ANALYSIS OF MULTI-WAY FUNCTIONAL DATA UNDER WEAK SEPARABILITY, WITH APPLICATION TO BRAIN CONNECTIVITY STUDIES

by

Brian C. Lynch

B.A. in Physics and Mathematics, Washington University in St. Louis, 2013

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Brian C. Lynch

It was defended on

November 29, 2018

and approved by

Kehui Chen, PhD, Department of Statistics

Yu Cheng, PhD, Department of Statistics

Zhao Ren, PhD, Department of Statistics

Jing Lei, PhD, Department of Statistics, Carnegie Mellon University

Dissertation Director: Kehui Chen, PhD, Department of Statistics

ANALYSIS OF MULTI-WAY FUNCTIONAL DATA UNDER WEAK SEPARABILITY, WITH APPLICATION TO BRAIN CONNECTIVITY STUDIES

Brian C. Lynch, PhD

University of Pittsburgh, 2018

We develop statistical methods for two-way functional data, in which we observe a sample of functions of two continuous variables, for example space and time. Analysis of two-way functional data presents complexities not found in traditional one-way functional data, and this analysis can serve as a starting point in understanding multi-way functional data. Motivated by the concept of factorizing the signal into separate spatial and temporal components, we develop the concept of weak separability of the underlying random process. Compared to the traditional strong separability assumption, which models the covariance structure as the product of the space and time covariances, weak separability is more flexible yet still interpretable, modeling the covariance structure as a weighted sum of strongly separable components.

We propose asymptotic and bootstrap testing procedures for weak separability, and their performance is studied in simulations. We apply the testing procedures to brain imaging data, in which functional connectivity between two brain regions is measured as a function of frequency and time. We illustrate how, under weak separability, the functional process can be understood in terms of products of basis functions for frequency and time. We go on to develop methods to approximate the covariance structure using L-separability, defined as a class of decompositions of the covariance structure under weak separability, and show its relationship to nonnegative matrix factorization. Using psychiatric data as a case study, we illustrate the L-separable decomposition, as well as two-way localization methods for the

basis functions.

Keywords: asymptotics, functional principal component, Human Connectome Project, hypothesis testing, marginal kernel, schizophrenia, separable covariance, spatio-temporal data, tensor product.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 ONE-WAY AND MULTI-WAY FUNCTIONAL DATA

Functional data analysis (FDA) is the field of statistics concerned with analyzing curves or surfaces that can be thought of as arising from some underlying random process. The data are recorded over some continuum, often time, though other possibilities include frequency, age, or multivariate quantities such as three-dimensional spatial location. With advances in recent decades in data collection and storage, functional data have become increasingly common, appearing in diverse fields such as neuroscience, economics, climatology, public health, and physics.

Subjects' observed functional data are regarded as independent realizations of some random process $X(t)$, where t may be a scalar or vector with entries taking values from a continuous spectrum. That is, we observe realizations $X_i(t)$, $i = 1, \ldots, n$, of $X(t) : \mathcal{T} \to \mathbb{R}$, $\mathcal{T} \subseteq \mathbb{R}^d$, $d \geq 1$. When $d = 2$, there is sometimes special interest in analyzing the two arguments of the vector t, in which case we consider the data to be realizations $X_i(s,t)$ of a random process $X(s,t) : \mathcal{S} \times \mathcal{T} \to \mathbb{R}$, $\mathcal{S} \subseteq \mathbb{R}$, $\mathcal{T} \subseteq \mathbb{R}$. We call realizations from $X(s,t)$ *two-way* functional data, whereas the conventional functional data from $X(t)$, $t \in \mathbb{R}$, is called *one-way* functional data. Although s and t could be combined into a single vector argument, we keep them separate due to computational and interpretational issues. That is, there is interest in how the variables s and t individually relate to the observed functions, and statistical modeling should give insight into each of their effects separately. *Multi-way* functional data extends the concept of two-way functional data by considering two or more arguments, i.e., denoting the process $X(t_1, \ldots, t_k)$ for some $k \geq 2$. This thesis will develop certain aspects of the theory of two-way functional data as a starting point in understanding

multi-way functional data.

