X^\top Y w_Y^1}{\sqrt{w_X^{1\top} X^\top X w_X^1 \cdot w_Y^{1\top} Y^\top Y w_Y^1}} \tag{2.1}$$

where ρ_1 indicates the first canonical correlation and w_X^1, w_Y^1 the first canonical weight pairs. Successive canonical correlations and weight pairs are similarly defined such that

$$\rho_p = \max_{w_X^p, w_Y^p} \frac{w_X^{p\top} X^\top Y w_Y^p}{\sqrt{w_X^{p\top} X^\top X w_X^p \cdot w_Y^{p\top} Y^\top Y w_Y^p}}$$

additionally satisfying

$$\begin{aligned} \text{corr}\left(Xw_X^i, Xw_X^j\right) &= 0 \\ \text{corr}\left(Yw_Y^i, Yw_Y^j\right) &= 0, \quad i \neq j \\ \text{corr}\left(Xw_X^i, Yw_Y^j\right) &= 0. \end{aligned} \tag{2.2}$$

Further, $p \leq \min(q_x, q_y)$ but often in practice $p << \min(q_x, q_y)$. Essentially, each canonical weight pair from CCA projects the two multivariate sets of N observations to two univariate sets of N observations, called canonical components, from which correlation can be obtained in the usual way.

2.2.2 Kernel CCA

Kernel CCA (Hardoon et al., 2004b) extends CCA to allow for nonlinear combinations of the variables, and it also remains numerically stable while CCA can become unstable with large numbers of variables. Importantly, even though kernel CCA applies CCA to a transformation of the data determined by the nonlinearity of interest, all calculations can be performed using inner products of the transformed data matrices. These inner products are defined by a kernel function and are collected in a matrix often called the kernel matrix. Because the resulting computations are efficient, and avoid explicit calculation of the nonlinear transformations, this is usually called the "kernel trick." It has been studied extensively (Christopher, 2016). We next make this explicit in our context.

Following the example of Hardoon et al. (2004b), kernel CCA maps observations into a "feature space"

$$\phi : x \longrightarrow (\phi_1(x), \phi_2(x), \ldots, \phi_M(x))$$

on which CCA is performed. Using the fact that the weights $w_{\phi(X)}$ and $w_{\phi(Y)}$ lie in the row space of $\phi(X)$ and $\phi(Y)$ respectively,

$$\begin{aligned} w_{\phi(X)} &= \phi(X)^\top \alpha \\ w_{\phi(Y)} &= \phi(Y)^\top \beta \end{aligned} \tag{2.3}$$

and substituting equation 2.3 into equation 2.1, we have

$$\rho_1 = \max_{\alpha_1,\beta_1} \frac{\alpha_1^\top \phi(X)\phi(X)^\top \phi(Y)\phi(Y)^\top \beta_1}{\sqrt{\begin{array}{l} \alpha_1 \phi(X)\phi(X)^\top \phi(X)\phi(X)^\top \alpha_1 \\ \quad \cdot \beta_1^\top \phi(Y)\phi(Y)^\top \phi(Y)\phi(Y)^\top \beta_1 \end{array}}}. \tag{2.4}$$

Rewriting $\phi(X)\phi(X)^\top$ as K_X and $\phi(Y)\phi(Y)^\top$ as K_Y, we can express equation 2.4 as

$$\rho_1 = \max_{\alpha_1,\beta_1} \frac{\alpha_1^\top K_X K_Y \beta_1}{\sqrt{\alpha_1 K_X^2 \alpha_1 \cdot \beta_1^\top K_Y^2 \beta_1}} \tag{2.5}$$

with successive canonical correlations found as in plain CCA. The Gram matrices K_X and K_Y require regularization, and the regularization parameter is typically set using cross-validation.

Instead of calculating the canonical weights, canonical components can be calculated directly using

$$\begin{aligned} \phi(X) w^1_{\phi(X)} &= K_X \alpha_1 \\ \phi(Y) w^1_{\phi(Y)} &= K_Y \beta_1. \end{aligned} \tag{2.6}$$

In this paper we use the linear kernel, which returns weights that can be interpreted like they are in plain CCA. Even with the linear kernel, kernel CCA provides benefits when the number of variables in the data matrices is larger — in our case, much larger — than the number of observations. Kernel CCA requires the estimation and use of the kernel matrices of size $N \times N$, rather than the much larger matrices of size $q_x \times q_x$, $q_y \times q_y$, and $q_x \times q_x$. Further, as we describe below, the Gram matrices used in kernel CCA permit a decomposition specific to our method that is not possible with plain CCA, and that allows for substantial improvement in computational speed.

2.2.3 Dynamic Canonical Correlation Analysis

We first describe DCCA and DKCCA. Then, we show how to integrate multiple components of CCA analysis into the DCCA framework, and determine significance of observed canonical correlations based on a permutation test.

To motivate the problem and our solution, we first describe the procedure for assessing correlation structure in the case of two univariate signals. Suppose we have two simultaneously recorded univariate time series of length T over N repeated trials collected into matrices $X \in \mathbb{R}^{N\times T}$ and $Y \in \mathbb{R}^{N\times T}$. Let $X(i,s)$ and $Y(i,t)$ denote the i^{th} trial at times s and t in X and Y, then the the cross-correlation between times s and t is

$$\text{cc}(s,t) = \frac{\sum_{i=1}^{N}\big(X(i,s)-\bar{X}(s)\big)\big(Y(i,t)-\bar{Y}(t)\big)}{\sqrt{\sum_{i=1}^{N}\big(X(i,s)-\bar{X}(s)\big)^2\sum_{i=1}^{N}\big(Y(i,t)-\bar{Y}(t)\big)^2}} \tag{2.7}$$

where $\bar{X}(s) = \frac{1}{N}\sum_{i=1}^{N} X(i,s)$ and $\bar{Y}(t) = \frac{1}{N}\sum_{i=1}^{N} Y(i,t)$. Our goal is to extend this to a multivariate version

$$\begin{aligned} \text{ccc}(s,t) &= \text{max}_{w_{X(s,t)},w_{Y(s,t)}} \frac{w_{X(s,t)}^{\top}\mathbf{X}(s)^{\top}\mathbf{Y}(t)w_{Y(s,t)}}{\sqrt{\chi(s,t)\cdot\gamma(s,t)}} \qquad (2.8)\\ \text{with}\ \ \chi(s,t) &= w_{X(s,t)}^{\top}\mathbf{X}(s)^{\top}\mathbf{X}(s)w_{X(s,t)} \\ \gamma(s,t) &= w_{Y(s,t)}^{\top}\mathbf{Y}(t)^{\top}\mathbf{Y}(t)w_{Y(s,t)}. \end{aligned}$$

where arrays $\mathbf{X} \in \mathbb{R}^{q_x\times T\times N}$ and $\mathbf{Y} \in \mathbb{R}^{q_Y\times T\times N}$ collect the N multivariate time series of length T of dimensions q_x and q_Y respectively, and $\mathbf{X}(s) \in \mathbb{R}^{N\times q_x}$ and $\mathbf{Y}(t) \in \mathbb{R}^{N\times q_y}$ denote slices of arrays $\mathbf{X}$ and $\mathbf{Y}$ corresponding to times s and t. The challenge is to identify linear combinations $w_{X(s,t)}$ and $w_{Y(s,t)}$ for all s,t. As mentioned in the introduction, a natural first thought is to solve for $w_{X(s,t)}$ and $w_{Y(s,t)}$ by using CCA for every pair of times s and t, but this leads to correlations and weights that are difficult to interpret. Alternatively we could find fixed $w_{X(s,t)}$ and $w_{Y(s,t)}$ for all s,t, but this does not take into account the nonstationarity of signals in the brain. DCCA, which we now describe, solves these problems by inferring a single set of weights that modulate over time without knowing a priori the lagged correlation structure of the system.

We start by creating extended observations at each time point that are a concatenation of observations in a local window of time. Fix local window g, and let $\mathbf{X}(s') \in \mathbb{R}^{N\times q_x\cdot(2g+1)}$ and $\mathbf{Y}(s') \in \mathbb{R}^{N\times q_y\cdot(2g+1)}$ be matrices defined as

$$
\begin{aligned}
\mathbf{X}(s') = \big[&X(s-g), X(s-g+1), \ldots, X(s), \\
&\ldots, X(s+g-1), X(s+g)\big] \\
\mathbf{Y}(s') = \big[&Y(s-g), Y(s-g+1), \ldots, Y(s), \\
&\ldots, Y(s+g-1), Y(s+g)\big]
\end{aligned}
$$

where $\big[A, \ldots Z\big]$ indicates concatenation of matrices $A, \ldots, Z$ along columns. For each time point s, we run a CCA between $\mathbf{X}(s')$ and $\mathbf{Y}(s')$, which yields linear combination weights $w_{X(s')}$ and $w_{Y(s')}$ of lengths $q_x \cdot (2g+1)$ and $q_y \cdot (2g+1)$ respectively (see Equation 2.1). Since the CCA is run using concatenated observations matrices $\mathbf{X}(s')$ and $\mathbf{Y}(s')$, we can express the concatenated weights as

$$
w_{X(s')} = \begin{bmatrix} w'_{X(s-g)} \\ \vdots \\ w'_{X(s)} \\ \vdots \\ w'_{X(s+g)} \end{bmatrix} \in \mathbb{R}^{q_x \cdot (2g+1) \times 1}
$$

$$
w_{Y(s')} = \begin{bmatrix} w'_{Y(s-g)} \\ \vdots \\ w'_{Y(s)} \\ \vdots \\ w'_{Y(s+g)} \end{bmatrix} \in \mathbb{R}^{q_y \cdot (2g+1) \times 1} \tag{2.9}
$$

where the w' emphasizes that these weights are not the same weights that would be obtained from a CCA with observations that are not concatenated. We then set the canonical weights for $X(s)$ and $Y(s)$ equal to $w'_{X(s)}$ and $w'_{Y(s)}$ in equation 2.9 and calculate matrices $X^{\text{proj}} \in \mathbb{R}^{N \times T}$ and $Y^{\text{proj}} \in \mathbb{R}^{N \times T}$ where

$$
\begin{aligned}
X^{\text{proj}}(s) &= X(s) w'_{X(s)} \qquad (2.10) \\
Y^{\text{proj}}(s) &= Y(s) w'_{Y(s)}
\end{aligned}
$$

and the superscript label is used to denote projection. Finally, we calculate the matrix ccc as in Equation 2.7 using matrices X^{proj} and X^{proj}.

In this approach, exact lagged relationships need not be established a priori. Instead, if there is a strong lagged correlation between X and Y at times s and t, then as long as $|s-t| < g$, this lagged relationship will be a driving factor in setting the linear combination weights for both X at time s and Y at time t (though these will not be set simultaneously). Further, the strong correlation between X and Y at times s and t will drive setting the linear combination weights at time s' where $0 < |s' - s| < g$ and $0 < |s' - t| < g$, instead of $w_X(s')$ and $w_Y(s')$ being set to greedily maximize correlation at time s'.

Extended observations are the key to discovering a priori unknown lagged cross-correlation, as they make observations within the specified local window visible to each other. To see this, if array $\mathbf{X}$ has observations that are independent of each other across time, and $\mathbf{Y}$ is a *shifted* copy of $\mathbf{X}$, so $\mathbf{Y}(t) = \mathbf{X}(t-k)$, then without creating extended observations, this procedure would only return false correlations since in the shifted case the instantaneous observations are independent (and hence uncorrelated) by design. Recovery of the actual lagged correlation without extended observations would require knowledge of the true lag, which in the above example would be k.

While DCCA with CCA might work well in some scenarios, as described above it has some disadvantages. First, in many cases, $q_x \cdot (2g+1) >> N$ and $q_y \cdot (2g+1) >> N$, which causes numerical instability. Second, as a result of forming concatenated matrices $\mathbf{X}(s')$ and $\mathbf{Y}(s')$, we must estimate a covariance matrix for all observation pairs $X(s)$ and $X(t)$, $Y(s)$ and $Y(t)$, and $X(s)$ and $Y(t)$ for $|s - t| < 2g + 1$. Finally, for inference involving trial permutations or bootstrapping, these covariance matrices must be recalculated for each simulated data set.

The first and third problems are well understood from the literature, and the solution is to use kernel CCA (KCCA) instead of CCA. In the next section, we describe DCCA with KCCA and show that not only is KCCA a solution to the first and third problems with respect to the DCCA procedure, but also the second. Further, KCCA allows for nonlinear combinations over signals.

2.2.4 Dynamic Kernel Canonical Correlation Analysis (DKCCA)

In theory, DKCCA proceeds as described above, but with KCCA instead of CCA. Observations are concatenated over a local window of time, $s - g$ to $s + g$, and kernel matrices are computed from these concatenated observations. Weights for the concatenated observations (see Equation 2.9) are derived according to Equation 2.3, from which the weights at time s are extracted and matrices X^{proj} and Y^{proj} are created as in Equation 2.10.

As indicated in Equation 2.5, the KCCA algorithm finds optimal $\alpha(s)$ and $\beta(s)$, which from Equation 2.3 we can interpret as coefficients for the linear combinations over the N replications in $X(s)$ and $Y(s)$ respectively (see, for instance, Hardoon et al. (2004b) for details). Because the concatenated observation matrix $\mathbf{X}((s+1)')$ contains all the observations in the matrix $\mathbf{X}(s')$ with the exception of one, and likewise for matrices $\mathbf{Y}((s+1)')$ and $\mathbf{Y}(s')$, $\alpha(s)$ and $\beta(s)$ modulate smoothly over time in the DKCCA procedure.

As a consequence, using Equation 2.3, we have that $w_{X(s)}$ and $w_{Y(s)}$ evolve smoothly over time as well. Note that the degree of smoothness depends on the smoothness of the original time series.

Calculating the canonical weights in Equation 2.3, while informative in many scientific settings, is not strictly necessary. In cases where the focus lies solely on the temporal lagged correlations, $X^{\text{proj}}(s)$ and $Y^{\text{proj}}(s)$ in Equation 2.10 can be calculated directly using Equation 2.6. This is especially useful when ϕ maps observations into high-dimensional or infinite-dimensional space, where the weights become difficult to interpret. Calculating $X^{\text{proj}}(s)$ and $Y^{\text{proj}}(s)$ directly allows for a wide range of nonlinear combinations of the time series to be considered with low computational cost.

In addition to extending CCA to nonlinear combinations, KCCA allows for numerical stability when the number of signals multiplied by concatenated time points is larger than the number of trials. Also, bootstrap and permutation testing procedures are much faster with KCCA, since they amount to selection or permutation of the rows and columns of the kernel matrices. Further, KCCA solves the second problem mentioned above, that of needing to estimate a covariance matrix for all observation pairs $X(s)$ and $X(t)$, $Y(s)$ and $Y(t)$, and $X(s)$ and $Y(t)$ for $|s-t| < 2g+1$ under DCCA with CCA. Focusing for now on the case where ϕ is the identity map, we calculate the kernel matrix from Equation 2.5, $K_X(s') \in \mathbb{R}^{N\times N}$, as $K_X(s') = \mathbf{X}(s')\mathbf{X}^\top(s')$ (likewise with the kernel matrix $K_Y(s')$). This calculation can be decomposed as

$$K_X(s') = K_X(s-g) + \ldots + K_X(s) + \ldots + K_X(s+g) \tag{2.11}$$

and since in Equation 2.5 only K_X and K_Y need to be computed (whereas in CCA the cross-covariance terms between X and Y must be computed), then for each time point s, $K_X(s)$ and $K_Y(s)$ need be computed only once, with $K_X(s')$ and $K_Y(s')$ computed as the sum of the relevant kernel matrices. For arbitrary ϕ, the decomposition in Equation 2.11 does not necessarily hold, as interactions between time points are potentially allowed to occur. To take advantage of the decomposition, we can impose a constraint on ϕ that allows for nonlinearities across time series dimensions, but not across time.

2.2.5 Incorporating multiple components

Typically in CCA-based methods, it suffices to describe the procedure using only the first canonical correlation, as canonical correlations associated with components $p \geq 2$ can be studied independently, or can be added to get the total canonical correlation over all components. This is because the canonical components satisfy the constraints in equation 2.2. However, in DKCCA, we don't have these guarantees. While the constraints in Equation 2.2 do apply to the canonical weights calculated from the concatenated observations (Equation 2.9), they do not apply to the canonical components $X^{\text{proj}}(s)$ and $Y^{\text{proj}}(s)$ in Equation 2.10 as these are derived from *extracted* weights. As an analogy, a set of orthogonal bases over the interval $(0, 1)$ are not necessarily orthogonal when restricted to an arbitrary interval $(a, b) \subset (0, 1)$. In order to incorporate canonical correlations across multiple components in DKCCA, we therefore cannot add the *ccc* matrices calculated from each component.

We now describe a way of combining the top k correlations between components at times s and t in X and Y, respectively. Let $w^i_{X(s)}$ and $w^i_{Y(s)}$ be the i^{th} canonical weights for X and Y at time s, and similarly to Equation 2.10, for $i = 1 \ldots k$, let $X^{\text{proj}}_i(s) = X(s)w^i_{X(s)}$ and $Y^{\text{proj}}_i(t) = Y(t)w^i_{Y(t)}$ be the k components for $X(s)$ and $Y(t)$. Then,

1. Set $\gamma_1 = \text{corr}(X^{\text{proj}}_1(s), Y^{\text{proj}}_1(t))$.

2. Decompose $X^{\text{proj}}_2(s)$ and $Y^{\text{proj}}_2(t)$ into a projection onto $X^{\text{proj}}_1(s)$ and $Y^{\text{proj}}_1(t)$, respectively, and the associated residual:

$$\begin{aligned} X^{\perp}_2(s) &= \frac{X^{\text{proj}}_2(s)^\top X^{\text{proj}}_1(s)}{||X^{\text{proj}}_1(s)||^2} X^{\text{proj}}_1(s) \\ X^{\text{resid}}_2(s) &= X^{\text{proj}}_2(s) - X^{\perp}_2(s) \\ Y^{\perp}_2(t) &= \frac{Y^{\text{proj}}_2(t)^\top Y^{\text{proj}}_1(t)}{||Y^{\text{proj}}_1(t)||^2} Y^{\text{proj}}_1(t) \\ Y^{\text{resid}}_2(t) &= Y^{\text{proj}}_2(t) - Y^{\perp}_2(t). \end{aligned}$$

3. Set

$$\gamma_2 = \frac{\text{cov}\big(X^{\text{resid}}_2(s), Y^{\text{resid}}_2(t)\big) + \text{cov}\big(X^{\text{resid}}_2(s), Y^{\perp}_2(t)\big) + \text{cov}\big(X^{\perp}_2(s), Y^{\text{resid}}_2(t)\big)}{\sqrt{\text{var}\big(X^{\text{proj}}_2(s)\big)}\sqrt{\text{var}\big(Y^{\text{proj}}_2(t)\big)}}.$$

   This is almost $\text{cov}\big(X^{\text{proj}}_2(s), Y^{\text{proj}}_2(t)\big)$, however we have removed the covariance component $\text{cov}\big(X^{\perp}_2(s), Y^{\perp}_2(t)\big)$, as this is redundant. Importantly, we still allow for correlations across components, where strict orthogonalization would not.

4. For the i^{th} canonical components, decompose $X^{\text{proj}}_i(s)$ into a projection onto the subspace spanned by components $\{X^{\text{proj}}_1(s), \ldots, X^{\text{proj}}_{i-1}(s)\}$ and its orthogonal residual, and $Y^{\text{proj}}_i(t)$ into a projection onto the subspace spanned by components $\{Y^{\text{proj}}_1(t), \ldots, Y^{\text{proj}}_{i-1}(s)\}$ and the orthogonal residual. Calculate γ_i as in step 3.

5. Finally, set $TC(s,t) = \sum_{i=1}^k \gamma_i$.

This procedure has the nice property that if it were performed on output from a vanilla CCA, for canonical correlations ρ_i, we have that $\gamma_i = \rho_i$.

2.2.6 Identifying significant regions of cross-correlation

We generate a null distribution for cross-correlations between regions with a permutation test, combined with an excursion test. We create B new data sets by randomly permuting the order of the trials of Y, and rerunning DKCCA on each of the permuted data sets. This results in matrices ccc_i^* for $i \in 1 : B$. For each entry (s, t), we have B sampled correlations $\mathrm{ccc}_i^*(s, t)$ from the null distribution, and say that point (s, t) is $\alpha_{\mathrm{pw}}(s, t) -$ significant if its correlation value is greater than $1 - \alpha_{\mathrm{pw}}$ percent of the B sampled correlations. As mentioned above, the relevant kernel matrices need not be recalculated for each permuted data set. Only the order of the rows and columns of kernel matrices $K_Y(s)$ are affected.

Because we are interested in broad temporal regions of correlated activity, and not isolated time point pairs, we use the B data sets from the permutation test to perform an excursion test (see Xu et al. (2011)). We say that two α_{pw} significant points (s, t) and (s', t'), with either $s \neq s'$, $t \neq t'$, or both, belong to the same *contiguous region* if there exists a connected path between them such that each point along the path is also α_{pw} significant. To perform the excursion test, we identify contiguous regions in the ccc_i^* matrices, and for each contiguous region $C_i^{m_i}$ with $m_i \in [1 \dots M_i]$ where M_i is the number of contiguous regions in the i^{th} bootstrapped dataset, we record the sum of the excess correlation above the α_{pw} cutoff values,

$$k(i, m_i) = \sum_{(s,t) \in C_i^{m_i}} \mathrm{ccc}_i^*(s, t) - \alpha_{\mathrm{pw}}(s, t). \tag{2.12}$$

The collection $\{k(i, m_i) : i \in [1 \dots B], m_i \in [1 \dots M_i]\}$ defines a null distribution over the total excess correlation for contiguous regions, from which an α_{region}-level cutoff value can be calculated. We consider as significant any contiguous regions C^m whose total excess correlation exceeds the α_{region}-level cutoff value. Instead of total excess correlation, we can also use the size of each contiguous region as a statistic (Maris and Oostenveld, 2007).

2.2.7 Simulations

We ran several simulation scenarios to evaluate the effectiveness of the DKCCA procedure[*]. Each scenario simulated an experiment with two multivariate time series of dimensions 96 and 16, with 100 repeated time-locked trials. Each time series for each trial was generated from a two-dimensional latent variable model

$$X(t) = A(t)H(t) + \epsilon(t) \tag{2.13}$$

[*] Code for the DKCCA algorithm and this simulation is provided at `https://github.com/jrodu/DKCCA.git`

where $H(t) \in \mathbb{R}^2$ is the latent trajectory with the first and second components uncorrelated, and $A(t) \in \mathbb{R}^{q\times 2}$ (where $q = 96$ for time series 1 and $q = 16$ for time series 2) is a mapping from the latent variable to the signal space, and $\epsilon(t)$ is a noise term. $H(t)$ and $\epsilon(t)$ were both resampled for each trial, while $A(t)$ was fixed across trials, though as is clear by the notation, was allowed to vary across time. We refer to the collection of $A(t)$ for all t as the *A-operator*.

We introduced correlation between time series 1 and 2 through the first latent dimension for each time series. Let H_1^1 and H_2^1 be the sample paths of the first latent dimension for time series 1 and 2, respectively. Further, let $\rho(k)$ be a "ramp function" that is piecewise linear with $\rho(0) = 0$, ramps up to a plateau where it stays for a length of time, then returns to 0. Then starting at time s and for desired lag l,

$$H_2^1(s+k) = \rho(k)H_1^1(s+k-l) + (1-\rho(k))H_2^1(s+k). \tag{2.14}$$

While we kept lag l fixed across trials, we allowed the induced correlation starting time s to vary uniformly within 10 time steps in order to simulate small differences in the time-course of brain signals with respect to the time-locked stimulus.

We sampled each latent trajectory of H for each time series, and the trajectories for each entry in the time-dynamic A-operators for each time series, as zero mean Gaussian processes with covariance kernels of the form

$$\Sigma(i,j) = \sigma^2 \exp\left(-.5 * \left(\frac{i-j}{\lambda}\right)^2\right). \tag{2.15}$$

Let A_1 be the A-operator for time series 1 and A_2 the A-operator time series 2. For A_1, A_2, H_1, and H_2, we let $\sigma = 1$ in Equation 2.15. For H_1 and H_2 we let λ be 40 and 20, respectively, and for A_1 and A_2 we set $\lambda = 100$.

Finally, $\epsilon(t)$ for each time series is simulated as a two-dimensional latent state model as in Equation 2.13 in order to control spatial-dependence and long-range temporal-dependence in the noise. For each A_ϵ-operator (where the subscript emphasizes that it is used to simulate the noise terms $\epsilon(t)$), we let $\sigma = 1$ and $\lambda = 30$ in Equation 2.15. For H_ϵ we let $\lambda = 80$ and, across various simulation conditions on the signal to noise ratio, fix $\sigma_{\text{noise}} = \sigma$ in Equation 2.15 to be between 0.2 and 2.

2.2.8 Experimental Methods

The experiment consisted of the presentation of one of four cue objects, each mapped to one of two associated objects. A rhesus macaque was required to fixate on a white center dot, after which one of the four object cues was shown, followed a correct or incorrect associated object. A correct object presentation required an immediate saccade to a target, while an incorrect object presentation required a delay until a correct object was presented. The presentations of fixation, object cue, and associated objects were separated by blank intervals. Multi-electrode recordings were made from both Hippocampus (HPC) and lateral prefrontal cortex (PFC) in order to examine the roles of HPC and PFC in non-spatial declarative memory.

Each day over 1000 trials were conducted. Electrodes were daily re-implanted into the macaque brain, hence the exact location, and number, of electrodes in the HPC and PFC varied from day to day. For the analysis of the DKCCA algorithm we only considered days in which there were at least 8 electrodes in each section, with typical numbers being between 8 and 20 per brain region. The sampling rate for each trial was 1000Hz. For further details on the experimental setup and data collection, see Brincat and Miller (2016a). All procedures followed the guidelines of the MIT Animal Care and Use Committee and the US National Institutes of Health. We expand upon details relevant to analysis of the DKCCA algorithm here.

Past work has shown that for multi-electrode recordings such as LFP and EEG, the use of a common pickup and the presence of volume conduction can adversely affect functional connectivity analysis. See, for example, the work of Bastos and Schoffelen (2016) and Trongnetrpunya et al. (2016) for a description of the problem and steps that can be taken to mitigate the effect. These same issues impact DKCCA, especially when the instantaneous correlation of the contaminating signal is almost as strong as or stronger than the lagged correlation of interest, and we recommend following the guidance proposed by past work. For the current data, to minimize pickup of any specific signals through the reference, collection was made using a common reference via a low-impedance connection to animal ground.

For the analyses reported here, we filtered each trial using a Morlet wavelet centered around 16 Hz, where this frequency was determined from previous studies. It is critical in correlation based analyses to isolate the frequency band of interest, and 16 Hz was chosen for this analysis because it is in this region (the beta band) that we see the strongest LFP power and coherence effects (see Brincat and Miller (2016a) for details). Finally, we downsampled the signal to 200Hz.

2.3 Results

The goal of DKCCA was to extend dynamic cross-correlation analysis to two multivariate time series with repeated trials. The algorithm infers a linear or nonlinear combinations over the signal dimensions that (a) are interpretable, (b) are allowed to change over time, and (c) do not require prior knowledge of the lagged correlation structure. Because it uses sliding windows, the combinations can adapt to the underlying nonstationarity of the brain regions and their interactions. DKCCA can thereby recover correlation structure where averaging methods fail. In this section, we evaluate DKCCA on both simulated and real data. In simulations, we show that DKCCA is able to recover the underlying correlation structure between two highly dynamic multivariate time series under varying noise conditions, and simulate a realistic example where DKCCA recovers the correlation structure but averaging methods do not. In the real data, DKCCA detected a novel cross-correlation while the averaging methods did not.

2.3.1 DKCCA analysis of simulated data

To illustrate DKCCA's ability to recover dynamic cross correlation from two multivariate signals, we ran several simulations (see Equations 2.13, 2.14, and 2.15) under varying noise conditions, with lag $l = 20$ and simulation start time s sampled for each trial uniformly between min = 310 and max = 320. We designed our simulations to verify three important properties of the algorithm: (a) that it can recover cross-correlation between highly dynamic time series, (b) that it is robust against spurious correlation, and (c) that it can recover cross-correlation in adversarial conditions, where averaging methods fail. We subsequently illustrate these properties in our analysis of the real data.

Despite the highly dynamic nature of the simulated time series, DKCCA correctly identifies the temporal region in which lagged correlation exists, and indicates no correlation outside of that region. Figure 2.1 shows the raw cross-correlograms for various noise levels generated from one such set of simulations. Figure 2.2 shows the cross-correlograms after accounting for the inference step. The magnitude displayed is excess pairwise canonical correlation after inference. As noise increases, cross-area correlated multivariate signals are more difficult to find, but DKCCA is successful even in high noise regimes. Further, our simulations suggest that the method is robust to finding spurious correlations, since in Figures 2.1 and 2.2 almost no cross-correlation is indicated outside of the regions of induced correlation.

To see this more directly, we ran simulations with no cross-correlation structure, again under varying noise conditions. Results are in Figure 2.3. For all analyses, we set our window size parameter, g, to be 20.

To asses DKCCA's recovery of the correct lag, we computed a standard cross correlogram from the ccc matrix generated by the DKCCA algorithm by averaging the lagged correlations of one region with respect to the other, within the time period of induced lagged correlation. The results in Figure 2.4 show that our algorithm correctly recovers the location of the maximal correlation at a lag of 20. As is expected, the accuracy of the method decreases as the signal-to-noise ratio decreases.

Finally we compared the performance of DKCCA to two common methods, averaging the pairwise correlations between signals (APC) and taking the correlation between the average of signals in each region (CAS). Figure 2.5 shows the results of the comparison with the same simulation setup as before, but with signal present in only a portion of the simulated electrodes. We achieved this by permuting the remaining electrodes across trials, thereby leaving the temporal characteristics of each electrode intact, but breaking their dependence with other electrodes in a given trial. While APC and CAS do not capture the simulated correlation structure, DKCCA does capture the correct correlation dynamics.

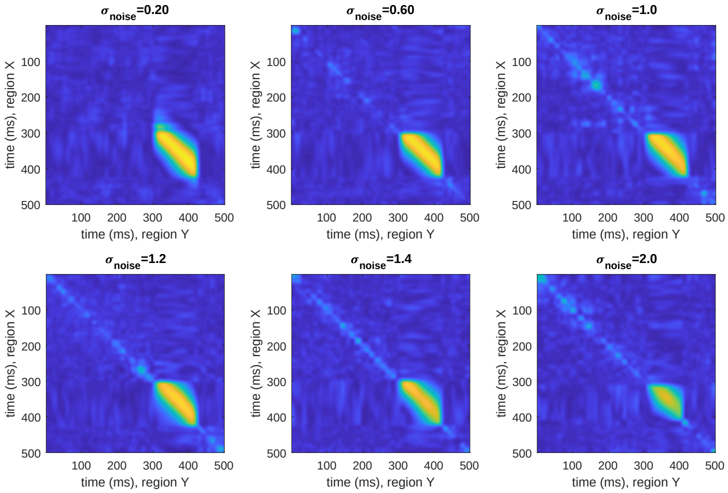

Figure 2.1: DKCCA captures lagged correlation across a variety of noise levels. Representative results for running DKCCA on simulated data with $\sigma_{\text{noise}} = .2, .6, 1, 1.2, 1.4$ and 2 (left to right, top to bottom). y-axis is *time (ms)* in section X, x-axis is *time (ms)* in section Y. Trial length: 500ms. True lag of 20ms begins in each trial between 300 and 310 in section Y (280 and 290 in section X) and lasts about 80ms. DKCCA recovers the true correlation structure well even for relatively small signal to noise ratio.

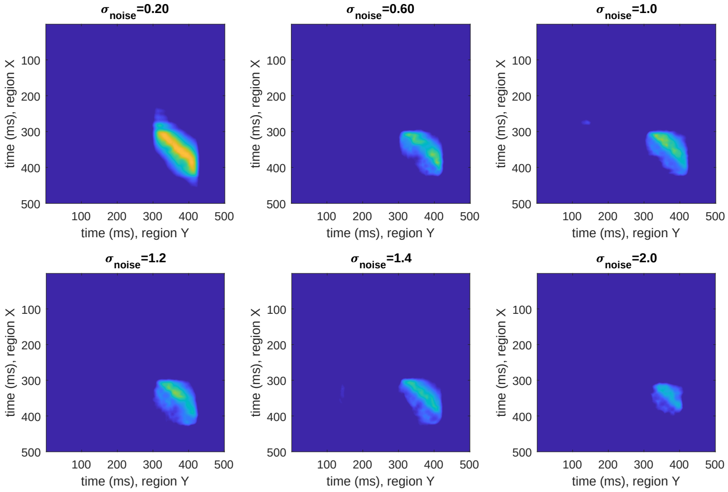

Figure 2.2: Excursion test for DKCCA correctly identifies significant regions of cross-correlation. Representative results after bootstrap and excursion test for trials from simulations with $\sigma_{\text{noise}} = .2, .6, 1, 1.2, 1.4$ and 2 (left to right, top to bottom). y-axis is *time (ms)* in section X, x-axis is *time (ms)* in section Y. Trial length: 500ms. True lag of 20ms begins in each trial between 300 and 310 in section Y (280 and 290 in section X) and lasts about 80ms. Nonzero values are excess above cutoff. Zero values indicate time-pair correlation that was at cutoff or below.

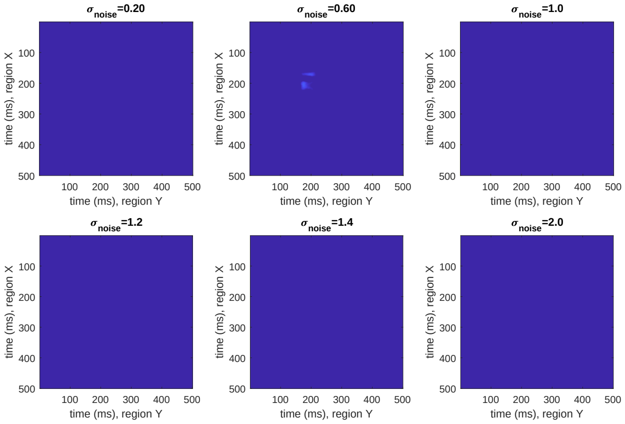

Figure 2.3: Excursion test for DKCCA avoids identifying spurious cross-correlation. Representative results for running DKCCA on simulated data with $\sigma_{\text{noise}} = .2, .6, 1, 1.2, 1.4$ and 2 (left to right, top to bottom). y-axis is *time (ms)* in section X, x-axis is *time (ms)* in section Y. Trial length: 500ms. No true lead-lag relationships present.

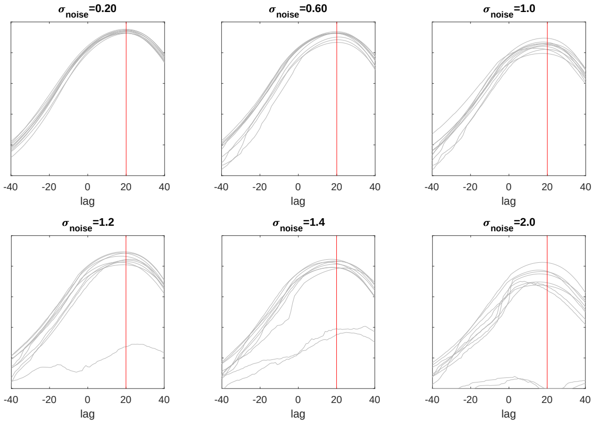

Figure 2.4: DKCCA accurately identifies the correct lag length. Each grey line is the averaged lagged correlation of area of interest for trials from simulations (10 simulations for each noise level) with $\sigma_{\text{noise}} = .2, .6, 1, 1.2, 1.4$ and 2. Lag $\tau = 20$ is indicated with a vertical line.

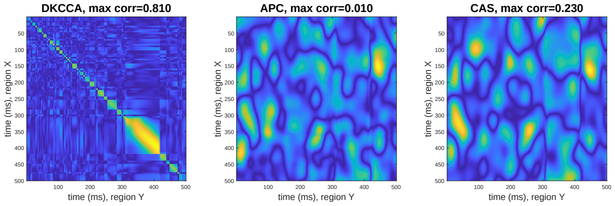

Figure 2.5: DKCCA identifies the correct region of interest where other methods fail. Comparison of DKCCA, APC, and CAS on simulated data with signal present in only a portion of the electrodes. DKCCA correctly identifies signal, as above. APC and CAS are unable to identify signal. Note that colors are not comparable across methods as all correlations in APC and CAS are comparatively small, so correlations have been scaled for visual presentation.

2.3.2 DKCCA analysis of LFPs from PFC and HPC

We applied DKCCA to analyze local field potential (LFP) data from the paired-association task described in the Experimental Methods section. In the experiment, a cue was presented, followed by a series of objects, one of which had a learned association with the cue. A rhesus macaque was required to make a saccade to a target following the display of the correct associated object, and a reward was provided as feedback following a correct saccade. In the original study, Brincat and Miller (2015b) concentrated their analysis on the feedback period of each trial. We focused our analysis on the period between the initial presentation of the cue object and the appearance of the first potential associated object. In this portion of the experiment, Brincat and Miller (2015b) showed there was high power in a band around 16Hz in both HPC and PFC, and that, broadly, signal in HPC led the signal in PFC (HPC → PFC). In our analysis we (a) verify that DKCCA also identified that, broadly, HPC → PFC, and (b) compare DKCCA to standard averaging methods to show that DKCCA is better able to detect nuanced cross-correlation. In addition to (a) and (b), our analysis uncovered an interesting reversal in the lead-lag relationship between HPC and PFC that was not found by traditional averaging methods.

In our analysis we used a block of 200 trials, and a total sliding window size of 21 (g=10) time steps, or 105ms. We used kernels generated by setting ϕ to be the identity map. The first column of Figure 2.6 shows the result of DKCCA generated from a representative set of trials. The figure shows both the ccc matrix for the full trial length (top panel) as well as the ccc matrix zoomed in to the time period corresponding to the portion of the trial between the presentation of the cue and the presentation of the associated pair (bottom panel). In the full trial, DKCCA shows strong correlation before the presentation of the cue (time -.5s to 0s), and during the presentation of the cue (time 0s to .5s). Lag time varies throughout the interval before and during presentation of the cue, but generally suggests that HPC leads PFC. Correlation is weaker and more intermittent between the presentation of the cue and the presentation of the associated pair, with a noticeable temporary increase in correlation just before the associated pair presentation. While not as distinct as during and just before cue presentation, correlation during the associated pair presentation (timing 1.25s to 1.75s) is persistent. Although neither APC nor CAS show significant lagged activity in the time period between the presentation of the cue and the presentation of the associated pair, the zoomed cross-correlogram for DKCCA shows a burst of correlation just after time=1s, and in particular suggests that PFC leads HPC in that time period. The effect is significant in the magnitude of the excursion with $p << .001$.

The highly significant correlation between PFC and HPC found by DKCCA in the bottom left panel of Fig. 2.6 appears to be asymmetric about the diagonal. We found the point of maximal cross correlation to occur when PFC leads HPC by approximately 65ms, as marked with a red cross in that figure. The within-task timing of a switch in lead-lag relationship from HPC → PFC to PFC → HPC would be roughly

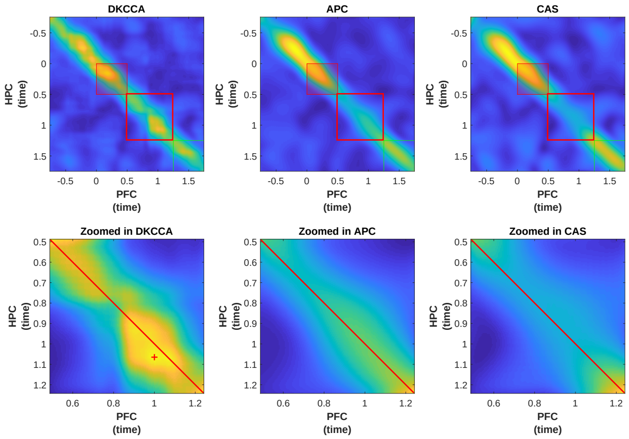

Figure 2.6: DKCCA finds lagged cross-correlation where traditional methods fail. Top left: cross-correlogram created by the DKCCA algorithm on full trial from -.75 seconds to about 1.75 seconds. The small red and green boxes show timing of display of cue and associated pairs, respectively. Bottom left: cross-correlogram from trial zoomed in to the period between the end of the cue presentation (time = .5 seconds) and the beginning of the associated pair presentation (time = 1.25 seconds), as indicated by the large red box on the full cross-correlogram. The red diagonal line shows time in HPC = time in PFC, and the red cross in the bottom left panel indicates the location of the maximum cross correlation in the excursion of interest. Results based on DKCCA are contrasted to the cross-correlogram and zoomed cross-correlogram created by averaging the absolute value of the pairwise cross-correlograms (APC, middle column) and those created by averaging the signals prior to calculating the cross-correlogram (CAS, right column). In both APC and CAS, we do not see any activity in the zoomed cross-correlograms. For display purposes, correlation magnitudes are not comparable between DKCCA, APC, and CAS as APC and CAS cross-correlations tend to be small compared to the DKCCA, so they have been rescaled.

consistent with the timing of the lead-lag switch found in Place et al. (2016), who studied the theta band in rats. However, with our descriptive method we are unable to provide a statistical test of the lead-lag relationship. We also do not contribute to the as of yet inconclusive discussion of why the theta band is predominant in hippocampal-cortical interaction in rodents, while beta is predominant in primates. But despite the difference in frequencies, in both rats and primates, interactions with HPC → PFC directionality seem to be prominent early in trials, during the preparatory and cue presentation periods, but switch to PFC → HPC directionality near the time period when the cue is eliciting memory retrieval. This suggests (in both situations) that PFC may be involved in guiding memory retrieval in the HPC.

This figure confirms on real data that DKCCA can effectively recover cross-correlation, and that it has the power to detect subtle changes in the lead-lag relationship between time series, such as the reversal seen between object presentations.

We next compare the results of DKCCA to results from APC. In Figure 2.6, DKCCA reports a much richer correlation structure throughout the trial than does APC. There are two places in particular where APC differs from DKCCA: first in the period before the presentation of the cue (time -.5s to 0s), and second between the presentation of the cue and the associated pair (time .5s to 1.25s), where APC indicates that correlation disappears between HPC and PFC. The reason is that with APC there is a single set of weights used when combining the pairwise cross-correlations, regardless of the temporal location of those cross correlations in the trial. However, because of the dynamic nature of the brain, signal pairs do not contribute to population-level correlation activity in a consistent manner over time. In Figure 2.6, even where APC does indicate correlation, it does not capture the optimal correlation structure. DKCCA, on the other hand, accommodates these changing dynamics by modulating the weights of the signals over time.

CAS has comparable performance to APC. In particular, the CAS method is unable to fully capture the dynamic cross-correlation structure both before the presentation of the cue and between the presentation of the cue and the presentation of the associated pair. APC and CAS both treat all signals with the same weight, which does not optimize for, and therefore potentially misses, dynamic cross-correlation between the two regions.

2.4 Discussion

In this paper we derived and studied a pair of descriptive methods for assessing dynamic cross-correlation between two multivariate time series. DCCA extends canonical correlation analysis to three-way arrays indexed by signals, time, and trials, where the canonical weights over signals are allowed to evolve slowly over time. Because it is based on sliding windows, DCCA accommodates nonstationary time series while avoiding strong assumptions on the dynamics of the lagged correlation over time. We prefer to use the kernelized variant, DKCCA because it scales well with a growing number of signals, incorporates nonlinear combinations, and provides computational efficiencies not available in DCCA.

Even though nonlinear kernels might provide powerful results, they are difficult to interpret. In our analyses, we have used the linear kernel, which yields weights that are interpreted identically to those in CCA while providing more stable and efficient computation.

DKCCA requires a few parameters to be set, or inferred. The first is a regularization parameter used in implementing KCCA. In one run of the DKCCA algorithm, KCCA is called multiple times, and in order to allow for comparisons between different entries in the *ccc* matrix, it is important to use a single regularization value for all KCCA calls. In our experiments, we set this global regularization value through cross-validation. The second parameter is the half-window size g. This value should be set according to scientific context, though there are a few considerations to take into account. A large g heavily smooths the kernel matrices over time. This leads to a slow evolution of $\alpha(s)$ and $\beta(s)$ in equation 2.5 and could possibly obscure the non-stationary characteristic of the time series. On the other hand, small g might miss important lagged-correlation structure. In this paper, we did not explore optimizing g in the absence of intuition about the maximal lag of interest. We leave this for future work.

We also suggest a possible method for computational savings when the time series under consideration are long and prohibit DKCCA (or DCCA) from running in the desired amount of time. As presented in this paper, the algorithms set the weights for a single time point for each window of the sliding window. Instead, weights for j time points can be set per window. We advise that j should be much smaller than the length of the window, $2g + 1$, in order to ensure the smoothness of weights over time. For the analyses in this paper, we set $j = 1$.

We validated DKCCA on both simulated and real data, comparing the results to methods commonly used to compute cross-correlation. On simulated data, DKCCA successfully captured the dynamic cross-correlation structure, even under adversarial conditions where traditional methods failed. On real data, in addition to showing more intricate dynamics where traditional methods also captured some correlation, DKCCA found cross-correlation where those methods did not, suggesting a switch in the lead-lag relationship between the hippocampus (HPC) and prefrontal cortex (PFC) in the rhesus macaque.

Signals captured from the brain are highly dynamic, demanding new statistical tools to characterize them. DKCCA provides one such tool to describe dynamic cross-correlation.

Chapter 3

Torus graphs for multivariate phase coupling analysis

This work was submitted in 2019 to *Annals of Applied Statistics.* I collaborated with co-first author Josue Orellana, data collaborators Scott Brincat and Earl K. Miller, and advisor Robert E. Kass.

3.1 Introduction

New technologies for recording electrical activity among large networks of neurons have created great opportunities to advance neurophysiology, and great challenges in data analysis (e.g., Steinmetz et al., 2018). One appealing idea, which has garnered substantial attention, is that under certain circumstances, long-range communication across brain areas may be facilitated through coordinated network oscillations (Buzsáki and Draguhn, 2004; Ching et al., 2010; Fell and Axmacher, 2011; Sherman et al., 2016). To demonstrate coordination among oscillatory networks, computational neuroscientists have examined *phase coupling* across repetitions (trials) of the experiment. That is, when the phase of an oscillatory potential at a particular location, and a particular latency from the beginning of the trial, is measured repeatedly it will vary; phase coupling refers to the tendency of two phases, measured simultaneously at two locations, to vary together, i.e., to be associated, across trials. The data we analyze here consist of 24 phase angles recorded simultaneously, on each of many trials, from several brain regions known to play a role in memory formation and recall (Brincat and Miller, 2015a, 2016b), prefrontal cortex (PFC) and three sub-areas of the hippocampus, the dentate gyrus (DG), subiculum (Sub) and CA3. Being angles, phases may be considered circular random variables. A commonly-applied measure of phase coupling, known as Phase Locking Value (PLV), is an estimator of the natural circular analogue of Pearson correlation under certain assumptions (which we review). PLV, however, like correlation, can not distinguish between direct association and indirect association via alternative pathways. Thus, a large PLV between PFC and DG does not distinguish between direct coupling and indirect coupling via a third area, such as via CA3 (neural activity in PFC could be coupled directly with that in CA3, and that in CA3 with that in DG). To draw such a distinction we need, instead, a circular measure that is analogous to partial correlation. More generally, we wish to construct circular analogues of Gaussian graphical models. Because key properties of Gaussian graphical models are inherited by exponential families, and the product of circles is a torus, we consider exponential families on a multidimensional torus and call the resulting models *torus graphs.* We used torus graphs to provide a thorough description of associations among the 24 repeatedly-measured phases in the Brincat and Miller data and, in particular, we found strong evidence that the association between activity in PFC and DG is indirect, via both CA3 and Sub, rather than direct.

When circular random variables are highly concentrated around a central value, there is little harm in ignoring their circular nature, and multivariate Gaussian methods could be applied. However, in most of the neurophysiological data we've seen, including those analyzed here, the marginal distributions of phases are very diffuse, close to uniform, so the topological distinction between the circle and the real line is important. The torus topology is consequential not only for computation of probabilities but also for the interpretation of association. Figure 3.1 displays the inability of rectangular coordinates to preserve the clustering of points around a diagonal line under strong positive association. Furthermore, unlike the Gaussian case where a single scalar, correlation, describes both positive and negative association, on the torus, positive and negative association each have both an amplitude and a phase, so each pairwise association is, in general, described by 2 complex numbers. Also, in the Gaussian case, it is possible to interpret the association of two variables without knowing their marginal concentrations. This is no longer true for torus graphs.

After defining torus graphs and providing a few basic properties in Section 2, in Section 3 we consider several important alternative families of distributions for multivariate circular data that have appeared in the literature, and show that they are all special cases of torus graphs. In Section 4, we step through the interpretation of phase coupling in torus graphs by considering in detail the bivariate and trivariate cases. In Section 5, we provide estimation and inference procedures and, in Section 6, document via simulation studies the very good performance of these procedures in realistic settings. Our analysis of the data appears in Section 7 and we make a few closing remarks in Section 8.

3.2 Torus graph (TG) model

Suppose $\mathbf{X}$ is a d-dimensional random vector with jth element X_j being a circular random variable, which may be expressed as an angle in $[0, 2\pi)$, though other choices of angular intervals, such as $[-\pi, \pi)$, are equivalent. When $d = 2$, $\mathbf{X}$ lies on the product of two circles (a torus), and in general it lies on a multidimensional torus. When considering phase coupling in neural data, $\mathbf{X}$ represents a vector of phase angle values extracted

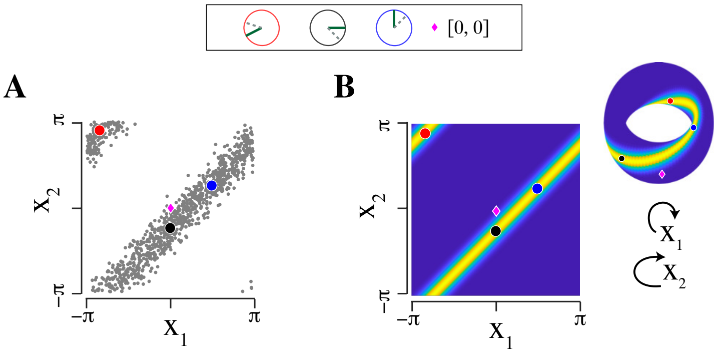

Figure 3.1: Rectangular coordinates are unable to accurately represent strong positive association between two circular random variables. (A) Scatter plot of simulated observations from a pair of dependent circular variables in rectangular coordinates, with three observations highlighted in red, black, and blue (simulated plot is similar to real data plots, but somewhat more concentrated for visual clarity; see Figure A.5). The highlighted observations are shown on circles at the top of the figure (one angle as a dashed line, one as a solid line; positive dependence implies a consistent offset between the angles). While the black and blue points follow the diagonal line, the red point falls near the upper left corner due to conversion to rectangular coordinates. (B) Probability density representing the two variables, plotted both in rectangular coordinates and on the torus, with the same three points marked. On the torus, there is a single band with high probability, which wraps around and connects to itself as a Möbius strip, and all three points fall on this strip.

from oscillatory signals for a single time point from each of d signals, with repeated trials providing multiple observations.

The torus graph model may be developed by analogy to the multivariate Gaussian distribution, a member of the exponential family that models dependence between d real-valued variables. In general, for a random vector $\mathbf{Y}$, an exponential family distribution is specified through a vector of natural parameters $\boldsymbol{\eta}$ that multiply a vector of sufficient statistics $\mathbf{S}(\mathbf{y})$ summarizing information from the data that is sufficient for the parameters (Wainwright et al., 2008) and has a density of the form :

$$p(\mathbf{y}|\boldsymbol{\eta}) \propto \exp\left(\boldsymbol{\eta}^T \mathbf{S}(\mathbf{y})\right).$$

In the bivariate Gaussian distribution, $\mathbf{Y} \in \mathbb{R}^2$ and the sufficient statistics corresponding to the natural parameters are $\mathbf{y}$ and $\mathbf{y}\mathbf{y}^T$, which describe the first- and second-order behavior of the variates. For a vector of angular variables $\mathbf{X} \in [0, 2\pi)^2$, we follow Mardia and Patrangenaru (2005) by representing the angles using rectangular coordinates on the unit circle as $\mathbf{Y}_1 = [\cos X_1, \sin X_1]$ and $\mathbf{Y}_2 = [\cos X_2, \sin X_2]$. The first-order sufficient statistics, $\mathbf{y}_1$ and $\mathbf{y}_2$, contain information both about the mean direction and the concentration around the mean direction. The second-order behavior is described by:

$$\mathbf{y}_1\mathbf{y}_2^T = \begin{bmatrix} \cos x_1 \cos x_2 & \cos x_1 \sin x_2 \\ \sin x_1 \cos x_2 & \sin x_1 \sin x_2 \end{bmatrix}.$$

This choice of sufficient statistics leads to the following natural exponential family density parameterized by $\boldsymbol{\eta} = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_{12}]$:

$$p(\mathbf{x}|\boldsymbol{\eta}) \propto \exp\left(\boldsymbol{\eta}_1^T \begin{bmatrix} \cos x_1 \\ \sin x_1 \end{bmatrix} + \boldsymbol{\eta}_2^T \begin{bmatrix} \cos x_2 \\ \sin x_2 \end{bmatrix} + \boldsymbol{\eta}_{12}^T \begin{bmatrix} \cos x_1 \cos x_2 \\ \cos x_1 \sin x_2 \\ \sin x_1 \cos x_2 \\ \sin x_1 \sin x_2 \end{bmatrix} \right) \tag{3.1}$$

The first two terms correspond to marginal circular means and concentrations of each variable, while the third term is a pairwise coupling term describing dependence between the variables. Under no pairwise coupling, the marginal distributions are all von Mises, and if $d = 1$, the torus graph model is itself von Mises. Extending Equation 3.1 to $d > 2$ yields

$$p(\mathbf{x}|\boldsymbol{\eta}) \propto \exp\left(\sum_{j=1}^{d} \boldsymbol{\eta}_j^T \begin{bmatrix} \cos x_j \\ \sin x_j \end{bmatrix} + \sum_{j<k} \boldsymbol{\eta}_{jk}^T \begin{bmatrix} \cos x_j \cos x_k \\ \cos x_j \sin x_k \\ \sin x_j \cos x_k \\ \sin x_j \sin x_k \end{bmatrix} \right). \tag{3.2}$$

The normalization constant is intractable, though numerical approximations may be used in the bivariate case (Kurz and Hanebeck, 2015).

Applying trigonometric product-to-sum formulas to the pairwise coupling terms of Equation 3.2 yields an equivalent, alternative parameterization in terms of natural parameters $\boldsymbol{\phi}$:

$$p(\mathbf{x}|\boldsymbol{\phi}) \propto \exp\left(\sum_{j=1}^{d} \boldsymbol{\phi}_j^T \begin{bmatrix} \cos x_j \\ \sin x_j \end{bmatrix} + \sum_{j<k} \boldsymbol{\phi}_{jk}^T \begin{bmatrix} \cos(x_j - x_k) \\ \sin(x_j - x_k) \\ \cos(x_j + x_k) \\ \sin(x_j + x_k) \end{bmatrix} \right). \tag{3.3}$$

We define a *d-dimensional torus graph* to be any member of the family of distributions specified by Equations 3.2 or 3.3. In the form of Equation 3.3, the sufficient statistics involving only a single angle are

$$\mathbf{S}_j^1(\mathbf{x}) = [\cos(x_j), \sin(x_j)]^T,$$

and the sufficient statistics involving pairs of angles are

$$\mathbf{S}^2_{jk}(\mathbf{x}) = [\cos(x_j - x_k),\, \sin(x_j - x_k),\, \cos(x_j + x_k),\, \sin(x_j + x_k)]^T.$$

We will use $\boldsymbol{\phi}$ and $\mathbf{S} \equiv [\mathbf{S}^1, \mathbf{S}^2]$ to refer to the full vectors of parameters and sufficient statistics for all angles. The natural parameter space is given by

$$\boldsymbol{\Phi} = \left\{ \boldsymbol{\phi} \,:\, \int_{[0,2\pi)^d} \exp\left(\boldsymbol{\phi}^T \mathbf{S}(\mathbf{x})\right)\, d\mathbf{x} \,<\, \infty \right\}$$

which implies that $\boldsymbol{\phi} \in \mathbb{R}^{2d^2}$ (because each angle has two marginal parameters, and each unique pair of angles has four coupling parameters, leading to $2d + 4[d(d-1)/2] = 2d^2$ parameters).

We prefer the parameterization of Equation 3.3 because it offers a simple interpretation: the sufficient statistics containing phase differences correspond to positive rotational dependence between the angles, while the sufficient statistics containing phase sums correspond to negative rotational (or reflectional) dependence. Positive rotational dependence occurs when phase differences are consistent across observations, that is, $X_j - X_k \approx \xi$ or $X_j \approx X_k + \xi$, for some angle ξ. Then conditionally on $X_j = x_j$, X_k is obtained by rotating from x_j by approximately ξ. Reflectional dependence instead refers to consistency in the phase sums so that $X_k \approx -X_j + \xi$, meaning that conditionally on $X_j = x_j$, X_k is obtained by rotating from $-x_j$ by approximately ξ. To demonstrate how each type of dependence might arise in repeated observations of neural oscillations, we show pairs of phase angles under each type of dependence in Figure A.2; in addition, bivariate torus graph densities dominated by each type of dependence are displayed in Figure A.1. While we have observed both kinds of dependence in neural phase angle data, rotational dependence appears to dominate in the data we analyze in this paper (see Section 3.7).

Because the natural parameter space is $\mathbb{R}^{2d^2}$ and the d-dimensional torus is compact, the full $2d^2$-dimensional exponential family is regular (see Brown 1986, p. 2). We call the full family a *torus graph model* and we summarize its properties, given above, in the following theorem.

Theorem 3.1 (Torus graph model)**.** *The d-dimensional torus graph model is a full regular exponential family. Equation (2.3) provides a reparameterization of the family in Equation (2.2) in which the expectations of the sufficient statistics are the first circular moments and (for $d \geq 2$) the second circular moments representing rotational and reflectional dependence between pairs of variables. In Equation (2.3), the natural parameter has components $\boldsymbol{\phi}_j \in \mathbb{R}^2$ corresponding to the first circular moment of X_j and $\boldsymbol{\phi}_{jk} \in \mathbb{R}^4$ corresponding to the second circular moments representing dependence between X_j and X_k.*

We prove Theorem 3.1 in Section A.1 by writing the angles as complex numbers and considering the complex first moments and complex-valued covariances between the variables.

Because the torus graph is an exponential family distribution with sufficient statistics corresponding to first circular moments and to pairwise interactions between variables, it is similar to the multivariate Gaussian distribution, and, as in a Gaussian graphical model, the parameters correspond to a conditional independence graph structure. Specifically, as we state in Corollary 3.1.1 and prove in Section A.3, the pairwise coupling parameters $\boldsymbol{\phi}_{jk}$ correspond to the structure of an undirected graphical model, where an edge is missing if, and only if, the corresponding pair of variables are conditionally independent given all the other random variables. This suggests that an undirected graphical model structure may be learned through inference on the pairwise interaction parameters, or by applying regularization in high dimensions to shrink the pairwise interaction parameters.

Corollary 3.1.1 (Torus graph properties)**.** *The d-dimensional torus graph model has the following properties:*

1. *It is the maximum entropy model subject to constraints on the expected values of the sufficient statistics.*
2. *In the torus graph model, the random variables X_j and X_k are conditionally independent given all other variables if and only if the pairwise interaction terms involving X_j and X_k vanish (that is, if the entire vector $\boldsymbol{\phi}_{jk} = \mathbf{0}$ in the density of Equation 3.3).*

Another interesting property of the torus graph model, given in Theorem 3.2 and proven in Section A.6, is that the univariate conditional distributions of one variable given the rest are von Mises, enabling Gibbs sampling to be used to generate samples from the distribution. In addition, Theorem 3.2 shows torus graphs are similar to other recent work in graphical modeling in which the joint distribution is specified through univariate exponential family conditional distributions (Chen et al., 2014; Yang et al., 2015).

Theorem 3.2 (Torus graph conditional distributions)**.** *Let $\mathbf{X}_{-k}$ be all variables except X_k. Under a torus graph model, the conditional density of X_k given $\mathbf{X}_{-k}$ is von Mises; specifically,*

$$p(x_k|\mathbf{x}_{-k};\boldsymbol{\phi}) = \frac{1}{2\pi I_0(A)} \exp(A\cos(x_k + \Delta))$$

where A and Δ are defined as

$$A = \sqrt{\left(\sum_m L_m \cos(V_m)\right)^2 + \left(\sum_m L_m \sin(V_m)\right)^2},$$
$$\Delta = \arctan\left(\frac{\sum_m L_m \sin(V_m)}{\sum_m L_m \cos(V_m)}\right),$$

where

$$L = [\kappa_k, \boldsymbol{\phi}_{\cdot k}]\,,\; V = [\mu_k, \mathbf{x}_{-k}, \mathbf{x}_{-k} + \mathbf{h}\tfrac{\pi}{2}, -\mathbf{x}_{-k}, -\mathbf{x}_{-k} + \tfrac{\pi}{2}],$$

and $\boldsymbol{\phi}_{\cdot k}$ denoting all coupling parameters involving index k, $\mathbf{h}_j = -1$ if $j < k$ and $\mathbf{h}_j = 1$ otherwise.

3.3 Important subfamilies of the torus graph model

In this section, we discuss some important subfamilies of the torus graph model that are particularly relevant to the application to neural data. In particular, for neural phase angle data, the marginal distributions are often nearly uniform, prompting consideration of a *uniform marginal model* in which the parameters $\boldsymbol{\phi}_j$ corresponding to the first moments of each variable are set to zero, resulting in a model with uniform marginal distributions. In addition, while in our experience neural phase angle data may exhibit both rotational and reflectional covariance, in many data sets, the primary form of dependence is rotational, prompting us to consider a submodel with parameters corresponding to reflectional dependence set to zero, which we will call the *phase difference submodel* since all pairwise relationships are described by the sufficient statistics involving $x_j - x_k$. A subfamily that combines these two (that is, restricts the torus graph to have marginal uniform distributions and no reflectional dependence) coincides with the models of Zemel et al. (1993) and Cadieu and Koepsell (2010), which were developed from Boltzmann machines and coupled oscillators, respectively.

The uniform marginal model, phase difference model, and a phase difference model with uniform margins all correspond to affine restrictions on the parameter space. This implies (see Section A.2) that each is itself a regular exponential family, so that each inherits many nice properties, such as concavity of the loglikelihood function, as a function of the natural parameter. Most previous work in multivariate circular distributions has focused on the so-called *sine model* (e.g., Mardia et al., 2007), which is again a subfamily, but it is not itself a full regular exponential family and does not, in general, have a concave loglikelihood function. As a result, estimation and inference are less straightforward than for either the torus graph model or the full regular exponential family submodels (Mardia et al., 2016). We summarize properties of these subfamilies in Theorem 3.3, which is proven in Section A.2.

Theorem 3.3 (Torus graph subfamilies)**.** *The uniform marginal model, the phase difference model, and a model combining both parameter space restrictions, form full regular exponential families but the sine model does not.*

We note that the sine models may provide parsimonious fits to data for which the marginal distributions appear unimodal. Even though the torus graph is a full regular exponential family, and is therefore identifiable, when the data are highly concentrated it may be hard to estimate all four coupling parameters, a phenomena we explore with simulations in Figure A.3. Neural phase angle data, however, often tend to have low concentrations while still exhibiting strong pairwise dependence (Figure 3.9). As shown in Mardia et al. (2007), when the concentration is low relative to the pairwise interaction strength, the sine model fitted density enters a regime of multimodality. In Section 3.7 we demonstrate lack of fit of the sine model to our neural data.

3.4 Phase coupling in torus graphs

In this section, we discuss the distinction between bivariate measures of phase coupling, such as PLV, and multivariate measures. In Section 3.4.1, we briefly review bivariate phase coupling measures based on the marginal distributions of pairwise phase differences. In Sections 3.4.2 and 3.4.3 we investigate bivariate and trivariate examples analytically. We show that when a trivariate distribution of angles follows a torus graph, the marginal distributions of pairwise phase differences may be influenced by the marginal distributions of each variable and by indirect coupling through other nodes. This fundamental limitation of bivariate phase coupling measures can produce inaccurate phase coupling descriptions in multivariate systems. For the special case of phase difference models with uniform margins, in Section 3.4.4 we propose a transformation of the torus graph parameters that produces a generalization of PLV to multivariate data (where coupling between two variables is measured conditionally on all other variables), having the nice feature that, like PLV, it falls between 0 and 1.

3.4.1 Bivariate phase coupling measures

The most common bivariate phase coupling measure between angles X_j and X_k is the Phase Locking Value (PLV) (Lachaux et al., 1999b), defined by

$$\hat{P}_{jk} = \left| \frac{1}{N} \sum_{n=1}^{N} \exp\left\{ i \left(x_j^{(n)} - x_k^{(n)} \right) \right\} \right| \tag{3.4}$$

where $x_j^{(n)}$ is the nth observation of X_j. We have used the notation $\hat{P}_{jk}$ to indicate it may be viewed as an estimator of its theoretical counterpart P_{jk}. In Section A.7, we show that P_{jk} corresponds to a measure of positive circular correlation under the assumption of uniform marginal distributions. The value of P_{jk} falls between 0 and 1, with 0 indicating no consistency in phase differences across trials and 1 indicating identical phase differences across trials. One way to assess significance of $\hat{P}_{jk}$ is Rayleigh's test for uniformity of the phase differences (Kass et al., 2014, p. 268); other assessments of significance typically involve permutation tests or comparison to non-task recording periods (Rana et al., 2013).

A similar approach to characterizing bivariate phase coupling follows from considering the univariate random variable $Y_{jk} = X_j - X_k$. If Y_{jk} is distributed as von Mises with concentration parameter κ, then

$$\hat{P}_{jk} = \frac{I_1(\hat{\kappa})}{I_0(\hat{\kappa})} \tag{3.5}$$

where $\hat{\kappa}$ is the maximum likelihood estimator for κ and I_m denotes the modified Bessel function of the first kind of mth order (Forbes et al., 2011, p. 191). More generally, any measure of the concentration of the marginal distribution of phase differences around a mean direction may be used as a measure of bivariate phase coupling (Aydore et al., 2013). We will refer to measures based on the marginal distribution of phase differences as *bivariate phase coupling measures.*

3.4.2 Marginal distribution of phase differences in a bivariate torus graph

Because bivariate phase coupling measures are based on the marginal distributions of phase differences, we investigate here the form of the marginal phase difference distributions in a bivariate torus graph model to determine how the torus graph parameters influence the phase differences. For the most straightforward and analytically tractable exposition, we consider the bivariate phase difference model. We will use the notation

$$\boldsymbol{\phi}_{jk} = [\alpha_{jk}, \beta_{jk}, \gamma_{jk}, \delta_{jk}]^T$$

to refer to elements of the pairwise coupling parameter vector, and use trigonometric identities to write the marginal terms in terms of κ and μ (see Section A.2 for details). Then the bivariate phase difference model density is

$$p(x_1, x_2) \propto \exp\left\{\alpha_{12}\cos(x_1 - x_2) + \beta_{12}\sin(x_1 - x_2) + \sum_{j=1}^{2}\kappa_j \cos(x_j - \mu_j)\right\}.$$

Let $W = X_1 - X_2 \,(\mathrm{mod}\, 2\pi)$ be the phase differences wrapped around the circle so that $W \in [0, 2\pi]$. As shown in Section A.4, the unnormalized theoretical distribution of W is a product of two functions:

$$p_W(w) \propto f(w; \boldsymbol{\kappa}, \boldsymbol{\mu}) \cdot g(w; \alpha_{12}, \beta_{12}), \tag{3.6}$$

where

$$f(w; \boldsymbol{\kappa}, \boldsymbol{\mu}) = I_0\left(\sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2 \cos(w - (\mu_1 - \mu_2))}\right)$$

and

$$g(w; \boldsymbol{\phi_{12}}) = \exp\left\{\sqrt{(\alpha_{12}^2 + \beta_{12}^2)}\cos\left(w - \arctan\left(\frac{\beta_{12}}{\alpha_{12}}\right)\right)\right\}. \tag{3.7}$$

The first factor, f, is proportional to the density of the sum of two independent von Mises random variables with concentrations κ_1, κ_2 and means μ_1, μ_2 (Jammalamadaka and Sengupta, 2001, p. 40) and reflects the

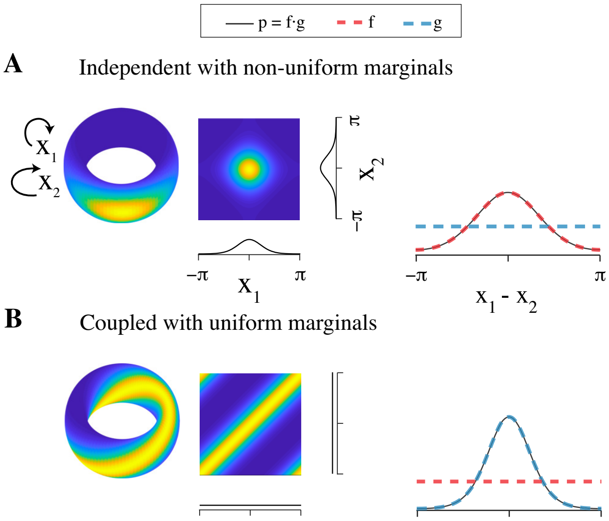

Figure 3.2: Examples of bivariate torus graph densities and the resulting marginal distributions of phase differences upon which bivariate phase coupling measures would be based. As shown in Equation 3.6, the density of phase differences, p, is affected not only by coupling (through g) but also by marginal concentration (through f). As a result, bivariate phase coupling measures like PLV could give misleading results. (A) Left: bivariate torus graph density with independent angles and non-uniform marginal distributions; density is shown on the torus and flattened on $[-\pi, \pi]$ with marginal densities on each axis. Right: analytical phase difference density (p, black) which is a product of a direct coupling factor (g, blue) and a marginal concentration factor (f, red). Here, p is concentrated solely through the marginal concentration factor f, implying bivariate phase coupling measures would indicate coupling despite the independence of X_1 and X_2. (B) Similar to A, but with coupling between angles and uniform marginal distributions; only in this case does p correctly reflect the coupling.

influence of the marginal distributions of X_1 and X_2 on the phase differences. Such convolved densities are unimodal on $[0, 2\pi)$ with mode $\mu_1 - \mu_2 \pmod{2\pi}$ and concentration increasing with the argument of $I_0(\cdot)$. The second factor, g, is proportional to a von Mises density that depends only on the phase difference and the coupling parameters.

The functional forms of f and g show that the distribution of phase differences is influenced both by the coupling parameters and by the marginal concentration parameters, which implies that bivariate phase coupling measures reflect both coupling and marginal concentration. In Figure 3.2, we illustrate effects on $\hat{P}_{jk}$ of pairwise dependence and marginal concentration. Even when the variables are independent, if the marginal distributions are not uniform, the distribution of phase differences will have nonzero concentration due to the influence of f. Thus, PLV is only appropriate when the marginal distributions are uniform. In contrast, torus graph parameters can separate the influence of marginal concentration and phase coupling to provide a measure of the dependence between angles.

3.4.3 Marginal distribution of phase differences in a trivariate torus graph model

While we have shown that torus graph models are preferable to bivariate phase coupling measures in the bivariate case, the biggest advantage of using torus graphs comes from the ability to work with multivariate data and determine unique associations between each pair of variables after conditioning on the other variables. For instance, in a trivariate torus graph model with direct coupling only from nodes 1 to 3 and nodes 2 to 3 (Figure 3.3.A), if we were to apply bivariate phase coupling measures to all pairwise connections, we would likely infer a connection between 1 and 2 because we would be measuring the bivariate association between phase angles without taking into account node 3.

To demonstrate analytically how this happens, we consider a trivariate phase difference model with marginal concentrations equal to zero for simplicity, which has density

$$p(x_1, x_2, x_3) \propto \exp\left\{ \sum_{(j,k)\in E} \begin{bmatrix} \alpha_{jk} \\ \beta_{jk} \end{bmatrix}^T \begin{bmatrix} \cos(x_j - x_k) \\ \sin(x_j - x_k) \end{bmatrix} \right\}, \tag{3.8}$$

where $E = \{(1,2),\ (1,3),\ (2,3)\}$. Letting $W = X_1 - X_2 \pmod{2\pi}$ be the phase difference between nodes 1 and 2, we show in Section A.4 that the unnormalized density of W is given by the product of two factors:

$$p_W(w) \propto g(w; \boldsymbol{\phi_{12}}) \cdot h(w; \boldsymbol{\phi_{13}}, \boldsymbol{\phi_{23}}). \tag{3.9}$$

The first factor is the same as g in Equation 3.7 and reflects direct connectivity between X_1 and X_2 as it depends only on the coupling parameters for the pair, $\boldsymbol{\phi}_{12}$. The second factor reflects indirect connectivity through the other nodes, as it depends on the coupling parameters for the other pairs:

$$h(w; \boldsymbol{\phi_{13}}, \boldsymbol{\phi_{23}}) \propto I_0\left(\sqrt{s + 2t\cos(w - u)}\right)$$

where

$$\begin{aligned} s &= \alpha_{13}^2 + \beta_{13}^2 + \alpha_{23}^2 + \beta_{23}^2, \\ t &= \sqrt{(\alpha_{13}^2 + \beta_{13}^2)(\alpha_{23}^2 + \beta_{23}^2)}, \\ u &= \arctan\left(\tfrac{\beta_{13}}{\alpha_{13}}\right) - \arctan\left(\tfrac{\beta_{23}}{\alpha_{23}}\right). \end{aligned}$$

Therefore, h is proportional to the density of the sum of two independent von Mises random variables with concentrations $\alpha_{13}^2 + \beta_{13}^2$ and $\alpha_{23}^2 + \beta_{23}^2$ and mean directions $\arctan(\beta_{13}/\alpha_{13})$ and $-\arctan(\beta_{23}/\alpha_{23})$, respectively.

Equation 3.9 implies that the density of the phase differences for one pair of variables depends on all of the coupling parameters, so a bivariate phase coupling measure such as PLV will be unable to distinguish between the effects of direct coupling and indirect coupling through other nodes. Consequently, bivariate phase coupling measures will accurately represent the direct coupling between 1 and 2 only when there is no indirect path between 1 and 2 through the other nodes. In the most extreme case, bivariate phase coupling measures could indicate coupling even when there are *only* indirect connections between two nodes through the rest of the network. In Figure 3.3, we show examples to demonstrate how the phase difference distribution is affected by both direct and indirect connections, which may result not only in contributions to the observed phase difference concentration but also in shifts in the mean phase difference. This demonstrates that bivariate phase coupling measures generally reflect both direct and indirect coupling; in contrast, torus graph parameters identify direct coupling.

3.4.4 Interpreting phase difference model parameters

An appealing feature of PLV is that it always falls between 0 and 1, so it is easy to interpret its magnitude and to compare PLV values for different pairs of variables. Unfortunately, the torus graph parameters lack these qualities. However, for the special case of the phase difference model with uniform margins, we propose a generalization of PLV based on a transformation of the torus graph model parameters that offers increased interpretability, and that, unlike PLV, measures pairwise relationships conditional on the other nodes.

As shown in Equation 3.5, if the marginal phase difference is distributed according to a von Mises distribution, then PLV corresponds to a function of the maximum likelihood estimator of the marginal concentration parameter. Under the phase difference model with uniform margins, we showed the marginal density of the phase difference $X_1 - X2$ factors into terms corresponding to direct and indirect connections; the direct connectivity term $g(w; \boldsymbol{\phi_{12}})$ of Equation 3.7 has the form of a von Mises density depending only on the parameters $\boldsymbol{\phi_{12}}$. Therefore, in analogy to the definition of PLV for von Mises-distributed phase differences, we propose the following transformation of the parameters:

$$\tilde{P}_{jk} = \frac{I_1\left(\sqrt{\alpha_{jk}^2 + \beta_{jk}^2}\right)}{I_0\left(\sqrt{\alpha_{jk}^2 + \beta_{jk}^2}\right)}.$$

Like PLV, the measure always falls between 0 and 1 and therefore may be used to compare relative edge strengths in the phase difference model with uniform margins.

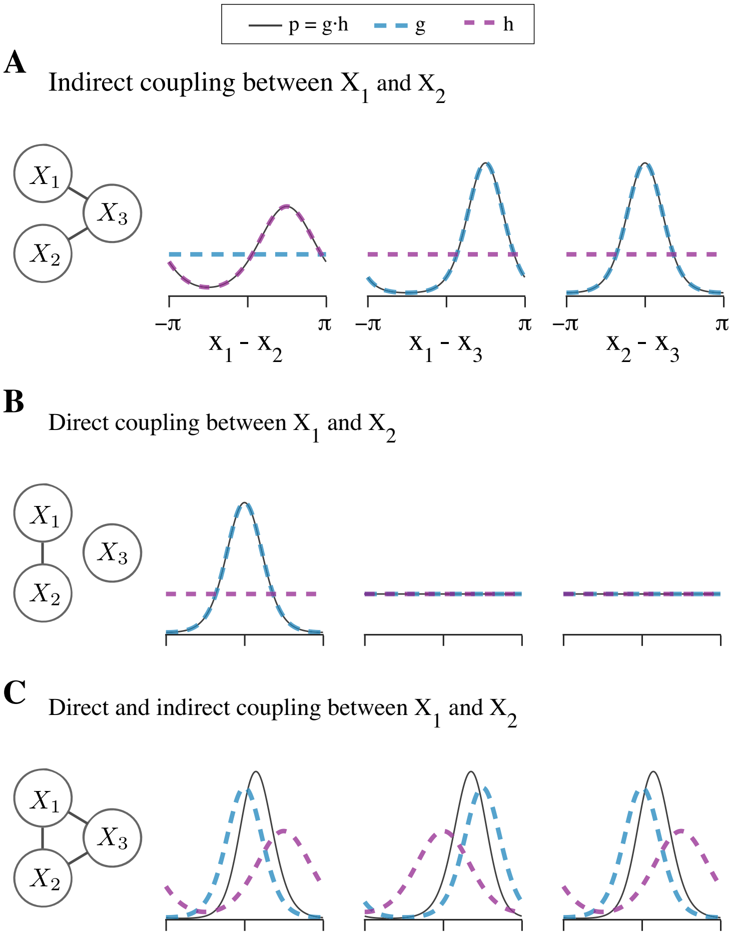

Figure 3.3: Examples of trivariate torus graph densities and the resulting densities of phase differences for each pair of variables. As shown in Equation 3.9, in general, the density of phase differences, p, is affected not only by direct coupling (through g) but also by indirect connections (through h). As a result, bivariate phase coupling measures like PLV will generally reflect both direct and indirect coupling. (A) Left: ground truth graphical model, with no direct connection between X_1 and X_2 but an indirect connection through X_3. Right: analytical phase difference densities for each pair of angles (p, black) which are each a product of a direct coupling factor (g, blue) and an indirect coupling factor (h, purple). The concentration in the phase difference $X_1 - X_2$ arises solely due to indirect connections. (B) Similar to A, but with direct connection only between X_1 and X_2; in this case, p reflects the direct coupling. (C) Similar to A, but with direct connections between all nodes. Notice that indirect connections (h) still influence the distribution of phase differences $X_1 - X_2$ by multiplying with the direct connection term (g), which increases the concentration of p and shifts the mean (compared to g).

3.5 Torus graph estimation and inference

Because the normalization constant is intractable for the torus graph density and it cannot easily be approximated even for moderate dimension, estimation and inference are not straightforward. In particular, maximum likelihood estimation is not feasible. Instead, we turn to an an alternative procedure for estimation and inference called *score matching*. In Section 3.5.1, we establish the applicability of the score matching estimator, originally defined for densities on $\mathbb{R}^d$, to multivariate circular densities like torus graphs, then give the explicit form of the objective function and derive closed-form estimators that maximize the objective function. Section 3.5.2 discusses two main approaches for determining a graph structure, one based on the asymptotic distributions of score matching estimators and the second based on regularization, which is particularly relevant for high-dimensional problems.

3.5.1 Estimation via score matching

Score matching is an asymptotically consistent estimation method that does not require computation of the normalization constant and is based on minimizing the expected squared difference between the model and data *score functions* (gradients of the log-density functions), which leads to a tractable objective function for estimating the parameters (Hyvärinen, 2005b). It can be seen as analogous to maximum likelihood estimation, which uses the log likelihood as a scoring rule; score matching instead uses the gradient of the log density (with respect to the data) as a scoring rule (Dawid and Musio, 2014). In addition, for real-valued exponential family distributions, the estimator comes from an unbiased linear estimating equation, so asymptotic inference is straightforward (Forbes and Lauritzen, 2015; Yu et al., 2018). However, the original score matching estimator requires the density to be supported on $\mathbb{R}^d$ and the proof of consistency relies on tail properties of such densities. We show that score matching estimators applied to circular densities such as the torus graph model retain the same form and therefore remain consistent. Score matching estimators have been considered previously for the phase difference model with uniform margins (Cadieu and Koepsell, 2010) and the sine model (Mardia et al., 2016), where the procedure requires modification because the sine model is a curved exponential family distribution (Theorem 3.3).

The score matching objective function is the expected squared difference between the log gradients:

$$J(\boldsymbol{\phi}) = \frac{1}{2}\int p_{\mathbf{X}}(\mathbf{x})||\nabla_{\mathbf{x}} \log q(\mathbf{x};\boldsymbol{\phi}) - \nabla_{\mathbf{x}} \log p_{\mathbf{X}}(\mathbf{x})||_2^2 \, d\mathbf{x}. \tag{3.10}$$

The objective function depends on the unknown data density $p_{\mathbf{X}}(\mathbf{x})$ in a nontrivial way, but we show in Theorem 3.4, using techniques similar to Hyvärinen (2005b, 2007), that the objective function may be simplified to depend on the data density only through an expectation, allowing it to be estimated as an average over the sample (proof in Section A.5).

Theorem 3.4 (Score matching estimators for torus graphs)**.** *Under some mild regularity assumptions (given in Section A.5), the score matching objective function for the torus graph model takes the form*

$$J(\boldsymbol{\phi}) = E_{\mathbf{x}}\left\{\frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\phi} - \boldsymbol{\phi}^T\mathbf{H}(\mathbf{x})\right\}$$

where

$$\mathbf{H}(\mathbf{x}) = [\mathbf{S}^1(\mathbf{x}),\, 2\mathbf{S}^2(\mathbf{x})]^T$$

is a vector with dimension $2d^2$ that is a simple function of the sufficient statistics and

$$\boldsymbol{\Gamma}(\mathbf{x}) = \mathbf{D}(\mathbf{x})\mathbf{D}(\mathbf{x})^T$$

where

$$\mathbf{D}(\mathbf{x}) = \nabla_{\mathbf{x}}\mathbf{S}(\mathbf{x})$$

is the $2d^2 \times d$ Jacobian of the sufficient statistic vector. Specific expressions for the Jacobian elements are given in Section A.5.

Theorem 3.4 shows that the score matching objective may be estimated empirically by averaging over N observed samples. The empirical objective function is

$$\tilde{J}(\boldsymbol{\phi}) = \frac{1}{2}\boldsymbol{\phi}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{\phi} - \boldsymbol{\phi}^T\hat{\mathbf{H}} \tag{3.11}$$

where

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\Gamma}\left(\mathbf{x}^{(n)}\right),\ \hat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}\left(\mathbf{x}^{(n)}\right)$$

with $\mathbf{x}^{(n)}$ denoting sample n. Taking the derivative of Equation 3.11 with respect to the parameter vector yields an unbiased estimating equation (Dawid and Musio, 2014), which has a unique solution when $\hat{\boldsymbol{\Gamma}}$ is invertible:

$$\hat{\boldsymbol{\Gamma}}\boldsymbol{\phi} - \hat{\mathbf{H}} = 0 \;\longrightarrow\; \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}^{-1}\hat{\mathbf{H}}.$$

The number of parameters for a d-dimensional torus graph is $2d^2$ so sample sizes may not be sufficient for $\hat{\boldsymbol{\Gamma}}$ to be invertible. In particular, $\hat{\boldsymbol{\Gamma}}$ is a sum with Nd terms, so N must be greater than $2d$ for $\hat{\boldsymbol{\Gamma}}$ to be invertible.

In practice, the variance of estimated parameters will be high if N is not much larger than $2d$, leading to less accurate point estimates and inference. We investigate the effect of sample size on the resulting inferences, using simulated data, in Section 3.6.

For higher-dimensional problems, Equation 3.11 is a convex objective function that may be minimized numerically with regularization. In torus graphs, a group ℓ_1 penalty may be placed on the groups of pairwise coupling parameters $\boldsymbol{\phi}_{jk}$ to enforce sparsity in the estimated edges, yielding the following objective function:

$$\tilde{J}_\lambda(\boldsymbol{\phi}) = \tilde{J}(\boldsymbol{\phi}) + \lambda \sum_{j<k} ||\boldsymbol{\phi}_{jk}||_2.$$

Here, λ is a tuning parameter that may be selected by criteria such as cross-validation or extended BIC (Lin et al., 2016). Other structured penalties may be used to encourage the model toward specific submodels (such as the phase difference model or the uniform marginal model). For instance, separate group ℓ_1 penalties could be applied to the pairs of coupling parameters corresponding to positive and negative dependence, or an ℓ_2 penalty on the marginal parameters $\boldsymbol{\phi}_j$ could encourage low concentration. This type of penalization may improve behavior of the objective function, and could be especially useful when certain subfamilies appear reasonable based on exploratory data analysis. The computational burden of calculating $\hat{\boldsymbol{\Gamma}}^{-1}$ may be reduced using the conditional independence structure of the graph, as we may estimate each four-dimensional group of parameters $\boldsymbol{\phi}_{jk}$ using score matching on the conditional distribution $p\left(x_j, x_k | \mathbf{x}_{-jk}\right)$, which involves only $8(d-1)$-dimensional sufficient statistics and thus uses only a subset of the rows of $\mathbf{D}(\mathbf{x})$, lessening the computational burden of matrix inversion (Yu et al., 2016).

3.5.2 Inference for graphical structure

In our setting, the goal of inference is to determine a graph structure by determining which pairs of variables $\{j,k\}$ have nonzero $\boldsymbol{\phi}_{jk}$, indicating an edge between nodes j and k. As shown in previous works, score matching estimators are asymptotically Normal (Dawid and Musio, 2014; Forbes and Lauritzen, 2015; Yu et al., 2018), that is,

$$\sqrt{N}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right) \tag{3.12}$$

where the asymptotic variance is given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}_0^{-1}\mathbf{V}_0\boldsymbol{\Gamma}_0^{-1}$$

where

$$\mathbf{\Gamma}_0 = E[\mathbf{\Gamma}(\mathbf{x})],\ \mathbf{V}_0 = E[(\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\phi} - \mathbf{H}(\mathbf{x}))(\mathbf{\Gamma}(\mathbf{x})\boldsymbol{\phi} - \mathbf{H}(\mathbf{x}))^T].$$

Sample averages may be substituted for the expectations to obtain an estimate of the asymptotic variance, and because the true value of $\boldsymbol{\phi}$ is unknown, we may substitute either our estimate $\hat{\boldsymbol{\phi}}$ or a null hypothetical value.

By considering the marginal Gaussian distribution of each element of $\boldsymbol{\phi}$, confidence intervals may be constructed in a standard way. However, in the torus graph model, there are four parameters per edge, so individual parameters are not of primary interest. In addition, we may be interested in testing hypotheses about groups of edges (for example, the null hypothesis might be that there are no edges between regions A and B). Fortunately, inference on groups of edges is also straightforward, as specified in the following lemma.

Lemma 3.4.1 (Asymptotic distribution for groups of torus graph parameters.)**.** *A vector of parameters indexed by an index set E of size $|E|$, denoted $\boldsymbol{\phi_E}$, satisfies*

$$\sqrt{N}\left(\hat{\boldsymbol{\phi}}_{\boldsymbol{E}} - \boldsymbol{\phi}_{\boldsymbol{E}}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma_E}\right)$$

where $\mathbf{\Sigma_E}$ is the corresponding submatrix of the asymptotic variance $\mathbf{\Sigma}$ of Equation 3.12. Then

$$N\left(\hat{\boldsymbol{\phi}}_{\boldsymbol{E}} - \boldsymbol{\phi}_{\boldsymbol{E}}\right)^T \mathbf{\Sigma_E}^{-1}\left(\hat{\boldsymbol{\phi}}_{\boldsymbol{E}} - \boldsymbol{\phi}_{\boldsymbol{E}}\right) \xrightarrow{d} \chi^2(|E|).$$

Lemma 3.4.1 enables computation of p-values for single edges (if E indexes the four parameters for a single edge) or for groups of edges. In particular, if E indexes the four parameters corresponding to a single edge, then under the null hypothesis that $\boldsymbol{\phi_E} = 0$,

$$X_E^2 \equiv N\hat{\boldsymbol{\phi}}_{\boldsymbol{E}}^T \mathbf{\Sigma_E}^{-1}\hat{\boldsymbol{\phi}}_{\boldsymbol{E}} \xrightarrow{d} \chi^2(4),$$

so for an observed value of the test statistic $\hat{X}_E^2$, the probability statement $P\left(X_E^2 \geq \hat{X}_E^2\right)$, which gives a p-value for the edge, may be evaluated using the χ^2 distribution with 4 degrees of freedom. Similarly, a χ^2 distribution with two degrees of freedom may be used to test for only rotational or only reflectional covariance, or to test for nonzero marginal parameters.

Inference after regularization is less straightforward. Recent work has addressed inference for score matching estimators when using an ℓ_1 penalty on each parameter (Yu et al., 2016), which could potentially be extended to torus graphs with a group ℓ_1 penalty. Other approaches for inference in high dimensions include the bootstrap or stability selection (Meinshausen and Bühlmann, 2010). One parametric bootstrap approach is as follows. Assume we are interested in testing the null hypothesis that some particular subset of edges is missing from the graph. We may first fit a null torus graph model in which the coupling parameters corresponding to the edge set of interest are set to zero, selecting the regularization parameter by cross-validation of the score matching objective function. Next, B times, we would draw samples of the same size as the data from the null torus graph model, re-select the regularization parameter by cross-validation, and fit the unrestricted torus graph model to the samples using this regularization parameter. By computing the distribution of a suitable statistic (such as the number of nonzero edges or the maximal edgewise parameter vector norm) from these fitted null models, we obtain an empirical estimate of the null distribution of the statistic, which can then be used to judge the size of the same statistic computed on the original data.

3.6 Simulation study

As our analytical results of Section 3.4 show, torus graphs can separate the effects of pairwise coupling and marginal concentration and have pairwise coupling parameters that represent direct connections between nodes. In contrast, bivariate phase coupling measures like PLV are sensitive to the marginal distribution of the variables and can reflect not only direct connections but also indirect paths through the other nodes. We conducted simulations to demonstrate these results. In addition, we explored the performance of torus graphs in recovering graph structures in simulated data similar to real data to determine how well we expect torus graphs to perform in the real data. Section 3.6.1 gives the simulation details and Section 3.6.2 provides the results.

3.6.1 Simulation methods

When comparing PLV to torus graphs, we chose to generate data using the notion of positive rotational dependence discussed in Section 3.2. This was done to demonstrate that torus graphs recover interactions of this type even when data were not directly generated from a torus graph model. To generate bivariate data with rotational dependence and nearly uniform marginal distributions, we first drew N trials of phase angles x_1 from a von Mises distribution with low circular concentration κ_1. Then, for each trial, we let $x_2 = x_1 + \xi + \epsilon$ where ξ is a fixed phase offset and ϵ is noise drawn from a concentrated mean-zero von Mises distribution with concentration κ_ϵ (where, on a small number of trials, we used less concentrated noise to emulate the noisiness present in real data). Extending to more than two nodes follows a similar process, where data for an additional node is generated based on data from a neighbor in the graph.

We generated synthetic data with two ground truth phase coupling structures that are intended to reflect realistic scenarios (and with parameters chosen to produce samples that emulate real neural phase angle data; Figure A.10 compares the simulated data to real data, showing similar first- and second-order behavior and similar observed pairwise PLV values). First, we constructed five-dimensional data meant to emulate the effects of spatial dependence, such as dependence between electrodes on a linear probe situated within a single functional region, which, under a nearest-neighbor Markov assumption, would induce sparse conditional independence graph structures (because each node would be directly dependent only on its nearest neighbors on the probe). We coupled nodes in a linear chain and chose $\xi = \pi/100$ and $\kappa_\epsilon = 40$ (with 15 of 840 trials contaminated with extra noise with concentration 0.1). Second, we constructed three-dimensional data meant to emulate the effects of indirect connections, which may occur when electrodes are in different regions, but not all of the regions are communicating directly. In particular, x_2 was concentrated at $\kappa_2 = 0.01$, x_1 and x_3 had phase offsets of $\xi = \pi/6$ and $\xi = \pi/100$, respectively, from x_2, and the coupling noise had concentration $\kappa_\epsilon = 2$ (with 75 of 840 trials contaminated with extra noise with concentration 0.1). For each scenario, we simulated data of sample size 840 (to match the sample size of the real data). Then, for each data set, we fit a torus graph and selected edges based on Lemma 3.4.1; we also used the Rayleigh test of uniformity to construct a graph based on PLV (Kass et al., 2014, p. 268). For both tests, we used an alpha level of $\alpha = 0.001$ with Bonferroni correction for multiple tests.

To gain intuition on how well torus graphs could be expected to perform in the real LFP data we analyze in Section 3.7, we investigated how well torus graphs recover the edges for varying dimensions, sample sizes, and underlying levels of sparsity in the edges . For this simulation, we generated data from a torus graph model of varying dimension with zero marginal concentration and with either 25% or 50% of edges present in the generating distribution. By varying the threshold on the edgewise X^2 statistics (Lemma 3.4.1), we computed an ROC curve for each simulated data set. The ROC curves were averaged across 30 data sets, then the area under the curve (AUC) was calculated as a measure of performance.

3.6.2 Simulation results

For the first set of simulations, Figure 3.4 shows that in both the three-dimensional and five-dimensional cases, the torus graph recovered the correct structure while the PLV graph recovered a fully connected graph. Although the performance of PLV may be better for other graph structures, our analytical results in Section 3.4 suggest that graph structures with indirect paths between nodes are likely to induce excess edges in the PLV graph. To follow up on this result, we further explored the False Positive Rate (FPR) and False Negative Rate (FNR) for PLV and torus graphs by repeating the simulations. We found that PLV graphs have very high FPR (near 1) but also have a high FNR, so PLV likely won't miss a true edge but will also add many additional edges Figure A.11. This result agrees with the notion that hypothesis testing based on PLV, even when corrected for multiple comparisons, cannot be reliably used to control FPR for multivariate graphs because PLV measures both direct as well as indirect connectivity and thus tends to overestimate connectivity. In contrast, torus graphs are more conservative in assigning edges and control the FPR at the nominal level (though they tend to have higher FNR, especially for low sample sizes).

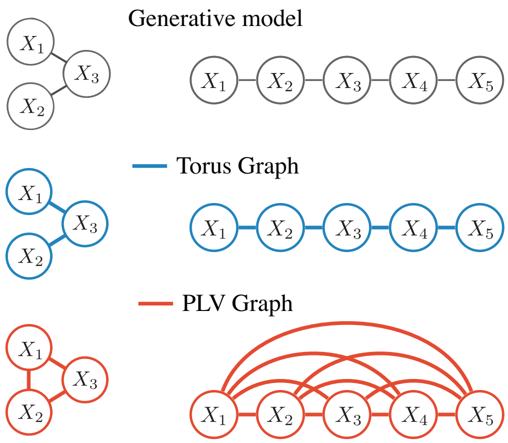

Figure 3.4: The torus graph recovers the ground truth graph structures (top row) from realistic simulated data sets while a bivariate phase coupling measure, phase locking value (PLV), does not (edges shown for $p < 0.001$). Left column: a 3-dimensional simulated example of cross-area phase coupling where regions X_1 and X_2 are not directly coupled, but are both coupled to region X_3. Right column: a 5-dimensional simulated example of a graph structure that could be observed for channels on a linear probe with nearest-neighbor spatial dependence. In both cases, PLV infers a fully-connected graph due to indirect connections.