In recent decades, many analogs of classical statistical methods have been developed for one-way functional data. These include PCA, regression, classification, clustering, quantile analysis, and various inference procedures. Additionally, statistical methods have been developed to analyze data from only one realization of some process $X(s,t)$. In this case the data are usually regarded as space-time (or spatio-temporal) data, which can include weather patterns such as precipitation, wind speed, or temperature as functions of spatial location and time. Examples of spatio-temporal analysis include Fuentes (2006), Nerini et al. (2010), and Gromenko et al. (2012). Other analyses in which a single realization of $X(s,t)$ is observed include Hyndman & Ullah (2007), Huang et al. (2009), and Hyndman & Shang (2009). Several works in the area of longitudinal or multilevel functional data analyze what could be considered as i.i.d. realizations of $X(s,t)$, but in these cases one of the arguments takes only a few values, such as visits by a patient, and the focus is on modeling the data as a function of the other argument (Morris & Carroll, 2006; Di et al., 2009; Greven et al., 2011; Chen & Müller, 2012). Our interest is on data $X_i(s,t)$, $i = 1, \ldots, n$, where s and t are treated as continuous arguments, as is often the case when both s and t are observed on a dense grid.

1.2 MOTIVATION AND STRUCTURE OF THIS THESIS

We motivate our study of two-way functional data with an example of data arising from brain imaging studies. Consider a study where n subjects undergo magnetoencephalography (MEG), in which a large number of sensors around the head detect magnetic fields generated within the brain. Each sensor produces an oscillatory signal with high temporal resolution, and thus power and connectivity between signals are measured as continuous functions of time and frequency, i.e., $X_i(s,t)$, where i denotes the subject, s denotes frequency, and t denotes time. An example of connectivity between two distinct regions of the brain for a single subject is shown in Figure 1. In this thesis we analyze MEG data from the Human Connectome Project (to be discussed in Chapter 3) and the University of Pittsburgh's Clin-

ical Neurophysiology Research Laboratory (to be discussed in Chapter 5). In Chapter 3 we discuss details of how we calculate connectivity from the MEG signals.

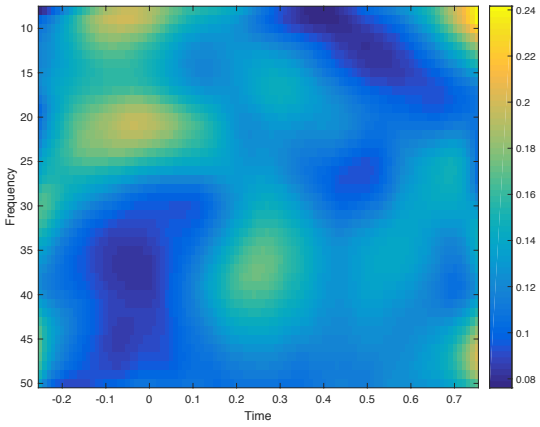

Figure 1: Connectivity between two regions of the brain for one MEG subject.

In the above scenario, we are able to record each subject's connectivity on a dense grid of p evenly spaced frequency points and q evenly spaced time points. Hence, we can store subjects i's connectivity $X_i(s,t)$ as a $p \times q$ matrix. In the example plotted in Figure 1, $p = 43$ and $q = 101$. To measure how brain connectivity varies among subjects, we wish to estimate the covariance structure $C(s,t;u,v) = \text{cov}(X(s,t), X(u,v))$, which is also essential in subsequent modeling such as functional principal component analysis (FPCA). When the data are recorded on a dense grid, we can obtain the empirical estimator of C by stacking the data into vectors of length $p \times q$, calculating the sample covariance, and reorganizing back to get the 4-dimensional covariance structure. However, the possibly large size of this structure can present problems with slow computing. Additionally, this method does not fully bring out the two-way nature of the data, and we would like to be able to interpret the connectivity in terms of frequency and time.

The separable covariance assumption $C(s,t;u,v) = aC_1(s,u)C_2(t,v)$, which we call *strong separability*, in contrast to the *weak separability* that will be introduced, is a common dimension reduction assumption. In the context of the example above, factorizing $X(s,t)$, the underlying process for the subjects' connectivity, into frequency and time components can be justified under strong separability, where the covariance function factorizes into the

product of the frequency covariance function and time covariance function. The eigenfunctions of these covariance functions are in many ways the optimal choices of basis functions for the frequency and time components of $X(s,t)$. In general, factorizing the two-way process into its two components is common, often being justified using a vague notion of "separability" of s and t, and has seen empirical success. However, rigorous justification of this type of factorization has been limited to the scope of strong separability, which can be an overly restrictive assumption. To address these concerns, we introduce the concept of *weak separability* for the process X. The analysis of weak separability shows connections to the usual strongly separable covariance structure, and provides insights into tensor methods for multi-way functional data.