Figure 3.5 displays the results of the second set of simulations, which investigated the ability of torus graphs to recover the true structure as a function of true edge density, sample size, and data dimension. Importantly, for simulated data of dimension 24 and sample size of 840 (matching the real LFP data), the torus graph model is able to achieve 0.9 AUC as long as the graph is sufficiently sparse (around 25% of all possible edges present). In the real data results of Figure 3.8.B, we in fact observe approximately 25% of edges present, suggesting that this graph density may be reasonable for the real data. A more detailed investigation of the ROC curves and precision curves by dimension with fixed sample size 840 is given in Figure A.4, which again demonstrates that for a sufficiently sparse underlying graph structure, the torus graph method is expected to perform well for the sample size and dimension in the real LFP data. However, prior beliefs about the sparsity of the underlying graph will play a role in judging the likely accuracy of results.

3.7 Analysis of neural phase angles

We demonstrate torus graphs in a set of local field potentials (LFPs) collected from 24 electrodes in the prefrontal cortex (PFC) and hippocampus (HPC) of a macaque monkey during a paired-associate learning task. Previous analysis of these data in Brincat and Miller (2016b) found that beta-band (16 Hz) phase coupling between PFC and HPC peaked during the cue presentation and also increased with learning after the subject received feedback on each trial. Here, we sought a more fine-grained description of the phase

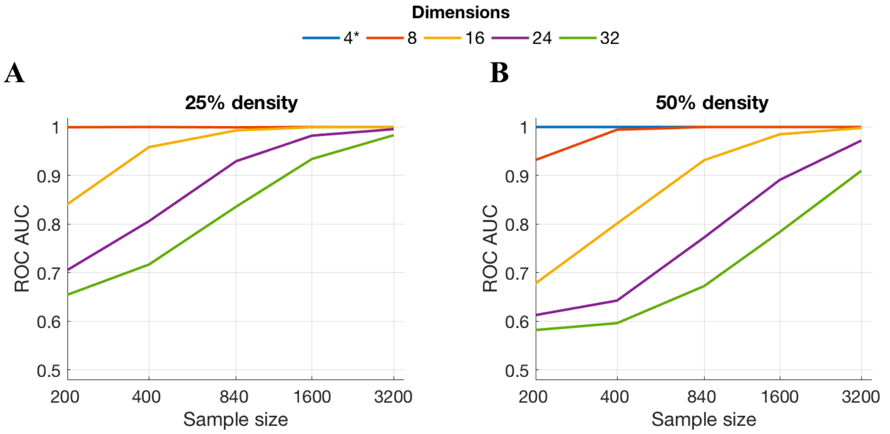

Figure 3.5: In simulated data with two different underlying edge densities, the average ROC curve area under the curve (AUC) was computed across 30 simulated data sets as a function of sample size (shown along the horizontal axis). The dimension of the data is marked by line color. Panel A demonstrates that if the true underlying graph has only 25% of all possible edges present, then even for 24 dimensional data, a sample size of 840 (the size of our real LFP data set) is sufficient to reach AUC above 0.9. While performance degrades when the underlying graph is more dense, panel B shows that performance is still reasonable (AUC near 0.8) for 24 dimensional data with 840 samples.

coupling between PFC and HPC during the cue presentation period, and focused on describing relationships between PFC and three distinct subregions of HPC: subiculum (Sub), dentate gyrus (DG), and CA3.

First, we applied torus graphs to two different low-dimensional subnetworks: (i) a subnetwork consisting of five electrodes arranged linearly along a probe within CA3 and (ii) a collection of all trivariate subnetworks consisting of an electrode in each of the regions Sub, DG, and PFC. The five-dimensional subnetwork was chosen as a proof of concept, because electrodes in the same region and with a linear spatial arrangement ought to exhibit a nearest-neighbor conditional independence structure. We chose to examine connectivity between Sub, DG, and PFC because the patterns of connectivity between these three regions could be informative about whether hippocampal activity is leading prefrontal activity and because torus graphs should be able to disentangle the effect of direct and indirect connections to give a more informative connectivity structure than could bivariate phase coupling measures. Second, we applied torus graphs to the full 24-dimensional data set by first testing for the presence of any cross-region edges between PFC, Sub, DG, and CA3, and then following up with post-hoc tests of individual cross-region and within-region edges to construct a full 24-dimensional graph. Finally, we used a subset of the PFC electrodes to examine the goodness-of-fit of the torus graph model to the data and to investigate whether any torus graph subfamilies appeared to be appropriate for this data set.

We describe the data and preprocessing in Section 3.7.1 and give an outline of our data analysis methods in Section 3.7.2. Section 3.7.3 presents the results and we discuss implications of the results in Section 3.7.4.

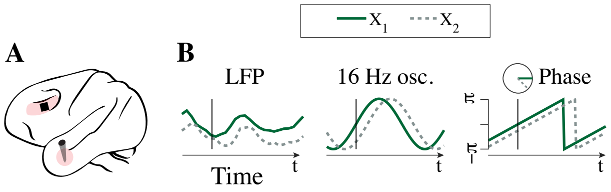

Figure 3.6: (A) Depiction of recording sites in ventrolateral prefrontal cortex (PFC) and hippocampus. (B) Preprocessing to obtain phase angles: local field potential (LFP) signals are filtered using Morlet wavelets to extract phase angles from 16 Hz oscillations at a time point of interest (two signals are shown for a single trial; repeated observations of phase angles are collected across repeated trials).

3.7.1 Experiment and data details

The experimental design and data collection procedures are described thoroughly in Brincat and Miller (2015a, 2016b). We use data from a single animal in a single session comprising 840 trials in which a correct response was given. (The sample size here is 840; a very small number of animals, usually 1 or 2, is standard practice in nonhuman primate neurophysiology because, even though there is large subject-to-subject variability in the fine details of brain structure and function, the overall structure and function of major brain regions is conserved, as are, typically, the primary scientific conclusions, though it is common to replicate in a second animal results found in a single animal; we also note that while part of the purpose of the original experiment involved learning, we are here ignoring any transient learning effects, which take place rapidly.) Briefly, four images of objects were randomly paired; the monkey learned the associations between pairs through repeated exposure to the pairs followed by a reward for correctly identifying a matching pair. In each trial, the pairs of images were presented sequentially with a 750 ms delay period between the images, during which a fixation mark was shown. All procedures followed the guidelines of the MIT Animal Care and Use Committee and the US National Institutes of Health. (The experimental procedures were painless to the animals, as all forms of sensation originate outside the brain.) The data used in this paper contain 8 single-channel electrodes in PFC and a linear probe with 16 channels in HPC, with HPC channels categorized based on neural spiking characteristics into three subregions: dentate gyrus (DG), CA3, and subiculum (Sub). Recording regions and data processing steps are depicted in Figure 3.6. We focus on a time point at 300ms after initial cue presentation, as PLV pooled across all sessions identified phase coupling peaking near 16Hz at this time point (Brincat and Miller, 2016b, Supplementary Figure 5); we verified that the session we used showed the same overall phase coupling relationship. After downsampling the data to 200Hz and subtracting the average (evoked) response, we used complex Morlet wavelets (cycle number 6) to extract the instantaneous phase of each channel at 16Hz in each trial.

3.7.2 Data analysis methods

To examine whether torus graphs could recover the spatial features we would expect along the linear probe, we first applied torus graphs to the 5-dimensional network containing channels on a linear probe that are all within CA3, likely to exhibit strong spatial dependence between neighboring channels, and used a hypothesis test for each edge with $\alpha = 0.0005$. We chose a stringent threshold because, based on the first simulation study of Section 3.6, we expected PLV to add extraneous edges, yet we wanted to demonstrate that torus graphs and PLV give very different results even when a small threshold is used. Then, to examine whether torus graphs appeared to disentangle the effects of direct and indirect edges, we focused first on a trivariate network containing Sub, DG, and PFC where there is a simple interpretation of direct edges because CA3 and Sub send output signals from hippocampus while DG receives input signals to the hippocampus. Therefore, prominent connections between CA3 and PFC and/or Sub and PFC would suggest hippocampal activity may be leading PFC activity during this period of the task, while dominance of connections between DG and PFC would suggest the opposite. Because there are multiple channels in each of the three regions, we aggregated results across all possible triplets of channels from the three regions by inferring edges by majority vote across all possible trivariate graphs (using an alpha level of $p < 0.0005$ for each edgewise test).

We also investigated the graphical structure between the four regions (PFC, DG, CA3, and Sub) using a 24-dimensional torus graph on all electrodes. First, we assessed between-region connectivity between each pair of regions using the results of Lemma 3.4.1 to test the set of null hypotheses that there were no edges between each pair of regions. For example, there are 40 total possible edges between CA3 and PFC with four parameters for each edge, so the hypothesis test for the existence of any edges between CA3 and PFC was based on a $\chi^2(160)$ distribution. For each pair of regions, we obtained a p-value for the entire group of edges between the two regions under the null hypothesis that there are no edges between the two regions. Because we were interested only in further investigating cross-region interactions with strong evidence, for this set of hypothesis tests, we chose a stringent alpha level of $\alpha = 0.001$ with a Bonferroni correction across all 6 between-region tests to control for multiple tests. To better understand the individual connections driving this across-region connectivity, and to investigate within-region connectivity patterns, we then applied post-hoc tests on the individual edges. That is, for each possible edge between pairs of regions that were identified as having some connection by the first step, we obtained a p-value using a $\chi^2(4)$ distribution. Similarly, we calculated a p-value for each possible edge connecting electrodes within the same region. In this case, we assigned edges using a less stringent alpha level of $\alpha = 0.05$ without correcting for multiple tests, as we expected the evidence for any specific edge would be weaker than the evidence for across-region connections and because multiple comparisons were already taken into account in the first set of between-region tests.

Finally, to assess the appropriateness of torus graphs for this data set and to assess whether any submodels were appropriate, we explored the first- and second-order behavior of the LFP phase angles. To determine whether the uniform marginal model should be fitted, we tested for uniform marginal distributions using a Rayleigh test on each electrode, marginally, then obtained an overall decision regarding the null hypothesis that all distributions were uniform using Fisher's method to combine p-values (Kass et al., 2014, p. 301). Equation 3.9 showed that the marginal distribution of phase differences in a torus graph model depends on the coupling parameters for all possible direct and indirect paths between two nodes, so that any observed concentration of phase differences, marginally, could indicate the presence of *some* nonzero coupling parameters (possibly corresponding to indirect paths). Therefore, we applied the Rayleigh test to the observed phase differences for each pair of variables, then combined p-values using Fisher's method to test the null hypothesis that all marginal phase difference distributions were uniform; we also performed a similar procedure on the observed pairwise phase sums. For these exploratory tests, we used a p-value threshold of $p < 0.05$ so that we would be sensitive to departures from uniformity in either the marginal distributions or the distributions of phase differences/sums. Finally, after fitting the chosen model, we used Kolmogorov-Smirnov (KS) goodness-of-fit tests to determine whether there was any evidence that LFP angles were drawn from a different distribution than the fitted theoretical model. In particular, we tested for differences in the marginal distributions of the angles, then tested for differences in the marginal distributions of pairwise phase differences and pairwise phase sums, and used Fisher's method to combine p-values within each of the three groups of tests. We used an alpha level of 0.05 for each test.

3.7.3 Data results

In the low-dimensional subnetworks, we found that the torus graph recovered a structure consistent with nearest-neighbor coupling along the linear probe (Figure 3.7.B), while PLV suggested a fully-connected graph; Figure A.7.B shows p-values for the five-dimensional graph for both torus graphs and PLV. In addition, Figure 3.7.A shows that torus graphs appear to capture an interesting trivariate network with no coupling between PFC and DG, but with each of PFC and DG coupled with Sub. Figure A.7.C shows p-values for all of the individual trivariate graphs, indicating that in the majority of individual trivariate graphs, there was no evidence for an edge between PFC and DG, while again, PLV suggested a fully connected graph.

For the 24-dimensional torus graph applied to all electrodes, we found, first, using the overall tests for the presence of any edges between each pair of regions, that there was apparent connectivity between all hippocampal subregions (CA3, DG, and Sub) and connectivity from CA3 to PFC and Sub to PFC (Figure 3.8.A). In the follow-up post-hoc tests of individual edges, we observed dense connectivity within regions and somewhat sparser connectivity between regions (as judged by the number of edges out of the possible number that could be present; graph and adjacency matrix shown in Figure 3.8 panels B and C). Interestingly, none of the individual CA3 to Sub connections were significant (the smallest p-values were around 0.1), suggesting that the aggregate effect of several weak edges led to the edge between CA3 and

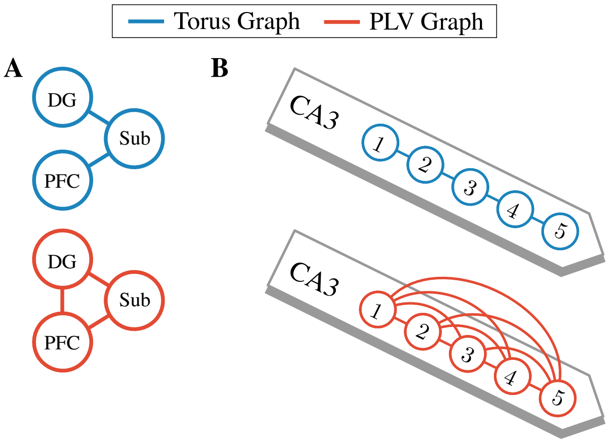

Figure 3.7: Torus graphs and PLV graphs from low-dimensional networks of interest in LFP data. (A) Across-region connectivity between dentate gyrus (DG), subiculum (Sub), and PFC, where the torus graph (blue) indicated that DG and PFC are each coupled to Sub; in contrast, PLV (red) inferred a fully-connected graph. (B) Within-region connectivity in CA3, where the torus graph (blue) indicated a spatial dependence structure which reflects the placement of channels along a linear probe, while PLV (red) inferred a fully-connected graph.

Sub in Figure 3.8.A. In the adjacency matrix corresponding to the 24-dimensional graph, which is ordered to respect the ordering of channels on the hippocampal linear probe, entries are colored by p-value (with white indicating non-significant entries at level $\alpha = 0.05$); notably, torus graphs recovered the linear probe structure across the entire hippocampus without prior knowledge of this structure being used in the model or estimation procedure.

In our investigation of the appropriateness of submodels and goodness-of-fit, we show results for three electrodes from PFC (results on the full data set were similar). Based on a visual analysis of the marginal and pairwise behavior of the data, the full model, the uniform marginal model, and the phase difference model all appeared to be reasonable candidates for these data. We found evidence that neither the marginal distributions nor the phase differences were uniform (Rayleigh/Fisher's method, $p < 0.001$); however, there was no evidence for concentration in the phase sums (Rayleigh/Fisher's method, $p > 0.05$). As a result, we selected the phase difference model to investigate goodness-of-fit. In Figure 3.9.A, we show the data from the three PFC electrodes along with the theoretical (fitted) phase difference model; along the diagonal, histograms for the real data (blue) are shown with theoretical kernel density estimates (red), indicating similar marginal distributions. Below the diagonal are two-dimensional histograms from the real data and above the diagonal are two-dimensional kernel density estimates for the theoretical model, demonstrating that the torus graph appears capable of accurately representing the first- and second-order behavior present in the real data. Figure 3.9.B shows histograms of sufficient statistics from the real data (phase angles along the diagonal, phase sums for each pair above the diagonal, and phase differences for each pair below the

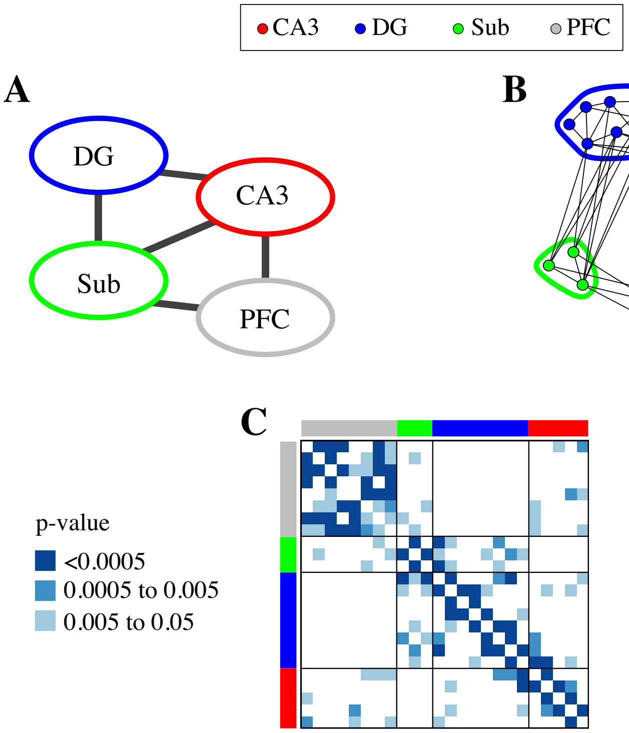

Figure 3.8: Torus graph analysis of coupling in 24-dimensional LFP data with four distinct regions: CA3 (red), dentate gyrus (DG; blue), subiculum (Sub; green), and prefrontal cortex (PFC; grey). (A) Across-region tests indicated evidence for edges between DG and CA3, DG and Sub, CA3 and Sub, PFC and CA3, and PFC and Sub (determined by testing the null hypothesis that there were no edges between a given pair of regions, corrected $p < 0.001$). (B) For the significant across-region connections, a post-hoc edgewise significance test examined the specific connections between regions, with the 24-dimensional graph containing edges for $p < 0.05$. Compared to the across-region graph in A, notice that no edges from Sub to CA3 were individually significant at $p < 0.05$. (C) The 24-dimensional adjacency matrix (with hippocampal electrodes ordered by position on linear probe) with non-significant across region connections in white and other entries colored by edgewise p-value (with $p > 0.05$ in white). Despite no built-in knowledge of spatial information, the torus graph recovered the linear probe structure within hippocampus.

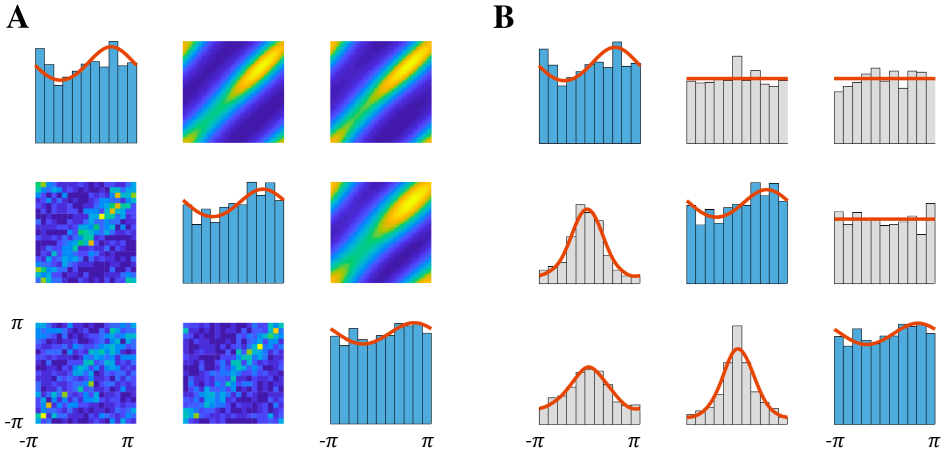

Figure 3.9: Comparison between phase angles from three LFPs located in PFC and the theoretical torus graph distribution demonstrates that torus graphs capture the salient first- and second-order behavior present in the LFP phase angles. In contrast, the sine model fails to fit the data accurately as shown in Figure A.6. (A) Along the diagonal are the marginal distributions of the phase angles. The real data are represented by blue histograms and the theoretical marginal densities from the torus graph model are overlaid as red lines. Two-dimensional distributions (off-diagonal) show bivariate relationships, with theoretical densities above the diagonal and real data represented using two-dimensional histograms below the diagonal. (B) Plots along the diagonal same as panel A. Below the diagonal are distributions of pairwise phase differences and above the diagonal are distributions of pairwise phase sums, represented by histograms for the real data and by red density plots for the theoretical torus graph model. Both the real data and theoretical distributions exhibit concentration of phase differences but not phase sums, suggesting prevalence of rotational covariance, and the sufficient statistics are very similar for the theoretical and real data.

diagonal), along with kernel density estimates (red) for the statistics from the theoretical model. There is visual similarity between the distributions of sufficient statistics, and, in both cases, we observed the concentration of the pairwise phase differences indicating a prevalence of rotational dependence; this pattern held in the observed sufficient statistics for all 24 LFP channels shown in Figure A.9. The KS tests comparing the sufficient statistics of the fitted torus graph model to the data failed to reject the null hypothesis (KS/Fisher's method, $p > 0.05$, for each group of statistics), indicating no evidence that the data were drawn from a different distribution than the fitted model.

3.7.4 Summary and discussion of results

In the 24-dimensional analysis, PFC to hippocampal connections appeared to be driven by a relatively small number of significant connections from Sub to PFC and CA3 to PFC. However, the results did not show evidence for connections between PFC and DG, which coincides with the analysis of the trivariate subnetwork. In contrast, the PLV edgewise adjacency matrix (shown in Figure A.8) was very densely connected and shows little resemblance to the structure recovered by torus graphs; for any reasonable p-value threshold, the PLV graph would be nearly fully connected, suggesting that PLV was unable to distinguish between direct and indirect connections.

Because the patterns of connectivity between hippocampal subregions and PFC recovered by these analyses suggested that hippocampal activity was leading prefrontal activity during this time period in the task, as a follow-up analysis, we considered whether lead-lag relationships could be detected in the between-region distributions of phase differences. In fact, in this data set, most of the edgewise dependence appears to correspond to positive rotational dependence (as judged by the relative magnitude of the torus graph coupling parameters and the concentration of phase differences but not phase sums in the observed sufficient statistics shown in Figure A.9). Thus, the distributions of phase differences between PFC and Sub and between PFC and CA3 summarize the overall PFC-hippocampus coupling. Figure 3.10.A displays a circular histogram of the between-region phase differences with the mean phase difference and a 95% confidence interval, pooling across all significant ($p < 0.05$) pairwise phase differences to compute an overall circular mean phase offset. Between-region phase differences were centered at $-8.5°$ (95% CI: $[-10.4°, -6.6°]$), indicating that on average, hippocampal phase angles led PFC phase angles, providing more evidence that hippocampal activity was leading PFC activity. The between-region phase differences agree with those displayed in Brincat and Miller 2015a, Figure 5.a (though we examined a different time period of the trial and pooled only across significant PFC to hippocampus connections). In contrast, Figure 3.10.B displays within-region phase differences tightly clustered around $0°$ (mean: -0.04, 95% CI: $[-0.2°, 0.2°]$), indicating that within-region phase coupling may be driven mostly by spatial correlations in the recordings. These results suggest that due to the CA3 to PFC and Sub to PFC connections identified in both the trivariate and 24-dimensional analyses and the overall negative phase difference (PFC - hippocampus), hippocampal activity leads PFC activity during this period of the task. Importantly, while aggregated pairwise phase differences may have given some evidence of directionality, torus graphs provided detailed information about direct connections between hippocampal output subregions and PFC.

In the low-dimensional subnetworks, we found that torus graphs yielded intuitive results while PLV did not. In particular, for the five-dimensional subnetwork consisting of electrodes along a linear probe, torus graphs inferred a nearest-neighbor conditional independence structure which we would expect for electrodes arranged linearly in space, while PLV inferred a fully connected graph. In the trivariate subnetworks, torus graphs suggested connectivity between PFC and Sub and DG and Sub, but not between PFC and DG, while again, PLV inferred a fully connected graph. Based on our analytic derivations in Section 3.4 and simulation study in Section 3.6, PLV is likely not reflecting the correct dependence structures in either low-dimensional network, and is instead reflecting both direct and indirect connections between the nodes.

When we investigated possible use of submodels, we found that a phase difference model appeared reasonable due to the lack of concentration in the phase sums, but that a uniform marginal model was not warranted. On the full 24-dimensional data set, we applied the full torus graph, but found that we would have

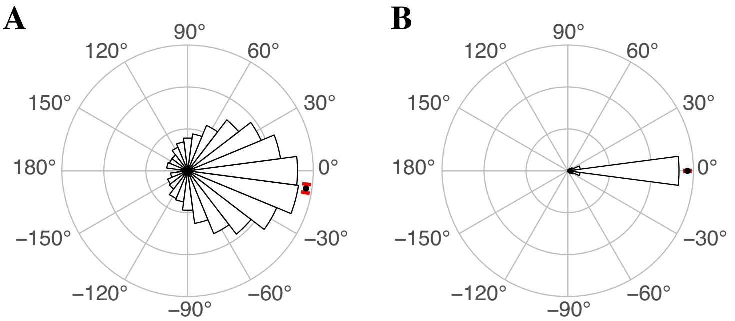

Figure 3.10: Circular histograms of phase differences, in degrees, from 24-dimensional LFP data for (A) significant connections between PFC and hippocampus (specifically, CA3 and Sub) and (B) significant PFC, CA3, and Sub within-region connections, with mean phase offset (black dot) and 95% confidence interval (red). Observations were pooled across all significant edges within or between the regions. The within-region phase differences were tightly concentrated around zero while the PFC-hippocampus phase differences were centered below zero, indicating a possible lead-lag relationship with hippocampus leading PFC.

obtained nearly the same results using a phase difference submodel, with discrepancies for only a few edges in the post-hoc edgewise test results shown in Figure 3.8.B (which do not change the overall conclusions). We found that torus graphs appeared to fit the data reasonably well, with both visual summaries and KS tests suggesting that the marginal distributions of each angle and of the pairwise sum and difference sufficient statistics were similar in the data and the fitted model. In contrast, when we followed up by fitting a sine model to the same data (using code from Rodriguez-Lujan et al., 2017), we observed multimodality in the bivariate densities of the fitted model and a poor correspondence between the fitted and observed sufficient statistics, leading us to conclude that, as discussed in Section 3.3, the sine model fails to match the second-order dependence structure in the neural data due to low marginal concentration Figure A.6. KS tests comparing the fitted sine model phase sums and differences to the data also suggested the data distribution does not match the sine model distribution (KS/Fisher's method, $p < 0.0001$, for both phase sums and phase differences).

Torus graphs provided a good description of the neural phase angle data and provided substantive conclusions that could not have been obtained using bivariate phase coupling measures like PLV.

3.8 Discussion

We have argued that torus graphs provide a natural analogue to Gaussian graphical models: Theorem 3.1 and Corollary 3.1.1 show that starting with a full torus graph, which is an exponential family with two-way interactions, setting a specific set of interaction coefficients $\boldsymbol{\phi}_{jk}$ to zero results in conditional independence of the jth and kth circular random variables. We provided methods for fitting a torus graph to data, including identification of the graphical structure, i.e., finding the non-zero interaction coefficients, corresponding to edges in the graph (code and data are available at `https://github.com/natalieklein/torus-graphs`). We also demonstrated that previous models in the literature amount to special cases, and therefore make additional assumptions that may or may not be appropriate for neural data. In particular, while the uniform marginal model or the phase difference model may be reasonable for neural phase angle data, the most widely studied model in multivariate circular statistics, the sine model, is less well-behaved and does not appear to be capable of matching the characteristics of neural phase angle data. In addition, we showed that PLV is a measure of positive circular correlation under the assumption of uniform marginal distributions of the angles, but that PLV is unable to recover functional connectivity structure that takes account of multi-way dependence among the angles. In our analysis of LFP phases 300 ms after cue presentation in an associative memory task, the fitted torus graph correctly identified the apparent dependence structure of the linear probe within CA3; it suggested Sub may be responsible for apparent phase coupling between PFC and DG; and it led to the conclusion that, at this point in the task, hippocampus phases lead those from PFC (by $8.5°$ with $\mathrm{SE} = 0.95°$).

Here, our torus graphs were based on phases of oscillating signals, with no regard to their amplitudes. This is different than phase amplitude coupling in which the phase of one oscillation may be related to the amplitude of an oscillation in a different frequency band (Tort et al., 2010). Also, like other graph estimation methods, interpretations based on torus graphs assume that all relevant signals have been recorded, while in reality, they could be affected by unmeasured confounding variables (e.g., activity from other brain regions). In addition, in applications such as phase angles in LFP, several preprocessing steps (referencing, localization, and filtering) are needed to extract angles from the signals. The torus graph implementation we have described here ignores these steps, and takes well-defined angles as the starting point for analysis. Furthermore, local field potentials tend to be highly spatially correlated, suggesting that inclusion of spatial information might be helpful for identifying structure.

Future work could include further investigation and theoretical analysis of how well torus graphs perform when the sample size is smaller relative to the dimension of the data. In some data sets, the full torus graph with $2d^2$ parameters may be overparameterized, and estimation and inference may be more accurate using one of the subfamilies; we demonstrated a model selection approach in our analysis of three electrodes from PFC which indicated that the phase difference submodel would be reasonable for this data set. Furthermore, even when a full torus graph model is used, interpretability of the results could be enhanced by assessing evidence for reflectional and rotational dependence separately; that is, instead of putting a single edge based on the test of all four coupling parameters, we could construct a graph based only on the two parameters corresponding to reflectional (or rotational) dependence. Finally, in the uniform marginal phase difference model, the strength of coupling for each type of dependence could be quantified using the measure we introduced in 3.4.4, which falls between 0 and 1, facilitating comparison of relative strengths of the connections.

By extending existing models, torus graphs are able to represent a wide variety of multivariate circular data, including neural phase angle data. Extensions to this work could study changes in graph structure across time or across experimental conditions, and could investigate latent variable models involving hidden states, or a spatial hierarchy of effects. We anticipate a new line of research based on torus graphs.

Chapter 4

Gaussian process current source density estimation

I am preparing this work for submission in a computational neuroscience journal. I worked with data collaborators Tobias Teichert (University of Pittsburgh) and Josh Siegle (Allen Institute) and with advisor Robert E. Kass.

4.1 Introduction

Electrical signals recorded from electrodes placed in brain tissue reflect time-varying neural activity. The high-frequency content of the recorded signal indicates spikes of individual neurons in the vicinity of the electrode, while the lower-frequency content (consisting of timescales slower than about 500Hz) is termed the *local field potential* (LFP) and reflects mostly post-synaptic potentials primarily from nearby populations of neurons (Buzsáki et al., 2012; Einevoll et al., 2013). However, unlike spiking activity that can typically be attributed to a small collection of individual cells very close to the electrode, the LFP is an indirect and aggregated measure of the activity of many neurons and may contain not only activity in the vicinity of the electrode but also activity from further away sources (Lindén et al., 2011; Kajikawa and Schroeder, 2011; Herreras, 2016). That is, despite the name, LFP is not necessarily local because all transmembrane currents in the brain contribute linearly to the voltage measurement at any given point; though the weight of any particular contribution to the voltage decays as a function of distance, strong but distant current sources may still contribute substantially to the LFP. In addition, a collection of nearby current sources will superimpose to give rise to the LFP voltage at any given measurement point, complicating interpretation of the signal and its relationship to other nearby voltage signals. In this chapter, I develop novel methodology to infer measures of the local transmembrane current flow and I use these measures to assess information flow from the LFP.

That is, the neural activity of interest is the aggregate current flow in and out of specific populations of neurons near the recording electrode; the volume density of the net current flow is termed the *current source density* (CSD). The CSD differs from local spiking activity in that it mostly reflects post-synaptic potentials rather than spiking activity. Biophysical models relate the latent CSD to the measured LFPs through a *forward model*; attempting to invert the forward model to infer the latent CSD from measured LFPs, also known as solving the *inverse problem*, is the goal of CSD estimation methods (Pitts, 1952; Nicholson and Llinas, 1971).

Estimating the CSD is important because it provides a measure of neural activity that better represents the synaptic activity of specific neural populations near the electrodes, and thereby it should be better suited for understanding information flow within a neural circuit. As an example, linear probes oriented perpendicular to the cortex are often used to measure activity from different cortical layers (as in one of the data sets I study in this chapter, in which linear probes are used to measure activity in auditory cortex). We may be interested in assessing correlations between activity in different layers over time, but the LFP alone is unlikely to give an accurate sense of correlated activity across layers due to the superposition of nearby current sources at each electrode location. To give some intuition on this point, after introducing the physical models relating the CSD to the LFP, I use a small theoretical case study to demonstrate that the correlation between LFPs at two spatial locations does not necessarily correspond to the correlation between current sources at those locations (Section 4.2.3). This suggests that recovering the CSD will in general give a more accurate assessment of correlated neural activity than the LFP.

In addition to correlation-based or phase-based measures of information flow (such as those discussed in Chapters 2 and 3), trial-to-trial variation in responses to a stimulus (known as *evoked responses*) are of great interest (Arieli et al., 1996). Previous work in the auditory system suggests that trial-to-trial variation in stimulus-evoked responses is better understood using the CSD than the LFPs (Szymanski et al., 2011). In the auditory cortex data set, I use the CSD to explore trial-to-trial variation in the timing and amplitude of specific extracted components of the CSD stimulus response. To illustrate why the CSD stimulus response may be more useful for assessing trial-to-trial variation, Figure 4.1.A shows an example trial of real spatiotemporal LFP data recorded in primary auditory cortex (details in Section 4.5). In the model I developed to analyze this data set, the latent noiseless LFP (panel B) decomposes into the sum of an evoked response (middle) and ongoing activity (right). Likewise, the latent CSD (panel C) also decomposes into a sum of evoked and ongoing activity, with the forward model describing how a given CSD generates an LFP. The CSD evoked response, consisting of several separate bumps, more likely reflects stimulus-related activity of individual cell populations which are difficult to infer from the spatially blurred LFP evoked response. As a result, trial-to-trial variation of the CSD evoked responses should be more informative about relationships between cell populations than trial-to-trial variation of the LFP evoked responses.

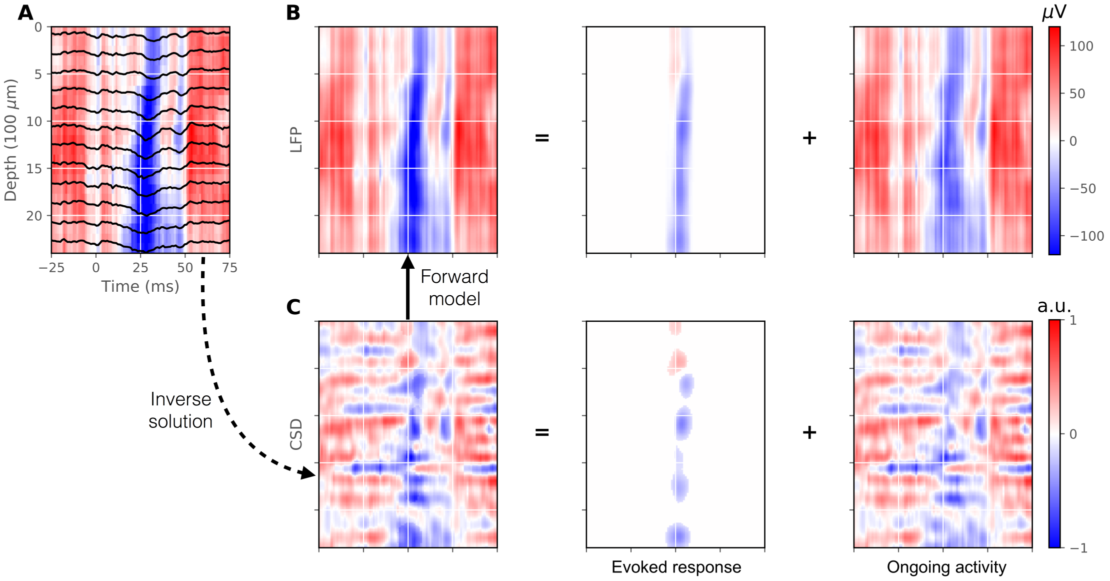

Figure 4.1: A) A single trial of real auditory LFP data (details in Section 4.5), with LFP voltage traces overlaid on a heat map describing the variation in LFPs across electrode depth (vertical axis) and time (horizontal axis). B) Illustration of the model for the LFP, which is decomposed into an evoked response (middle column) and ongoing activity (right column). C) Model for the latent CSD which, like the LFP model, decomposes each trial into an evoked response and ongoing activity and is related to the LFP through the forward model (a biophysical model describing how a given CSD generates an LFP). The proposed GPCSD statistical model provides the inverse solution (recovers the CSD from the observed LFPs).

In this chapter, I propose GPCSD, a novel method for estimating the CSD that is based on a spatiotemporal Gaussian process model for the latent CSD combined with an appropriate forward model. In contrast to existing CSD estimation methods, GPCSD models temporal correlations and borrows strength across trials to estimate hyperparameters governing not only the frequency of spatial and temporal fluctuations, but also the properties of the forward model. I demonstrate that GPCSD outperforms existing CSD estimation methods in simulated data, and I use it to assess information flow in two neural data sets: one measuring activity in primate auditory cortex and one measuring activity in mouse visual, hippocampal, and thalamic areas. In the auditory data set, GPCSD uncovers phase coupling networks from 10 Hz oscillations with connections between two simultaneously recorded probes not only at similar cortical depths, but also across different cortical layers; crucially, this network was not recovered using the LFPs directly. An analysis of trial-to-trial variation in the timing and amplitude of evoked response components in the CSD revealed a similar pattern of connectivity between probes, with most of the timing and amplitude correlations occurring between the early evoked responses at similar cortical depths but with a few correlations between input and output layers. In the Neuropixels data set, GPCSD combined with PCA discovered detailed spatial patterns of current flow associated with different components of the time series. In particular, the principal components appeared to extract portions of the evoked response in both the GPCSD predictions and in the LFP, but with slightly different per-trial timecourses and more detailed spatial information about the origin of each component in the CSD. These preliminary results suggest that GPCSD analysis of Neuropixels data could be advantageous for analysis of correlated activity or time-lagged activity between differnt brain regions. Together, the simulation and real data results suggest that information flow in neural circuits is better assessed using the GPCSD predictions than the LFP.

In Section 4.2, I provide some notation and background information about CSD estimation and existing methods. Section 4.3 develops the GPCSD method. The GPCSD method is applied to simulated data in Section 4.4, to auditory LFPs in Section 4.5, and to Neuropixels LFPs in Section 4.6. A discussion of the GPCSD method and results and future work is given in Section 4.7.

4.2 Background

In this section, I give details of the biophysical models necessary for understanding previous CSD methods and for developing the Gaussian process CSD (GPCSD) method. Section 4.2.1 gives notation and definitions and Section 4.2.2 discusses biophysical models for LFPs. In Section 4.2.3, I use the one-dimensional biophysical model to demonstrate with a simple case study how correlations between LFPs do not necessarily reflect correlations between the underlying sources, underscoring the importance of using the CSD for analyzing association between neural populations. Section 4.2.4 reviews existing methods for estimating the CSD from

LFP recordings and, in highlighting some drawbacks of each method, demonstrates the need for improved CSD estimation methodology.

4.2.1 Notation and definitions

The local field potentials (LFPs) and current source densities (CSDs) may be conceptualized as spatiotemporal functions of a spatial coordinate s in $\mathbb{R}^d$ (with $d \in \{1, 2, 3\}$) and a time coordinate t in $\mathbb{R}^1$, so I will write the LFPs as a function $\phi : \mathbb{R}^{d+1} \to \mathbb{R}$ and the CSDs as a function $c : \mathbb{R}^{d+1} \to \mathbb{R}$. I will call the portion of the activity time-locked to a stimulus the *evoked response* and all other activity the *ongoing activity.* A single *trial* captures neural activity in a fixed time epoch relative to a single stimulus presentation. The average across trials, aligned in time to the stimulus onset, provides an estimate of the evoked response from all trials and is called the *average evoked response.*

To denote a realization of a function or parameter on the nth trial, I will use the notation $f^{(n)}$. The scalar output of a function $f^{(n)} : \mathbb{R}^{d+1} \to \mathbb{R}$ at input point (s, t) will be denoted $f^{(n)}(s, t)$. If the function is evaluated at a grid of space-time points corresponding to D spatial locations and T temporal locations, represented by a matrix $\mathbf{s} \in \mathbb{R}^{D \times d}$ and a vector $\mathbf{t} \in \mathbb{R}^T$, then $f^{(n)}(\mathbf{s}, \mathbf{t})$ will be a $D \times T$ matrix of function values. Unless specified otherwise, I will use $\mathbf{s} \in \mathbb{R}^{D \times d}$ and $\mathbf{t} \in \mathbb{R}^T$ to refer to the spatial locations and time points of the observed LFPs. Often, such matrices will be vectorized; I will use $\mathbf{f}^{(n)}_{\mathbf{s},\mathbf{t}} \equiv \text{vec}\left[f^{(n)}(\mathbf{s}, \mathbf{t})\right]$ to represent this DT-vector. I will use $\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}}$ to refer to the observed LFPs to distinguish them from function values $\boldsymbol{\phi}_{\mathbf{s},\mathbf{t}}$ representing latent noiseless LFPs evaluated at the same space-time points.

When considering Gaussian process models, I will refer to a trial-specific spatiotemporal mean function $\mu^{(n)} : \mathbb{R}^{d+1} \to \mathbb{R}$ and to spatiotemporal covariance functions $k : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$. The scalar value of a covariance function at a pair of input points $\{(s, t), (s', t')\}$ is denoted $k(s, t; s', t')$. A covariance matrix is obtained by evaluating the covariance function on all pairs of multiple input points, which I will denote as $\mathbf{K}_{\mathbf{s},\mathbf{t};\mathbf{s}',\mathbf{t}'} \in \mathbb{R}^{DT \times D'T'}$ for inputs $\{\mathbf{s} \in \mathbb{R}^{D \times d}, \mathbf{t} \in \mathbb{R}^T, \mathbf{s}' \in \mathbb{R}^{D' \times d}, \mathbf{t}' \in \mathbb{R}^{T'}\}$. I will use *separable* covariance functions across time and space that take the form $k(s, t; s', t') = k^s(s, s')k^t(t, t')$; evaluating such a covariance function at multiple input points then yields a covariance matrix that decomposes as a Kronecker product,

$$\mathbf{K}_{\mathbf{s},\mathbf{t};\mathbf{s}',\mathbf{t}'} = \mathbf{K}^s_{\mathbf{s},\mathbf{s}'} \otimes \mathbf{K}^t_{\mathbf{t},\mathbf{t}'}, \tag{4.1}$$

where $\mathbf{K}^s_{\mathbf{s},\mathbf{s}'} \in \mathbb{R}^{D \times D'}$ and $\mathbf{K}^t_{\mathbf{t},\mathbf{t}'} \in \mathbb{R}^{T \times T'}$ are matrices formed by evaluating the space and time covariance functions, respectively, on all pairs of inputs. When pairs of inputs are the same, the expression $\mathbf{K}_{\mathbf{t}}$ will be used as shorthand for $\mathbf{K}_{\mathbf{t},\mathbf{t}}$.

Linear operators will be denoted $\mathcal{A}_s$ where s is the input acted upon by the operator; that is, $\mathcal{A}_s f(s) = h(s)$. When applying operators to functions with more than one input, I will use the following conventions. By $\mathcal{A}_s f(s, t)$, I mean to apply the operator to f as a function of s with the other argument, t, held fixed.

To apply a linear operator to both arguments of a two-input function, as in Särkkä (2011), I will use the following notation: $\mathcal{A}_s f(s, s')\mathcal{A}_{s'}^T$.

4.2.2 Biophysical models of LFPs

Biophysical models give the relationship between the LFPs and the underlying CSD at any instant in time. In particular, current flow across a single cell membrane creates a current source or sink and the resulting field potential can be derived using volume conductor theory. Including the contributions of all such current sources or sinks in space results in a biophysical *forward model* relating LFPs to the CSD. Because it is not possible to estimate the contribution of individual transmembrane currents from measured LFPs, the CSD may be conceptualized as a continuous function in three-dimensional space that reflects average transmembrane current in a small area (Einevoll et al., 2013). In the following discussion, I first give an overview of the three-dimensional forward model, then detail the use of additional assumptions, which I call *a priori physical models*, to adapt the forward model to one-dimensional and two-dimensional LFP measurements.

Three-dimensional biophysical models In this section, the spatial location s is three-dimensional and will be indexed by three coordinates, x, y, and z. Using the quasi-static assumption and assuming an isotropic, homogeneous medium with scalar conductivity σ, the relationship between the CSD c and the LFP ϕ is governed by the Poisson equation (Pitts, 1952):

$$\sigma\nabla \cdot (\nabla\phi(x,y,z)) = \sigma\left(\frac{\partial^2\phi(x,y,z)}{\partial x^2} + \frac{\partial^2\phi(x,y,z)}{\partial y^2} + \frac{\partial^2\phi(x,y,z)}{\partial z^2}\right) = -c(x,y,z) \tag{4.2}$$

While this appears to give a formula for computing the CSD from the LFP, it requires detailed, accurate knowledge of the LFP in three dimensions, without which it fails to accurately recover the CSD (Nicholson and Freeman, 1975). Instead, the differential equation in Equation 4.2 may be inverted to an integral equation which gives ϕ in terms of an integral operator on c. Assuming an infinite volume conductor with negligible boundary conditions leads to the following integral operator (Nicholson and Llinas, 1971):

$$\phi(x,y,z) = -\frac{1}{4\pi\sigma}\int\int\int \frac{c(x',y',z')}{\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}}\,dx'\,dy'\,dz'. \tag{4.3}$$

Equation 4.3 is often called a *forward model* since it describes how to generate the LFP from the underlying CSD. One advantage of this formulation is that when observations are only available in $d < 3$ dimensions, we must provide some prior beliefs about the behavior of the CSD in the unmeasured dimensions through an *a priori* physical model. The forward model then dictates how prior beliefs about the CSD influence the model of the measured LFPs. I will discuss common one-dimensional and two-dimensional *a priori* physical models in the following paragraphs.

One-dimensional biophysical models In this section, the spatial location s is one-dimensional and will be indexed by a single coordinate z, since it usually represents cortical depth along a linear probe inserted perpendicular to the cortex. As in previous works, I use an *a priori* physical model in which the CSD is assumed constant in the dimensions perpendicular to the linear probe on a cylinder of radius R around the probe and zero elsewhere; previous work has shown deviations from this shape do not have a large impact on the results (Nicholson, 1973; Potworowski et al., 2012). Additionally, as linear probes are typically inserted to cover all layers of cortex, it is reasonable to assume the CSD is nonzero only on an interval $a \leq z \leq b$, leading to the following *a priori* physical model that describes the variation of the CSD in the z direction through $g(z)$:

$$c(x, y, z) = g(z)\mathbb{1}(x^2 + y^2 \leq R)\mathbb{1}(a \leq z \leq b). \tag{4.4}$$

I will assume $x = y = 0$ corresponds to the probe. Under this *a priori* physical model, as shown in Appendix B.1, Equation 4.3 reduces to

$$\phi(0, 0, z) \equiv \phi(z) = \mathcal{A}_z g \equiv -\frac{R}{2\sigma}\int_a^b g(z') \underbrace{\left[\sqrt{\left(\frac{r}{R}\right)^2 + 1} - \sqrt{\left(\frac{r}{R}\right)^2}\right]}_{b(r;R),\ \text{where } r=z-z'} dz'. \tag{4.5}$$

Thus, under this *a priori* physical model, the one-dimensional LFP is the result of applying a linear operator $\mathcal{A}_z$ to g, where the weighting function $b(r; R)$ decreases with distance r but also depends on the radius R; see Figure 4.2 for an illustration of the weight function and the effect of R on the LFPs. Equation 4.5 gives a forward model relating CSDs that vary along depth z, and are constant on a cylinder of radius R in the other dimensions, to measured LFPs along the probe.

Two-dimensional biophysical models Two-dimensional LFP measurements are commonly collected using microelectrode arrays such as the Neuropixels probe (Jun et al., 2017; Steinmetz et al., 2018). In the Neuropixels data I will analyze, the two-dimensional spatial locations represent depth and width along the probe, so I will index the spatial location s by coordinates y for width and z for depth. In the case of probes like these, unlike Utah arrays, it appears that sources behind the face of the probe are unlikely to make contributions to the measured LFPs (Buccino et al., 2019). Therefore, the *a priori* physical model will assume that the CSD is constant in front of the face of the probe in the unmeasured x direction for some

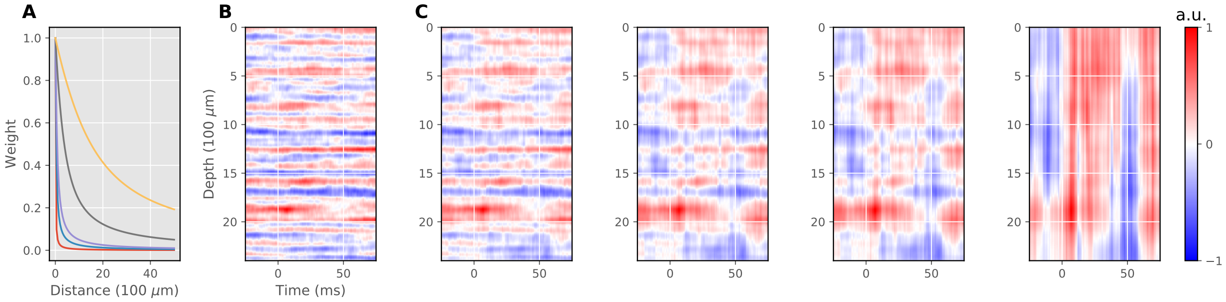

Figure 4.2: A) Plot of the one-dimensional forward model weight function $b(r; R)$ of Equation 4.5 as a function of distance r between current source and measured LFP; the different lines represent different R values, where smaller R values lead to faster decay of the weight function with distance. B) Ground truth CSD generated from a zero-mean Gaussian process with one spatial dimension (vertical axis) and one temporal dimension (horizontal axis). C) LFPs generated from the ground truth CSD with four different R values (increasing from left to right); larger R leads to increasingly spatially smooth LFPs. (LFP values in arbitrary units since R affects amplitude and smoothness.)

distance R; that is, assuming $x = 0$ is the face of the probe and positive x are in front of the probe,

$$c(x, y, z) = g(y, z)\mathbb{1}(x \leq R)\mathbb{1}(a_z \leq z \leq b_z)\mathbb{1}(a_y \leq y \leq b_y). \tag{4.6}$$

Similar to the two-dimensional kCSD method (Potworowski et al., 2012), Equation 4.3 reduces to (see Appendix B.2 for details)

$$\phi(0, y, z) \equiv \phi(y, z) = \mathcal{A}_{yz} g(y, z) \equiv -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} g(y', z') \underbrace{\operatorname{arcsinh}\left(\frac{R}{\sqrt{(y - y')^2 + (z - z')^2}}\right)}_{b(r_z, r_y; R),\ \text{where } r_z = z - z',\, r_y = y - y'} dy'\, dz'. \tag{4.7}$$

Note that the weight function b has a singularity when the denominator of the argument to arcsinh is zero; as done by previous authors, I avoid this singularity by truncating small values in the denominator to 10^{-7} (Särkkä, 2011; Potworowski et al., 2012). Equation 4.7 gives a forward model relating CSDs that vary along depth z and width y and are constant on rectangle of depth R in the other dimension to LFPs along the face of probe; the form of the forward model is actually nearly the same as two-dimensional forward models for slightly different measuring devices such as Utah arrays, but with a factor of 2 missing inside the arcsinh function (because those forward models assumed symmetric currents around the electrode tips).

4.2.3 Distortion of correlation in LFPs: one-dimensional case study

As argued in Section 4.1, the CSD should be preferable to the measured LFPs for assessing correlation between neural populations because it localizes the activity. To demonstrate that the LFPs may not correctly reflect the correlation of individual sources, I consider here a simplified one-dimensional situation and show that correlation between the LFPs at two spatial locations can be either attenuated or inflated compared to the underlying correlation of the sources at those locations. I consider a simple situation with d point sources evenly spaced at locations $x_1, ..., x_d$ and where the CSD values $c(x)$ at all but two of these locations are uncorrelated and iid with mean zero and variance 1. At two locations x_k and x_ℓ with $k, \ell \in \{1, ..., d\}$, the CSD values are still unit variance but are also correlated with nonzero correlation coefficient ρ. In summary,

$$\mathrm{Corr}(c(x_i), c(x_j)) = \begin{cases} \rho, & i = k, j = \ell \text{ or } j = k, i = \ell \\ 1, & i = j \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

The LFPs at the locations $x_1, ..., x_d$ are then generated using the forward model of Equation 4.5; because there are a discrete number of point sources, the integral becomes a sum, so that

$$\phi(x_j) = -\frac{R}{2} \sum_{i=1}^{d} c(x_i) b(x_i - x_j; R)$$

where I have assumed without loss of generality that the conductivity constant $\sigma = 1$. Then the correlation between the LFPs $\phi(x_k)$ and $\phi(x_\ell)$ is

$$\begin{aligned} \mathrm{Corr}(\phi(x_k), \phi(x_\ell)) &= \frac{\sum_{i=1}^{d} \sum_{j=1}^{d} b(x_i - x_k) b(x_j - x_\ell) \mathrm{Corr}(c(x_i), c(x_j))}{\sqrt{\mathrm{Var}(\phi(x_k)) \mathrm{Var}(\phi(x_\ell))}} \\ &= \frac{\rho[b(x_k - x_\ell)^2 + 1] + 2 \sum_{i=1}^{d} b(x_i - x_k) b(x_i - x_\ell)}{\sqrt{\mathrm{Var}(\phi(x_k)) \mathrm{Var}(\phi(x_\ell))}}. \end{aligned} \tag{4.9}$$

The variances in the denominator may be expressed

$$\mathrm{Var}(\phi(x_k)) = \sum_{i=1}^{d} \sum_{j=1}^{d} b(x_i - x_k) b(x_j - x_k) \mathrm{Corr}(c(x_i), c(x_j)) = 2\rho b(x_k - x_\ell) + 2 \sum_{i=1}^{d} b(x_i - x_k)^2$$

and

$$\mathrm{Var}(\phi(x_\ell)) = 2\rho b(x_k - x_\ell) + 2 \sum_{i=1}^{d} b(x_i - x_\ell)^2.$$

Therefore, the correlation of the LFPs at locations x_k and x_ℓ is a function of ρ, the correlation between $c(x_k)$ and $c(x_\ell)$, but is also a function of the other point sources through the forward model weight function b. Equation 4.9 shows that, in general, the LFP correlation is not equal to the CSD correlation, but it is difficult to assess directly as it depends implicitly on the forward model parameter R and the distances between x_k, x_ℓ, and the other source locations.

Similarly, if we query the LFP correlation at other locations x_m and x_n for which the CSD values are uncorrelated, we find

$$\mathrm{Corr}(\phi(x_m), \phi(x_n)) = \frac{\rho[b(x_k - x_m)b(x_\ell - x_n) + b(x_\ell - x_m)b(x_k - x_n)] + 2\sum_{i=1}^{d} b(x_i - x_m)b(x_i - x_n)}{\sqrt{\mathrm{Var}(\phi(x_m))\mathrm{Var}(\phi(x_n))}}$$

where

$$\mathrm{Var}(\phi(x_m)) = 2\rho b(x_k - x_m)b(x_\ell - x_m) + 2\sum_{i=1}^{d} b(x_i - x_m)^2,$$

$$\mathrm{Var}(\phi(x_n)) = 2\rho b(x_k - x_n)b(x_\ell - x_n) + 2\sum_{i=1}^{d} b(x_i - x_n)^2.$$

Therefore, in general, even though the current sources at x_m and x_n have zero correlation, the LFPs will exhibit nonzero correlation influenced by ρ and the forward model.

To give some intuition on these expressions, I consider 500 discrete sources spaced evenly on the interval $[0, 10]$ with the correlation between $c(x_{100})$ and $c(x_{400})$ equal to 0.5 and forward model parameter R varying between 1^{-20} and 10; for each setting of R, I calculated the correlation between $\phi(x_{100})$ and $\phi(x_{400})$ to see how it compared to the known ground truth correlation of $c(x_{100})$ and $c(x_{400})$. I also computed the correlation of the LFPs for another pair of locations, x_{200} and x_{300}, where $c(x_{200})$ and $c(x_{300})$ are uncorrelated. Figure 4.3 shows the resulting LFP correlations in red dashed lines as a function of R, with the CSD correlations in solid blue. The left plot is for the pair $\phi(x_{100})$ and $\phi(x_{400})$, where the correlation of the underlying sources is 0.5. For $R \approx 0$, the LFPs reflect the correct correlation as the LFP is not spatially blurred, but as R increases, the correlation is attenuated as nearby uncorrelated sources begin to influence the LFP. As R continues to increase, the correlation actually becomes inflated because R is large enough for both of the correlated sources to influence the LFPs at both locations. The right plot is for the pair $\phi(x_{200})$ and $\phi(x_{300})$, where the correlation of the underlying sources is 0.0 but the correlation of the LFPs is greater than zero for $R > 0$ and increases monotonically with R, approaching a correlation of 1 as R approaches 10. This happens because with increasing R, the correlated sources influence the LFP at locations x_{200} and x_{300}, leading us to infer correlation even though the sources at those locations were not correlated.

This case study demonstrates the problematic interpretation of correlation between LFP locations; in this scenario, two correlated sources may have the correlation of the LFPs at their locations either attenuated or amplified due to the influence of the other uncorrelated sources, meaning that the LFP correlation is unlikely to correctly reflect the association between the underlying sources. Perhaps more problematic is that two locations for which the sources are independent can exhibit arbitrarily large correlations between the LFPs due to the influence of other correlated sources. Of course, in more complex situations, such as those with additional pairs of correlated sources or variation in correlation over time, the ability to assess

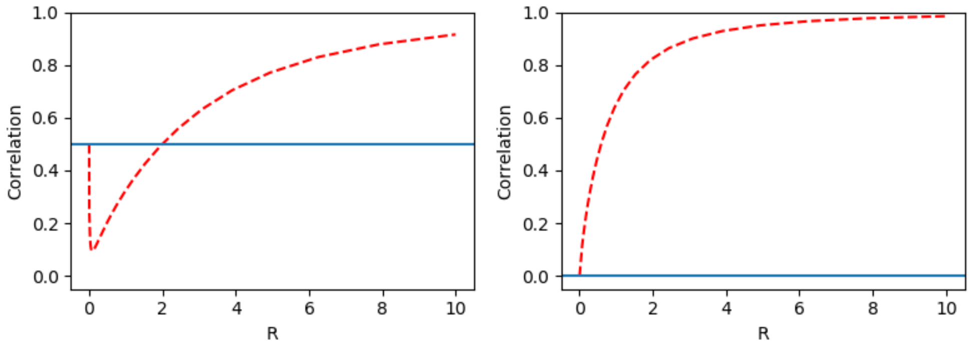

Figure 4.3: Left: for a pair of CSD sources with correlation 0.5 (blue solid line) surrounded by uncorrelated sources, the LFPs at the same locations as the correlated sources have correlations either attenuated or inflated, depending on the value of R (red dashed line). At $R \approx 0$, the LFP is actually representative of the underlying sources, so the correlation is correct, but as R increases, the correlation is first attenuated due to the influence on the LFPs of nearby uncorrelated sources, and then increases as both LFPs begin to be influenced simultaneously by the pair of correlated sources and by common influences from other sources. Right: similar to the left plot, but for a pair of CSD sources with correlation 0.0, where the LFPs have nonzero correlation for all $R > 0$; the correlation continues to increase as R increases because the LFPs are influenced by the pair of correlated sources and by common influences from other sources.

source correlations from the LFPs is unlikely to increase. Only by inferring the latent CSD can we hope to correctly identify patterns of association between distinct current sources representing the coordinated aggregate activity of neural populations.

4.2.4 Existing CSD estimation methods

As discussed in Section 4.2.2, in principle, Equation 4.2 may be used to infer the CSD if the LFPs are observed densely in three-dimensional space; in practice, this is generally not the case, and even if such recordings were available, numerical second derivatives are sensitive to noise. Nevertheless, the *traditional CSD* (tCSD) method is based on a direct application of Equation 4.2 to one-dimensional recordings under the assumption that neural activity is constant in the directions perpendicular to the probe (so that those elements of the second derivative are zero). In contrast, a second class of techniques, which I will call *inverse CSD* methods, uses Equation 4.3 along with *a priori* physical models to estimate the CSD by inverting the resulting linear operator (e.g., inverting $\mathcal{A}_z$ in Equation 4.5 or $\mathcal{A}_{yz}$ in Equation 4.7).

Traditional CSD (tCSD) method The traditional CSD method is typically applied to one-dimensional LFPs as follows. Ignoring the x, y directions in Equation 4.2 yields

$$\sigma \left(\frac{\partial^2 \phi}{\partial z^2} \right) = -c(z). \tag{4.10}$$

This suggests estimating the CSD by the second spatial derivative of the recorded LFPs, assuming equal spacing Δz between electrodes (Nicholson and Llinas, 1971):

$$\hat{c}(z_i) = \frac{\tilde{\phi}(z_{i+1}) - 2\tilde{\phi}(z_i) + \tilde{\phi}(z_{i-1})}{(\Delta z)^2}, \; \forall\, i \,|\, i \in \{2, ..., D-1\}. \tag{4.11}$$

However, implicit in this interpretation of Equation 4.2 is the assumption that the LFPs have zero curvature in the x and y directions, which has been shown to correspond to assuming $R \to \infty$ in the *a priori* physical model of Equation 4.4 (Pettersen et al., 2006). As a result, tCSD has been shown to work poorly when the CSD is actually confined in a cylinder of radius R around the recording probe (Nicholson and Freeman, 1975; Einevoll et al., 2013). A similar approach could be used for two-dimensional data, though in Neuropixels data it seems more intuitive to apply the one-dimensional tCSD estimator on each column of electrodes.

The utility of tCSD on single-trial recordings is questionable, as tCSD does not account for measurement noise, and the numerical calculation of second derivatives is sensitive to noise. Perhaps for this reason, tCSD is typically applied to many trials, then averaged across trials and further smoothed for visualization. In other words, tCSD is usually used only to address average evoked responses, not ongoing activity. While spatial and/or temporal smoothing may be used before or after tCSD, it is unclear how to choose appropriate smoothing parameters; I have not seen it done, but a cross-validation approach comparing the observed LFP to the reconstructed LFP (calculated by passing the estimated CSD through the forward model) could be used to select among a grid of spatial and temporal smoothing parameters. Furthermore, tCSD can only provide estimates of the CSD at the same locations where the LFP is measured (excluding the edge electrodes), and tCSD must be applied separately on each trial and at each time point.

Inverse CSD methods While ideally Equation 4.3 could be inverted analytically (resulting in Equation 4.2), once an *a priori* physical model is incorporated to describe variation in unmeasured dimensions, this is no longer possible. Additionally, viewing this problem through the lens of inverse theory suggests that it is an ill-posed inverse problem, meaning inverse solutions are highly sensitive to noise and may not be unique (Kropf and Shmuel, 2016), suggesting some kind of regularization is necessary.

For the one-dimensional case with the *a priori* physical model of Equation 4.4, Potworowski et al. (2012) developed *kernel CSD* (kCSD) which models the CSD $g(z)$ as a sum of finitely many spatial basis functions (and includes so-called iCSD by Pettersen et al. (2006) as a special case); in Potworowski et al. (2012), kCSD was also extended for two-dimensional recordings taken from devices similar to Utah arrays, but I will discuss the one-dimensional version here for simplicity. In kCSD, the CSD is modeled using M known basis functions $\tilde{b}_j(z)$:

$$g(z) = \sum_{j=1}^{M} a_j \tilde{b}_j(z). \tag{4.12}$$

Then, applying the forward model, one may obtain basis functions for the LFPs, denoted $b_j(z)$. Then minimum-norm inverse solution (see Appendix B.3), in terms of kernel functions $k(z, z') \equiv \sum_{j=1}^{M} b_j(z) b_j(z')$ and $\tilde{k}(z, z') \equiv \sum_{j=1}^{M} \tilde{b}_j(z) b_j(z')$, is

$$\hat{\mathbf{c}}_{\mathbf{z}'} = \tilde{\mathbf{K}}_{\mathbf{z}',\mathbf{z}} \mathbf{K}_{\mathbf{z},\mathbf{z}}^{-1} \tilde{\boldsymbol{\phi}}_{\mathbf{z}}, \tag{4.13}$$

where, unlike tCSD, kCSD can provide predictions at new spatial locations $\mathbf{z}'$. Additionally, the inverse solution may be regularized by replacing $\mathbf{K}_{\mathbf{z},\mathbf{z}}^{-1}$ with $[\mathbf{K}_{\mathbf{z},\mathbf{z}} + \lambda \mathbf{I}]^{-1}$ (necessary in cases where $\mathbf{K}_{\mathbf{z},\mathbf{z}}$ is not invertible).

For one-dimensional LFPs, inverse CSD methods are preferable to tCSD because the forward model parameter R is not implicitly assumed to be infinite and because regularization reduces sensitivity to noise and ensures a unique inverse solution. However, like tCSD, kCSD must be applied separately to each trial and each time point, precluding sharing of information across adjacent timepoints or across trials. In addition, kCSD requires choosing the functional form of the basis functions in addition to the placement and number of basis functions, for which Potworowski et al. (2012) provide only heuristics. While cross-validation is suggested to choose the regularization parameter λ and the width of the basis functions, no suggestion is made for selecting R, and including it in the cross-validation procedure would increase the computational burden considerably.

4.3 Gaussian process CSD (GPCSD) method

In this section, I develop my contribution to CSD estimation, which is based on modeling the latent CSD as a Gaussian process. The Gaussian process CSD (GPCSD) model starts with an *a priori* hypothesis that the underlying CSD on any given trial is distributed according to a spatiotemporal Gaussian process; as I will demonstrate in Section 4.3.1, this also implies that the LFP follows another spatiotemporal Gaussian process in which the forward model is incorporated into the mean and covariance functions. Spatiotemporal Gaussian process models in this context have several advantages over previous CSD estimation methods, including the ability to handle LFP data at irregular spatial locations (and to predict both CSD and LFP at arbitrary space-time locations), robustness to noise through an explicit noise model on the LFP, principled hyperparameter tuning, and pooling of information across time and trials. As I discuss in Section 4.3.2, different prior beliefs about the mean of the CSD across trials, the smoothness and variation of the CSD in space and time, and the possible existence of different spatial or temporal components can all be incorporated through choices for the Gaussian process mean and covariance functions. Section 4.3.3 describes tuning of hyperparameters corresponding to the mean and covariance functions. Finally, the CSD can be predicted from the LFP using the tuned Gaussian process model, as detailed in Section 4.3.4.

4.3.1 Gaussian process models of CSDs and LFPs

Under the *a priori* physical models of Equations 4.4 and 4.6, the time-varying CSD on trial n can be described as a spatiotemporal function $g^{(n)}(s,t)$. In the data sets I will explore, s could be one- or two-dimensional. In general, I will model g as the sum of a mean function μ representing the evoked response and a mean-zero Gaussian process η representing the ongoing activity (though, as is typical in applications of Gaussian process regression, one may assume the mean function is zero, potentially after subtracting out the average evoked response). That is, $g^{(n)}(s,t)$ will be decomposed as

$$g^{(n)}(s,t) = \mu^{(n)}(s,t) + \eta^{(n)}(s,t), \qquad \eta^{(n)} \sim \mathrm{GP}\left(0, k(s,t;s',t')\right) \tag{4.14}$$

where the Gaussian process η is assumed iid across trials and models for the mean and covariance functions μ and k will be discussed in Section 4.3.2. I will now show how combining this model for the CSD with Equation 4.5 or Equation 4.7 and putting iid additive noise on the observed LFPs then yields a joint Gaussian process model for the CSD and LFPs.

For simplicity, in the following equations, I consider a spatial process $g(s)$, suppressing the time and trial indices; when I discuss choices for spatiotemporal covariance functions in Section 4.3.2, I will demonstrate how the forward operator applies to a spatiotemporal process. Applying the linear operator $\mathcal{A}_s$ to g results in another Gaussian process (Särkkä, 2011); that is, if

$$\mathcal{A}_s g(s) = \phi(s),$$

then $\phi(s)$ is jointly, with $g(s)$, a Gaussian process. The marginal mean and covariance functions for $\phi(s)$ are influenced by the linear operator:

$$\phi(s) \sim GP(\mathcal{A}_s \mu(s), \mathcal{A}_s k^s(s,s') \mathcal{A}_s^T).$$

If we then suppose that $\tilde{\phi} = \phi + \epsilon$ where ϵ has variance σ^2, we obtain the joint distribution of vectors $\mathbf{g_s}$ and $\tilde{\boldsymbol{\phi}}_\mathbf{s}$ (for simplicity, at the same $\mathbf{x}$, though this is not required) as:

$$\begin{bmatrix} \mathbf{g_s} \\ \tilde{\boldsymbol{\phi}}_\mathbf{s} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\mathbf{s} \\ \mathcal{A}_s \boldsymbol{\mu}_\mathbf{s} \end{bmatrix}, \begin{bmatrix} \mathbf{K_{s,s}} & \mathcal{A}_s \mathbf{K_{s,s}} \\ \mathbf{K_{s,s}} \mathcal{A}_s^T & \mathcal{A}_s \mathbf{K_{s,s}} \mathcal{A}_s^T + \sigma^2 I \end{bmatrix} \right). \tag{4.15}$$

So, if we have observations of the process $\tilde{\phi}(s)$ but want to predict values of $g(s)$, it is natural to propose mean and covariance functions μ and k for g; application of the linear operator induces mean and covariance functions for the observed process that are influenced by $\mathcal{A}_s$ (and recall that $\mathcal{A}_s$ involves integration over an interval, and as the integral is generally not available in closed form, I use Gauss-Legendre quadrature.)

4.3.2 Specification of mean and covariance functions

I consider stationary covariance functions under the assumption that ongoing neural activity within a single trial can be treated as approximately stationary in time and space, though nonstationary covariance functions could certainly be used. Because we expect the CSD to reflect the temporal evolution of neural activity in spatially fixed neural populations, it appears reasonable to model the covariance as separable in space and time:

$$k(s, t; s', t') = k^s(s, s') k^t(t, t'). \tag{4.16}$$

This allows simpler specification of the spatiotemporal covariance function and also yields considerable computational advantages as the full covariance matrix becomes a Kronecker product between spatial and temporal covariance matrices. If there is no additional additive noise on the Kronecker product covariance, the inverse may be easily computed in $O(T^3 + D^3)$ time using the property that $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (Schacke, 2004). However, if additive noise is included, other properties of Kronecker products must be used to obtain some computational speed-ups compared to naively inverting the matrix (see Section B.5 for details). In general, the spatial and temporal stationary covariance functions may be sums of multiple covariance functions, each with its own functional form and hyperparameters, allowing processes with multiple temporal or spatial scales. In addition, predictions can be made at each time or spatial scale separately (Duvenaud, 2014).

To apply the forward model to the Gaussian process $\eta^{(n)}$, note that the integral operator is linear and applies only to the spatial part of the covariance function, so that

$$\mathcal{A}_s \mathbf{K}_{\mathbf{s},\mathbf{t};\mathbf{s}',\mathbf{t}'} \mathcal{A}_{s'}^T \equiv \mathcal{A}_s \mathbf{K}^s_{\mathbf{s},\mathbf{s}'} \mathcal{A}_{s'}^T \otimes \mathbf{K}^t_{\mathbf{t},\mathbf{t}'}. \tag{4.17}$$

In some of the simulations and application to real data, I use a zero-mean Gaussian process, but in analyzing the one-dimensional auditory LFPs, I use a mean function composed of a sum of Gaussian bump functions (described in detail in Section 4.5.2). In all simulations and data analysis, I use the same covariance function structure: the spatial covariance function is a unit variance squared exponential, or SE, with either one or two lengthscales, depending on the spatial dimension $D \in \{1, 2\}$:

$$k^s_{SE}(s, s') = \exp\left(-\sum_{d=1}^{D} \frac{(s_d - s'_d)^2}{2\ell_d^2}\right).$$

Note that the spatial covariance function has unit variance to avoid identifiability issues with the temporal variances. The temporal covariance function is a sum of an SE covariance function meant to capture slower, smoother variations, and an exponential covariance function meant to capture rougher, faster variations:

$$k^t(t, t') = \sigma^2_{SE} \exp\left(-\frac{(t - t')^2}{2\ell^2_{SE}}\right) + \sigma^2_E \exp\left(-\frac{|t - t'|}{\ell_E}\right).$$

Therefore, the overall temporal covariance function has a marginal variance for each process and a lengthscale for each process, and separate predictions can be made for each process which then sum to give the entire process. Other common choices for covariance functions and properties of different covariance functions are discussed in detail in Rasmussen and Williams 2006, Chapter 4.

4.3.3 Hyperparameter tuning

Using the marginal likelihood function (that is, Equation 4.15 marginalized over latent values g) is an attractive way to tune hyperparameters because, as discussed in Rasmussen and Williams 2006, Ch. 5.2, it incorporates a trade-off between model complexity and fit and can be interpreted as penalizing the complexity of the underlying function. In particular, suppose we have noisy observations $\tilde{\phi}$ at spatial locations $\mathbf{s}$ and temporal locations $\mathbf{t}$. Equation 4.15 shows that the joint distribution of $\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}}$ and $\mathbf{g}_{\mathbf{s}',\mathbf{t}'}$, for arbitrary $\mathbf{s}'$ and $\mathbf{t}'$, is still multivariate Gaussian. Then using properties of multivariate Gaussian distributions, the marginal density of the observations $\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}}$,

$$p(\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}}) = \int p(\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}} | \mathbf{g}_{\mathbf{s}',\mathbf{t}'}) p(\mathbf{g}_{\mathbf{s}',\mathbf{t}'})\, d\mathbf{g}_{\mathbf{s}',\mathbf{t}'}, \tag{4.18}$$

is available in closed form:

$$\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}} \sim \mathcal{N}(\mathcal{A}_s \boldsymbol{\mu}_{\mathbf{s},\mathbf{t}},\, \mathcal{A}_s \mathbf{K}^s_{\mathbf{s}} \mathcal{A}_s^T \otimes \mathbf{K}^t_{\mathbf{t}} + \sigma^2 \mathbf{I}). \tag{4.19}$$

The resulting marginal likelihood can be maximized with respect to hyperparameters of the mean and covariance functions. To avoid potential identifiability issues between the mean and ongoing activity (van den Boogaart and Brenning, 2001), I will use a two-stage procedure when fitting a nonzero mean function: I will first estimate the covariance parameters on pre-stimulus data, then, with the covariance parameters fixed, estimate the mean parameters on the portion of the trial containing the evoked response.

The procedure can also be modified to include prior information on the hyperparameters, important when there are potentially difficult to identify parameters or physical constraints on reasonable values for the hyperparameters (Zhang, 2004; Betancourt, 2017). For instance, lengthscales that are larger than the range of observed data or smaller than the minimum distance between observed data points are not identifiable from the data and may lead to problems during likelihood optimization or may result in overfitting, so we would want to discourage too-small or too-large lengthscales. Similarly, we likely have some intuition on reasonable values for the forward model parameter R, which influences both the spatial smoothness and variance of the observed LFPs and therefore may exhibit weak identifiability issues with the spatial lengthscale and marginal variances. When using multiple timescales, prior information can also be used to encourage different components toward longer or shorter lengthscales.

If we denote the collection of hyperparameters as $\boldsymbol{\theta}$, we can specify a prior on $\boldsymbol{\theta}$ to incorporate prior knowledge of the reasonable ranges of the parameters, then combine the prior with the marginal likelihood to obtain a posterior distribution over the hyperparameters given the observed data:

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{\phi}}_{\mathbf{s}}) \propto p(\tilde{\boldsymbol{\phi}}_{\mathbf{s}})p(\boldsymbol{\theta}). \tag{4.20}$$

While a fully Bayesian approach such as MCMC could be used to obtain a distribution over the hyperparameters, for computational reasons, I will instead use maximum *a posteriori* (MAP) estimation by maximizing the logarithm of Equation 4.20 with respect to $\boldsymbol{\theta}$; in this case, the use of priors can be seen as a form of regularization on the log marginal likelihood that discourages unrealistic or unidentifiable hyperparameter values. In all simulations and data analysis, lengthscales had inverse Gamma priors in which the parameters were chosen so that the 1% and 99% quantiles would fall at specific values, typically corresponding to the minimum and maximum observed distances in space or time (Betancourt, 2017). The forward model parameter R also had an inverse Gamma prior with quantiles chosen to be uninformative but to encourage what seemed to be reasonable values, and marginal variances and the noise variance had uninformative half-Normal priors.

4.3.4 Prediction of the CSD using the Gaussian process

Given hyperparameter values, predictions conditional on the observed values $\tilde{\phi}$ can be made for any $\mathbf{s}', \mathbf{t}'$ for either ϕ or g; typically, we will be mostly interested in predicting the latent CSD described by g. Using properties of multivariate Gaussians, we have the following conditional distribution:

$$\mathbf{g}_{\mathbf{s}',\mathbf{t}'} \Big| \tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}} \sim \mathcal{N}(\boldsymbol{\mu}^*, \mathbf{K}^*)$$

where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_{\mathbf{s}',\mathbf{t}'} + \left(\mathcal{A}_s \mathbf{K}^s_{\mathbf{s}',\mathbf{s}} \otimes \mathbf{K}^t_{\mathbf{t}',\mathbf{t}}\right) [\mathcal{A}_s \mathbf{K}^s_{\mathbf{s}} \mathcal{A}_s^T \otimes \mathbf{K}^t_{\mathbf{t}} + \sigma^2 \mathbf{I}]^{-1} (\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}} - \mathcal{A}_s \boldsymbol{\mu}_{\mathbf{s},\mathbf{t}}) \tag{4.21}$$

$$\mathbf{K}^* = \mathbf{K}^s_{\mathbf{s}'} \otimes \mathbf{K}^t_{\mathbf{t}'} - \left(\mathcal{A}_s \mathbf{K}^s_{\mathbf{s}',\mathbf{s}} \otimes \mathbf{K}^t_{\mathbf{t}',\mathbf{t}}\right) [\mathcal{A}_s \mathbf{K}^s_{\mathbf{s}} \mathcal{A}_s^T \otimes \mathbf{K}^t_{\mathbf{t}} + \sigma^2 \mathbf{I}]^{-1} \left(\mathcal{A}_s \mathbf{K}^s_{\mathbf{s}',\mathbf{s}} \otimes \mathbf{K}^t_{\mathbf{t}',\mathbf{t}}\right)^T \tag{4.22}$$

so that point predictions for $\mathbf{g}_{\mathbf{s}',\mathbf{t}'}$ given $\tilde{\boldsymbol{\phi}}_{\mathbf{s},\mathbf{t}}$ are naturally given by the conditional mean $\boldsymbol{\mu}^*$; the diagonal of $\mathbf{K}^*$ may also be used as a measure of uncertainty around the point prediction. When comparing CSD methods on simulated and real data, I will use GPCSD to refer to the process of using this conditional mean to predict the CSD at some set of spatial locations (typically, after selecting hyperparameters using the regularized log marginal likelihood, i.e., the posterior mode).

4.4 Application of GPCSD to simulated data

In this section, I use simulated data both to validate the GPCSD method and to compare it against existing CSD methods (traditional CSD and kernel CSD). The general idea for the simulations was to generate known ground-truth CSDs, then apply the forward model to obtain LFPs; as in real data, the LFPs were assumed to be observed at fewer spatial locations than the underlying CSDs. Then the goal of any CSD prediction method is to predict the ground-truth CSD based only on the LFPs, so the performance of each method was assessed by comparing the predicted CSD to the ground truth CSD. In Section 4.4.1, I describe the simulation methods and in Section 4.4.2, I describe the results.

4.4.1 Simulation methods

The first simulation aimed to demonstrate the ability of the GPCSD method to recover realistic CSD patterns even in the presence of noise, and to show qualitative differences between the GPCSD method and the existing CSD methods (tCSD and kCSD). I used a simple CSD template spanning a spatial dimension with minimum value -2 and maximum value 26 and a temporal dimension with 50 integer-valued time points. The CSD template was made up of two positive-valued unit magnitude Gaussian-shaped bumps (means: 2 and 16 in spatial dimension, 25 and 30 in temporal dimension; SDs: 1.5 in spatial dimension, 3 and 4 in temporal dimension) and two negative-valued unit magnitude Gaussian-shaped bumps (means: 8 and 22 in spatial dimension, 25 and 30 in temporal dimension; SDs: 1.5 in spatial dimension, 3 and 4 in temporal dimension).

The CSD template, evaluated at 50 time points and 100 spatial locations, was passed through the one-dimensional forward model with parameter $R = 2$ (using the trapezoid rule to compute the integral) to generate a noiseless LFP observed at 24 spatial locations between 0 and 24 across the 50 time points; noisy LFPs were also generated by adding white noise (variance 0.001). Then, GPCSD, kCSD, and tCSD were applied to the generated LFPs. The GPCSD method was applied with the mean function assumed to be zero. GPCSD hyperparameters, including the forward model parameter R, were selected by maximization of the marginal likelihood with priors (as described in Section 4.3.3). In particular, the spatial SE lengthscale prior was inverse Gamma with parameters selected so that the 1% and 99% quantiles would fall at the minimum distance between LFP observations and the maximum distance between LFP observations. The temporal covariance was exponential with inverse Gamma prior such that the quantiles were 0.1 and 30. The prior for R was inverse Gamma such that the quantiles were 0.1 and 4.0. The marginal variance and the noise variance were given noninformative half-Normal priors with standard deviations 2 and 0.5, respectively. For kCSD, the forward model parameter R was set to the ground truth value, and the other tuning parameters (basis width and noise variance) were chosen by cross-validation over a two-dimensional grid of values using the Python toolbox `elephant`, which includes an implementation of both one- and two-dimensional kCSD as described in Potworowski et al. (2012). The grid of basis function widths consisted of ten values between 0.1 and 10 and the grid of noise variances consisted of ten values logarithmically spaced between 10^{-25} and 10^{-2}. Traditional CSD was applied directly with no smoothing beforehand. Different CSD methods may only recover the CSD pattern up to some multiplicative constant, so to compare the results of different methods to the true CSD, the true CSD and the CSD predictions were both rescaled to have standard deviations equal to 1.

Second, I simulated a detailed spatiotemporal CSD pattern from a Gaussian process, then used different R values in the forward model to generate high-resolution noiseless LFPs, with the goal of illustrating that tCSD cannot accurately recover the CSD if the true R is small even when high-resolution noiseless LFPs are available. The CSDs were simulated from a Gaussian process model with hyperparameters set to the values fitted to real auditory LFPs and with spatial and temporal range the same as the auditory LFPs. Then, simulated LFPs with high spatial resolution (300 points) were generated using the one-dimensional forward model of Equation 4.5 but with varying $R \in \{0.1, 0.5, 1.0, 5.0, 20.0\}$. The traditional CSD method was applied to each LFP, and GPCSD was also applied using the ground truth hyperparameters to demonstrate that the ability of GPCSD to use different R values gives it an advantage over tCSD.

Finally, to quantify the performance of GPCSD relative to other methods, I generated multiple realizations of spatially one-dimensional and two-dimensional CSDs from spatiotemporal Gaussian process models, then passed these CSDs through the forward model to obtain LFPs. For the one-dimensional simulation, the spatial and temporal coordinates for the CSD and LFP were the same as the first simulation. For the generative process, I used an SE spatial covariance with lengthscale 2 and a temporal covariance made up of a sum of an SE with lengthscale 20 and variance 0.5 and an exponential with lengthscale 5 and variance 0.5. The LFP white noise variance was 0.0001 and the forward model parameter was $R = 0.5$. The generated LFP trials were split into training and testing sets, each of size 50; the training set was used for selecting GPCSD and kCSD hyperparameters and the test set was used to evaluate the performance of each method in reconstructing the true CSD (tCSD has no tuning parameters, so it was simply applied to the test set). In particular, kCSD used cross-validation in the same manner as the first simulation (with R set to the ground truth value), and because the kCSD method applies independently to each time step and does not explicitly make use of multiple trials as independent realizations, trials were concatenated before performing cross-validation.

For the one-dimensional simulation, GPCSD used MAP estimation where the spatial lengthscale prior was inverse Gamma with parameters selected so that the 1% and 99% quantiles would fall at the minimum distance between observations and the maximum distance between observations. The temporal SE component had inverse Gamma prior such that the quantiles were 10 and 50, and the temporal exponential component had inverse Gamma prior such that the quantiles were 1 and 30. The prior for R was inverse Gamma such that the quantiles were 0.1 and 3.0. The two marginal variances and the noise variance were given noninformative half-Normal priors with standard deviations 2, 2, and 0.5, respectively. Because tCSD can only estimate the CSD at the interior electrode positions, kCSD and GPCSD were also used to estimate the CSD at the interior electrode locations so that the mean squared error (MSE) across space and time would be based on the same number of points for all methods. For each test set trial, I computed the MSE across all space-time points (after rescaling the predicted CSDs and the true CSDs to be on the same scales), then used the distribution of MSEs across the test set as a measure of performance. To summarize the performance of tCSD and kCSD relative to GPCSD, paired differences in per-trial MSEs between GPCSD and the other methods were computed. The null hypothesis that the mean MSEs across trials were the same for each pair of methods (GPCSD and tCSD, GPCSD and kCSD) were tested using paired t-tests.

For the two-dimensional simulation, the generative process spanned spatial coordinates $y \in [0, 48]$, $z \in [1900, 2500]$ which are similar to the real Neuropixel data coordinates (in microns). The generated CSDs were calculated at 15 by 100 spatial points and 20 time points, with 50 training and 50 test trials. The generative parameters were $R = 75$, noise variance 0.0001, spatial SE lengthscales 25 and 100, temporal SE lengthscale 15, temporal exponential lengthscale 1, and marginal variances both equal to 4. For GPCSD, due to optimization difficulties, I simply used the prior mean hyperparameters to demonstrate the method; the priors for the marginal variance and noise variance were the same as the one-dimensional case, the priors for spatial lengthscales were similar to the one-dimensional case but adjusted for the new spatial coordinates, the prior on R was inverse Gamma with quantiles at 50 and 100, the prior on the temporal SE lengthscale was inverse Gamma with quantiles at 10, 20, and the prior on the temporal exponential lengthscale was

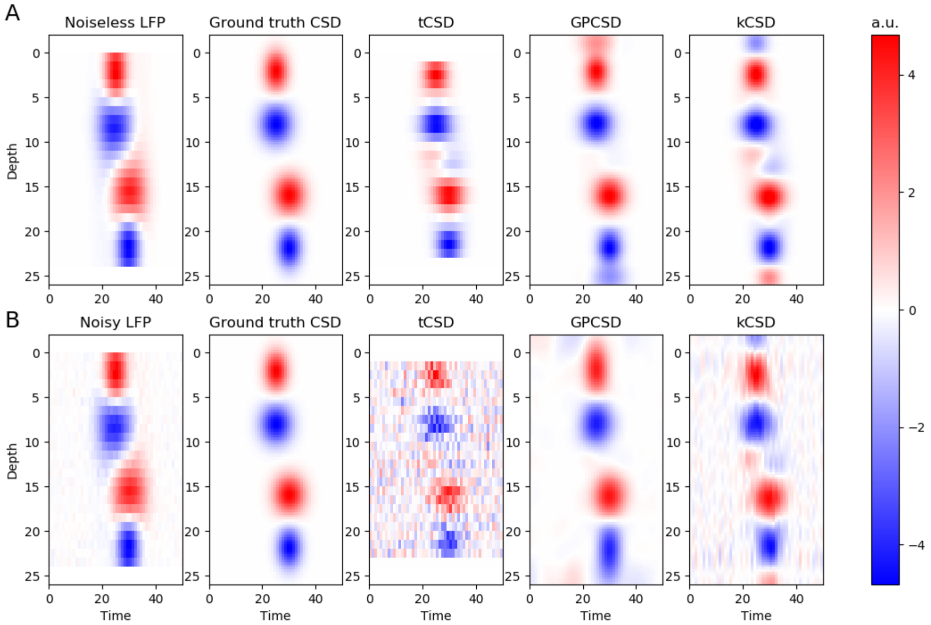

Figure 4.4: Row A: from left to right, the generated noiseless LFP, the underlying true CSD pattern, the tCSD prediction, the GPCSD prediction, and the kCSD prediction. Row B: same as row A, but with additive noise on the LFP. In both cases, GPCSD appears to do better than tCSD and kCSD in recovering the true CSD pattern with minimal artifacts, and GPCSD appears more robust to noise than the other methods.

inverse Gamma with quantiles at 0.5 and 10. The per-trial MSEs were then compared between kCSD and GPCSD.

4.4.2 Simulation results

For the first simulation with a simple one-dimensional CSD template, the true CSD, true LFP, and CSD predictions for tCSD, GPCSD, and kCSD are shown in Figure 4.4, with predictions based on the noiseless LFP in row A and based on the noisy LFP in row B. In both cases, GPCSD appears to reconstruct the ground truth pattern well, though there are some small-amplitude artifacts that likely arise partly because I used stationary covariance functions in both space and time while the actual CSD was nonstationary. However, such artifacts appear much more severe in both tCSD and kCSD, where they are of much larger amplitude relative to the signal, even in the noiseless case. In addition, the performance of both tCSD and kCSD clearly degrades when white noise is added to the LFP, while the performance of GPCSD remains similar to the noiseless case.

In the second simulation with LFPs generated from a Gaussian process CSD using different values of R, we find that even if tCSD was applied to noiseless, high spatial resolution LFPs, it cannot recover the

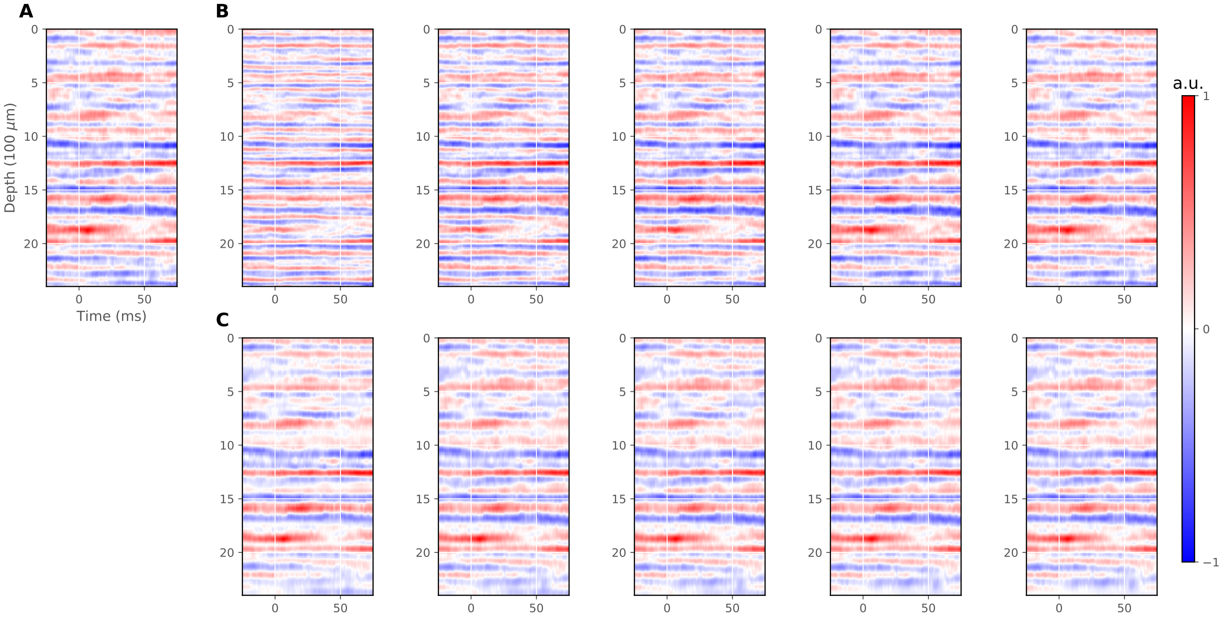

Figure 4.5: A) Ground truth CSD generated from a zero-mean Gaussian process. B) Across: tCSD estimates based on noiseless, high spatial resolution LFPs generated with different R (increasing from left to right). For small R, the tCSD estimates do not match the ground truth pattern, while for larger R, they do, as tCSD implicitly assumes $R \to \infty$. C) Across: GPCSD estimates for the same LFPs, showing that when the correct R is used as part of the GPCSD model, the ground truth CSD pattern can be recovered from each LFP. (All values in arbitrary units for comparison between tCSD and GPCSD.)

correct ground truth CSD pattern if the true R is actually small; this is because the tCSD method implicitly assumes $R \to \infty$ (Section 4.2.4). Figure 4.5.A shows the ground truth CSD, while B shows tCSD estimates for each value of R, demonstrating that when R is small, tCSD fails to recover the correct pattern, though as R increases, the tCSD pattern appears to match the ground truth fairly well. In contrast, Figure 4.5.C shows that GPCSD using the ground truth R can recover the CSD in each case.

For the test set error analysis from one-dimensional CSDs simulated from a Gaussian process model, the distributions of per-trial differences in mean squared error (MSE) on the test set for tCSD-GPCSD and kCSD-GPCSD are shown in Figure 4.6. It appears that GPCSD achieves smaller MSE across the test set than the other methods, which was confirmed by paired t-tests ($p << 0.0001$ for both pairwise comparisons). While the connection between Gaussian processes and RKHS regression imply that in theory kCSD and GPCSD could be tuned to produce similar results (Appendix B.4), it is possible that the shape, number, and spacing of basis elements or the range of values considered for cross-validation of the basis width and noise variance were not appropriate for this data. In addition, kCSD does not use any temporal information, while the ability of GPCSD to model the spatiotemporal process should result in more accurate predictions since nearby timepoints are modeled not as independent but as correlated. Interestingly, tCSD

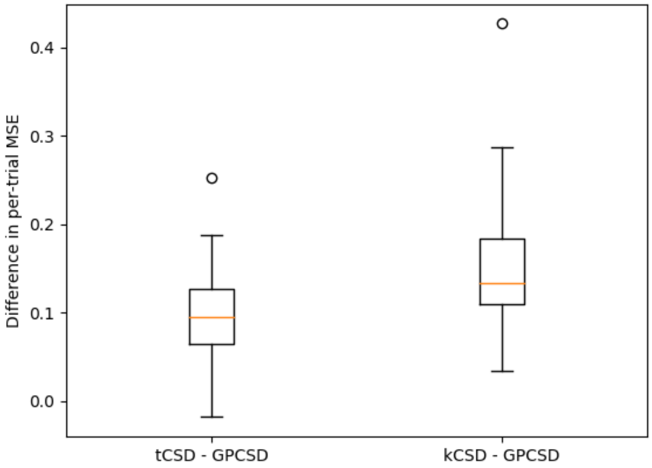

Figure 4.6: Boxplot representing the distribution, across 50 trials in the test set, of differences in per-trial MSE over space and time for tCSD-GPCSD and kCSD-GPCSD on simulated one-dimensional data (errors evaluated by comparing the true CSD and estimated CSD at the interior electrode positions). It appears that GPCSD obtains a smaller MSE than the other methods (confirmed using paired t-tests, $p << 0.0001$ for both pairwise comparisons). While tCSD appears to be the second best method in this simulation, it is important to keep in mind that it suffers from some drawbacks not taken into account in this comparison; for instance, it is limited to only predicting at the interior electrode locations, while the other two methods can predict at arbitrary spatial locations.

appears to be the second best method in this simulation, but recall that it is limited in that it can only predict CSDs at the interior electrode locations, while the other two methods can predict CSDs at arbitrary spatial locations. For two-dimensional CSDs simulated from a Gaussian process model, the distributions of per-trial MSE on the test set for kCSD and GPCSD, shown in Figure B.11, also suggest that GPCSD outperforms kCSD, despite the fact that the GPCSD hyperparameters were simply set to reasonable prior means instead of being optimized.

These simulations demonstrate that GPCSD appears to outperform existing CSD methods, even when the underlying CSD is not generated from a stationary spatiotemporal Gaussian process (as in Figure 4.4).

4.5 Application of GPCSD to auditory LFPs

In analyzing the auditory LFPs, I used a mean function model to separate out the evoked response from the ongoing activity, and in this section I present some results related to the ongoing activity and some related to the evoked response. In Section 4.5.1, I describe the experiment and data. Data analysis methods are described in Section 4.5.2 and the results are given in Section 4.5.3. Finally, I discuss implications of the results in Section 4.5.4.

4.5.1 Experiment and LFP data

The auditory LFP recordings consist of multiple sessions of LFPs from 24-electrode linear probes (V-Probes from Plexon) inserted in primary auditory cortex of two macaque monkeys. The probes were inserted in various locations, approximately perpendicular to the orientation of the left superior temporal plane. The spacing between electrodes was 100 microns so that the probe spanned 2,300 microns; however, I sometimes refer to the spatial locations by electrode index (1 through 24) rather than actual depth. The treatment of the animals was in accordance with the guidelines set by the U.S. Department of Health and Human Services (NIH) for the care and use of laboratory animals, and all methods were approved by the Institutional Animal Care and Use Committee at the University of Pittsburgh. The sampling rate of the LFPs was 1000 Hz and there were at least 2000 trials (stimulus presentations) per session.

Stimuli consisted of short tones at different frequencies and with different latencies between tones across the session. The tones were 80 dB and lasted 55ms, with 11 different frequencies spaced linearly in $\log_2$ space (starting at 257 Hz and increasing by 0.32740 octaves each step). The times since the last tone onset (the *inter-stimulus intervals*, or ISIs) ranged between 0.2 and 18.0 seconds and followed a box-car distribution in $\log_2$ space. Initial experiments had used an exponential distribution with flat hazard rates in order to minimize the ability of animals to anticipate the time of the next tone-onset. However, that had led to an under-sampling of tones with long ISIs. The box-car $\log_2$ distribution provides a compromise between flat hazard rates and equal sampling of short and long ISIs. Both the frequency of the tone and the ISI have been shown to affect the neural activity (see Figure B.4 for an illustration of the effect in this data set). That is, the particular neural populations being recorded in each session will be tuned to respond to particular preferred frequencies, and trials with longer ISIs tend to elicit larger responses (Pereira et al., 2014; Teichert et al., 2016), though the mechanisms underlying this well-known variation in response amplitude with ISI are currently still being debated. Earlier work has shown that the amplitudes of tone-evoked neural responses can be approximated as linear functions of the natural logarithm of the ISI. Hence, $\log_2$ of ISI was used as a covariate for the analysis.

The auditory paradigm was designed to test whether neural responses in auditory cortex are modulated by whether or not the time or identity of a tone can be anticipated. To that aim, the paradigm alternated between blocks of trials in which the identity and ISIs between tones was fixed for long sequences, and blocks of trials in which they were not. Detailed analyses that are not presented here have revealed subtle effects of predictability on neural responses in auditory cortex. However, because the effect of predictability is small, this analysis includes all trials from the predictable as well as the unpredictable blocks.

I focused on a single session in a single animal in which simultaneous recordings were made using two probes, spaced 3mm apart and with similar frequency tuning, so that relationships between the probes could be studied. Trials containing large artifacts were not used in the analysis; in particular, trials in which the difference between the voltage at channels 1 and 24 exceeded 500 microvolts in absolute value in either probe were removed, resulting in 2,509 trials for the two-probe simultaneous recording. While in principle the conductivity scalar σ of Equation 4.2 should be measured experimentally, in this work, I focused on recovering the spatiotemporal pattern of the CSDs and was not concerned with the exact amplitude values, so I simply used $\sigma = 1$ and treated all CSD estimates as having arbitrary units. Before fitting the Gaussian process model, all LFPs were rescaled by a factor of 0.01 for numerical reasons.

4.5.2 Auditory data analysis methods

In this section, I describe two separate analyses of the two-probe auditory data set. First, I give details on GPCSD estimation of the zero-mean ongoing activity. With a focus on comparing results obtained using the CSD instead of the LFPs, I analyzed spectral power and phase coupling using the estimated CSD ongoing activity at each electrode location. Second, I describe an analysis of trial-to-trial variation in CSD evoked responses, with a focus on finding spatial locations in the CSD in which correlated timing and/or amplitude variability in the evoked responses across trials occurred either between- or within-probes.

Ongoing activity To estimate the GPCSD hyperparameters for the ongoing activity, I first subtracted the LFP average evoked response (mean across trials) from each trial, then used the baseline period (100 ms before tone onset until tone onset) to obtain MAP estimates of the hyperparameters separately for each probe. The spatial SE lengthscale prior was inverse Gamma with 1% and 99% quantiles set to the minimum and maximum electrode spacing. The temporal SE lengthscale prior had quantiles at 50ms and 100ms and the temporal exponential lengthscale prior had quantiles at 1ms and 40ms, to encourage the SE component to extract slow-timescale variation and the exponential component to extract fast-timescale variation. The forward model parameter R had an inverse Gamma prior with quantiles chosen to represent a prior belief that the radius of the cylinder in the forward model was between 100 and 400 microns. The marginal variances and noise variances had uninformative half-Normal priors with standard deviations 2, 2, and 0.5.

After obtaining point estimates of the hyperparameters by using the Nelder-Mead method to optimize the regularized log marginal likelihood, I then predicted both the CSD and noiseless LFP at the original electrode positions and at time points from tone onset to 500ms after tone onset, including separate predictions for slow-timescale SE and fast-timescale exponential processes which add up to give the entire process. I computed a periodogram for each trial separately for the fast and slow timescales, then averaged the resulting periodograms across trials. Based on the periodograms, I selected a frequency range with a dominant oscillation, then filtered each channel and trial using a fourth-order Butterworth bandpass filter centered at the frequency of interest plus or minus 2 Hz; the filter was applied forward and backward to prevent phase distortion, then the Hilbert transform of the filtered signals was used to extract instantaneous phases at each time point. Based on timecourses of Phase Locking Value (PLV; see Section 3.4) aggregated across all possible between- and within-probe electrode pairs (Figure B.5), I selected a time point of interest to further investigate phase coupling. Phase coupling at this time point was assessed using torus graphs (Chapter 3) on all between- and within-probe electrode pairs. As described in Section 3.7, I performed model selection to determine whether a torus graph submodel would be preferable to the full torus graph model, then I first tested whether there appeared to be any across-probe connections using a stringent alpha level of 0.0001 and followed this test with edgewise tests for all within- and between- connections (edgewise alpha level of 0.01 with Bonferroni correction for all edges tested).

Evoked response With the ongoing activity hyperparameters fixed to the values estimated from the pre-stimulus baseline data, I estimated an evoked response function across all trials as the mean function of the Gaussian process (assuming the mean function was the same for all trials). To parameterize the latent CSD mean function, I used a mixture of unit amplitude Gaussian-shaped components with possibly negative scaling factors α_j:

$$\mu(z,t) = \sum_{j=1}^{J} \alpha_j \exp\left(-\frac{(z-\mu_{z,j})^2}{2\sigma_{z,j}^2}\right) \exp\left(-\frac{(t-\mu_{t,j})^2}{2\sigma_{t,j}^2}\right). \tag{4.23}$$

I chose this formulation in order to obtain surfaces that are potentially nonstationary in space and time, using prior information on the temporal location and duration of the evoked response to select starting points for $\mu_{t,j}$ and $\sigma_{t,j}^2$ and uniformly initializing the spatial locations and scaling factors α_j. While fitting mixtures of Gaussians can be unstable, sufficiently large J initialized evenly across space should be able to approximate arbitrary evoked response shapes, and the large number of trials used to estimate $\mu(z,t)$ should make the resulting surface fairly stable to different initializations. In this analysis, I used $J = 200$ Gaussian-shaped bumps to model the evoked response over space and time and optimized the marginal likelihood function; I found that starting with different initializations for the parameters and different choices of J resulted in similar evoked response surfaces.

While assuming a single mean function shared across all trials is common (as it is, in fact, what is implicitly assumed when the average evoked response is used as the mean), it is also possible that the evoked response varies from trial to trial in either timing or amplitude, and assessing such variation in the CSD should give more detailed spatial information about the variation than in the LFP. In this section, I present one potential approach to modeling trial-to-trial variation in the evoked responses, but I discuss some alternative models in Section B.6. To assess trial-to-trial variation, I separated the fitted CSD mean function into multiple CSD components by applying image segmentation techniques. To segment the CSD evoked response, local maxima were detected in the absolute value of the CSD evoked response, then a watershed algorithm was applied to find clusters around each maxima; maxima that were above or below the range of electrode locations were excluded. Space-time points not belonging to any component were kept the same for each trial and had no amplitude or shift variation. For each component, I estimated a per-trial time shift and per-trial amplitude scaling factor in a two-step fashion. First, the time shifts were estimated separately for each trial and each component by maximizing the likelihood of the data for that trial conditional on the estimated component shapes and estimated ongoing activity Gaussian process covariance function. To discourage physiologically implausible shift values, the likelihood was regularized to encourage shifts near zero; that is, truncated Normal priors were added so that components occurring before 80ms after stimulus could not shift more than 10ms either direction (with SD 5ms) and later components could not shift more than 20ms either direction (with SD 10ms); this was done because unrestricted optimization resulted in large, physiologically improbable shifts. Using the estimated time shifts, per-trial mean functions were constructed using time shifts for each component and each trial. Conditional on these per-trial mean functions and on the estimated ongoing activity Gaussian process covariance function, the per-trial amplitude scales for the components of the per-trial mean functions were modeled as depending linearly on log ISI with a latent per-trial residual. That is, given the estimated ongoing activity covariance matrix $\mathbf{K}$ and the per-trial estimated mean functions $\mu^{(n)}(s,t) \equiv \sum_{c=1}^{C} \mu_c^{(n)}(s,t)$, the amplitude model was

$$g^{(n)}(s,t) = \left[\sum_{c=1}^{C} (\beta_0^c + \beta_1^c x^{(n)} + w_c^{(n)}) \mu_c^{(n)}(s,t)\right] + \eta(s,t)$$

where x was the $\log_2$ ISI, $\eta \sim \mathrm{GP}(0, \mathbf{K})$, C was the total number of clusters, and $\mu_c^{(n)}(s,t)$ represents the time-shifted cluster c on trial n. Applying the forward model then yields an expression for the observed LFPs which gives the likelihood of the data for a single trial as

$$\tilde{\boldsymbol{\phi}}_{\boldsymbol{s},\boldsymbol{t}}{}^{(n)} \sim N\left(\sum_{c=1}^{C} (\beta_0^c + \beta_1^c x^{(n)} + w_c^{(n)}) \mathcal{A}_s \boldsymbol{\mu}_{\boldsymbol{c},\boldsymbol{s},\boldsymbol{t}}{}^{(n)},\, \Sigma\right)$$

where $\Sigma = \mathcal{A}_s \mathbf{K}_{\mathbf{s},\mathbf{s}}{}^{s} \mathcal{A}_s^T \otimes \mathbf{K}_{\mathbf{t},\mathbf{t}}{}^{T} + \sigma^2 I$ is treated as known.

For the amplitude analysis, I focused only on a single tone which had the largest evoked responses (as judged by the maximum absolute value of the per-tone average evoked responses; this resulted in 178 trials for tone 2, which had a frequency of approximately 700 Hz). I used a Bayesian model to infer a posterior distribution not only for the linear relationship between amplitude and log ISI (parameterized by β_0^c and β_1^c), but also for each per-trial residual $w_c^{(n)}$; for computational reasons, the posterior distribution was approximated using automatic differentiation mean-field variational inference (ADVI) (Kucukelbir et al., 2017) as implemented in the software `PyMC3` (Salvatier et al., 2016). The log ISI slope and intercept for each component's amplitude scale were given Normal priors with mean zero and standard deviation 2, and the per-trial amplitude scale residuals $w_c^{(n)}$ were modeled as independently Normal with standard deviation having a half-Normal prior with standard deviation 1. The variational optimization was run for 50,000 iterations and checked for convergence before using the variational approximation to the posterior distribution.

To assess within- and between-probe relationships in the per-trial scaling factors and shifts of different CSD components, correlations between these values across trials were calculated as follows. Observed correlations between the point estimates of the shifts of different components, computed across trials, were transformed using Fisher's z transform, then p-values were obtained using Normal quantiles; significant correlations were determined using an alpha level of 0.01 with Bonferroni correction for all pairs tested. To obtain a distribution of correlations for the amplitudes based on the posterior distributions, 1,000 samples were drawn from the posterior distribution. For each sample, the log ISI relationship was subtracted and a correlation was computed across trials between the residuals of the per-trial amplitude scale factors for each pair of components. Repeating this for 1,000 samples resulted in a posterior distribution of correlations for each pair; to obtain tail probabilities, a Gaussian distribution was used to approximate each distribution, and significant correlations were determined using Bonferroni-corrected 99% high posterior density (HPD) intervals. To determine components which exhibited both amplitude and shift correlation, I took as significant any pairwise correlation that was significant in both the shift and amplitude.

Finally, to investigate goodness-of-fit of the GPCSD model under different mean models to this data set, I compared the error in the predicted LFPs using four different mean models with the same zero-mean GP ongoing activity hyperparameters estimated from the baseline period. One model used no mean function, so the CSD was estimated using a zero-mean GP on the raw post-stimulus LFP data. A second model used the empirical across-trial mean (the average evoked response, or AERP) as the mean function, which is equivalent to first subtracting the AERP, then estimating a zero-mean GP on the residuals. A third model used the fitted shared mean function of Equation 4.23. Finally, the last model used the estimated per-trial and per-component shifts to compute a trial-specific mean function for each trial; the per-trial and per-component amplitudes were not used since they were not computed for all trials, though it is expected that including the amplitudes should result in a better fit. For each method, the RMSE (across space and time) for the LFP predictions was computed for each trial, and the overall RMSE was summarized by taking the mean and standard error of the RMSEs across all trials.

4.5.3 Auditory data analysis results

Ongoing activity First, I show a representative trial from one of the single-probe recordings to illustrate how the fitted GPCSD model separated the ongoing activity into two distinct timescales (which were modeled using two separate temporal covariance components: one smooth with prior biasing it toward slow variation, and one rougher with a prior biasing it toward fast variation). In particular, Figure 4.7.B shows the estimated CSD for one trial decomposed into slow and fast components; Figure 4.7.D shows

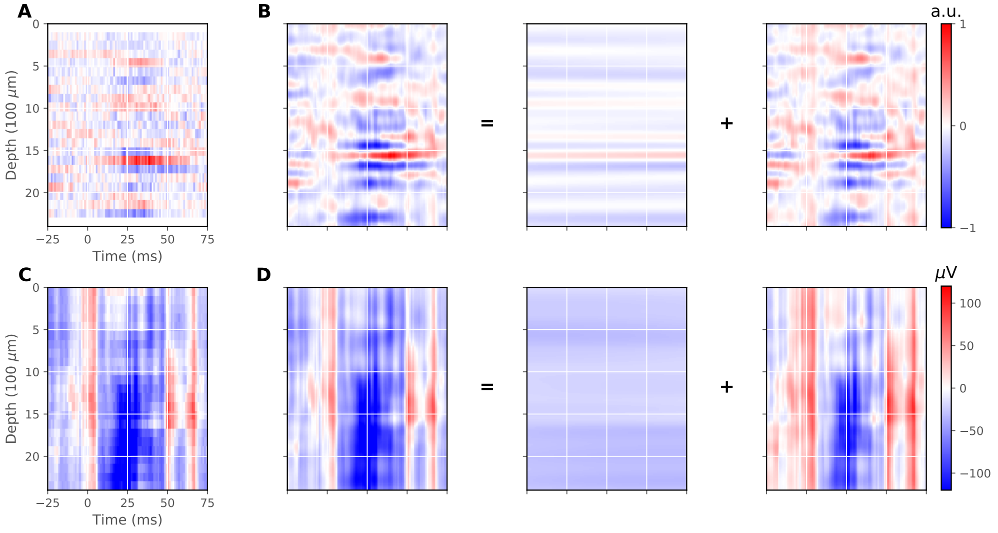

Figure 4.7: A) The tCSD prediction of the CSD for the ongoing activity in a single trial of real LFP data from one of the single-probe recordings, where the vertical axis is depth along the probe and the horizontal axis is time. B) The GPCSD prediction of the CSD for the same trial, which decomposes into slow- and fast-timescale components (middle and right). Notice the GPCSD prediction bears some resemblance to the tCSD prediction, but the latter has poorer spatial resolution and a lack of temporal smoothness. (CSDs are in arbitrary units for comparison between tCSD and GPCSD.) C) The observed ongoing LFP for this trial. D) The GPCSD prediction of the LFP for this trial, which again decomposes into slow- and fast-timescale components (middle and right). The GPCSD model LFP prediction resembles the real data, and the slow-timescale component appears to capture a baseline shift effect. Results were similar for other trials.

the corresponding predicted LFP components. It appears that the slow component captures some kind of baseline shift as it is nearly constant over time, and, in the LFP, it is also nearly flat over space; a similar result was also observed in other trials and other recordings. To show the fidelity of GPCSD in reconstructing the LFP, the observed LFP is shown in Figure 4.7.C, and the predicted LFP appears to match it very well. The estimated CSD using tCSD, shown in Figure 4.7.A, bears some resemblance to the GPCSD estimate, but with poorer spatial resolution and more noise over time.

Next, I show the results of estimating the power spectra for both the CSD and LFP, separately for the fast and slow components. Figure 4.8 shows the trial-averaged periodograms for each timescale separately for both the CSD and LFPs for probe 1; colors represent different electrode depths (dark blue is the top of the probe, yellow is the bottom of the probe). For comparison, the periodograms for the raw LFPs are also shown; results were similar for the other probe. The periodograms for the CSD and LFPs show some similarities; in the fast components, both CSD and LFP periodograms show evidence of oscillations near 5 Hz, 10 Hz, and 20 Hz, while in the slow components, both have power concentrated below 5 Hz. However, the distribution of power across depth along the probe differs between the CSD and LFPs. In the LFP periodograms, all channels tend to follow similar profiles, and changes to the profile across depth appear to happen smoothly

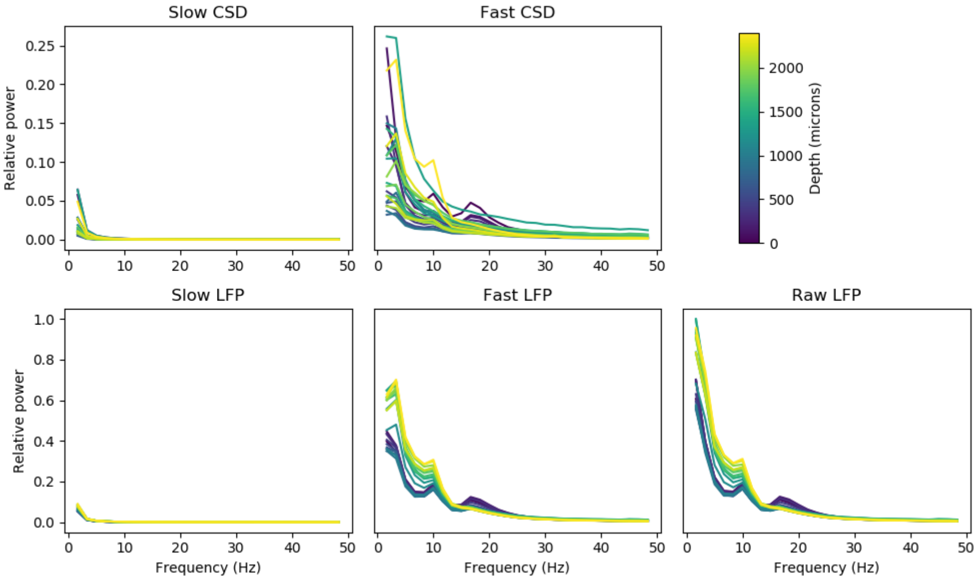

Figure 4.8: Trial-averaged periodograms for CSD (top row) and LFP (bottom row), split into slow-timescale predictions (left) and fast-timescale predictions (right) using the GPCSD model. The far bottom right is the trial-averaged periodogram from the raw LFP data. For each subplot, frequency is shown on the horizontal axis, while relative power (to the maximum for either timescale) is shown on the vertical axis. The color of each line represents the electrode depth (increasing depth from dark blue to yellow). The LFP spectra change smoothly over space in both the raw LFP and model predictions, while the CSD profiles are more differentiated across electrodes, with a less clear relationship between electrode depth and changes to the profile. Both the CSD and LFP fast components contain bumps indicating oscillatory components near 5 Hz, 10 Hz, and 20 Hz, while the slow components are concentrated at below 5 Hz. Results are shown for one probe and were similar for the other probe.

over space, which is consistent with the intuition that the forward model spatially smooths the underlying source signals. In the CSD periodograms, the profiles are more differentiated between electrodes and a smooth pattern with electrode depth is less apparent. To make the change in power as a function of depth more clear, I selected four dominant frequencies (near 0, 5, 10, and 16 Hz) and plotted the power across depth for probe 1 in Figure B.7, which shows that the CSD power in each frequency band tends to peak sharply in specific cortical layers while the LFP power tends to be smooth across space. In particular, the higher frequency oscillations show a clear peak in layer 3 which is not apparent from the LFP.

Because the periodograms for the fast-timescale process in both probes exhibited strong oscillations near 10 Hz, I used a bandpass filter centered at 10 Hz to extract instantaneous phases from the fast-timescale processes to assess phase coupling. In both the CSD and LFP, the mean pairwise PLV across all electrode locations (both within- and between- probe) appeared to increase monotonically from before the stimulus until 100 ms after stimulus, then remained high and nearly constant until about 300 ms after stimulus (Figure B.5), so I selected a time point during the later period (250 ms after stimulus) to investigate phase coupling using torus graphs on 48 nodes (24 from each probe). The torus graph model selection procedure indicated that a marginal uniform phase difference model was appropriate for both the CSD and LFP phases. The overall test of torus graph edges between the two probes was significant for both the LFP and CSD phases ($p < 0.0001$); the results of the follow-up edgewise tests using corrected $\alpha = 0.01$ are shown in Figure 4.9, with significant edges colored by edgewise $\log_{10}$ p-value and non-significant edges colored white. Within-probe, most connections tended to be the near the diagonal, particularly for the LFP, which only had connections along the diagonal (except for a few off-diagonal connections within layer 3 of probe 2). In the CSD, there were some connections between further-away locations (most notably within probe 2, which exhibits a cluster of phase correlations between layers 1 and 2 and layer 6). Across-probes, the torus graph based on LFP had no edges, while CSD torus graph had many edges, with a noticeable spatial structure to the edge pattern. There were three particularly noticeable clusters in the across-probe CSD torus graph which suggested phase coupling between CSDs in layer 1 of probe 2 and layer 2 of probe 1, between layer 3 in both probes, and between layers 5 and 6 in both probes. There are also more sporadic connections near the diagonal indicating connections between the same layers across probes; a graphical version showing the edges by depth and layer is given in Figure 4.10. The graph reveals that most connections between probes occur between the same layers on each probe, though there is also a noticeable cluster of edges from layer 2 of probe 2 to layers 5 and 6 of probe 1.

Evoked response Figure 4.11 shows the estimated evoked response function for each probe along with the components identified by the segmentation algorithm. While the evoked responses and components are similar between the two probes, there are some differences in the spatial and temporal profiles. It appears that the probe 2 evoked responses start slightly earlier than the probe 1 responses, which is also consistent with the trial-averaged multi-unit spiking activity. Both probes exhibit a current inversion near a depth of 1000 microns that appears to persist even as the activity fluctuates between positive and negative current over time. It is possible that some clusters are actually part of the same evoked event and could be combined for increased power, but for this analysis we used the clusters as-is.

The estimated per-trial shifts for all components across both probes had across-trial means near 0 ms and across-trial standard deviations typically near 1 ms (though a few components had higher standard deviations, near 3 ms). The estimated relationships between amplitude scale and $\log_2$ ISI were generally positive, with 13 out of 22 components in probe 1 and 12 of 21 components in probe 2 having 99.9% one-sided high posterior density (HPD) intervals for the slopes, estimated from 1,000 samples from the posterior distribution, falling above zero. The amount of time shift variation and the positive relationship between $\log_2$ ISI and amplitude were similar when using alternative models that did not rely on segmentation into distinct clusters (Section B.6)

In the analysis of across-trial shift and amplitude correlations, the patterns of significant correlations for both the shifts and amplitudes were quite similar (Figure B.8), so I focus on showing pairs for which both the

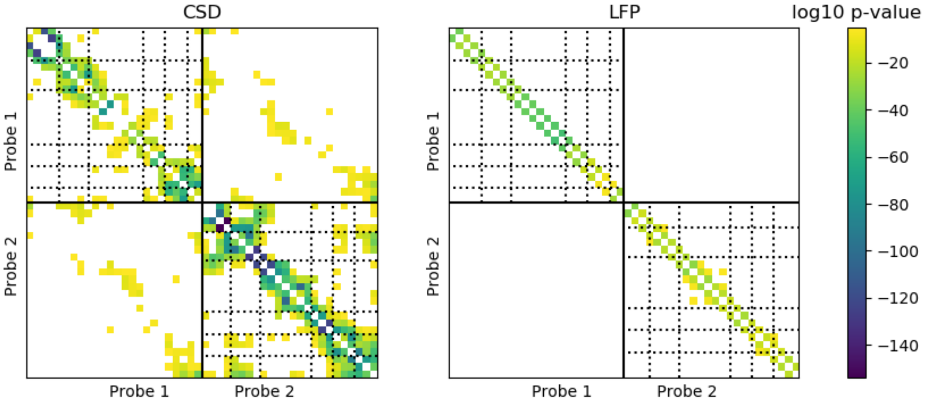

Figure 4.9: Results of edgewise torus graph inference both within- and across-probe for CSD (left) and LFP (right); depth along each probe is increasing left to right and top to bottom, with putative layer boundaries in dashed lines. Colored entries correspond to edges, with color representing the $\log_{10}$ of the p-value (edges shown for corrected $p < 0.01$). Within-probe, the LFP has edges for connections primarily along the diagonal; in probe 2, there are some off-diagonal edges within layer 3. The pattern of within-probe CSD edges is much richer, with some across-layer connections present in both probes. Between probes, the CSD torus graph contains edges while the LFP torus graph does not. Especially notable is the spatial structure that nearly follows the diagonal, including clusters of connections between probes which reflect coupling between the CSDs at the top of each probe, the middle of each probe, and the bottom of each probe. In particular, it appears layer 1 of probe 2 is connected to layer 2 of probe 1, that there are connections between the probes in layer 3, and that layers 5 and 6 are connected across probes. There are also some off-diagonal edges indicating connections from probe 2 layer 2 to the deep layers of probe 1; more detail is given in Figure 4.10, which shows the cross-probe connections in graphical form.

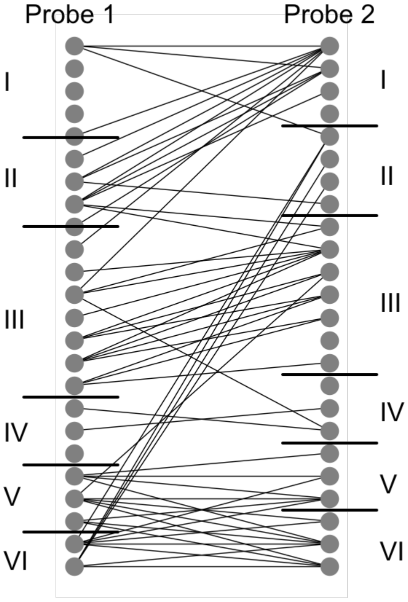

Figure 4.10: Graph showing significant between-probe CSD torus graph edges, with probe 1 nodes ordered by depth in the left column and probe 2 nodes ordered by depth in the right column; putative layer boundaries and labels are also shown to the side of each probe. Many of the cross-probe connections occur near the same depth or layer on both probes, though there is a noticeable group of edges from layer 2 of probe 2 to layers 5 and 6 of probe 1.

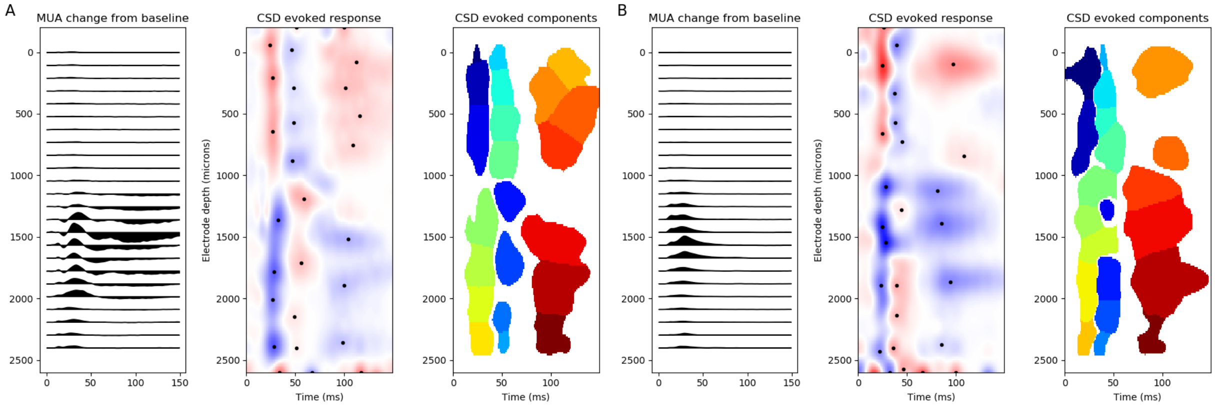

Figure 4.11: A) Trial-averaged multi-unit activity (MUA) relative to baseline (left) and CSD evoked response from probe 1 with local maxima/minima marked in black (middle), with depth along the probe on the vertical axis and time on the horizontal axis. On the right are the components returned by the image segmentation (colors correspond to arbitrary cluster number). B) Same as A, but for probe 2. The evoked responses for both probes have similar features but slightly different spatial and temporal properties. In particular, both the MUA and CSD evoked responses indicate that the evoked response begins earlier in probe 2 than probe 1.

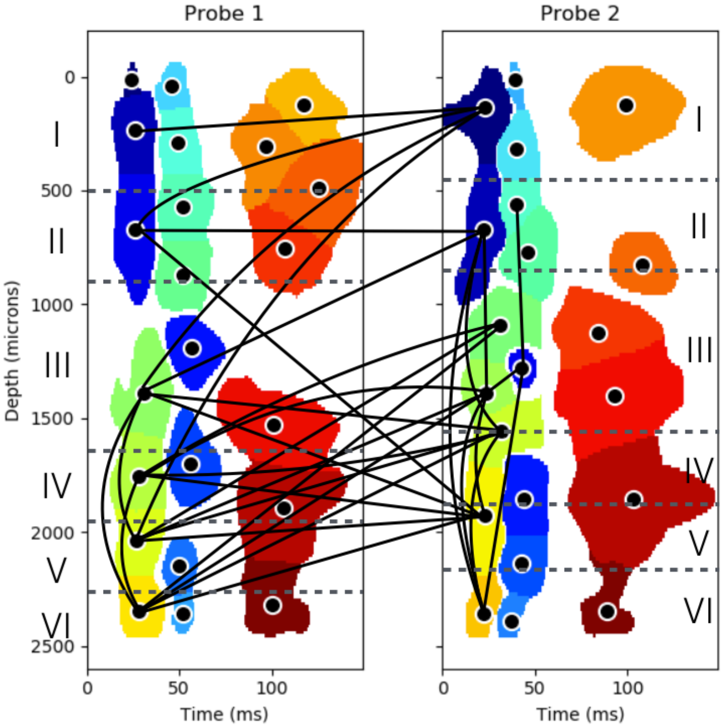

Figure 4.12: Spatiotemporal plots of probe 1 components (left) and probe 2 components (right) with putative layer boundaries as dashed lines (labels to the side); components are colored using an arbitrary numbering scheme. Edges between pairs of black nodes indicate that the components exhibited both shift and amplitude correlation across trials (significant shift correlation assessed using Fisher's transform of the observed correlation, $p < 0.01$, corrected; significant amplitude correlation assessed using 99% HPD intervals corrected for multiple comparisons). Within-probe connections in probe 1 are between the early ($\sim$ 25ms) evoked components of layers 3 through 6. In probe 2, these layers are also connected during the early evoked component, and there are also some connections to layer 2 during this time period, as well as connections between layers 2 and 3 near 50ms. Between-probe connections are predominantly between early evoked components in layers 3 through 6, but there are also connections between early evoked layers 1 and 2. In addition, there is a connection from layer 2 of probe 1 to layer 5 of probe 2, and multiple connections from layers 1 and 2 of probe 2 to layers 3 through 6 of probe 1.

amplitudes and shifts appeared to be significantly correlated. Figure 4.12 shows the CSD components for each probe side by side, with edges between black nodes inside each component representing a correlation graph for which both the shift and amplitude correlations were significant. Within each probe, the components that were correlated in amplitude and shift occurred primarily in the early evoked response ($\sim$ 25ms) and the correlations were present between layers 3, 4, 5, and 6; in probe 2, there are also some connections to layer 2 during this time period, as well as connections between layers 2 and 3 near 50ms. Between the probes, most of the connections occurred between early evoked response components at similar depths across layers 3 through 6. There were also a few connections between early evoked responses in layers 1 and 2, and from layer 2 of probe 1 to layer 5 of probe 2. Also noticeable are multiple connections from layers 1 and 2 of probe 2 to layers 3 through 6 of probe 1. No significant correlations were detected between the later evoked components. In the alternative models based only on the early evoked response, amplitude correlation between the probes was also present, and shift correlation appeared to occur most strongly in the deeper layers (Section B.6).

When I assessed the goodness-of-fit of different GPCSD models to the observed LFPs, I found that the GPCSD model with per-trial shifts included in the mean function obtained the lowest in-sample RMSE; Table 4.1 shows mean across-trial RMSEs with standard errors for both probes and for four different mean function models: no mean function, average evoked response (AERP) as mean, shared mean function across all trials, and shifted mean functions with per-trial and per-component shifts. Note that the values are not directly comparable between probes since the LFPs were on slightly different scales. Including per-component amplitudes also slightly improved the fit, but because I only estimated the amplitudes for a subset of trials corresponding to one specific tone, those results are not included here.

Probe 1		Mean RMSE	SE RMSE
	without mean function	0.029177	6.5150 $\times 10^{-5}$
	with AERP as mean	0.029164	6.5118 $\times 10^{-5}$
	with shared mean function	0.029174	6.5154 $\times 10^{-5}$
	with shifted mean function	**0.029131**	6.5151 $\times 10^{-5}$
Probe 2			
	without mean function	0.018432	1.0339 $\times 10^{-5}$
	with AERP as mean	0.018403	1.0209 $\times 10^{-5}$
	with shared mean function	0.018426	1.0311 $\times 10^{-5}$
	with shifted mean function	**0.018387**	1.0262 $\times 10^{-5}$

Table 4.1: The RMSE for the predicted LFP for each trial was computed across time and spatial coordinates; the mean RMSE across trials and its standard error are shown for different mean function models. Using GPCSD without a mean function and without accounting for the average evoked response gives the worst RMSE, while using a fixed mean function shared across all trials is only slightly better. Using the empirical mean (the AERP) does slightly better, but the best performance comes from the model using per-trial shifts in the CSD components. Note that the scales are not directly comparable across probes because the LFPs were not on the same scale.

4.5.4 Discussion of auditory data analysis results

In the analysis of ongoing activity, the greater differentiation of power spectra across depths in the CSD compared to the LFP suggests that the CSD may be helpful in isolating the spatial location of specific oscillatory components; in this data set, it appeared that most of the fast oscillations had a peak in power within layer 3. In contrast, it appears that the LFP spectra across depths are much more similar and change slowly across depth, a result that makes sense given the physical model dictating that LFPs are generally spatially blurred versions of the CSDs. Note that both the CSD and LFP tend to place high power at layers 1 and 6; however, it is possible some of this power comes from sources outside the probe. Overall, it seems that extracting oscillatory components in the CSD should give a more informative picture of phase coupling as a function of depth, and indeed, the torus graph analysis shows that torus graphs based on CSD phases appear qualitatively different from those based on LFP phases. While the within-probe phase coupling in both CSDs and LFPs generally follows the diagonal (indicating a prevalence of connections between nearby spatial locations), in both probes, the CSD torus graph contained some off-diagonal edges while the LFP torus graph did not. In particular, the CSD torus graph connections within probe 2 contain a cluster of connections from near the top of the probe to the bottom of the probe, which could appear in CSD and not LFP because the CSD correctly localizes the oscillations to their respective locations and LFP reflects multiple separate oscillatory components across all electrodes. Similarly, the between-probe LFP torus graph contained no edges while the between-probe CSD torus graph contained quite a few edges; particularly noticeable are the clusters of edges indicating phase coupling between locations near the top of each probe, the middle of each probe, the bottom of each probe. In addition to edges between probes at the same layer level, there were multiple edges from layer 2 of probe 2 to layer 5/6 of probe 1 which may indicate that probe 1 is projecting to probe 2 since layer 5 is an output layer. These results suggest that CSD enabled us to extract more meaningful phase information that allows us to see phase coupling between distinct depths along each probe, in contrast to the LFP that reflected only within-probe coupling following the spatial structure of the probe. For comparison, the raw PLV matrices corresponding to the same time point are shown in Figure B.6. Compared to the torus graph analysis, the PLV matrices typically have smaller values in the CSD than the LFP, but they exhibit somewhat similar spatial structures; however, the phase differences in the LFP are all concentrated near zero, a possible indication of volume conduction, while some of the CSD phase differences concentrate away from zero. Further investigation of the CSD locations with near-180° phase shifts is warranted as it is possible these are artifacts arising from problems with the LFPs at one or two spatial locations, though these particular locations do not appear to have many connections in the torus graph.

In the analysis of trial-to-trial variation of evoked responses, there was evidence of trial-to-trial variation in timing and amplitude of most of the detected CSD evoked components, and, for most components, there was evidence that the amplitude increased linearly with $\log_2$ of the ISI. Importantly, the amount of shift variation, the relationship of amplitude to $\log_2$ ISI, and the existence of some shift and amplitude correlation between probe were confirmed by alternative models (Section B.6), suggesting converging evidence for the trial-to-trial variation in shift and amplitude. There were similar patterns of correlation between components within and between probes when looking at both the shifts and amplitudes; focusing on pairwise relationships in which both shift and amplitude correlation appeared statistically different from zero revealed a network in which many of the early evoked components were correlated. The between-probe correlations occurred mostly at the same cortical depth in the early evoked response, but there were also some connections across depths, such as an edge from probe 1 layer 2 to probe 2 layer 5, and multiple edges from probe 2 layers 1/2 to probe 1 layers 3 and 5. Similar to the layer 2 to 5 coupling in the torus graph analysis, these edges may indicate some communication between probes occurring during the early evoked response, or they could also indicate the influence of common inputs. Based on the analysis of forward models in Section 4.2.2 and the case study of LFP and CSD correlations in Section 4.2.3, it is likely that these patterns of connectivity would not be the same if the procedure were applied on the LFP directly; in addition, the discovery of so many spatially-specific evoked components would not be possible in the LFP due to the spatially blurred nature of the evoked response. It is possible in this data set that the correlations are driven mostly by correlated inputs to the auditory cortex from thalamus and by the propagation of these inputs to different cortical layers, and the fact that none of the later components appeared to have strong correlations seems to conform with this hypothesis. Nevertheless, these results demonstrate the possibilities for more advanced modeling in the CSD domain to obtain detailed patterns of activity over time at specific cortical depths.

In investigating the impact of the mean function model on the reconstruction of the LFP, I saw that even when ignoring the presence of a possible mean function, one can still obtain relatively good fit using a zero-mean GPCSD model, but that including some kind of mean function likely does better (even if the AERP is directly used as the mean function). In addition, it appears that estimating per-component shifts does buy some accuracy in reconstructing the LFP, suggesting that the estimated shifts are picking up meaningful variation in the LFP that is not well-explained by the ongoing activity.

4.6 Application of GPCSD to Neuropixel LFPs

In the analysis of Neuropixels data, I do not attempt to separate out the evoked response and ongoing activity, but instead focus on recovering time-varying activity using a zero-mean GPCSD model. In Section 4.6.1, I describe the experiment and data. Data analysis methods are described in Section 4.6.2 and the results are discussed in Section 4.6.3.

4.6.1 Experiment and LFP data

The data were collected using a Neuropixels probe described in Jun et al. (2017). I use data from three different mice, with a collection of LFPs at different electrode locations on a Neuropixels probe inserted to cover visual areas (VISp), hippocampal (CA1/CA3, DG), and thalamic areas (TH, SC); the probe was identified as Probe C. From this probe, LFPs were recorded at up to 384 electrode locations spanning two spatial dimensions; the possible recording positions were arranged in four columns 16 microns apart, with electrodes within a column spaced 20 microns apart. Each electrode with an LFP recording was labeled by region if there was a spiking unit associated with the electrode that was identified as belonging to that region (based on manual inspection of spiking activity). LFP electrodes without region labels were not used in the analysis; Table 4.2 gives the number of LFPs in each region for each mouse, and Figure B.10 shows the recording positions for one mouse with spiking unit counts for each location. The visual stimulus was a simple full-field flash and the recorded data consisted of 150 repeated trials. The original temporal sampling rate was 2500Hz but was downsampled to 500Hz for computational reasons.

	Number of LFP channels			
Mouse ID	Visual	Hippocampal	Thalamic	Total
403407	40	43	25	279
412793	61	31	37	300
419117	57	51	7	278

Table 4.2: For each mouse ID, the number of identified LFP channels recording from visual areas, hippocampal areas, and thalamic areas. The total number refers to the total number of LFP channels recorded on the probe, but some of these were not identified with any particular region and were therefore not used in the analysis.

4.6.2 Neuropixels data analysis methods

In contrast to the auditory LFPs, here I do not focus on modeling the evoked responses and ongoing activity separately, and instead predict the CSD containing both portions of activity using a single zero-mean GPCSD model, which Table 4.1 suggests may perform slightly worse than using a mean function, but which greatly simplifies the analysis as it does not require parameterization of a nonstationary mean function in two spatial and one temporal dimension. This approach can be seen as more exploratory in nature than the approach on the auditory LFPs which started with a structured parametric mean function. The GPCSD hyperparameters were selected separately for each region and for each mouse by regularized maximum marginal likelihood using the first 100ms of data after the stimulus presentation. The prior for R was inverse Gamma with 1% and 99% quantiles at 1 and 50 microns, while the priors for the spatial lengthscales were set so the quantiles were at the minimum and maximum electrode spacing in each direction. The temporal SE lengthscale had inverse Gamma prior with quantiles at 2 and 40 milliseconds and variance had half-Normal prior with standard deviation 2. The temporal exponential lengthscale had inverse Gamma prior with quantiles at 50 and 100 milliseconds and variance had half-Normal prior with standard deviation 2. The noise variance had half-Normal prior with standard deviation 0.5. (As in the auditory data, the LFPs were rescaled by a common factor for numerical reasons.)

The numerical optimization for the regularized log marginal likelihood was slightly more involved compared to the one-dimensional case; in the one-dimensional case, the objective function could be optimized using standard numerical methods because the integrals in the forward operator could be computed quickly and accurately, but in the two-dimensional case, higher-dimensional integrals are required, so the objective function becomes much more expensive to evaluate and may also be noisy due to error in the numerical integration. That is, there is a trade-off in speed and accuracy for the numerical integration which becomes more obvious in the two-dimensional case; for this analysis, I chose to accept larger integration error to speed up each function evaluation. To handle optimization of the now potentially noisy objective function, I used a stochastic optimization technique based on approximating the objective function with gradient boosted regression trees as implemented in Python library `skopt`; such techniques are particularly well-suited to optimization of noisy and expensive objective functions. In this method, 300 random hyperparameter vectors are selected from a reasonable range defined by the priors, and gradient boosted regression trees are used to learn a mapping between the hyperparameters and the noisy objective values. Then, with the regression surface fixed, the objective function was evaluated 300 more times in an effort to find a local minimum in the regression function.

As a preliminary visualization of the data, I selected the visual region (because the LFPs showed the largest evoked responses) and computed the evoked response from the GPCSD predictions. To visualize and summarize other aspects of the estimated CSDs across space and time, I concatenated the estimated CSD timecourses across all trials, then applied PCA to extract five spatial components explaining the most variance across all time points and trials; the PCA scores then yield a time series for each trial describing the weight of the contribution of the component at each time point. The component scores were normalized so that each has variance equal to 1 for visualization. For comparison, the same PCA procedure was applied directly to the LFPs.

4.6.3 Neuropixels data analysis results

First, I show the evoked response from approximately 17ms to 77 ms for the visual region of one mouse that contained well-defined visual evoked responses (Figure B.9 shows the LFP average evoked responses for this mouse grouped by region). Figure 4.13 shows snapshots of the two-dimensional data across time, with time moving from left to right; the top row is the raw LFP (interpolated for visualization), the middle row is the GPCSD evoked response, and the bottom row is tCSD applied to the average evoked response (interpolated

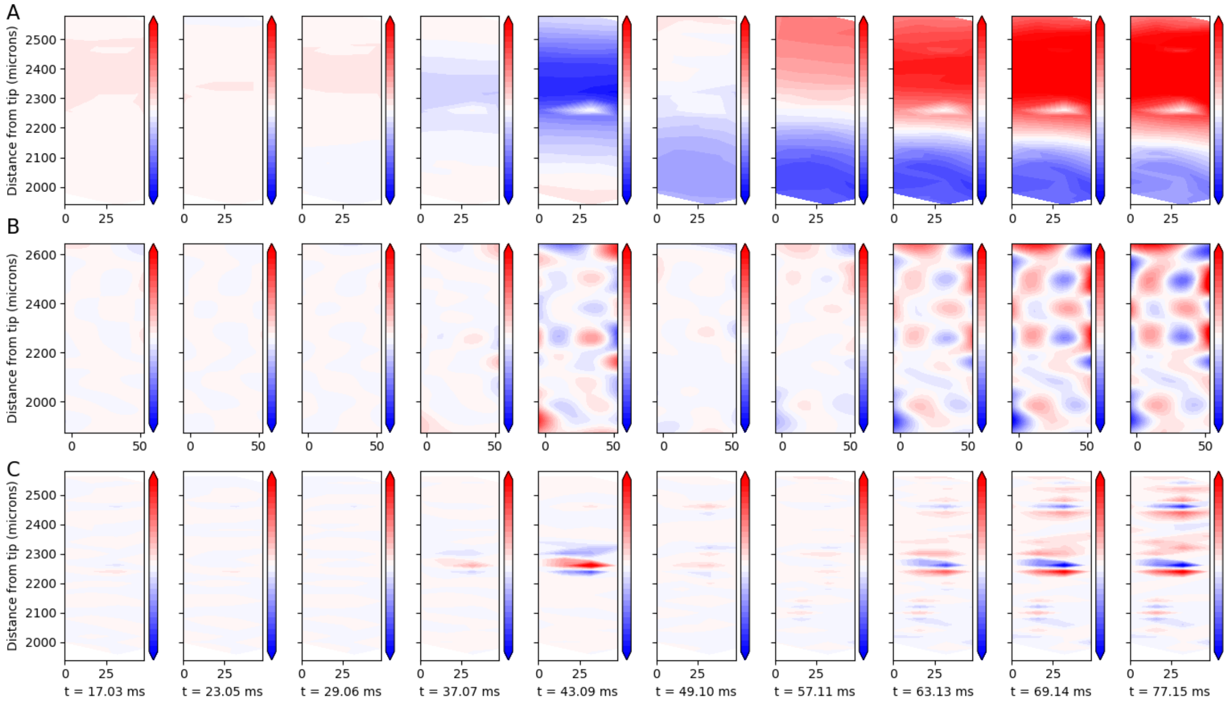

Figure 4.13: Each panel is a snapshot in time of two-dimensional Neuropixels data from the visual region of one mouse, with LFPs in row A, GPCSD predictions in row B, and tCSD predictions in row C. The temporal evolution of GPSD appears to match the amplitude changes in the LFP. The tCSD predictions bear some resemblance to the GPCSD prediction but have much less detail particularly in the horizontal direction.

for visualization). Each two-dimensional snapshot spans the width of the probe along the horizontal axis, and spans the vertical region of the probe corresponding to the visual brain region along the vertical axis; see Figure B.10 for a depiction of all recording channels on the probe with their region assignments and number of identified spiking units as a function of distance from the tip of the probe. While the tCSD prediction bears some resemblance to the GPCSD prediction, it includes much less detail, particularly across the width of the probe.

Example results of applying PCA to the CSD and LFPs for the visual region in the same mouse are shown in Figure 4.14. Examining the time series of the scores for ten trials (colored lines) and the across-trial means of the scores (black lines) shows that the PCA scores have similar types of trajectories in both the CSD and LFPs, but they are not identical for any given component. Furthermore, the spatial information present in the CSD components is much richer than that available in the LFP components, which are highly spatially smoothed compared to the CSDs and only available at an irregular grid of locations within the area of interest. For example, component 1 appears to capture the canonical visual evoked response in both the LFP and CSD, based on the black averaged score time series; in the LFP, the spatial distribution of component 1 is highly correlated across space, with a dominant pattern of negative values at the top of the depth range and positive values at the bottom of the depth range. In contrast, in the CSD, component 1 appears to take the form of several dipoles arranged in columns. It appears that component 1 matches well with the profile observed in the average evoked responses of Figure 4.13, suggesting that in this case, further analysis of component 1 could address to trial-to-trial variation in evoked responses. The other components shown in Figure 4.14 appear to correspond to other types of variation around the evoked response function, and again, the spatial information in the CSD components is more detailed than that in the LFP in each case. The correspondence of the LFP and CSD score time series in the other regions and mice was similar, though not all regions/mice had such clearly defined evoked responses.

4.7 Discussion

In this chapter, I argued that the current source density (CSD) is useful for understanding information flow between neural populations or regions based on LFP data. In contrast to the LFPs, which reflect summed post-synaptic currents from many neural populations, the CSD represents net current flow in specific spatial locations. By stepping through the forward models relating one- and two-dimensional CSDs to measured LFPs, I showed that the LFP will generally not reflect the correct association structure present in the underlying sources, meaning that the CSD is preferable for assessing correlation or other types of association. To estimate the CSD, I developed a novel current source density (CSD) estimation method, GPCSD, based on modeling the CSD as a spatiotemporal Gaussian process. Combining the CSD model with an appropriate forward model then yields a statistical model for the LFP that takes the forward model into account and allows principled hyperparameter tuning through maximization of the marginal likelihood. I demonstrated the GPCSD method on simulated data, where it outperformed existing methods in both one-dimensional and two-dimensional CSD estimation, even when the underlying CSD was not generated from a GPCSD model. Compared to existing CSD methods, GPCSD is most closely related to the inverse CSD methods (of which the most general version is kCSD). In addition, Gaussian process regression and the RKHS regression used in kCSD result in similar-looking estimators (Section B.4), so I expect that under some conditions, GPCSD and kCSD would perform similarly. However, the specification of the prior Gaussian process through mean and covariance functions appears more natural than the specification of some number and spacing of user-selected basis functions as in kCSD, and in my simulations, GPCSD clearly dominated both tCSD and kCSD. The dominance of GPCSD likely stems from its use of a spatiotemporal model which captures not only spatial but temporal correlations and therefore borrows strength across time points (sometimes called *multi-task* or *transfer learning* (Alvarez et al., 2012)). In contrast, tCSD and kCSD do not consider correlations over time, treating each time point as independent. In addition, tCSD implicitly assumes $R \to \infty$, while kCSD requires manual selection of R (as well as the shape, number, and spacing of basis functions and the grid of

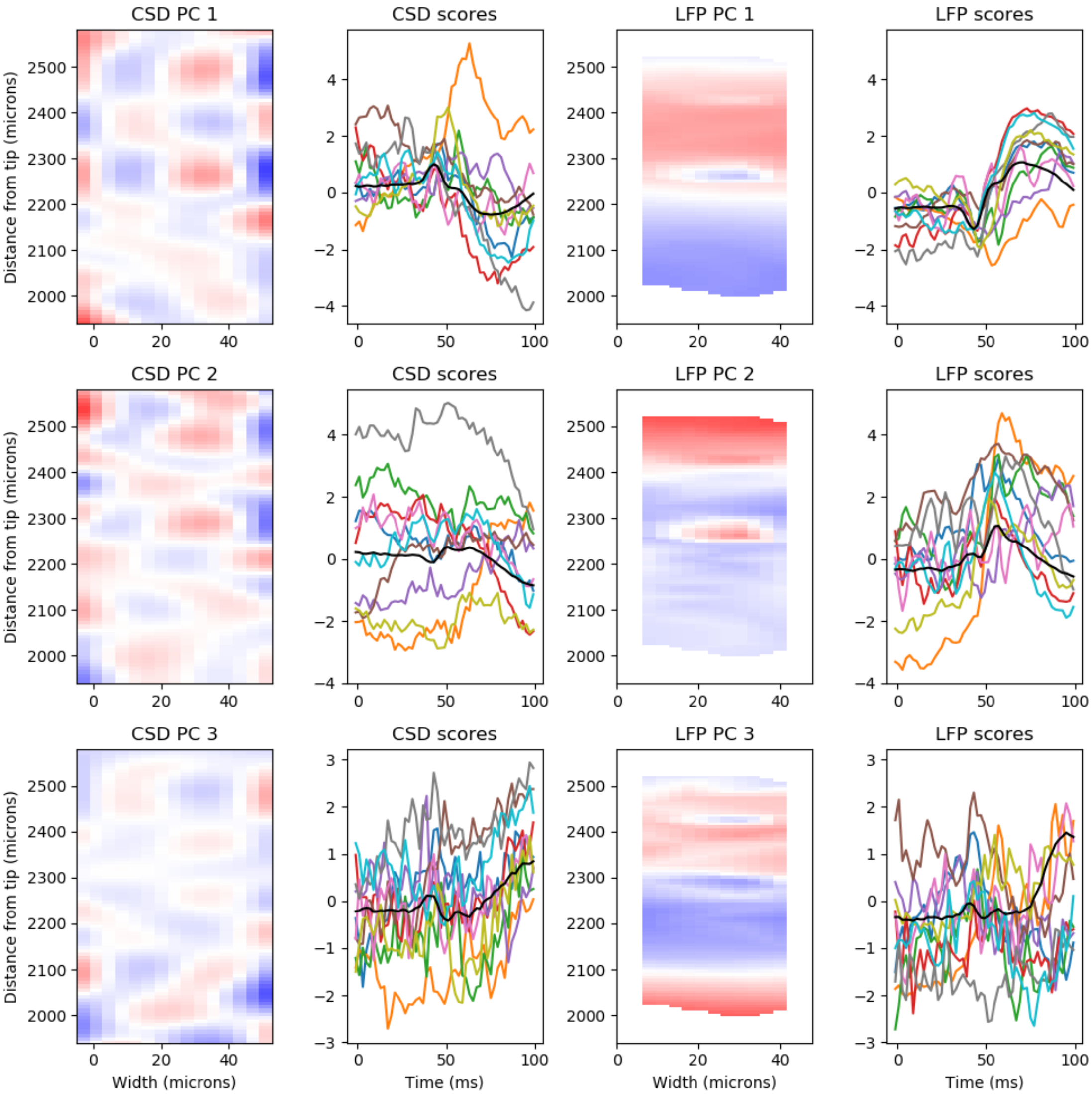

Figure 4.14: Top 3 principal components (rows) along with the corresponding score time series for ten trials (colored lines; across-trial mean in black), for CSDs (left two columns) and LFPs (right two columns). Spatial principal components were extracted after concatenating trials so they represent spatial patterns explaining the most variation across all trials and time points. While the components appear to capture similar temporal characteristics for both LFPs and CSDs, particularly in components 1 and 3, the score time series are not identical and the spatial information present in the CSD components is much more detailed than in the LFP components. In this case, the first component appears to capture a canonical visual evoked response, and the spatial pattern in the CSD gives a detailed picture of where the evoked activity occurs; in contrast, the LFP is very smooth over space with the only noticeable feature being an inversion near a depth of 2250 microns.

basis widths and λ values, all of which could greatly affect the solution). It seems that the good performance of GPCSD compared to the other methods in simulations is due to these factors.

In the auditory LFP data, the GPCSD method provided detailed spatial information about various oscillatory components that was not available from the LFPs. That is, the power in any particular frequency band tended to vary smoothly across depth in the LFP, making it unclear whether the oscillation was truly present at all electrodes or whether oscillations at a smaller number of specific locations were being picked up on all electrodes; in the CSD, the power in any particular frequency band tended to show peaks at specific locations, suggesting that oscillatory behavior may actually be local to these locations rather than present across the entire probe. Further lending support to this idea, torus graphs estimated from GPCSD values and observed LFPs filtered to one of the dominant oscillatory bands (centered at 10 Hz) showed that qualitatively different graph structures were inferred using the GPCSD predictions rather than the LFPs. Of particular interest were the across-probe connections in the CSD graph, with most connections between similar depths on each probe, but also some connections from layer 2 of probe 2 to layer 5 of probe 1 which may indicate information flow between the probes; these cross-probe connections were absent in the LFP graph. It seems that because the LFPs are highly correlated across space, even a localized oscillation appears simultaneously on multiple electrodes, leading torus graphs, which are based on conditional independence relationships, to be more sparse for the LFPs than the GPCSD because of conditioning on other locations essentially containing the same oscillatory signals. When analyzing the trial-to-trial variation in evoked response components, I found evidence of both within-probe and between-probe relationships in the shifts and amplitudes of the early evoked components; this lends credence to the idea that the early evoked response results from one dominant dipole generated by input coming from thalamus, and that the inputs are correlated between the two probes. Some alternative trial-to-trial variation models presented in Section B.6 also suggested similar results. In the Neuropixels LFP data, the GPCSD method was combined with PCA to identify spatial patterns capturing the most variance in the CSD across time and trials, and the resulting timecourses of the scores of these components appeared similar to PCA on the LFPs, but not identical. The spatial patterns in the LFPs, in contrast to the CSDs, were not very informative about potential current sources and were very spatially smooth. While cross-region communication was not yet addressed using GPCSD in the Neuropixels data, these results suggest that GPCSD could be used not only for a high-level description of cross-region connectivity, but also to locate specific spatial locations driving the connectivity (such as a certain layer of visual cortex communicating with a specific thalamic area). These results demonstrate the potential use of GPCSD not only as an accurate method for predicting the CSD, but also as part of downstream analyses of correlated neural activity.

However, GPCSD does suffer from some computational issues, as the numerical integration in the forward model must be re-done each time the hyperparameters change during numerical optimization, and this starts to become computationally challenging, and numerical integration error is potentially problematic for the optimization. This issue is particularly salient in the two-dimensional case, where four-dimensional integrals are required to evaluate the spatial covariance function which is part of the log marginal likelihood. In my current implementation, I found that using some non-standard optimization routines (Nelder-Mead in the one-dimensional case and stochastic optimization in the two-dimensional case) appeared to result in clearer convergence to a local minimum than more common quasi-Newton methods, which often appeared to get stuck near the initialization point. The current implementation could be improved by using analytic gradients of the log marginal likelihood and by using more specialized numerical integration routines. In particular, the function being integrated over in the forward model has a sharp peak in one-dimensional CSD and a pole in two-dimensional CSD, but the current Gauss-Legendre integration technique does not make use of this structure to improve the integration; it is possible that tanh-sinh quadrature would handle these regions of the function better. In addition, it seems that the forward model parameter R is potentially more problematic than the other hyperparameters due to weak identifiability issues with the other hyperparameters and due to its effect on the peaked-ness of the forward model weight function and therefore on the accuracy of the numerical integration. One option would be to simply fix R ahead of time, as is done for kCSD; this would greatly reduce the computational burden of fitting the other hyperparameters and would likely make the objective function more well-behaved. Another option would be to use an alternating optimization procedure to update R separately from the other hyperparameters, or to select a fixed grid of R values and use the one which achieves the largest log marginal likelihood after optimization of the other hyperparameters. One could also use low-rank Gaussian process approximations both to stabilize the calculations and to lower the cost of inverting the covariance matrix Solin and Särkkä (2014); I briefly explored this method and it appeared to work fairly well in the two-dimensional case, but I did not use it in the current implementation as it can cause the predicted CSDs to be overly smooth and it is not necessarily clear how to choose the rank of the spatial and temporal covariance matrices. Computational speed could also be increased by pre-whitening the data in time based on the temporal covariance estimated in the baseline data rather than estimating the temporal covariance function, allowing the spatial lengthscales and R to be estimated more quickly; however, this precludes the use of the GPCSD model to infer processes operating at separate time scales. While improvements to computational speed and accuracy in the optimization of the likelihood would be solid improvements, future work could also include expansion to non-separable or nonstationary covariance functions which may provide better descriptions of neural data. Though I demonstrated one analysis that used a parametric mean function to capture the evoked response, it seems somewhat difficult to directly specify suitable mean functions in the CSD space, so more investigation of appropriate parameterizations of the mean function, particularly in higher spatial dimensions, would be helpful. Another issue with the current implementation of GPCSD is that it uses point estimates for the hyperparameters and does not estimate or propagate uncertainty about the hyperparameters; a potential remedy would be to use a fully Bayesian approach to estimate the posterior distributions of the hyperparameters (via MCMC) rather than using optimization to find the posterior mode. Furthermore, I am using the posterior mean as a prediction of the CSD, but I have not taken into account the posterior variance of the predictions in downstream analyses. This is likely not a problem when predicting the CSD at the same space-time points as the observed LFPs (as was done in the auditory LFP data), but it is more important to consider uncertainty in the predictions in cases such as the Neuropixels data, where the observed LFPs are irregularly spaced and CSD predictions are desired in locations where the LFPs were not observed (as such predictions would have larger posterior uncertainty).

In summary, GPCSD appears to be a promising new method for CSD estimation that outperforms previous methods by pooling information across time and trials. After specifying some suitable ranges and/or priors for the hyperparameters, the GPCSD predictions can be obtained at any time and space locations of interest, and the resulting CSD will provide a better measure of localized neural activity than the LFP, important for downstream analyses of association between different neural populations or regions. Furthermore, the GPCSD model is flexible and extensible through the choice of mean and covariance functions. Though some computational issues remain for future work, GPCSD has been shown to offer a promising new approach for analyzing information flow based on LFPs.

Bibliography

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012). Kernels for vector-valued functions: a review. *Foundations and Trends® in Machine Learning*, 4(3):195–266. 97

Arieli, A., Sterkin, A., Grinvald, A., and Aertsen, A. (1996). Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283):1868–1871. 3, 58

Aydore, S., Pantazis, D., and Leahy, R. M. (2013). A note on the phase locking value and its properties. *NeuroImage*, 74:231–244. 33

Barnett, L. and Seth, A. K. (2011). Behaviour of granger causality under filtering: theoretical invariance and practical application. *J. Neurosci. Methods*, 201(2):404–419. 5

Bastos, A. M. and Schoffelen, J.-M. (2016). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175. 16

Betancourt, M. (2017). Robust Gaussian Processes in Stan, Part 3. Retrieved from `https://betanalpha.github.io/assets/case_studies/gp_part3/part3.html`. 73

Bießmann, F., Meinecke, F. C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N. K., and Mller, K.-R. (2010). Temporal kernel cca and its application in multimodal neuronal data analysis. *Machine Learning*, 79(1-2):5–27. 6

Brincat, S. L. and Miller, E. K. (2015a). Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature neuroscience*, 18(4):576. 2, 3, 25, 46, 52

Brincat, S. L. and Miller, E. K. (2015b). Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature neuroscience*, 18(4):576–581. 20

Brincat, S. L. and Miller, E. K. (2016a). Prefrontal cortex networks shift from external to internal modes during learning. *Journal of Neuroscience*, 36(37):9739–9754. 16

Brincat, S. L. and Miller, E. K. (2016b). Prefrontal cortex networks shift from external to internal modes during learning. *Journal of Neuroscience*, 36(37):9739–9754. 25, 44, 46

Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proc. Natl. Acad. Sci. U. S. A.*, 101(26):9849–9854. 5

Brown, L. D. (1986). Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims. 29

Buccino, A. P., Kuchta, M., Jæger, K. H., Ness, T. V., Berthet, P., Mardal, K.-A., Cauwenberghs, G., and Tveito, A. (2019). How does the presence of neural probes affect extracellular potentials? *Journal of neural engineering*, 16(2):026030. 63

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents: EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, 13(6):407. 1, 57

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929. 25

Cadieu, C. F. and Koepsell, K. (2010). Phase coupling estimation from multivariate phase statistics. *Neural computation*, 22(12):3107–3126. 31, 38

Chen, S., Witten, D. M., and Shojaie, A. (2014). Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64. 30

Ching, S., Cimenser, A., Purdon, P. L., Brown, E. N., and Kopell, N. J. (2010). Thalamocortical model for a propofol-induced α-rhythm associated with loss of consciousness. *Proceedings of the National Academy of Sciences*, 107(52):22665–22670. 25

Christopher, M. B. (2016). *PATTERN RECOGNITION AND MACHINE LEARNING.* Springer-Verlag New York. 8

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183. 38, 39, 40

Ding, M. and Wang, C. (2014). Analyzing MEG data with granger causality: Promises and pitfalls. In Supek, S. and Aine, C. J., editors, *Magnetoencephalography*, pages 309–318. Springer Berlin Heidelberg. 5

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes.* PhD thesis, University of Cambridge. 71

Einevoll, G. T., Kayser, C., Logothetis, N. K., and Panzeri, S. (2013). Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nature Reviews Neuroscience*, 14(11):770. 1, 57, 62, 68

Fell, J. and Axmacher, N. (2011). The role of phase synchronization in memory processes. *Nature Reviews Neuroscience*, 12(2):105. 25

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical distributions*. John Wiley & Sons. 33

Forbes, P. G. and Lauritzen, S. (2015). Linear estimating equations for exponential families with application to gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283. 38, 40, 121

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438. 5

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004a). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664. 2

Hardoon, D. R., Szedmak, S. R., and Shawe-taylor, J. R. (2004b). Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664. 7, 8, 11

Herreras, O. (2016). Local field potentials: myths and misunderstandings. *Frontiers in neural circuits*, 10:101. 57

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, pages 321–377. 7

Hyvärinen, A. (2005a). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709. 3

Hyvärinen, A. (2005b). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*. 38, 120, 121

Hyvärinen, A. (2007). Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512. 38

Jammalamadaka, S. R. and Sarma, Y. (1988). A correlation coefficient for angular variables. *Statistical theory and data analysis II*, pages 349–364. 125

Jammalamadaka, S. R. and Sengupta, A. (2001). *Topics in circular statistics*, volume 5. World Scientific. 33, 126

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232. 63, 94

Kajikawa, Y. and Schroeder, C. E. (2011). How local is the local field potential? *Neuron*, 72(5):847–858. 57

Kass, R. E., Eden, U. T., and Brown, E. N. (2014). *Analysis of neural data*, volume 491. Springer. 32, 43, 48

Kass, R. E. and Vos, P. W. (2007). *Geometrical foundations of asymptotic inference*. John Wiley & Sons. 114

Kropf, P. and Shmuel, A. (2016). 1D current source density (CSD) estimation in inverse theory: a unified framework for higher-order spectral regularization of quadrature and expansion-type CSD methods. *Neural computation*, 28(7):1305–1355. 68, 139

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474. 83

Kurz, G. and Hanebeck, U. D. (2015). Toroidal information fusion based on the bivariate von Mises distribution. *Multisensor Fusion and Integration for Intelligent Systems MFI, IEEE International Conference on*, IEEE. 28

Lachaux, J.-P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999a). Measuring phase synchrony in brain signals. *Human brain mapping*, 8(4):194–208. 2

Lachaux, J.-P., Rodriguez, E., Martinerie, J., and Varela, F. J. (1999b). Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208. 32

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press. 115

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854. 40

Lindén, H., Tetzlaff, T., Potjans, T. C., Pettersen, K. H., Grün, S., Diesmann, M., and Einevoll, G. T. (2011). Modeling the spatial reach of the LFP. *Neuron*, 72(5):859–872. 57

Lu, H. (2013). Learning canonical correlations of paired tensor sets via tensor-to-vector projection. In Rossi, F., editor, *IJCAI*. IJCAI/AAAI. 6

Mardia, K. V., Kent, J. T., and Laha, A. K. (2016). Score matching estimators for directional distributions. *arXiv preprint arXiv:1604.08470*. 31, 38

Mardia, K. V. and Patrangenaru, V. (2005). Directions and projective shapes. *The Annals of Statistics*, 33(4):1666–1699. 27

Mardia, K. V., Taylor, C. C., and Subramaniam, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63(2):505–512. 31

Marek, S., Tervo-Clemmens, B., Klein, N., Foran, W., Ghuman, A. S., and Luna, B. (2018). Adolescent development of cortical oscillations: Power, phase, and support of cognitive maturation. *PLoS Biology*, 16(11):e2004188. 2

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190. 14

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473. 41

Navarro, A. K., Frellsen, J., and Turner, R. E. (2017). The multivariate generalised von mises distribution: inference and applications. In *AAAI*, pages 2394–2400. 112

Nicholson, C. (1973). Theoretical analysis of field potentials in anisotropic ensembles of neuronal elements. *IEEE Transactions on Biomedical Engineering*, (4):278–288. 63

Nicholson, C. and Freeman, J. A. (1975). Theory of current source-density analysis and determination of conductivity tensor for anuran cerebellum. *Journal of neurophysiology*, 38(2):356–368. 62, 68

Nicholson, C. and Llinas, R. (1971). Field potentials in the alligator cerebellum and theory of their relationship to Purkinje cell dendritic spikes. *Journal of Neurophysiology*, 34(4):509–531. 3, 58, 62, 68

Pereira, D. R., Cardoso, S., Ferreira-Santos, F., Fernandes, C., Cunha-Reis, C., Paiva, T. O., Almeida, P. R., Silveira, C., Barbosa, F., and Marques-Teixeira, J. (2014). Effects of inter-stimulus interval (isi) duration on the n1 and p2 components of the auditory event-related potential. *International Journal of Psychophysiology*, 94(3):311–318. 80

Pettersen, K. H., Devor, A., Ulbert, I., Dale, A. M., and Einevoll, G. T. (2006). Current-source density estimation based on inversion of electrostatic forward solution: effects of finite extent of neuronal activity and conductivity discontinuities. *Journal of neuroscience methods*, 154(1-2):116–133. 68, 139

Pitts, W. (1952). Investigations on synaptic transmission. In *Cybernetics, Trans. 9th Conf. Josiah Macy, New York*, pages 159–162. 3, 58, 62

Place, R., Farovik, A., Brockmann, M., and Eichenbaum, H. (2016). Bidirectional prefrontal-hippocampal interactions support context-guided memory. *Nature Neuroscience*. 22

Potworowski, J., Jakuczun, W., Łski, S., and Wójcik, D. (2012). Kernel current source density method. *Neural computation*, 24(2):541–575. 63, 64, 68, 69, 75, 139

Rana, K. D., Vaina, L.-M., and Hamalainen, M. (2013). A fast statistical significance test for baseline correction and comparative analysis in phase locking. *Frontiers in neuroinformatics*, 7:3. 32

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian process for machine learning.* MIT press. 72, 140

Rodriguez-Lujan, L., Larrañaga, P., and Bielza, C. (2017). Frobenius norm regularization for the multivariate von Mises distribution. *International Journal of Intelligent Systems*, 32(2):153–176. 53

Rodu, J., Klein, N., Brincat, S. L., Miller, E. K., and Kass, R. E. (2018). Detecting multivariate cross-correlation between brain regions. *Journal of neurophysiology.* 2, 5

Saatçi, Y. (2012). *Scalable inference for structured Gaussian process models.* PhD thesis, University of Cambridge. 141

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55. 83

Särkkä, S. (2011). Linear operators and stochastic partial differential equations in Gaussian process regression. In *International Conference on Artificial Neural Networks*, pages 151–158. Springer. 62, 64, 70

Schacke, K. (2004). On the Kronecker product. *Master's thesis, University of Waterloo.* 71

Schreier, P. J. and Scharf, L. L. (2010). *Statistical signal processing of complex-valued data: the theory of improper and noncircular signals.* Cambridge university press. 112

Sherman, M. A., Lee, S., Law, R., Haegens, S., Thorn, C. A., Hämäläinen, M. S., Moore, C. I., and Jones, S. R. (2016). Neural mechanisms of transient neocortical beta rhythms: Converging evidence from humans, computational modeling, monkeys, and mice. *Proceedings of the National Academy of Sciences*, 113(33):E4885–E4894. 25

Solin, A. and Särkkä, S. (2014). Hilbert space methods for reduced-rank gaussian process regression. *arXiv preprint arXiv:1401.5508.* 100, 141

Steinmetz, N. A., Koch, C., Harris, K. D., and Carandini, M. (2018). Challenges and opportunities for large-scale electrophysiology with neuropixels probes. *Current opinion in neurobiology*, 50:92–100. 25, 63

Szymanski, F. D., Rabinowitz, N. C., Magri, C., Panzeri, S., and Schnupp, J. W. (2011). The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. *Journal of Neuroscience*, 31(44):15787–15801. 3, 58

Teichert, T., Gurnsey, K., Salisbury, D., and Sweet, R. A. (2016). Contextual processing in unpredictable auditory environments: the limited resource model of auditory refractoriness in the rhesus. *Journal of neurophysiology*, 116(5):2125–2139. 80

Tort, A. B., Komorowski, R., Eichenbaum, H., and Kopell, N. (2010). Measuring phase-amplitude coupling between neuronal oscillations of different frequencies. *Journal of neurophysiology*, 104(2):1195–1210. 54

Trongnetrpunya, A., Nandi, B., Kang, D., Kocsis, B., Schroeder, C. E., and Ding, M. (2016). Assessing granger causality in electrophysiological data: removing the adverse effects of common signals via bipolar derivations. *Frontiers in systems neuroscience*, 9:189. 16

van den Boogaart, K. G. and Brenning, A. (2001). Why is universal kriging better than irfk-kriging: Estimation of variograms in the presence of trend. In *Proceed. of 2001 Annual Conf. of the Intern. Assoc. for Math. Geology*. 73

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305. 27, 112, 115

Wang, X., Chen, Y., and Ding, M. (2008). Estimating granger causality after stimulus onset: a cautionary note. *Neuroimage*, 41(3):767–776. 5

Weisstein, E. W. (2017). Harmonic addition theorem. From MathWorld—A Wolfram Web Resource. Last visited on 2/6/2017. 115

Xu, Y., Sudre, G. P., Wang, W., Weber, D. J., and Kass, R. E. (2011). Characterizing global statistical significance of spatiotemporal hot spots in magnetoencephalography/electroencephalography source space via excursion algorithms. *Statistics in medicine*, 30(23):2854–2866. 14

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847. 30

Yu, M., Kolar, M., and Gupta, V. (2016). Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, pages 2829–2837. 40, 41

Yu, S., Drton, M., and Shojaie, A. (2018). Graphical models for non-negative data using generalized score matching. *arXiv preprint arXiv:1802.06340*. 38, 40, 121

Zemel, R. S., Williams, C. K. I., and Mozer, M. C. (1993). Directional-unit Boltzmann machines. *Advances in neural information processing systems*. 31

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261. 73

Appendix

Appendix A

Appendix to Chapter 3

A.1 Proof of Theorem 3.1 (Torus graph model)

To derive the appropriate first- and second-order sufficient statistics that correspond to first circular moments and to circular covariances, we first represent the angles as unit modulus complex random variables (where i is the imaginary unit):

$$Z_j = e^{iX_j}.$$

The first circular moment is defined as

$$E[Z_j] = R_j e^{i\mu_j}$$

where μ_j is the mean direction and R_j is the resultant length, so the corresponding complex first order sufficient statistic for a single observation is simply

$$z_j = \cos(x_j) + i\sin(x_j)$$

which may be described as a real-valued sufficient statistic vector

$$\mathbf{S}_j^1(x_j) = [\cos(x_j),\, \sin(x_j)]^T.$$

When considering second-order interactions between complex variables, there are two types of covariance (Schreier and Scharf, 2010, Ch. 2.2). Rotational covariance between Z_j and Z_k is described by

$$E[(e^{iX_j} - R_j e^{i\mu_j})\overline{(e^{iX_k} - R_k e^{i\mu_k})}] = E[e^{i(X_j - X_k)}] - R_j R_k e^{i(\mu_j - \mu_k)},$$

where $\overline{Z_k}$ is the complex conjugate, while reflectional covariance is described by

$$E[(e^{iX_j} - e^{i\mu_j})(e^{iX_k} - e^{i\mu_k})] = E[e^{i(X_j + X_k)}] - R_j R_k e^{i(\mu_j + \mu_k)}.$$

This shows that in addition to the first-order statistics, we additionally need two more complex sufficient statistics to describe the second-order behavior:

$$\begin{aligned} e^{i(x_j - x_k)} &= \cos(x_j - x_k) + i\sin(x_j - x_k) \\ e^{i(x_j + x_k)} &= \cos(x_j + x_k) + i\sin(x_j + x_k). \end{aligned}$$

These may be collected into a real-valued vector

$$\mathbf{S}^2_{jk}(x_j, x_k) = [\cos(x_i - x_j),\ \sin(x_i - x_j),\ \cos(x_i + x_j),\ \sin(x_i + x_j)]^T.$$

Therefore, the canonical exponential family distribution given the first circular moments of each variable and the complete second-order interactions (rotational and reflectional) between each variable coincides with that given in 3.3:

$$p(\mathbf{x}) \propto \exp\left\{ \sum_{j=1}^{d} \boldsymbol{\phi}_j^T \mathbf{S}^1_j(x_j) + \sum_{j<k} \boldsymbol{\phi}_{jk}^T \mathbf{S}^2_{jk}(x_j, x_k) \right\} \tag{A.1}$$

$$= \exp\left\{ \sum_{j=1}^{d} \boldsymbol{\phi}_j^T \begin{bmatrix} \cos(x_j) \\ \sin(x_j) \end{bmatrix} + \sum_{j<k} \boldsymbol{\phi}_{jk}^T \begin{bmatrix} \cos(x_j - x_k) \\ \sin(x_j - x_k) \\ \cos(x_j + x_k) \\ \sin(x_j + x_k) \end{bmatrix} \right\}. \tag{A.2}$$

Thus, the torus graph model is maximum entropy with respect to constraints on the expected values of the sufficient statistics, that is, the circular first moments and complex covariances Wainwright et al. (2008). We note that, similar to the multivariate Gaussian distribution, the torus graph model only contains sufficient statistics for circular first moments and covariances, but it does not contain sufficient statistics corresponding to the second circular moment of a single angle X_j (that is, it does not include interactions of the form $Z_j Z_j$); such a model was recently explored in Navarro et al. (2017).

The maximum entropy motivation for this model also offers some intuition for interpretation of the parameters; in particular, we see that the subvector $\boldsymbol{\phi}_{jk,1:2}$ corresponds to rotational covariance while the subvector $\boldsymbol{\phi}_{jk,3:4}$ corresponds to reflectional covariance. However, the magnitude of each parameter is difficult to interpret directly because it depends not only on the covariance but also the marginal concentration of each variable (which is related to the resultant lengths R_j and R_k) as well as the sum or difference of the mean directions.

A.2 Reparameterization of torus graphs to compare to previous work

While the canonical exponential family form in Equation A.2 is useful for understanding the maximum entropy constraints of the model and for deriving score matching estimators, it does not immediately appear similar to previous work in multivariate circular statistics (such as the sine model). To obtain another form that is easier to compare to previous work, we begin with an alternate parameterization that is similar to the sine model, then show how it can be transformed into our parameterization. Crucially, this transformation can also be reversed to potentially aid in interpretation of parameters.

Consider the *mean-centered parameterization*

$$p(\mathbf{x};\boldsymbol{\theta}) \propto \exp\left\{ \sum_{j=1}^{d} \kappa_j \cos(x_j - \mu_j) + \sum_{j<k} \begin{bmatrix} \lambda_{jk}^{cc} \\ \lambda_{jk}^{cs} \\ \lambda_{jk}^{sc} \\ \lambda_{jk}^{ss} \end{bmatrix}^T \begin{bmatrix} \cos(x_j - \mu_j)\cos(x_k - \mu_k) \\ \cos(x_j - \mu_j)\sin(x_k - \mu_k) \\ \sin(x_j - \mu_j)\cos(x_k - \mu_k) \\ \sin(x_j - \mu_j)\sin(x_k - \mu_k) \end{bmatrix} \right\}$$

where the parameters are $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\lambda}^{\boldsymbol{cc}}, \boldsymbol{\lambda}^{\boldsymbol{cs}}, \boldsymbol{\lambda}^{\boldsymbol{sc}}, \boldsymbol{\lambda}^{\boldsymbol{ss}}]$ with the interpretation that $\mu_j \in [0, 2\pi)$ is the mean direction of x_j, $\kappa_j > 0$ is the marginal concentration of x_j, and the λ parameters control interactions between angles.

In the univariate terms, we use trigonometric sum and difference formulas to rewrite

$$\kappa_j \cos(x_j - \mu_j) = \kappa_j \cos(\mu_j)\cos(x_j) + \kappa_j \sin(\mu_j)\sin(x_j)$$

so that in the parameterization of Equation A.2,

$$\boldsymbol{\phi}_j = \begin{bmatrix} \kappa_j \cos(\mu_j) \\ \kappa_j \sin(\mu_j) \end{bmatrix}.$$

Therefore, we can clearly calculate $\boldsymbol{\phi}_j$ for given κ_j and μ_j, and also, given $\boldsymbol{\phi}_j$, we have (using the Pythagorean theorem and definition of tangent)

$$\begin{aligned}\kappa_j &= \sqrt{\boldsymbol{\phi}_{j,1}^2 + \boldsymbol{\phi}_{j,2}^2} \\ \mu_j &= \arctan\left(\frac{\boldsymbol{\phi}_{j,2}}{\boldsymbol{\phi}_{j,1}}\right).\end{aligned}$$

Thus we have demonstrated a diffeomorphism between the two parameterizations for the marginal terms.

Similarly, for the pairwise coupling terms, we consider without loss of generality the pair $\{X_j, X_k\}$ but for simplicity drop subscripts on the λ parameters; using trigonometric sum and difference identities and simplifying, we find the pairwise coupling term is

$$\frac{1}{2}\begin{bmatrix}(\lambda^{cc}+\lambda^{ss})\cos(\mu_j-\mu_k)+(\lambda^{cs}-\lambda^{sc})\sin(\mu_j-\mu_k)\\ (\lambda^{sc}-\lambda^{cs})\cos(\mu_j-\mu_k)+(\lambda^{cc}+\lambda^{ss})\sin(\mu_j-\mu_k)\\ (\lambda^{cc}-\lambda^{ss})\cos(\mu_j+\mu_k)+(-\lambda^{cs}-\lambda^{sc})\sin(\mu_j+\mu_k)\\ (\lambda^{cs}+\lambda^{sc})\cos(\mu_j+\mu_k)+(\lambda^{cc}-\lambda^{ss})\sin(\mu_j+\mu_k)\end{bmatrix}^T \begin{bmatrix}\cos(x_j-x_k)\\ \sin(x_j-x_k)\\ \cos(x_j+x_k)\\ \sin(x_j+x_k)\end{bmatrix}$$

so that it is straightforward to calculate $\boldsymbol{\phi}_{jk}$ given $\mu_j, \mu_k, \kappa_j, \kappa_k$, and the four λ parameters.

The λ parameters may be recovered as follows, where for brevity we use $\mu_{jk}^- = \mu_j - \mu_k$ and $\mu_{jk}^+ = \mu_j + \mu_k$:

$$\begin{aligned}\lambda^{cc} &= \boldsymbol{\phi}_{jk,1}\cos(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,2}\sin(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,3}\cos(\mu_{jk}^+) + \boldsymbol{\phi}_{jk,4}\sin(\mu_{jk}^+)\\ \lambda^{cs} &= -\boldsymbol{\phi}_{jk,2}\cos(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,1}\sin(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,4}\cos(\mu_{jk}^+) - \boldsymbol{\phi}_{jk,3}\sin(\mu_{jk}^+)\\ \lambda^{sc} &= \boldsymbol{\phi}_{jk,2}\cos(\mu_{jk}^-) - \boldsymbol{\phi}_{jk,1}\sin(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,4}\cos(\mu_{jk}^+) - \boldsymbol{\phi}_{jk,3}\sin(\mu_{jk}^+)\\ \lambda^{ss} &= \boldsymbol{\phi}_{jk,1}\cos(\mu_{jk}^-) + \boldsymbol{\phi}_{jk,2}\sin(\mu_{jk}^-) - \boldsymbol{\phi}_{jk,3}\cos(\mu_{jk}^+) - \boldsymbol{\phi}_{jk,4}\sin(\mu_{jk}^+).\end{aligned}$$

This shows we have a diffeomorphism between the parameterizations.

Theorem 4.2.2 of Kass and Vos (2007) states that a subfamily of a regular exponential family is itself a (lower-dimensional) regular exponential family if and only if the subspace of the natural parameter space corresponding to the subfamily is an affine subspace of the natural parameter space. In the mean-centered parameterization, the sine model has parameter constraints $\lambda^{cc} = \lambda^{cs} = \lambda^{sc} = 0$. Given the equations above, the sine model corresponds to a restriction of the natural parameter space of the torus graph density of Equation A.2:

$$\boldsymbol{\phi}_{jk} = \frac{1}{2}\lambda^{ss}[\cos(\mu_{jk}^-),\, \sin(\mu_{jk}^-),\, -\cos(\mu_{jk}^+),\, -\sin(\mu_{jk}^+)]^T. \tag{A.3}$$

This implies that the pairwise interactions must follow a specific structured form in the sine model, where the magnitude of interactions is governed by λ^{ss} and the following relationship between the parameters is observed (regardless of μ_j and μ_k):

$$\phi_{jk,1}^2 + \phi_{jk,2}^2 = \phi_{jk,3}^2 + \phi_{jk,4}^2. \tag{A.4}$$

Because of the nonlinear relationship between the parameters, the subspace corresponding to the sine model is not an affine subspace of the natural parameter space. Therefore, the sine model is not itself a regular exponential family. On the other hand, the uniform marginal model and the phase difference model both are defined by setting components of the natural parameter to zero, as is the phase difference model with uniform marginals, so each of these families is itself a regular exponential family. This proves Theorem 3.3.

A.3 Proof of Corollary 3.1.1 (Torus graph properties)

1. Because exponential family models are maximum entropy models subject to constraints on the expected values of the sufficient statistics (Wainwright et al., 2008), the torus graph is the maximum entropy model subject to constraints on the first circular moments and complex covariances between angles (following the derivation in Section A.1 that relates the sufficient statistics to circular first moments and to complex covariances).

2. The torus graph density is positive and continuous on $[0, 2\pi]^d$ and factorizes into pairwise interaction terms as shown in Equation A.2. By the Hammersley-Clifford theorem (Lauritzen, 1996), the random variables X_j and X_k are conditionally independent given all other variables if and only if $\boldsymbol{\phi}_{jk} = \mathbf{0}$.

A.4 Derivations of phase difference densities from torus graph models

First, we state the Harmonic Addition Theorem which will be very useful throughout this set of derivations (see Weisstein (2017) for proof).

Theorem A.1 (Harmonic Addition Theorem)**.** *The weighted sum of cosine functions with the same period and arbitrary phase shifts is the cosine function*

$$\sum_{i=1}^{n} a_i \cos(x - \delta_i) = A \cos(x - \Delta)$$

where

$$
\begin{aligned}
b_x &= \sum_{i=1}^{n} a_i \cos(\delta_i) \\
b_y &= \sum_{i=1}^{n} a_i \sin(\delta_i) \\
A &= \sqrt{b_x^2 + b_y^2} \\
\Delta &= \arctan\left(\frac{b_y}{b_x}\right).
\end{aligned}
$$

Throughout these derivations, when we use the arctangent function $\arctan(\cdot)$, it is understood that the angular value is chosen to fall in the same interval as the random variables (in this case, $[0, 2\pi]$, though other intervals such as $[-\pi, \pi]$ could be chosen). We will use the notation $\boldsymbol{\phi}_{jk} = [\alpha_{jk}, \beta_{jk}, \gamma_{jk}, \delta_{jk}]^T$ to refer to elements of the pairwise coupling parameter vector.

For the bivariate torus graph model, we derive the distribution of phase differences to compare with bivariate phase coupling measures, which depend on the distribution of phase differences. Let $\theta = X_1 - X_2$ be a random variable with support $[-2\pi, 2\pi]$ (as $X_1 \in [0, 2\pi]$ and $X_2 \in [0, 2\pi]$) and let $p_{X_1,X_2}(x_1, x_2)$ denote the bivariate phase difference model density. Applying the change of variables $\theta = X_1 - X_2$ and trigonometric identity $\sin(\theta) = \cos\left(\theta - \frac{\pi}{2}\right)$ yields

$$
\begin{aligned}
p_{\theta,X_2}(\theta, x_2) &= p_{X_1,X_2}(\theta + x_2, x_2) \\
&\propto \exp\left\{\kappa_1 \cos(x_2 - (\mu_1 - \theta)) + \kappa_2 \cos(x_2 - \mu_2)\right\} \times \\
&\qquad \exp\left\{\alpha_{12} \cos(\theta) + \beta_{12} \cos\left(\theta - \tfrac{\pi}{2}\right)\right\}.
\end{aligned}
$$

Applying Theorem A.1 to each factor,

$$
p_{X_1,X_2}(\theta + x_2, x_2) \propto \exp\left\{A_1 \cos(\theta - \Delta_1)\right\} \exp\left\{A_2 \cos(x_2 - \Delta_2)\right\}
$$

where

$$
\begin{aligned}
A_1 &= \sqrt{\alpha_{12}^2 + \beta_{12}^2} \\
\Delta_1 &= \arctan\left(\tfrac{\beta_{12}}{\alpha_{12}}\right) \\
A_2(\theta) &= \sqrt{\kappa_1^2 + \kappa_2^2 + 2\kappa_1\kappa_2 \cos(\theta - (\mu_1 - \mu_2))} \\
\Delta_2(\theta) &= \arctan\left(\tfrac{\kappa_1 \sin(\mu_1 - \theta) + \kappa_2 \sin(\mu_2)}{\kappa_1 \cos(\mu_1 - \theta) + \kappa_2 \cos(\mu_2)}\right),
\end{aligned}
$$

and we use the notation $A_2(\theta), \Delta_2(\theta)$ to indicate that these are functions of θ.

To obtain the marginal density of θ, we need to integrate over x_2 and also wrap the resulting distribution back to the support $[0, 2\pi]$ (though we could choose a different support of length 2π, such as $[-\pi, \pi]$, if desired). Notice that $p_{X_1,X_2}(\theta + x_2, x_2)$ has constraints $X_1, X_2 \in [0, 2\pi]$, implying that $0 \leq \theta + X_2 \leq 2\pi$ so $-\theta \leq X_2 \leq 2\pi - \theta$. This means that when $\theta < 0$, $X_2 \in [-\theta, 2\pi]$ and when $\theta > 0$, $X_2 \in [0, 2\pi - \theta]$, so the marginal distribution of θ is defined piecewise:

$$p_\theta(\theta) \propto \begin{cases} \mathbb{1}_{(\theta\in[-2\pi,0])} g(\theta) \int_{-\theta}^{2\pi} \exp\left\{A_2(\theta)\cos(x_2 - \Delta_2(\theta))\right\} dx_2 \\ \mathbb{1}_{(\theta\in[0,2\pi])} g(\theta) \int_{0}^{2\pi-\theta} \exp\left\{A_2(\theta)\cos(x_2 - \Delta_2(\theta))\right\} dx_2 \end{cases}$$

where $g(\theta) = \exp\left\{A_1 \cos(\theta - \Delta_1)\right\}$.

Define $W = \theta \,(\mathrm{mod}\, 2\pi)$ to be the wrapped version of θ so that $W \in [0, 2\pi]$ has the wrapped distribution

$$\begin{aligned} p_W(w) &= p_\theta(w) + p_\theta(w - 2\pi) \\ &\propto g(w) \int_0^{2\pi - w} \exp\left\{A_2(w)\cos(x_2 - \Delta_2(w))\right\} dx_2 \\ &\quad + g(w - 2\pi) \int_{2\pi - w}^{2\pi} \exp\left\{A_2(w - 2\pi)\cos(x_2 - \Delta_2(w - 2\pi))\right\} dx_2 \end{aligned}$$

Using the fact that g, A_2, and Δ_2 are 2π-periodic functions and the definition of I_0 (the modified Bessel function of the first kind), we obtain

$$p_W(w) \propto g(w) \int_0^{2\pi} \exp\left\{A_2(w)\cos(x_2 - \Delta_2(w))\right\} dx_2 = g(w) I_0(A_2(w)).$$

Thus, $g(w)$ is the direct coupling term and $f(w) = I_0(A_2(w))$ is the marginal concentration term.

The derivation for phase differences from the trivariate torus graph model uses similar techniques. Consider a trivariate torus graph with $\kappa_1 = \kappa_2 = \kappa_3 = 0$ for simplicity; applying trigonometric identities yields

$$\begin{aligned} p(x_1, x_2, x_3) &\propto \exp\left\{ \sum_{(i,j)\in E} \begin{bmatrix} \alpha_{ij} \\ \beta_{ij} \\ \gamma_{ij} \\ \delta_{ij} \end{bmatrix}^T \begin{bmatrix} \cos(x_i - x_j) \\ \sin(x_i - x_j) \\ \cos(x_i + x_j) \\ \sin(x_i + x_j) \end{bmatrix} \right\} \\ &= \exp\left\{ \sum_{(i,j)\in E} \begin{bmatrix} \alpha_{ij} \\ \beta_{ij} \\ \gamma_{ij} \\ \delta_{ij} \end{bmatrix}^T \begin{bmatrix} \cos(x_i - x_j - 0) \\ \cos(x_i - x_j - \pi/2) \\ \cos(x_i + x_j - 0) \\ \cos(x_i + x_j - \pi/2) \end{bmatrix} \right\} \end{aligned}$$

where $E = \{(1,2),\ (1,3),\ (2,3)\}$. Apply the variable transformation $\theta_{12} = x_1 - x_2$ and expand the expression above:

$$p(\theta_{12}, x_2, x_3) \propto \exp\left\{ \begin{bmatrix} \alpha_{12} \\ \beta_{12} \\ \gamma_{12} \\ \delta_{12} \end{bmatrix}^T \begin{bmatrix} \cos(\theta_{12} - 0) \\ \cos(\theta_{12} - \pi/2) \\ \cos(\theta_{12} + 2x_2 - 0) \\ \cos(\theta_{12} + 2x_2 - \pi/2) \end{bmatrix} \right\} \times$$
$$\exp\left\{ \begin{bmatrix} \alpha_{13} \\ \beta_{13} \\ \gamma_{13} \\ \delta_{13} \end{bmatrix}^T \begin{bmatrix} \cos(x_1 - x_3 - 0) \\ \cos(x_1 - x_3 - \pi/2) \\ \cos(x_1 + x_3 - 0) \\ \cos(x_1 + x_3 - \pi/2) \end{bmatrix} \right\} \times$$
$$\exp\left\{ \begin{bmatrix} \alpha_{23} \\ \beta_{23} \\ \gamma_{23} \\ \delta_{23} \end{bmatrix}^T \begin{bmatrix} \cos(x_2 - x_3 - 0) \\ \cos(x_2 - x_3 - \pi/2) \\ \cos(x_2 + x_3 - 0) \\ \cos(x_2 + x_3 - \pi/2) \end{bmatrix} \right\}$$

To get the marginal distribution of θ_{12}, we need to integrate out other variables, which is not tractable analytically for the full torus graph model, so we consider the phase difference model which corresponds to setting $\gamma = \delta = 0$ for all pairs. This has the effect of making the density depend only on phase differences.

$$p(\theta_{12}, x_2, x_3) \propto \exp\left\{ \begin{bmatrix} \alpha_{12} \\ \beta_{12} \end{bmatrix}^T \begin{bmatrix} \cos(\theta_{12} - 0) \\ \cos(\theta_{12} - \pi/2) \end{bmatrix} \right\} \times$$
$$\exp\left\{ \begin{bmatrix} \alpha_{13} \\ \beta_{13} \end{bmatrix}^T \begin{bmatrix} \cos(x_1 - x_3 - 0) \\ \cos(x_1 - x_3 - \pi/2) \end{bmatrix} \right\} \times$$
$$\exp\left\{ \begin{bmatrix} \alpha_{23} \\ \beta_{23} \end{bmatrix}^T \begin{bmatrix} \cos(x_2 - x_3 - 0) \\ \cos(x_2 - x_3 - \pi/2) \end{bmatrix} \right\}.$$

Similarly to the bivariate case, we apply Theorem A.1 to each factor; the first factor, $g(\theta_{12}) = \exp\{A_1 \cos(\theta_{12} - \Delta_1)\}$, has the same form as in the bivariate case and represents the direct coupling between X_1 and X_2. The second factor may also be written as

$$\exp\{A_{13} \cos(x_1 - x_3 - \Delta_{13}\}$$

where $A_{13} = \sqrt{\alpha_{13}^2 + \beta_{13}^2}$ and $\Delta_{13} = \arctan(\beta_{13}/\alpha_{13})$, and the third factor may also be written as

$$\exp\left\{A_{23}\cos(x_2 - x_3 - \Delta_{23}\right\},$$

where $A_{23} = \sqrt{\alpha_{23}^2 + \beta_{23}^2}$ and $\Delta_{23} = \arctan(\beta_{23}/\alpha_{23})$. Next we apply Theorem A.1 again to combine the second and third factors:

$$p(\theta_{12}, x_3) \propto \exp\left\{A_1\cos(\theta_{12} - \Delta_1)\right\} \times \exp\left\{A_3(\theta_{12})\cos(x_3 - \Delta_3)\right\},$$

where $A_3(\theta_{12})$ is

$$\begin{aligned} A_3(\theta_{12}) = \Bigg[&\Big(A_{13}\cos(x_1 - \Delta_{13}) + A_{23}\cos(x_2 - \Delta_{23})\Big)^2 \\ &+ \Big(A_{13}\sin(x_1 - \Delta_{13}) + A_{23}\sin(x_2 - \Delta_{23})\Big)^2\Bigg]^{1/2}. \end{aligned}$$

Because we will integrate out x_3, the form of Δ_3 is not important (it will not affect the integral on this circular domain).

Expanding the squares, simplifying, and using a trigonometric sum identity yields

$$A_3(\theta_{12}) = \sqrt{\alpha_{13}^2 + \beta_{13}^2 + \alpha_{23}^2 + \beta_{23}^2 + 2t\cos(\theta_{12} - u)} \tag{A.5}$$

where

$$\begin{aligned} t &= \sqrt{(\alpha_{13}^2 + \beta_{13}^2)(\alpha_{23}^2 + \beta_{23}^2)} \\ u &= \Delta_{13} - \Delta_{23} = \arctan\left(\frac{\beta_{13}}{\alpha_{13}}\right) - \arctan\left(\frac{\beta_{23}}{\alpha_{23}}\right). \end{aligned}$$

Similar to the bivariate case, we integrate over x_3 and wrap the resulting distribution to obtain the marginal distribution of the wrapped phase difference $W = \theta_{12} \pmod{2\pi}$:

$$\begin{aligned} p(w) &\propto \exp\left\{A_1\cos\left(w - \Delta_1\right)\right\} \int_0^{2\pi} \exp\left\{A_3(w)\cos\left(x_3 - \Delta_3\right)\right\}\, dx_3 \\ &\propto \exp\left\{A_1\cos\left(w - \Delta_1\right)\right\} I_0(A_3(w)) \end{aligned}$$

where $A_1 = \sqrt{\alpha_{12}^2 + \beta_{12}^2}$, $\Delta_1 = \arctan(\beta_{12}/\alpha_{12})$, and $A_3(w)$ is given in Equation A.5. Thus, $g(w)$ is the direct coupling term, and $h(w)$ is the indirect coupling term.

A.5 Proof of Theorem 3.4 (Score matching estimators for torus graphs)

Let $p_{\mathbf{X}}(\mathbf{x})$ be the unknown d-dimensional circular data density and $p(\mathbf{x};\boldsymbol{\phi}) = \frac{1}{Z(\boldsymbol{\phi})} q(\mathbf{x};\boldsymbol{\phi})$ be a d-dimensional model density with parameter vector $\boldsymbol{\phi} \in \mathbb{R}^m$. Define the log model density gradient $\boldsymbol{\psi} : [0, 2\pi)^d \to \mathbb{R}^d$ as $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi}) = \nabla_{\mathbf{x}} \log q(\mathbf{x};\boldsymbol{\phi})$; similarly, let $\boldsymbol{\psi}_{\mathbf{X}}(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_{\mathbf{X}}(\mathbf{x})$.

To prove Theorem 3.4, make the following regularity assumptions:

A. For all $i \in \{1, ..., d\}$, $\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi})$ is differentiable with respect to $\mathbf{x}_i$ on $(0, 2\pi)$.

B. For all $\boldsymbol{\phi}$, $E_{\mathbf{x}}\left[||\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi})||^2\right]$ and $E_{\mathbf{x}}\left[||\boldsymbol{\psi}_{\mathbf{X}}(\mathbf{x})||^2\right]$ are finite.

These assumptions clearly hold for torus graphs as the log density is comprised of finite linear combinations of sine and cosine functions of $\mathbf{x}$, each of which is infinitely differentiable with derivatives bounded within $[-1, 1]$. Note that we need one less assumption than the original formulation of score matching in Hyvärinen (2005b) due to the circular nature of the density.

Proof of Theorem 3.4. First, we show that the score matching objective function only depends on the unknown data density through an expectation.

Expanding the squared difference gives

$$\begin{aligned} J(\boldsymbol{\phi}) = &\int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) \left[\tfrac{1}{2}||\nabla_{\mathbf{x}} \log p_{\mathbf{X}}(\mathbf{x})||_2^2\right] \, d\mathbf{x} \\ &+ \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) \left[\tfrac{1}{2}||\nabla_{\mathbf{x}} \log q(\mathbf{x};\boldsymbol{\phi})||_2^2\right] \, d\mathbf{x} \\ &- \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) [\nabla_{\mathbf{x}} \log q(\mathbf{x};\boldsymbol{\phi})]^T [\nabla_{\mathbf{x}} \log p_{\mathbf{X}}(\mathbf{x})] \, d\mathbf{x}. \end{aligned}$$

The first term does not depend on $\boldsymbol{\phi}$ and the second term is already in terms of an expectation over the data density, so we focus now on the third term (call it A):

$$\begin{aligned} A &= -\int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) \left[\sum_{i=1}^d \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi}) \boldsymbol{\psi}_{\mathbf{X},i}(\mathbf{x})\right] \, d\mathbf{x} \\ &= -\sum_{i=1}^d \int_0^{2\pi} \left[\int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi}) \boldsymbol{\psi}_{\mathbf{X},i}(\mathbf{x}) \, d\mathbf{x}_i\right] \, d\mathbf{x}_{-i} \\ &= -\sum_{i=1}^d \int_0^{2\pi} \left[\int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}_i} \log p_{\mathbf{X}}(\mathbf{x}) \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi}) \, d\mathbf{x}_i\right] \, d\mathbf{x}_{-i} \\ &= -\sum_{i=1}^d \int_0^{2\pi} \left[\int_0^{2\pi} \frac{\partial}{\partial \mathbf{x}_i} p_{\mathbf{X}}(\mathbf{x}) \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi}) \, d\mathbf{x}_i\right] \, d\mathbf{x}_{-i}. \end{aligned}$$

Applying integration by parts, the inner integral becomes

$$p_{\mathbf{X}}(\mathbf{x})\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\Big|_{\mathbf{x}_i=0}^{\mathbf{x}_i=2\pi} - \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x})\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right)\, d\mathbf{x}_i.$$

Notice that because the variables are circular on $[0, 2\pi]$,

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x})\big|_{\mathbf{x}_i=0} &= p_{\mathbf{X}}(\mathbf{x})\big|_{\mathbf{x}_i=2\pi} \\ \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\big|_{\mathbf{x}_i=0} &= \boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\big|_{\mathbf{x}_i=2\pi} \end{aligned}$$

Therefore, $p_{\mathbf{X}}(\mathbf{x})\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\big|_{\mathbf{x}_i=0}^{\mathbf{x}_i=2\pi} = 0$, so A becomes

$$\begin{aligned} A &= -\sum_{i=1}^{d}\int_0^{2\pi}\left[-\int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x})\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right)\, d\mathbf{x}_i\right]\, d\mathbf{x}_{-i} \\ &= \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x})\left[\sum_{i=1}^{d}\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right)\right]\, d\mathbf{x} \end{aligned}$$

Therefore, the score matching objective is

$$\begin{aligned} J(\boldsymbol{\phi}) &= C + \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x})\left[\tfrac{1}{2}||\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi})||^2\right]\, d\mathbf{x} \\ &\quad + \int_0^{2\pi} p_{\mathbf{X}}(\mathbf{x})\left[\sum_{i=1}^{d}\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right)\right]\, d\mathbf{x} \\ &= C + E_{\mathbf{x}}\left\{\tfrac{1}{2}||\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi})||^2 + \sum_{i=1}^{d}\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right)\right\} \end{aligned} \tag{A.6}$$

where C does not depend on $\boldsymbol{\phi}$ and may be ignored without affecting the maxima of the objective function. This coincides with the form of score matching given in Hyvärinen (2005b) except with the integral over the circular domain $[0, 2\pi]$.

Next, we show the explicit form of the score matching estimator for torus graphs. As shown in Forbes and Lauritzen (2015); Yu et al. (2018), for exponential families, this score matching estimator is quadratic in the parameters. Specifically, the torus graph density in Theorem 3.1 has a log density of the form

$$\log q(\mathbf{x};\boldsymbol{\phi}) = \boldsymbol{\phi}^T\mathbf{S}(\mathbf{x})$$

where $\boldsymbol{\phi}$ are vectors of length $m = 2d^2$ (the number of sufficient statistics).

Therefore,

$$\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi}) = \boldsymbol{\phi}^T\mathbf{D}(\mathbf{x})$$

where the Jacobian $\mathbf{D}(\mathbf{x})$ is $m \times d$ with i, jth element $\frac{\partial}{\partial \mathbf{x}_j} \mathbf{S}_i$. Thus the first term inside the expectation in the score matching objective of Equation A.6 may be written

$$\tfrac{1}{2}||\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\phi})||^2 = \tfrac{1}{2}\boldsymbol{\phi}^T \mathbf{D}(\mathbf{x})\mathbf{D}(\mathbf{x})^T \boldsymbol{\phi} \equiv \tfrac{1}{2}\boldsymbol{\phi}^T \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\phi}.$$

The elements of $\mathbf{D}(\mathbf{x})$ correspond to partial derivatives of the sufficient statistics with respect to the data. The derivatives of the univariate sufficient statistics $\mathbf{S}^1$ are given by

$$\frac{\partial}{\partial x_\ell}\cos(x_j) = \begin{cases} -\sin(x_j), & \ell = j \\ 0, & \ell \neq j \end{cases},$$

$$\frac{\partial}{\partial x_\ell}\sin(x_j) = \begin{cases} \cos(x_j), & \ell = j \\ 0, & \ell \neq j \end{cases}.$$

Similarly, the derivatives of the pairwise sufficient statistics $\mathbf{S}^2$ may be calculated as

$$\frac{\partial}{\partial x_\ell}\cos(x_j - x_k) = \begin{cases} -\sin(x_j - x_k), & \ell = j \\ \sin(x_j - x_k), & \ell = k \\ 0, & \ell \notin \{j,k\} \end{cases},$$

$$\frac{\partial}{\partial x_\ell}\sin(x_j - x_k) = \begin{cases} \cos(x_j - x_k), & \ell = j \\ -\cos(x_j - x_k), & \ell = k \\ 0, & \ell \notin \{j,k\} \end{cases},$$

$$\frac{\partial}{\partial x_\ell}\cos(x_j + x_k) = \begin{cases} -\sin(x_j + x_k), & \ell \in \{j,k\} \\ 0, & \ell \notin \{j,k\} \end{cases},$$

$$\frac{\partial}{\partial x_\ell}\sin(x_j + x_k) = \begin{cases} \cos(x_j + x_k), & \ell \in \{j,k\} \\ 0, & \ell \notin \{j,k\} \end{cases}.$$

Now we show that the second term inside the expectation in Equation A.6 may be written simply in terms of the sufficient statistics. Notice that the ith element of the gradient may be written in terms of columns of the Jacobian:

$$\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi}) = \boldsymbol{\phi}^T[\mathbf{D}(\mathbf{x})]_{\cdot,i}$$

so that

$$\frac{\partial}{\partial \mathbf{x}_i}\left(\boldsymbol{\psi}_i(\mathbf{x};\boldsymbol{\phi})\right) = \boldsymbol{\phi}^T \frac{\partial}{\partial \mathbf{x}_i}[\mathbf{D}(\mathbf{x})]_{\cdot,i}.$$

Therefore, the second term inside the expectation in the score matching objective may be written

$$\sum_{i=1}^{d} \frac{\partial}{\partial \mathbf{x}_i} \left(\boldsymbol{\psi}_i(\mathbf{x}; \boldsymbol{\phi})\right) = \boldsymbol{\phi}^T \left[\sum_{i=1}^{d} \frac{\partial}{\partial \mathbf{x}_i} [\mathbf{D}(\mathbf{x})]_{\cdot,i}\right] \equiv \boldsymbol{\phi}^T \mathbf{H}(\mathbf{x})$$

where

$$\mathbf{H}(\mathbf{x}) = [\mathbf{S}^1(\mathbf{x}),\, 2\mathbf{S}^2(\mathbf{x})]^T.$$

This relation holds because all nonzero elements of $\mathbf{D}(\mathbf{x})$ come from derivatives of sines and cosines; due to the relations $\frac{d}{dx}\cos(x) = -\sin(x)$ and $\frac{d}{dx}\sin(x) = \cos(x)$, taking the derivative again essentially converts the elements back to sufficient statistics. □

A.6 Proof of Theorem 3.2 (Conditional distributions in torus graphs)

We prove that the distribution of one angle conditional on the other angles is von Mises as stated in Theorem 3.2, enabling the use of Gibbs sampling for drawing samples from the distribution. We will use the notation $\boldsymbol{\phi}_{jk} = [\alpha_{jk}, \beta_{jk}, \gamma_{jk}, \delta_{jk}]^T$ to refer to elements of the pairwise coupling parameter vector.

Proof. Let $c_{ij}^- = \cos(x_i - x_j)$, $c_{ij}^+ = \cos(x_i + x_j)$, $s_{ij}^- = \sin(x_i - x_j)$, and $s_{ij}^+ = \sin(x_i + x_j)$. Factor the torus graph density into terms containing X_k and not containing X_k:

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\phi}) =& C(\boldsymbol{\phi}) \exp\left\{\sum_{i \neq k} \kappa_i \cos(x_i - \mu_i) + \sum_{i<j, j\neq k} \begin{bmatrix}\alpha_{ij}\\ \beta_{ij}\\ \gamma_{ij}\\ \delta_{ij}\end{bmatrix}^T \begin{bmatrix}c_{ij}^-\\ s_{ij}^-\\ c_{ij}^+\\ s_{ij}^+\end{bmatrix}\right\} \times \\ & \exp\left\{\kappa_k \cos(x_k - \mu_k)\right\} \times \qquad\qquad \text{(A.7)} \\ & \exp\left\{\sum_{i<k} \begin{bmatrix}\alpha_{ik}\\ \beta_{ik}\\ \gamma_{ik}\\ \delta_{ik}\end{bmatrix}^T \begin{bmatrix}c_{ik}^-\\ s_{ik}^-\\ c_{ik}^+\\ s_{ik}^+\end{bmatrix} + \sum_{i>k} \begin{bmatrix}\alpha_{ki}\\ \beta_{ki}\\ \gamma_{ki}\\ \delta_{ki}\end{bmatrix}^T \begin{bmatrix}c_{ki}^-\\ s_{ik}^-\\ c_{ik}^+\\ s_{ik}^+\end{bmatrix}\right\}. \end{aligned}$$

Let the first factor (including the normalization constant) be denoted by $g(\mathbf{x}_{-d};\boldsymbol{\phi})$ and the second and third factors be denoted by $f(\mathbf{x};\boldsymbol{\phi})$. Then the conditional distribution is

$$p(x_k|\mathbf{x}_{-k};\boldsymbol{\phi}) = \frac{g(\mathbf{x}_{-k};\boldsymbol{\phi})f(\mathbf{x};\boldsymbol{\phi})}{g(\mathbf{x}_{-k};\boldsymbol{\phi})\int f(\mathbf{x};\boldsymbol{\phi})\,dx_k} = \frac{f(\mathbf{x};\boldsymbol{\phi})}{\int f(\mathbf{x};\boldsymbol{\phi})\,dx_k}.$$

Applying trigonometric identities to the third factor of Equation A.7 and simplifying, we have

$$\begin{aligned}
&\exp\left\{\sum_{i<k}\begin{bmatrix}\alpha_{ik}\\ \beta_{ik}\\ \gamma_{ik}\\ \delta_{ik}\end{bmatrix}^T\begin{bmatrix}c_{ik}^-\\ s_{ik}^-\\ c_{ik}^+\\ s_{ik}^+\end{bmatrix}+\sum_{i>k}\begin{bmatrix}\alpha_{ki}\\ \beta_{ki}\\ \gamma_{ki}\\ \delta_{ki}\end{bmatrix}^T\begin{bmatrix}c_{ki}^-\\ s_{ik}^-\\ c_{ik}^+\\ s_{ik}^+\end{bmatrix}\right\}\\
&=\exp\left\{\sum_{i<k}\begin{bmatrix}\alpha_{ik}\\ \beta_{ik}\\ \gamma_{ik}\\ \delta_{ik}\end{bmatrix}^T\begin{bmatrix}\cos(x_k-x_i)\\ \cos(x_k-x_i+\pi/2)\\ \cos(x_k+x_i)\\ \cos(x_k+x_i-\pi/2)\end{bmatrix}\right\}\times\\
&\quad\exp\left\{\sum_{i>k}\begin{bmatrix}\alpha_{ki}\\ \beta_{ki}\\ \gamma_{ki}\\ \delta_{ki}\end{bmatrix}^T\begin{bmatrix}\cos(x_k-x_i)\\ \cos(x_k-x_i-\pi/2)\\ \cos(x_k+x_i)\\ \cos(x_k+x_i-\pi/2)\end{bmatrix}\right\}\\
&=\exp\left\{\sum_{i\neq k}\begin{bmatrix}\alpha_{ik}\\ \beta_{ik}\\ \gamma_{ik}\\ \delta_{ik}\end{bmatrix}^T\begin{bmatrix}\cos(x_k-x_i)\\ \cos(x_k-x_i+\operatorname{sgn}(i-k)\pi/2)\\ \cos(x_k+x_i)\\ \cos(x_k+x_i-\pi/2)\end{bmatrix}\right\}
\end{aligned}$$

where, with slight abuse of notation, we let, for instance, α_{ik} denote either α_{ik} if $i<k$ or α_{ki} if $i>k$, and $\operatorname{sgn}(\cdot)$ is the signum function. Now we see $f(\mathbf{x};\boldsymbol{\phi})$ is a sum of cosine functions with argument x_k, so applying Theorem A.1, we have

$$f(\mathbf{x};\boldsymbol{\phi}) = \exp(A\cos(x_k-\Delta))$$

where $A = \sqrt{b_x^2 + b_y^2}$, $\Delta = \arctan(b_y/b_x)$, defined as

$$
\begin{aligned}
b_x &= \sum_m L_m \cos(V_m) \\
b_y &= \sum_m L_m \sin(V_m) \\
L &= \left[\kappa_k, \boldsymbol{\alpha}_{\cdot,k}, \boldsymbol{\beta}_{\cdot,k}, \boldsymbol{\gamma}_{\cdot,k}, \boldsymbol{\delta}_{\cdot,k}\right] = [\kappa_k, \boldsymbol{\phi}_{\cdot k}] \\
V &= [\mu_k, \mathbf{x}_{-k}, \mathbf{x}_{-k} + \operatorname{sgn}(i-k)\tfrac{\pi}{2}, -\mathbf{x}_{-k}, -\mathbf{x}_{-k} + \tfrac{\pi}{2}] \\
&= [\mu_k, \mathbf{x}_{-k}, \mathbf{x}_{-k} + \mathbf{h}\tfrac{\pi}{2}, -\mathbf{x}_{-k}, -\mathbf{x}_{-k} + \tfrac{\pi}{2}],
\end{aligned}
$$

with, for example, $\boldsymbol{\alpha}_{\cdot,k}$ denoting all α parameters involving index k and $\mathbf{h}_j = -1$ if $j < k$ and $\mathbf{h}_j = 1$ otherwise. Then since $\int f(\mathbf{x}; \boldsymbol{\phi})\, dx_k = 2\pi I_0(A)$ we find that the conditional density is von Mises with concentration A and mean Δ. □

A.7 Measures of positive and negative circular dependence

The dependence between two circular variables can be measured using a correlation coefficient, ρ_c, analogous to the Pearson correlation coefficient for linear analysis Jammalamadaka and Sarma (1988),

$$
\rho_c = \frac{E\{\sin(X_i - \mu_i)\sin(X_j - \mu_j)\}}{\sqrt{\mathrm{Var}(\sin(X_i - \mu_i))\mathrm{Var}(\sin(X_j - \mu_j))}}
$$

where μ represents a mean circular direction. Using variance properties and trigonometric identities we have

$$
\rho_c = \frac{E\{\cos(X_i - X_j - (\mu_i - \mu_j)) - \cos(X_i + X_j - (\mu_i + \mu_j))\}}{2\sqrt{E\{\sin^2(X_i - \mu_i)\}E\{\sin^2(X_j - \mu_j)\}}}. \tag{A.8}
$$

The first component of the numerator measures the positive correlations from the concentration of $X_i - X_j - (\mu_i - \mu_j)$ and the second component measures the negative (or *reflectional*) correlations from the concentration of $X_i - (-X_j) - (\mu_i - (-\mu_j))$. Analogous to the real-valued data, it is important to note that both positive and negative circular correlations are possible and both are needed to fully define circular dependence between two variables.

The numerator of Equation A.8 can be rewritten as

$$
E\left\{\begin{bmatrix}\cos(X_i - X_j) \\ \sin(X_i - X_j)\end{bmatrix}^T \begin{bmatrix}\alpha \\ \beta\end{bmatrix} - \begin{bmatrix}\cos(X_i + X_j) \\ \sin(X_i + X_j)\end{bmatrix}^T \begin{bmatrix}\gamma \\ \delta\end{bmatrix}\right\}
$$

where $\alpha = \cos(\mu_i - \mu_j)$, $\beta = \sin(\mu_i - \mu_j)$, $\gamma = \cos(\mu_i + \mu_j)$, $\delta = \sin(\mu_i + \mu_j)$. This shows that the dependence between angles may be decomposed into a four-term linear combination involving the sines and cosines of the phase differences and phase sums, where the phase difference terms correspond to the positive correlation and the phase sum terms correspond to the negative correlation. This corresponds to the phase sum and phase difference terms that appear in the torus graph density, reinforcing the interpretation of the different pairwise coupling parameters as reflecting positive and negative rotational dependence. To illustrate the distinction between positive and negative rotational dependence, we show bivariate torus graphs with positive, negative, or both kinds of dependence in Figure A.1, and show what trial-to-trial rotational and reflectional covariance would look like in Figure A.2. For the case of uniform marginal distributions, the circular correlation coefficient becomes Jammalamadaka and Sengupta (2001):

$$\rho_c = \frac{R_{Xi-Xj} - R_{Xi+Xj}}{2\sqrt{E\{\sin^2(X_i - \mu_i)\}E\{\sin^2(X_j - \mu_j)\}}}$$

where $R_{Xi-Xj} \equiv |E\{\exp(\mathbf{i}(X_i - X_j))\}|$ corresponds to the positive correlation and $R_{Xi+Xj} \equiv |E\{\exp(\mathbf{i}(X_i + X_j))\}|$ corresponds to the negative correlation. The theoretical Phase Locking Value (PLV), for which an estimator is given in Equation 3.4, is equal to R_{Xi-Xj}. This shows that PLV is similar to a measure of positive circular correlation under the assumption of uniform marginal distributions (when the denominator of the circular correlation coefficient would be equal to 1).

A.8 Torus graph supplementary figures

This section contains a collection of supplementary figures referenced from within Chapter 3.

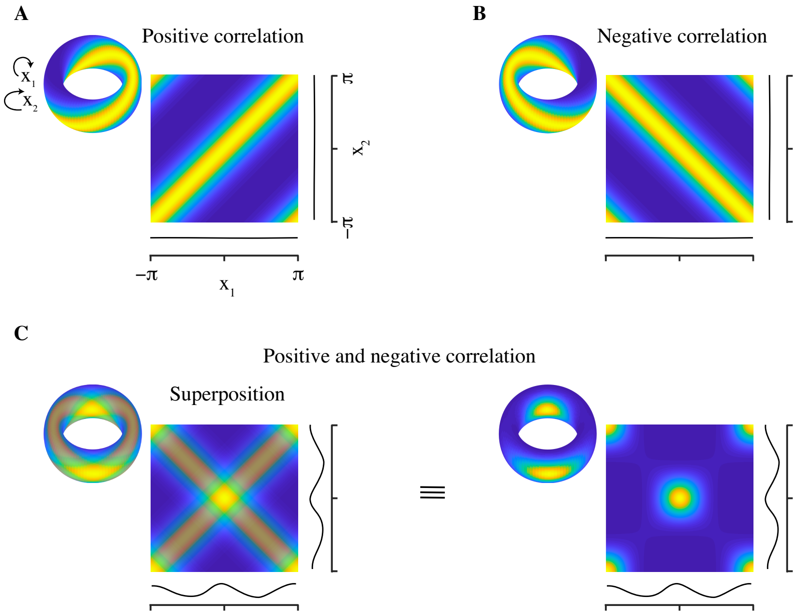

Figure A.1: Bivariate torus graph densities with uniform marginal distributions shown on the torus and flattened on $[-\pi, \pi]$ under positive, negative, or both kinds of rotational covariance. (A) Coupling parameters chosen to induce only positive correlation (coupling based on phase differences). (B) Coupling parameters chosen to induce only negative correlation (coupling based on phase sums). (C) Equal amounts of positive and negative coupling result in a distribution with two isotropic modes; the superposition figure (left) is shown to provide intuition about the resulting distribution (right).

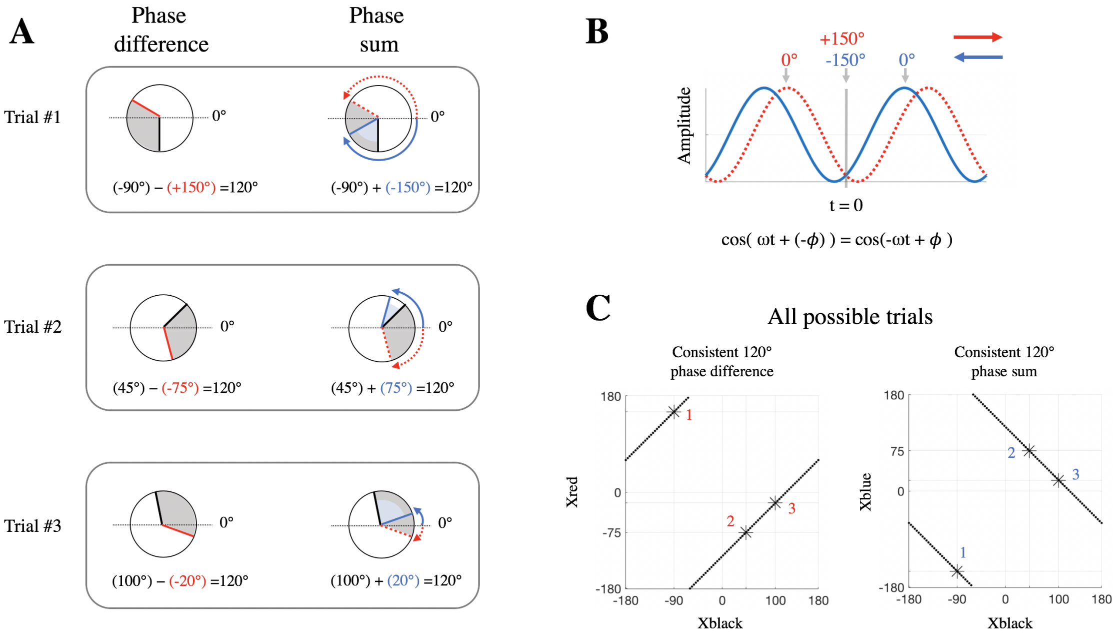

Figure A.2: Intuition about rotational and reflectional dependence. A) Illustration of 120° rotational (positive) and 120° reflectional (negative) dependence across three hypothetical observations. Rotational dependence implies a consistent phase offset between oscillations across trials (shaded gray angles) while reflectional dependence implies a consistent phase sum (shaded blue angles) that corresponds to a consistent phase offset between oscillations after one of the oscillations has been reflected with respect to 0°. B) Example of an oscillation (blue) and its reflection with respect to 0° (dashed red). The reflected signal is leading in time, i.e. $\phi = 150°$, whereas the blue signal is lagging in time, i.e. $\phi = -150°$. This demonstrates that we can think of the reflected signal as moving across time in the opposite direction. In neural data this phenomenon could arise, for example, if there was bidirectional communication. C) The lines indicate dependence between phase angles used in panel A, with rotational dependence on the left and reflectional dependence on the right and trials shown in panel A marked with stars. When phases have uniform marginal distributions across trials, but exhibit phase coupling, positive dependence is observed in the bivariate relationship on the left as a line with fixed orientation at 45°; rotation produces a shift along the anti-diagonal. On the other hand, negative dependence is observed in the bivariate relationship on the right as a line with fixed orientation at 135°; reflection produces a shift along the diagonal. In the current example, the respective shifts are at 120°; see Figure S1 to compare with the case of 0°. Weaker dependence blurs the relationship line but does not change the orientation.

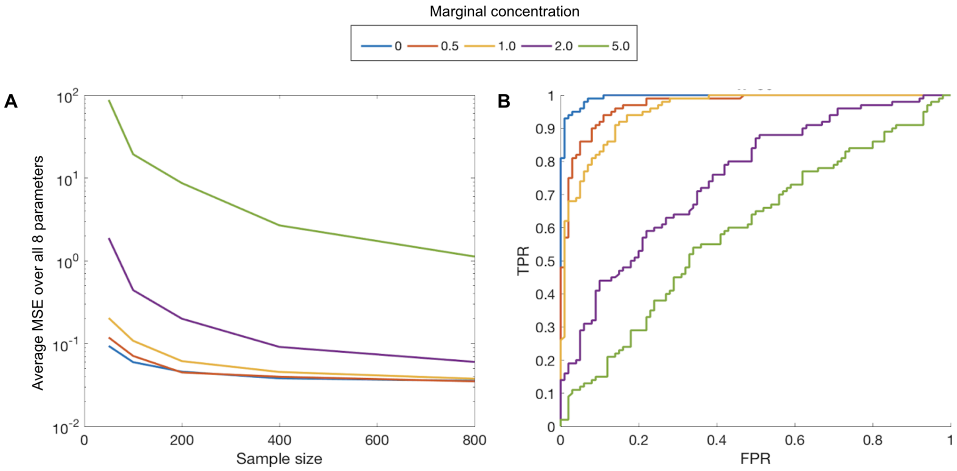

Figure A.3: In data simulated from a bivariate torus graph, the average MSE over all parameters is shown in panel (A) as a function of sample size and marginal concentration. The MSE is higher overall when marginal concentration is high. (B) ROC curves averaged over 200 simulations, half of which had no edge and half of which had an edge between the variables, as a function of marginal concentration for a fixed sample size (N=50), suggesting that structure recovery is also diminished when high marginal concentration is present.

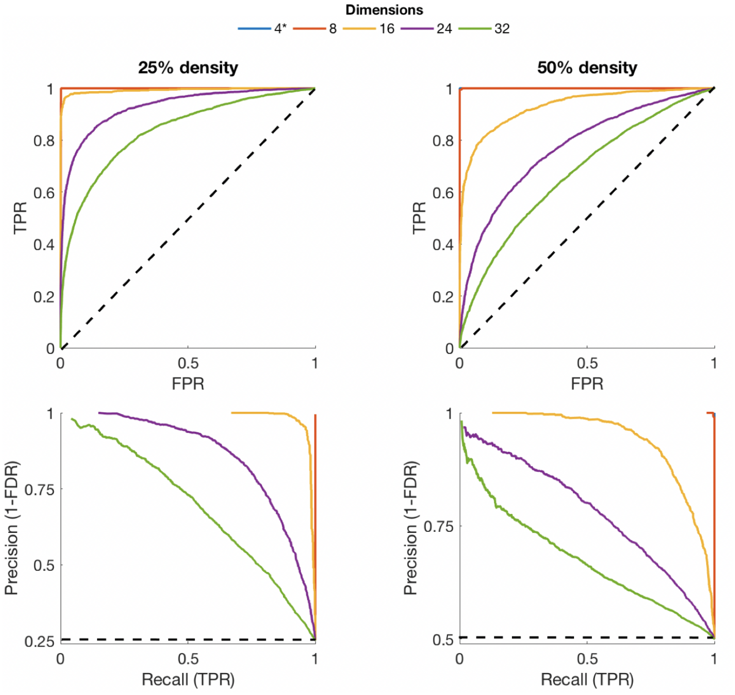

Figure A.4: Further detail on the simulation results shown in Figure 3.5 for a sample size of 840 (matching the real LFP data). Top row: ROC curves colored by dimension for two different underlying edge densities (averaged across 30 simulations). Bottom row: averaged precision curves corresponding to the same densities as the top row.

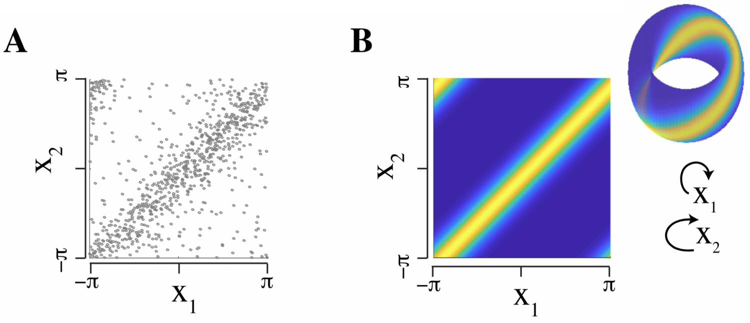

Figure A.5: (A) Scatter plot of data from two channels in dentate gyrus (DG). The angles follow a pattern of positive dependence similar to the simulated data of Figure 3.1, which was used to demonstrate the need for circular wrapping when modeling dependent phase angles. (B) Fitted torus graph density on the plane and torus.

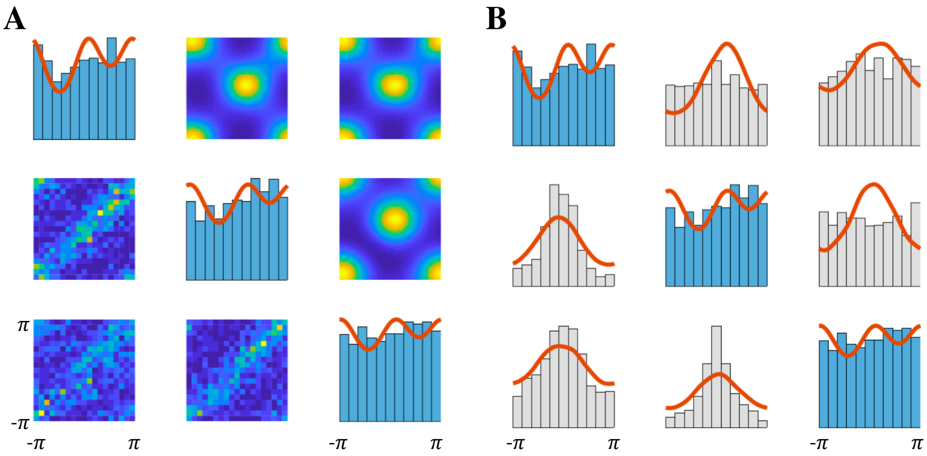

Figure A.6: Similar to Figure 3.9, but using the sine model as the theoretical distribution. The sine model fails to accurately fit this data set, which is evident in the bivariate dependence and sufficient statistics. (A) Along the diagonal are the marginal distributions of the three phase angles, where the real data is represented by a blue histogram and the theoretical marginal densities from the sine model are overlaid as a red line. Two-dimensional distributions (off-diagonal) show bivariate relationships, with theoretical densities above the diagonal and real data represented using a two-dimensional histogram below the diagonal. The multimodal behavior of the sine model is apparent in the two-dimensional distributions, which do not appear to match the real data. (B) Plots along the diagonal same as panel A. Below the diagonal are distributions of pairwise phase differences and above the diagonal are distributions of pairwise phase sums, represented by histograms for the real data and by red density plots for the theoretical torus graph model. In contrast to the torus graph model, the sine model fails to accurately capture the distributions of the sufficient statistics from these data.

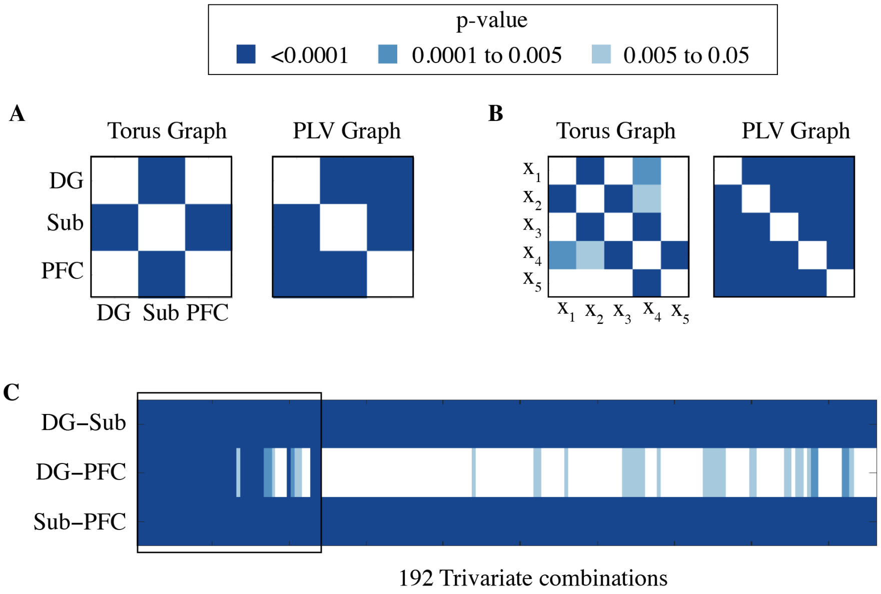

Figure A.7: (A) Adjacency matrices for the three-dimensional LFP analysis with entries colored by p-value. PLV p-values are very small for all connections, while torus graph p-values reflect finer structure, such as PFC-Sub coupling that is apparently more salient than PFC-DG coupling. The adjacency matrix for the trivariate network is a representative combination of channels reflecting the pattern that dominates in all trivariate combinations.(B) Same as (A) but for the five-dimensional LFP analysis, where the torus graph reveals nearest-neighbor structure along the linear probe in CA3 that PLV misses. (C) Edgewise p-values for each edge in each possible trivariate graph (composed of each combination of electrodes from each of the three areas). Note that channels 6 and 7 (boxed region) of DG are on the border between DG and CA3 and may be picking up signals from CA3; omitting these channels gives stronger evidence of an overall lack of connections between DG and PFC (corresponding to the adjacency matrix in (A)).

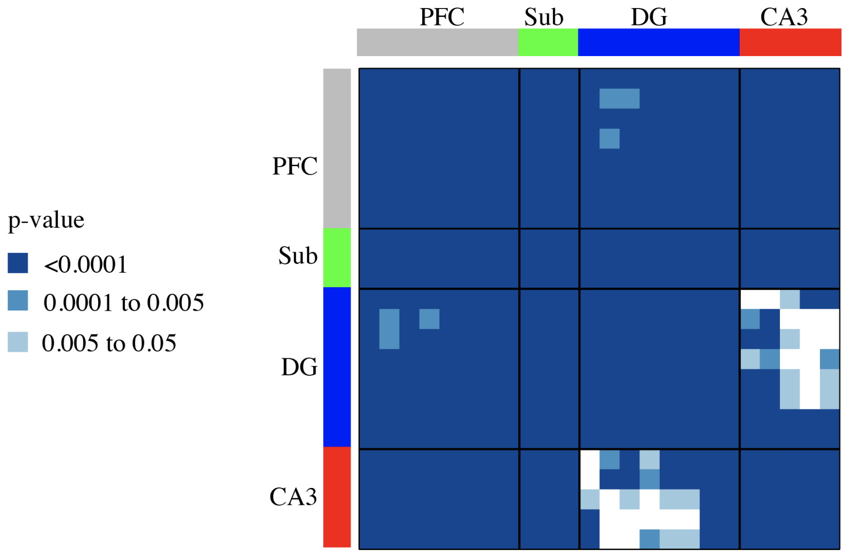

Figure A.8: Adjacency matrix for PLV graph with edgewise p-values determined using Rayleigh's test of uniformity on the circle for each pairwise phase difference. Entries are colored by p-value and, compared to the torus graph adjacency matrix (Figure 3.8.C), there is very little noticeable structure in the graph even for very small p-value thresholds (aside from a lack of edges between CA3 and DG).

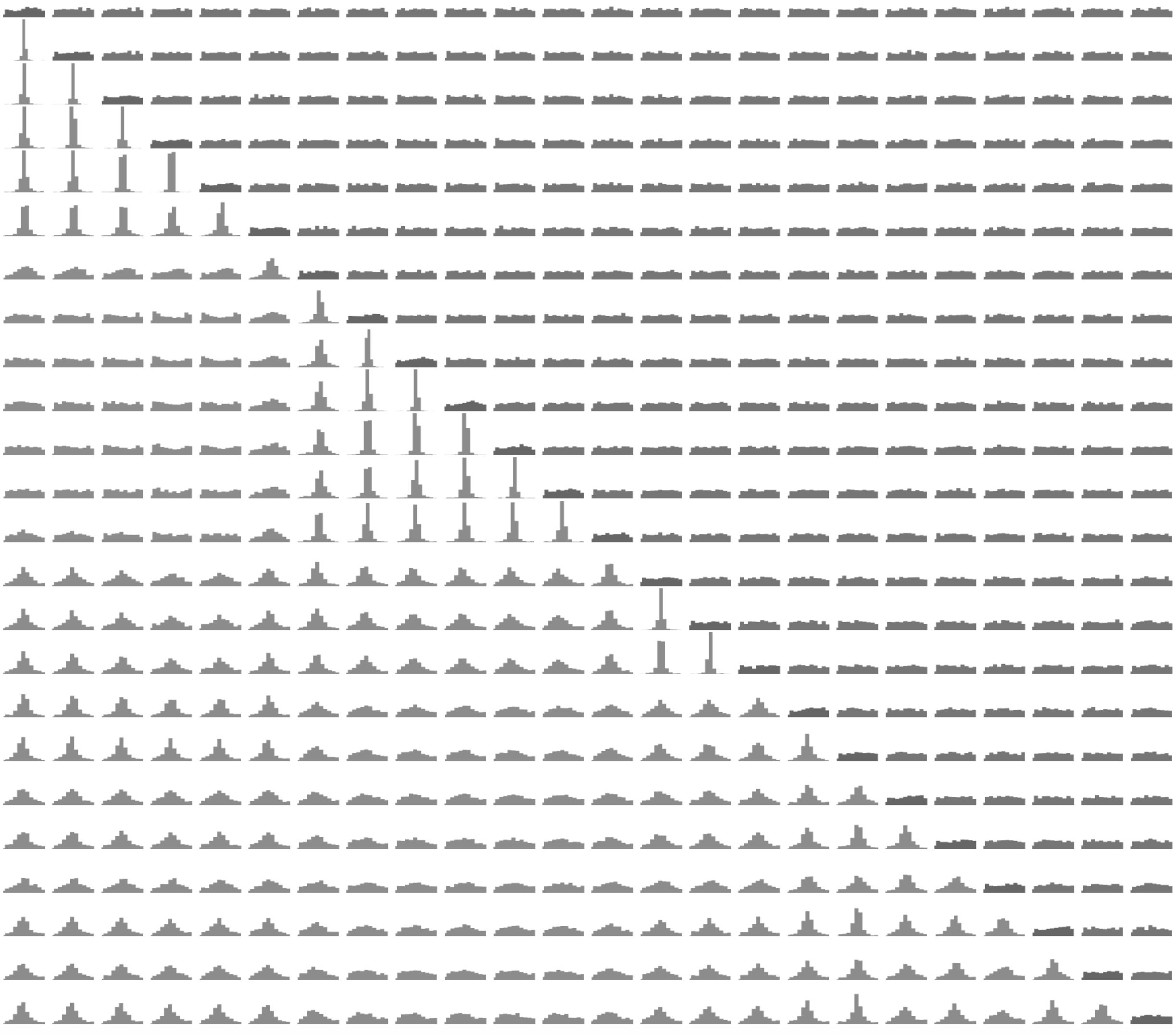

Figure A.9: For the 24-dimensional real LFP data, the diagonal shows univariate histograms which appear to have low concentration in all cases. Below the diagonal are histograms of phase differences between pairs of angles, showing some highly concentrated distributions suggesting rotational dependence; above the diagonal are histograms of phase sums, showing very little concentration, suggesting there is not strong evidence for reflectional dependence in these data.

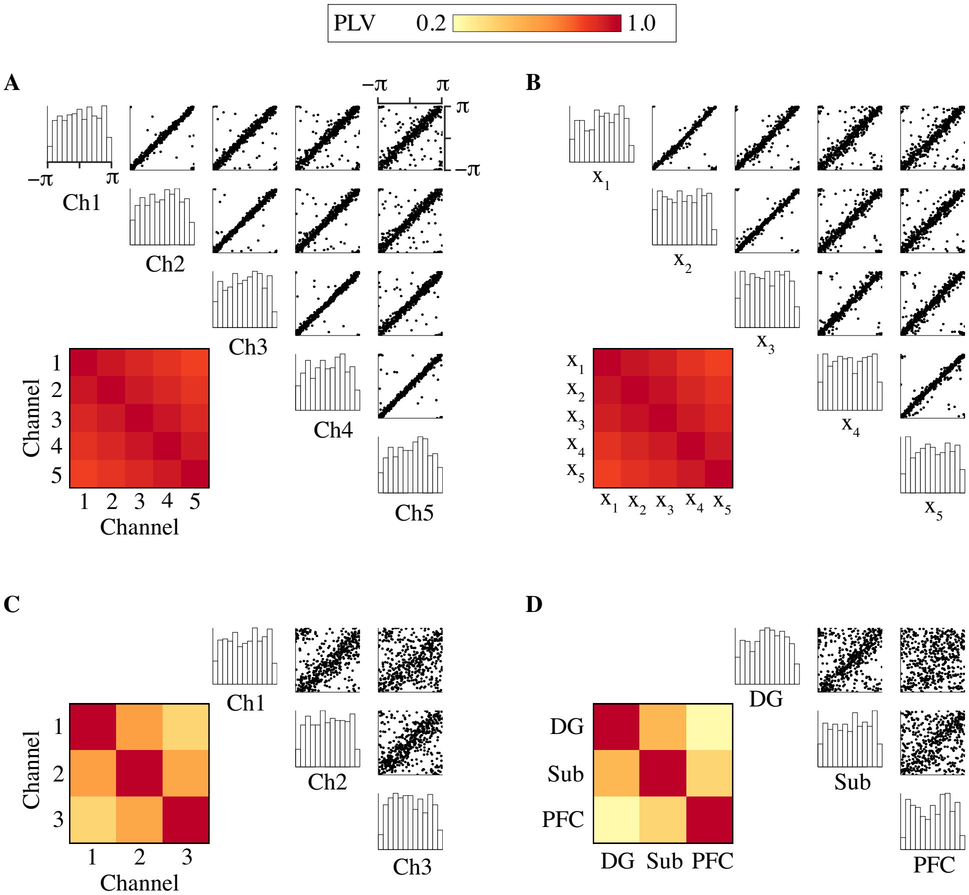

Figure A.10: Examples of simulated and real data to demonstrate the validity of the simulation process. Upper right: histograms and pairwise scatter plots, bottom left: estimated PLV matrices (color scale 0.2 to 1, with red indicating higher PLV values). (A) Simulated 5-channel data with linear probe structure. (B) Real 5-channel data from CA3. (C) Simulated 3-channel data. (D) Real 3-channel data from separate regions (DG, Sub, and PFC).

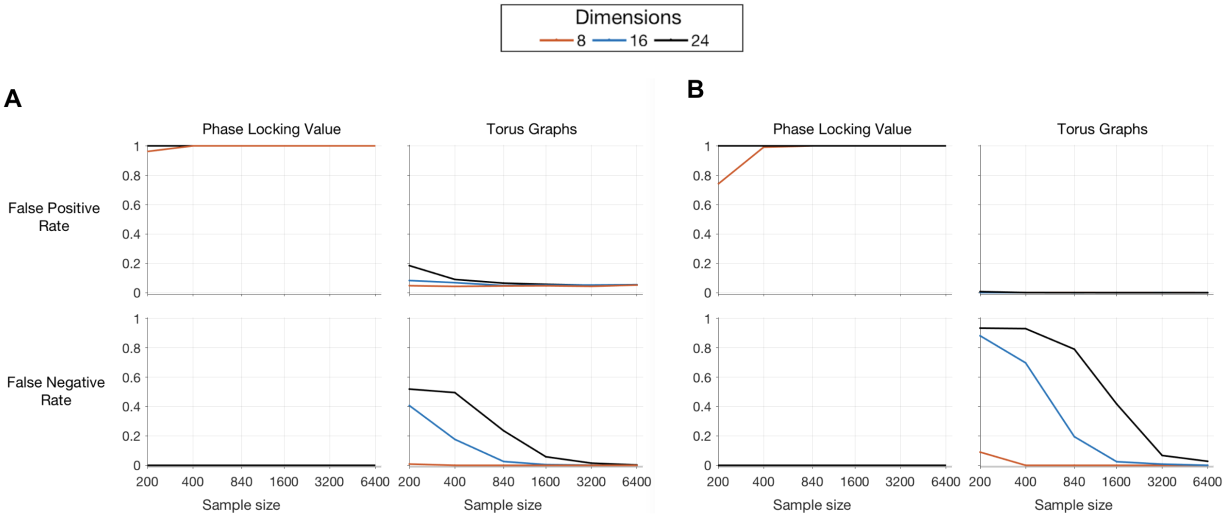

Figure A.11: Investigation of False Positive Rate (FPR) and False Negative Rate (FNR) for graphs of varying dimensions as sample size increases. (A) FPR and FNR for PLV (left) and torus graphs (right) using an alpha level of 0.05 for the edgewise hypothesis tests with no Bonferroni correction for the number of edges. PLV has high FPR for all sample sizes while torus graphs control the FPR; on the other hand, PLV has low FNR, but torus graphs is more conservative and for low sample sizes may be missing some edges. (B) Same as A, but with Bonferroni correction.

Appendix B

Appendix to Chapter 4

B.1 One-dimensional *a priori* physical model details

Substituting Equation 4.4 into Equation 4.3 yields

$$\phi(x,y,z) = -\frac{1}{4\pi\sigma}\int_a^b \int\int_{x^2+y^2\leq R} \frac{g(z')}{\sqrt{(x-x')^2+(y-y')^2+(z-z')^2}}\,dx'\,dy'\,dz'. \tag{B.1}$$

We will assume x and y are inside the cylinder (as typically, we assume we observe ϕ at the center of the cylinder). Changing to polar coordinates, we define $r^2 = (x-x')^2 + (y-y')^2$ as the variable radius inside the cylinder and use the substitution $dx'\,dy'\,dz' = r\,d\theta\,dr\,dz'$ to obtain

$$\begin{aligned}
\phi(x,y,z) &= -\frac{1}{4\pi\sigma}\int_a^b \int_0^R \int_0^{2\pi} \frac{rg(z')}{\sqrt{(z-z')^2+r^2}}\,d\theta\,dr\,dz' && \text{(B.2)}\\
&= -\frac{1}{2\sigma}\int_a^b g(z')\int_0^R \frac{r}{\sqrt{(z-z')^2+r^2}}\,dr\,dz' && \text{(B.3)}\\
&= -\frac{1}{2\sigma}\int_a^b g(z')\left[\sqrt{(z-z')^2+R^2}-\sqrt{(z-z')^2}\right]\,dz'. && \text{(B.4)}
\end{aligned}$$

Notice that after integration, this is no longer a function of x or y, so we can simply write

$$\phi(z) = -\frac{1}{2\sigma}\int_a^b g(z')\left[\sqrt{(z-z')^2+R^2}-\sqrt{(z-z')^2}\right]\,dz'. \tag{B.5}$$

To better understand how R affects the ϕ, I factor out R:

$$\phi(z) = -\frac{R}{2\sigma}\int_a^b g(z')\underbrace{\left[\sqrt{\left(\frac{r}{R}\right)^2+1}-\sqrt{\left(\frac{r}{R}\right)^2}\right]}_{b(r;R)}\,dz' \tag{B.6}$$

where $r = z - z'$ and $b(r; R)$ is a weight function with a maximum value of 1 when $r = 0$.

B.2 Two-dimensional *a priori* physical model details

Substituting Equation 4.6 into Equation 4.3 yields

$$\begin{aligned}
\phi(x, y, z) &= -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} \int_{x \le R} \frac{g(y', z')}{\sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}} \, dx' \, dy' \, dz' \\
&= -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} \int_{0}^{R} \frac{g(y', z')}{\sqrt{(x-x')^2 + m^2}} \, dx' \, dy' \, dz' \\
&= -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} g(y', z') \int_{0}^{R} \frac{1}{\sqrt{(x-x')^2 + m^2}} \, dx' \, dy' \, dz'
\end{aligned}$$

where $m = \sqrt{(y-y')^2 + (z-z')^2}$. Since we are interested in modeling the LFP at the face of the probe ($x = 0$), let $x = 0$ and integrate over x':

$$\begin{aligned}
\phi(0, y, z) &= -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} g(y', z') \int_{0}^{R} \frac{1}{\sqrt{(x')^2 + m^2}} \, dw \, dy' \, dz' \\
&= -\frac{1}{4\pi\sigma} \int_{a_z}^{b_z} \int_{a_y}^{b_y} g(y', z') \operatorname{arcsinh}\left(\frac{R}{\sqrt{(y-y')^2 + (z-z')^2}} \right) dy' \, dz'.
\end{aligned}$$

I will write the LFP as $\phi(y, z)$ where implicitly $x = 0$ when using the forward model.

B.3 Kernel CSD (kCSD) details

The observed potential corresponding to one source, which we will call b_i, is

$$b_i(x, y, z) = \mathcal{A}C(x, y, z) \equiv \frac{1}{4\pi\sigma} \int \int \int \frac{\tilde{b}_i(x, y, z)}{\sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}} dx' dy' dz' \tag{B.7}$$

and the overall potential from all sources adds linearly:

$$V(x, y, z) = \mathcal{A}C(x, y, z) = \sum_{i=1}^{M} a_i b_i(x, y, z) \tag{B.8}$$

Defining $\mathbf{x} = (x, y, z)$, the kernel function between LFP basis functions is:

$$K(\mathbf{x}, \mathbf{x}') \equiv \sum_{i=1}^{M} b_i(\mathbf{x}) b_i(\mathbf{x}')$$

This defines an RKHS where any function can be expressed as

$$f(\mathbf{x}) = \sum_{j=1}^{\ell} \alpha_j K(\mathbf{x}_j, \mathbf{x}) = \sum_{i=1}^{M} a_i b_i(\mathbf{x})$$

where $a_i = \sum_{j=1}^{\ell} \alpha_j b_i(\mathbf{x}_j)$. The minimum-norm solution that corresponds to interpolating the observed LFPs is given by

$$\boldsymbol{\beta} = \mathbf{K}^{-1}\mathbf{V}$$

where the estimated potentials are $V^*(\mathbf{x}) = \sum_{i=1}^{N} \beta_i K(\mathbf{x}_i, \mathbf{x})$. That is, we know from RKHS that $a_i = \sum_{k=1}^{N} \beta_k b_i(\mathbf{x}_k)$ such that

$$V^*(\mathbf{x}) = \sum_{i=1}^{M} a_i b_i(\mathbf{x}) = \sum_{i=1}^{M} b_i(\mathbf{x}) \sum_{k=1}^{N} \beta_k b_i(\mathbf{x}_k) = \sum_{k=1}^{N} \beta_k \sum_{i=1}^{M} b_i(\mathbf{x}_k) b_i(\mathbf{x}_k) = \sum_{k=1}^{N} \beta_k K(\mathbf{x}_k, \mathbf{x}).$$

Note that adding Tikhonov regularization simply adds λI inside the inverse and loosens the requirement that the LFPs are interpolated. Given this set of potentials V^* there is now a unique current distribution:

$$C^*(\mathbf{x}) = \sum_{j=1}^{M} a_j \tilde{b}_j(\mathbf{x}) = \sum_{i=1}^{N} \beta_i \sum_{j=1}^{M} b_j(\mathbf{x}_j) \tilde{b}_j(\mathbf{x}) \equiv \sum_{i=1}^{N} \beta_i \tilde{K}(\mathbf{x}_i, \mathbf{x})$$

where $\tilde{K}$ is the cross-kernel induced by the operator $\mathcal{A}$. Note that iCSD (Pettersen et al., 2006) is a special case of kCSD with either delta functions, piecewise-constant functions, or cubic spline functions used as the basis, but no regularization is applied, and that extensive comparisons of different kernels and regularization schemes was carried out in Kropf and Shmuel (2016).

B.4 Relating Gaussian process regression to RKHS regression

Under a zero-mean version of GPCSD, the form of the prediction in Equation 4.21 looks very similar to the prediction given by kCSD method discussed in Section 4.2.4, which happens because the two methods can be seen as two different frameworks for arriving at similar estimators. In kCSD, one specifies a set of basis functions, along with their centers and hyperparameters; in Potworowski et al. (2012), M Gaussian bumps with width ℓ were used, with a suggestion to use a large enough M to densely cover the electrode recording area, and to use wide enough ℓ so that the basis functions overlapped partially. From the basis functions, a kernel function was calculated in Equation 4.13 which functions in a similar manner as the Gaussian process covariance function of GPCSD. So the primary difference is in how the model is specified: through selection and placement of a number of basis functions or through direct specification of a suitable covariance function.

In fact, as shown in (Rasmussen and Williams, 2006, p. 84), the squared exponential covariance function is equivalent to using infinitely many densely-spaced Gaussian-shaped basis functions, and the width of the Gaussian basis functions is proportional to the characteristic lengthscale of the covariance function. This implies that kCSD would give similar results to a Gaussian process with squared exponential lengthscale if many densely-spaced basis functions were used. However, the Gaussian process approach precludes choosing M or using potentially computationally taxing cross-validation to select the width of the basis functions. Instead, maximum marginal likelihood can easily be used to select the lengthscale for the Gaussian process along with other hyperparameters. In addition, it is possible to specify more complex covariance functions, such as additive or product covariance functions, for which it may be difficult to directly specify the corresponding basis functions. The Gaussian process approach is also easily extendable to spatiotemporal processes, can incorporate a random mean function, can include multi-scale covariance functions, and provides a conditional distribution of CSD predictions given the LFP data (while kCSD only provides a point estimate).

B.5 GPCSD computational details

This section contains a few computational details used in the GPCSD implementation. For the two-dimensional zero-mean GPCSD model in my implementation, the hyperparameters $\boldsymbol{\theta}$ consist of the covariance hyperparameters and forward model parameter R, where

$$\boldsymbol{\theta} = [R,\, \ell_1,\, \ell_2,\, \sigma^2_{t,E},\, \ell_{t,E},\, \sigma^2_{t,SE},\, \ell_{t,SE},\, \sigma^2_{noise}].$$

In the above, ℓ_1 and ℓ_2 are the lengthscales for each spatial dimension (where only one of these would be needed for one-dimensional GPCSD), while $\sigma^2_{t,E}$, $\ell_{t,E}$, $\sigma^2_{t,SE}$, and $\ell_{t,SE}$ correspond to the temporal covariance functions and σ^2_{noise} is the white noise variance for the observed LFPs. Let $\mathbf{K}^s \in \mathbb{R}^{M\times M}$ be the LFP spatial covariance evaluated at the locations of the observed LFPs and $\mathbf{K}^t \in \mathbb{R}^{T\times T}$ be the temporal covariance evaluated at the observed time points; they are functions of $\boldsymbol{\theta}$ (including the forward model parameter R which is part of the LFP spatial covariance function through the forward operator). Let $\tilde{\boldsymbol{\phi}}^{(n)} \in \mathbb{R}^{M\times T}$ represent the matrix of observed LFPs on trial n (with N total trials). The traditional expression for the log marginal likelihood (excluding terms that don't depend on $\boldsymbol{\theta}$) takes the form

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{n=1}^{N} \log\left(\left|\mathbf{K}^s \otimes \mathbf{K}^t + \sigma^2_{noise}\mathbf{I}\right|\right) + \mathrm{vec}\left(\tilde{\boldsymbol{\phi}}^{(n)}\right)^T \left[\mathbf{K}^s \otimes \mathbf{K}^t + \sigma^2_{noise}\mathbf{I}\right]^{-1} \mathrm{vec}\left(\tilde{\boldsymbol{\phi}}^{(n)}\right).$$

However, this form relies on inversion of an $MT \times MT$ matrix which is clearly problematic for typical M and T observed in real data, so we instead leverage the special structure present in the matrix. In particular, I use the eigendecomposition of the covariance matrices, $\mathbf{K}^s = \mathbf{Q}_s \mathbf{\Lambda}_s \mathbf{Q}_s^T$ and $\mathbf{K}^t = \mathbf{Q}_t \mathbf{\Lambda}_t \mathbf{Q}_t^T$, where $\mathbf{\Lambda}_s$ and $\mathbf{\Lambda}_t$ are diagonal matrices. Let $\mathbf{D} = \mathbf{\Lambda}_s \otimes \mathbf{\Lambda}_t + \sigma^2_{noise}\mathbf{I}$ be a diagonal matrix, and let $\mathbf{q}$ be an MT-vector with elements $1/D_{ii}$. Note that a low-rank Gaussian process could be implemented by using truncated eigendecompositions (Solin and Särkkä, 2014). Using the eigendecomposition and properties of Kronecker products (as shown in more detail in Saatçi 2012), the log marginal likelihood may be rewritten:

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\sum_{i=1}^{MT}\log(D_{ii}) - \frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{MT}\left[\operatorname{vec}\left(\mathbf{Q}_s^T \tilde{\boldsymbol{\phi}}^{(n)} \mathbf{Q}_t\right) \circ \operatorname{vec}\left(\mathbf{Q}_s^T \tilde{\boldsymbol{\phi}}^{(n)} \mathbf{Q}_t\right) \circ \mathbf{q}\right]_i$$

where the Hadamard product $\circ$ indicates elementwise multiplication and $\mathbf{Q}_t$, $\mathbf{Q}_s$, $\mathbf{D}$, and $\mathbf{q}$ depend on $\boldsymbol{\theta}$. This form is faster and more stable to compute as it avoids direct inversion of an $MT \times MT$ matrix and makes use of elementwise operations and vectorization that are amenable to programming languages such as Python.

Given fixed $\boldsymbol{\theta}$, the log marginal likelihood may also be optimized over mean function parameters $\boldsymbol{\gamma}$; here I show the likelihood assuming a shared mean function across trials, though per-trial mean parameters could also be used (and if trials were assumed independent, this would result in a separate log marginal likelihood for each trial, which is what I used to optimize the per-trial time shifts in the auditory data). Let $\boldsymbol{\mu} \in \mathbb{R}^{M \times T}$ be the mean function evaluated at the observed LFP spatial and temporal points (where this function depends on $\boldsymbol{\gamma}$). I first calculate the inverse covariance matrix as

$$\mathbf{\Sigma}^{-1} = (\mathbf{Q}_s \otimes \mathbf{Q}_t)\operatorname{diag}(\mathbf{q})(\mathbf{Q}_s \otimes \mathbf{Q}_t)^T$$

and calculate the mean of the LFPs across trials as $\bar{\mathbf{y}} = \sum_{n=1}^{N} \tilde{\boldsymbol{\phi}}^{(n)}$. Then I use the following (rescaled) log marginal likelihood:

$$\log \mathcal{L}(\boldsymbol{\gamma}) = \operatorname{vec}(\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}\operatorname{vec}(\bar{\mathbf{y}}) - \frac{1}{2}\operatorname{vec}(\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}\operatorname{vec}(\boldsymbol{\mu}).$$

My current implementation does not include analytic forms of the gradients for optimization, though this should certainly be possible to incorporate and would improve optimization.

B.6 Alternative trial-to-trial variation models

In addition to the estimation of per-trial and per-component time shifts and amplitude modulations in the auditory LFPs discussed in Section 4.5, I considered some other models for trial-to-trial variation which did not explicitly separate the evoked response into separate components. For these analyses, I only used data from 0 to 75 ms after tone onset and fit CSD evoked response functions to each probe with covariance functions fixed from the baseline; the trial-to-trial variation was then modeled based on this fitted mean function.

In the first model, I let each trial's evoked response have a single amplitude scaling (where the amplitude scale had a separate $\log_2$ ISI intercept and slope for each tone). The posterior distributions of the amplitude scale factors were estimated using both ADVI and MCMC, with priors on amplitude scales similar to Section 4.5 (independent for each tone). The MCMC results showed that amplitude generally increased with $\log_2$ ISI, but that the response depended on the tone (Figure B.1); the posterior mean slopes per tone were similar between probes. In addition, the amplitude scales for each probe appeared to be correlated across trials, even after removing the effect of $\log_2$ ISI (Figure B.2), which appears to concur with the results found in Section 4.5 in which many of the components in the early evoked responses showed correlated amplitude variation. The comparison of ADVI and MCMC for this model showed that the same qualitative results would be obtained with either method, though ADVI tends to underestimate posterior variances.

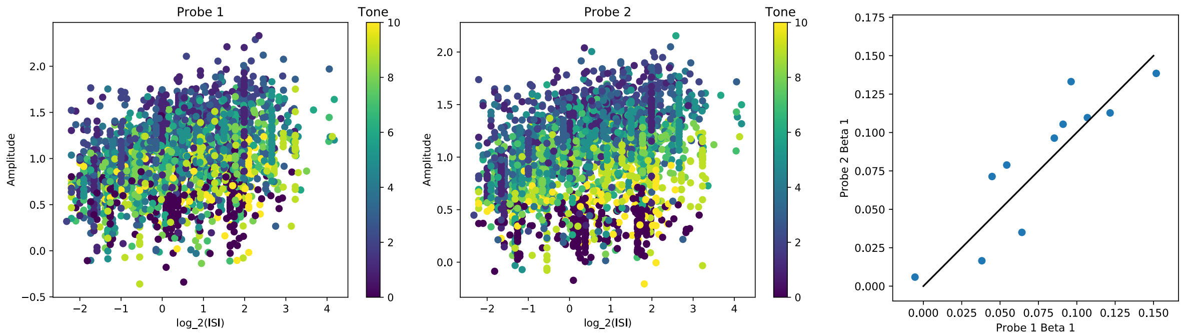

Figure B.1: Results are shown for a model with a single per-trial amplitude scale that depended linearly on $\log_2$ ISI with separate intercepts and slopes for each tone. The left two panels show the posterior mean amplitude scales for each trial as a function of $\log_2$ ISI, colored by tone. While there is generally an upward trend of amplitude with $\log_2$ ISI, the relationships are stronger for some tones than others. The relationships appear similar for both probes. The right panel shows the relationship between the posterior mean slopes of amplitude with $\log_2$ ISI for probe 1 (horizontal axis) and probe 2 (vertical axis), with one point for each of the 11 tones; it appears the amplitude with $\log_2$ ISI relationships for each tone are similar between the two probes.

In the second model, in addition to the amplitude model discussed above, I also incorporated time shift variation that could vary spatially. Instead of separating the evoked response into multiple components, I designed spatially-varying time shift function that would provide the time shift separately for each spatial point of interest. That is, if $\tau^{(n)}(s)$ is the time shift function for a single trial, then the CSD evoked response for a single trial would be

$$\mu^{(n)}(s,t) = \mu(s, t + \tau^{(n)}(s))$$

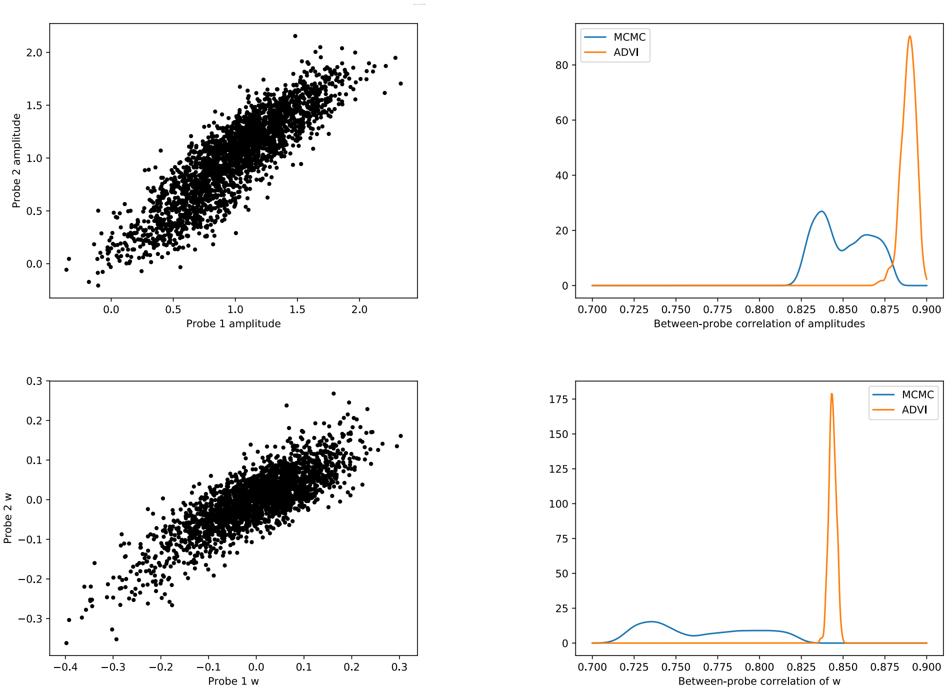

Figure B.2: Results are shown for a model with a single per-trial amplitude scale that depended linearly on $\log_2$ ISI with separate intercepts and slopes for each tone. Top row: scatter plot of posterior mean amplitude scales for probe 1 vs probe 2 which indicates correlated amplitudes (left), along with posterior distributions of the correlation between per-trial amplitude scales for each probe estimated by ADVI (orange) and MCMC (blue). ADVI slightly overestimates the correlation and with smaller uncertainty, but both methods indicate strong positive correlations. Bottom row: same but with the $\log_2$ ISI effect removed before calculating correlations. The correlations are slightly weaker, but still appear to be strong and positive.

To allow large jumps in the spatial function, I parameterized $\tau^{(n)}(s)$ as a cubic B-spline with a large number of knots (40 knots) and with priors designed to allow sudden jumps (adaptive smoothing). Specifically,

$$\tau^{(n)}(s) = \beta_0^{(n)} + \sum_{i=1}^{38} \beta_i^{(n)} B_i$$

where B_i is a B-spline basis function. The priors for $\beta_0^{(n)}$ were iid zero-mean Normal with standard deviation σ_0 where σ_0 had a Half-Normal prior with standard deviation 2. The rest of the coefficients were modeled using a random walk,

$$\beta_i^{(n)} = \sum_{j=1}^{i-1} \Delta_j^{(n)}$$

where $\Delta_j^{(n)}$ had a zero-mean Student t prior with standard deviation σ_j (which had a Half-Normal prior with standard deviation 2) and with degrees of freedom ν having a uniform prior on $[0, 30]$. Because the standard deviations of the increments $\Delta_j^{(n)}$ can change over space and because the prior is heavy-tailed, the variance of the spline coefficients can change suddenly to allow adaptive smoothing. For computational reasons, this model was estimated using ADVI instead of MCMC. Simulations using the fitted mean and covariance functions (results not shown) found that even if the true $\tau^{(n)}(s)$ were sharp, piecewise-constant functions, the spline function could recover the shift functions fairly well (and in particular recovered the shift well at all of the observed LFP spatial points). Posterior mean shift functions for each probe and each trial are shown in Figure B.3. The shift functions appear to be mostly constant over space at the top and bottom of each probe, with something like a change point appearing near the dominant inversion of current (near electrode position 12 in Probe 1 and position 15 in Probe 2); the posterior distributions of the σ_j standard deviations for the random walk also reflected larger values of σ_j near these points and small values elsewhere. Overall, it appears that most time shifts are small (within -2 ms and 2 ms). In addition, as shown in the right panel of Figure B.3, the shift functions appear to be correlated across-trials between probes, particularly at the upper and lower spatial locations. While this model is substantially different from that used in Section 4.5 to assess relationships between probes, the results are similar in a few important ways. This model suggests that the variation in shifts across trials is not large, and it also suggests that the strongest relationships between the shifts in each probe occur in the deeper layers during the early evoked response.

B.7 GPCSD supplementary figures

This section contains supplementary figures referred to from Chapter 4.

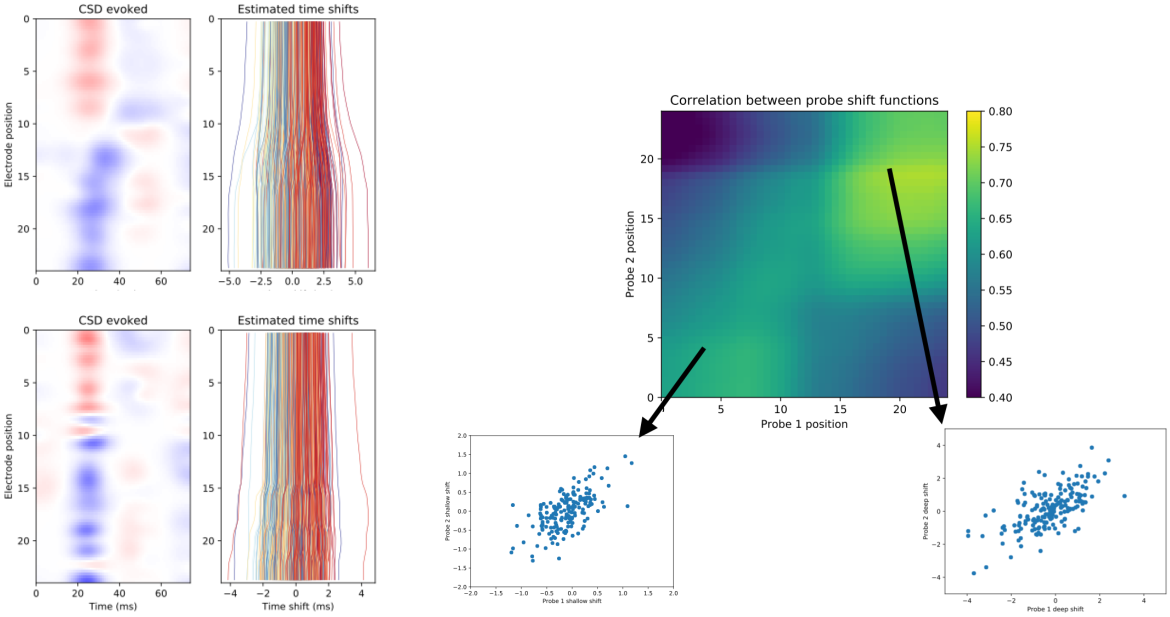

Figure B.3: Left panel: fitted CSD evoked functions for each probe alongside the estimated per-trial shift functions, colored by tone. There appears to be a possible relationship with tone, and in both probes, the shift functions are nearly flat near the top of each probe, with an apparent change point near the middle of each probe, below which the shifts have higher variance. Overall, the shifts are small (most between -2 ms and 2 ms). Right panel: correlation matrix computed across trials for the relationship between per-trial shift functions across probes, with scatter plots shown for two selected locations near the top and bottom of each probe. This indicates a suggestion that the shifts are correlated between probes at similar spatial locations.

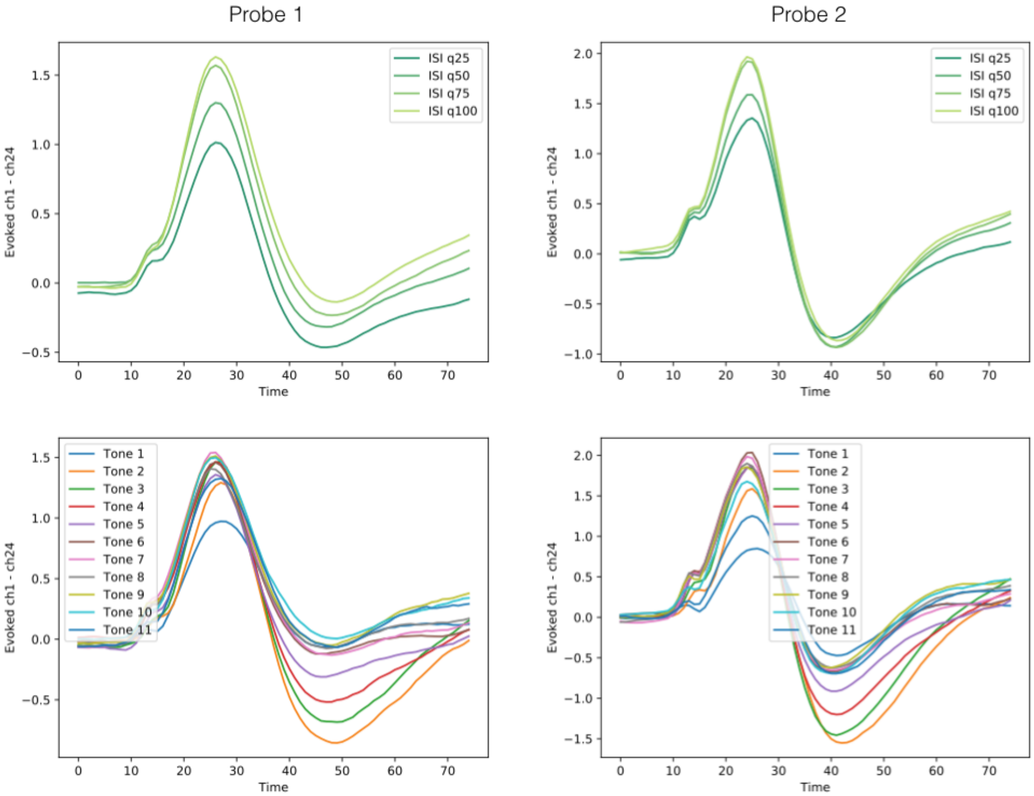

Figure B.4: For auditory LFPs from probe 1 (left column) and probe 2 (right column), the average evoked responses are shown for trial separated by quantiles of the $\log_2$ ISI (top row) and by the tone (bottom row). To summarize the evoked response across the entire probe, the difference in evoked responses between channels 1 and 24 is used. It appears in both probes that longer ISI leads to larger responses but with similar profiles. Tone also effects the amplitude and possibly the shape of the evoked response.

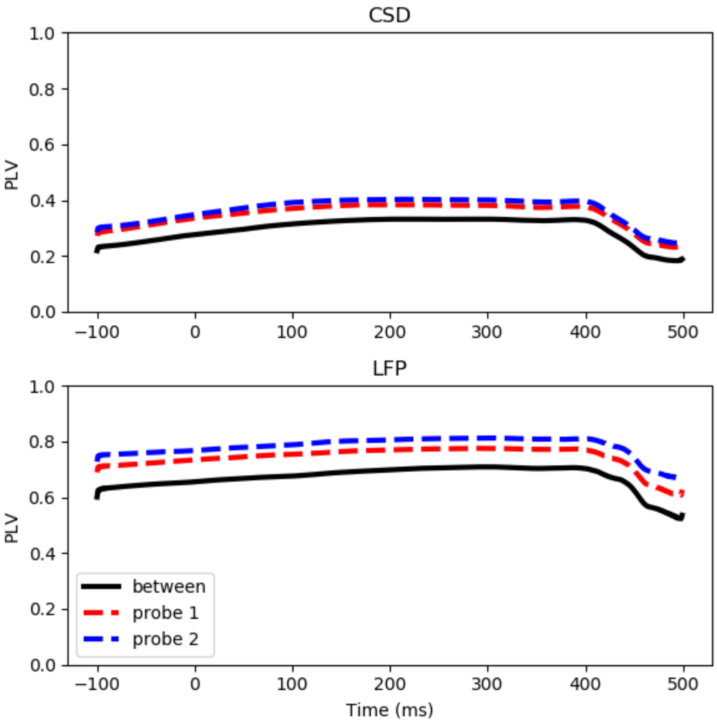

Figure B.5: Timecourses of aggregate PLV (averaged across all channel pairs) for within probe 1 (red dashed), within probe 2 pairs (blue dashed), and between probe pairs (black). Top panel is for the CSD and bottom panel is for the LFP. It appears that CSD generally has lower PLV values but a similar pattern over time, though the increase after stimulus may be more pronounced in the CSD than LFP.

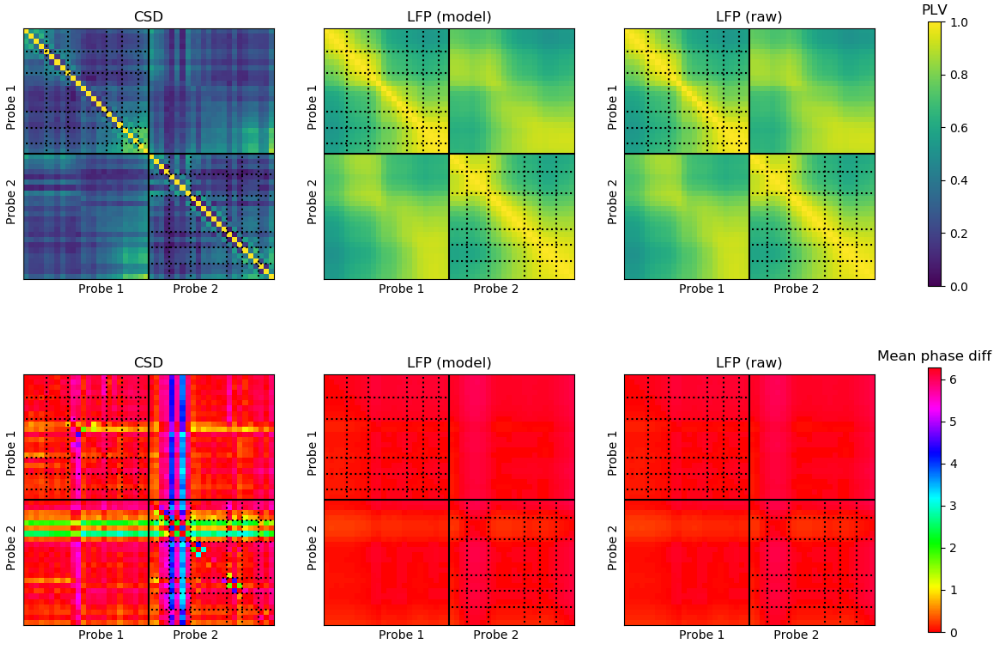

Figure B.6: Top row: analogous to Figure 4.9, but using PLV instead of torus graphs; color indicates the observed PLV value. The overall patterns are fairly similar though the LFP has much larger PLV values overall than the CSD. The rightmost plot shows PLV calculated from the raw LFPs rather than the model predictions. Bottom row: circular mean directions of pairwise phase differences corresponding to the PLV matrices; in the LFPs, phase differences are concentrated at 0 radians, while in the CSD, some of the phase differences are offset from 0; in particular, some electrodes in the top of layer 3 (probe 1) and in layer 2 (probe 2) appear to have nonzero phase offsets with many other locations, and some electrodes along the diagonal in probe 2 (in layers 1, 2, 3, and 4) have nonzero phase offsets with other locations in the layer.

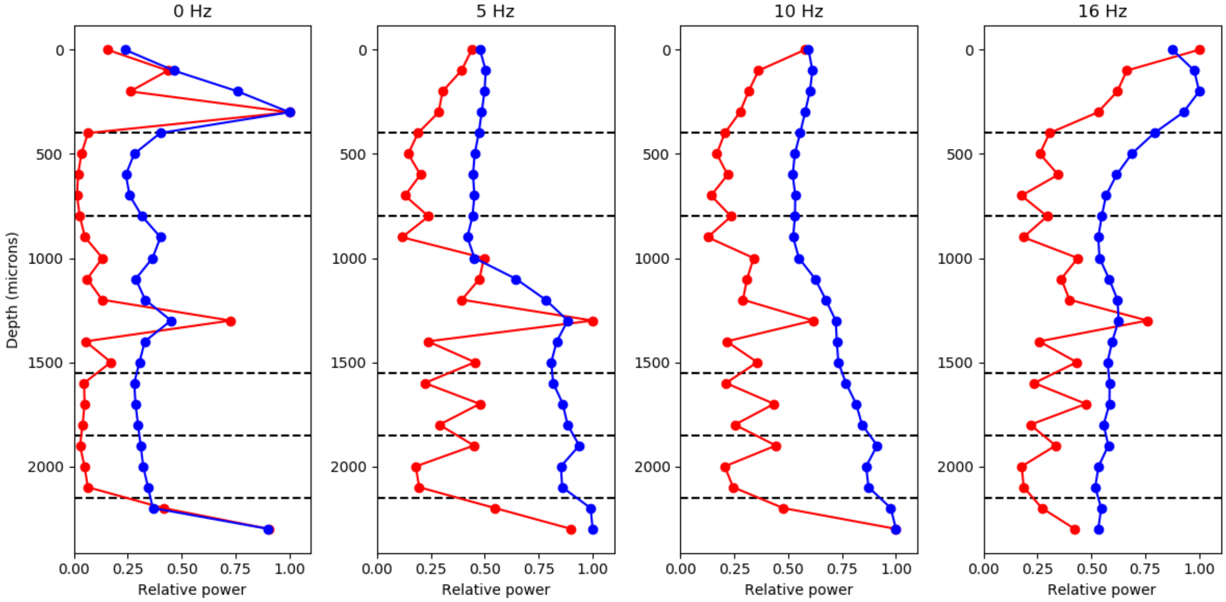

Figure B.7: For four different frequencies (left to right), the trial-averaged power (relative to maximum power for that frequency) is plotted as a function of depth for the LFP (blue) and the CSDs (red) for probe 1 (with spectra shown in Figure 4.8). The 0 Hz power was calculated using the slow-timescale GPCSD model predictions while the rest were calculated using the fast-timescale predictions. The putative cortical layer boundaries are shown as dashed lines. While the 0 Hz power profile with depth is fairly similar for the CSD and LFP, with peaks in layers 1, 3, and 6, the profiles for the higher frequencies are smoother in the LFP than the CSD. Particularly striking is that the CSD shows a marked peak in Layer 3, while the LFP shows a smoothly changing power which appears maximal in Layers 4, 5, and 6 for 5 and 10 Hz and maximal near layer 1 for 16 Hz. These results suggest that the oscillations may be generated from Layer 3, which is not readily apparent from the LFPs.

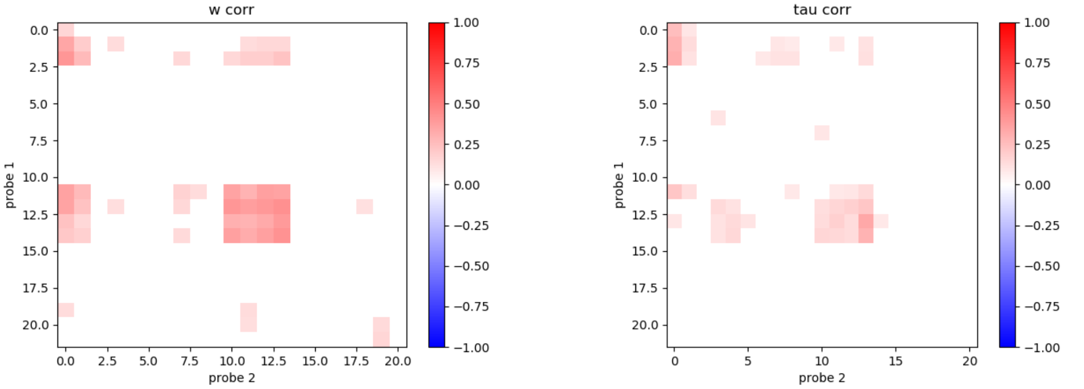

Figure B.8: Left: correlations in amplitude scale factors between the CSD components of probe 1 and the CSD components of probe 2, with white marking non-significant amplitude correlations. Right: correlations in shifts between the CSD components of probe 1 and the CSD components of probe 2, with white marking non-significant amplitude correlations. The patterns of correlation between components share some similar clusters, prompting the consideration of pairs that exhibited both shift and amplitude correlation.

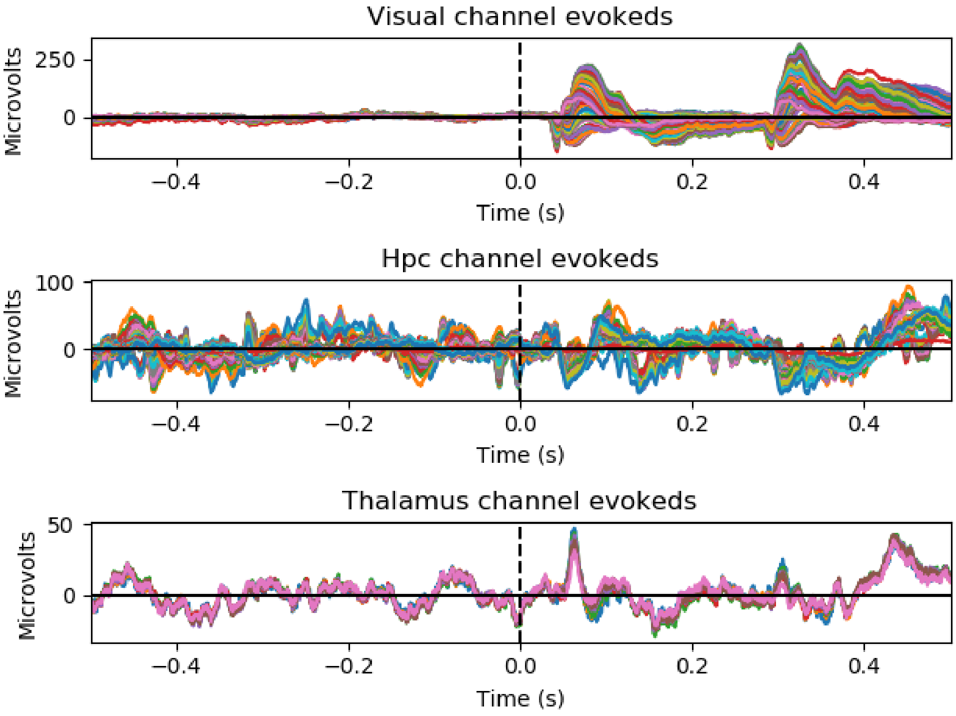

Figure B.9: Average evoked responses from Neuropixel LFPs, with channels grouped by region label; the visual area has the clearest evoked response.

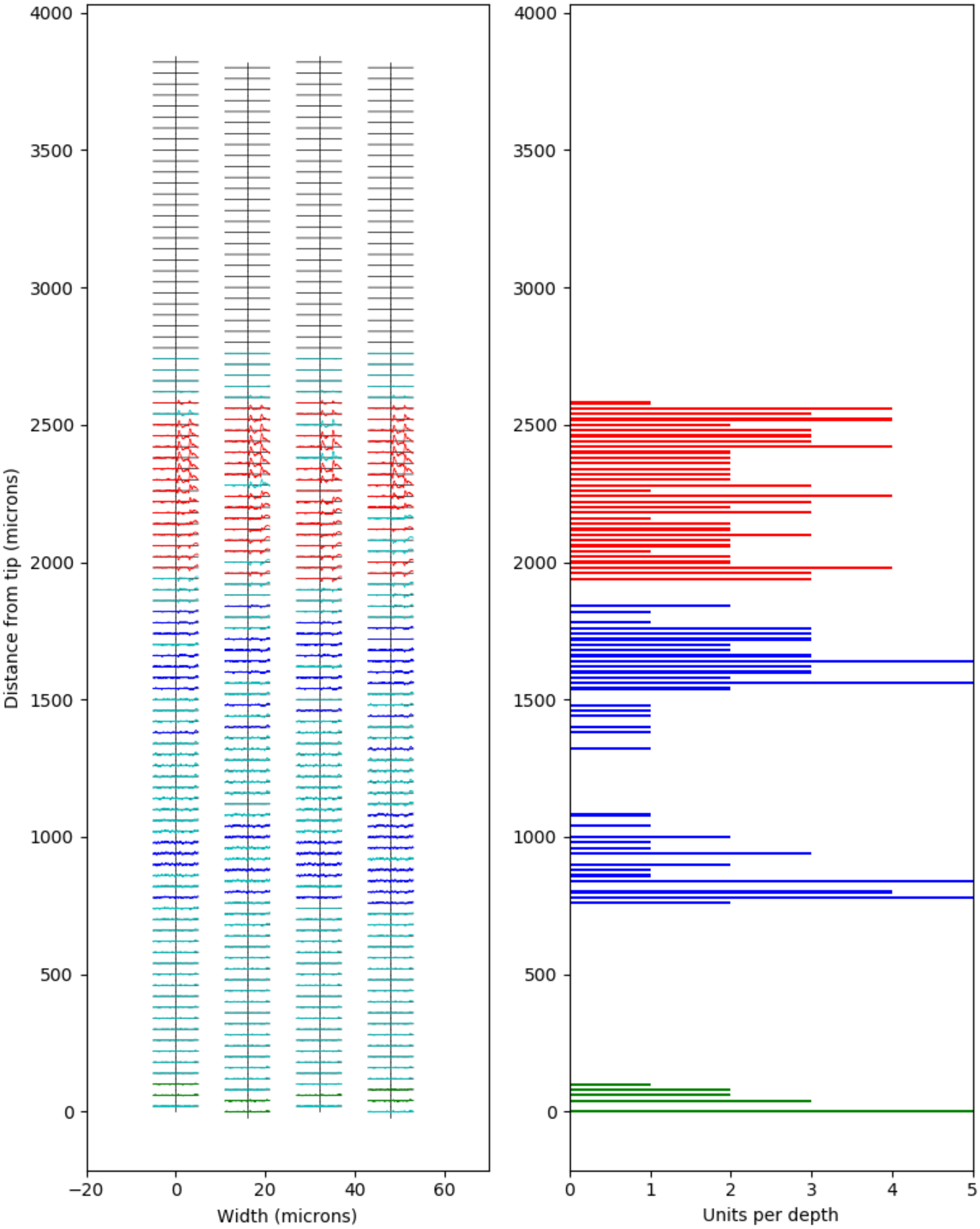

Figure B.10: Left panel: depiction of the Neuropixel probe recording locations, with evoked responses for all LFP channels colored according to region (red: visual, blue: hippocampal, green: thalamic, cyan: no region assigned). Right: bar plot of number of neurons/units identified for each depth along the probe.

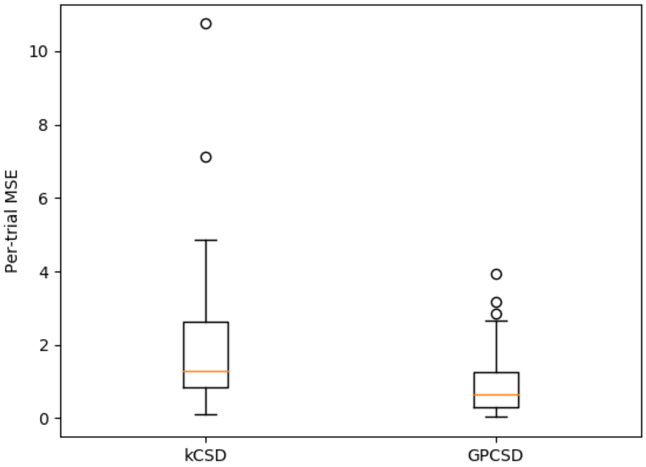

Figure B.11: Boxplot representing the distribution, across 50 trials in the test set, of per-trial MSE for kCSD and GPCSD on simulated two-dimensional data (error evaluated by comparing the true CSD and estimated CSD at the electrode positions). It appears that GPCSD obtains a smaller MSE than kCSD and has a smaller variance of the MSE across trials.