The rest of this thesis is organized as follows: Chapter 2 introduces the concept of weak separability, including its relationship to two-way FPCA, its flexibility in modeling the covariance structure, its test statistic, results of numerical experiments, and application to a mortality dataset. Chapter 3 applies the concept of weak separability to MEG data from the Human Connectome Project. We give an overview of the data, the calculation of functional connectivity from MEG, results of the weak separability test, and interpretation of the separable components of the product FPCA decomposition. Chapter 4 studies L-separability, a class of decompositions of the covariance structure under weak separability, and its relationship to nonnegative matrix factorization. Chapter 5 presents a case study of psychiatric data from the University of Pittsburgh's Clinical Neurophysiology Research Laboratory, for which we apply the methods of Chapter 4 as well as two-way localization methods. Proofs are given in Appendix A.

2.0 WEAK SEPARABILITY

2.1 MOTIVATION FOR WEAK SEPARABILITY

We observe two-way functional data $X_i(s,t)$, $s \in \mathbb{R}^{d_1}$, $t \in \mathbb{R}^{d_2}$, from a random process $X(s,t)$ with mean $\mu(s,t)$ and covariance structure $C(s,t;u,v) = \text{cov}(X(s,t), X(u,v))$. A central tool in analyzing functional data is functional principal component analysis (FPCA), which, when applied to the two-way process X, is based on the Karhunen–Loève representation $X(s,t) = \mu(s,t) + \sum_{l=1}^{\infty} \xi_l h_l(s,t)$, where ξ_l $(l = 1, 2, \ldots)$ are the (random) uncorrelated coefficients, and $h_l(s,t)$ $(l = 1, 2, \ldots)$ are the eigenfunctions of the covariance operator C. To alleviate the difficulties associated with modeling the $(2d_1 + 2d_2)$-dimensional full covariance structure $C(s,t,u,v)$ and characterizing its $(d_1 + d_2)$-dimensional eigenfunctions, one generally seeks dimension reduction through factorization of the signal $X(s,t)$ into its "spatial" s and "temporal" t components. This can be justified under a common assumption, used particularly in the field of spatio-temporal analysis, called "separability", which we refer to as *strong separability* in contrast to the *weak separability* that will be introduced. Strong separability imposes $C(s,t;u,v) = aC_1(s,u)C_2(t,v)$ for some nonnegative definite functions C_1 and C_2. In terms of spatio-temporal data, this means the covariance structure factorizes into a spatial covariance function and a temporal covariance function.

There is much literature on strong separability for functional data, as well as matrix and tensor data (Lu & Zimmerman, 2005; Fuentes, 2006; Srivastava et al., 2009; Hoff et al., 2011). Recently, tests have been proposed for strong separability given a sample of independent two-way functional data (Aston et al., 2017; Constantinou et al., 2017). Many works in spatio-temporal and image analysis have used a vague notion of "separability" to justify factorizing the spatial (row) and temporal (column) components of the functional or tensor data (Zhang

& Zhou, 2005; Lu et al., 2008; Huang et al., 2009; Hung et al., 2012; Chen & Müller, 2012; Allen et al., 2014; Chen et al., 2015, 2017). Although these methods have shown empirical success, their theoretical justification is mostly restricted to strong separability.

In many applications, strong separability is an overly restrictive model of the covariance, and several works in spatio-temporal analysis have proposed alternative models (Cressie & Huang, 1999; Gneiting, 2002; Stein, 2005). These methods use additional assumptions such as stationarity, or use specific parametric models. We propose a novel concept of *weak separability* for the process X, which can be tested. Under weak separability, the eigenfunctions of $C(s,t;u,v)$ are tensor products of the eigenfunctions of the *marginal kernels* $\int_{\mathcal{T}} C(s,t;u,t)dt$ and $\int_{\mathcal{S}} C(s,t;s,v)ds$, which allows a natural factorization of $X(s,t)$ into spatial and temporal components. Weak separability also allows the covariance structure to be approximated with a weighted sum of several strongly separable components, thereby being much more flexible than strong separability, while at the same time including strong separability as a special case.

2.2 CONCEPT AND PROPERTIES OF WEAK SEPARABILITY

Let $L^2(\mathcal{T})$ denote the space of square integrable functions defined on a domain $\mathcal{T}$, i.e., $L^2(\mathcal{T}) = \{f(t),\, t \in \mathcal{T} : \int_{\mathcal{T}} f^2(t)dt < \infty\}$, with inner product $\langle f, g\rangle = \int_{\mathcal{T}} f(t)g(t)dt$ and the corresponding norm $\|\cdot\|$. For one-way functional data, denote the individual realizations as $X_i(t)$, $i = 1,\ldots,n$, from an underlying random process $X(t) \in L^2(\mathcal{T})$, $t \in \mathcal{T} \subseteq \mathbb{R}^d$, with mean $\mu(t)$ and covariance function $C(s,t) = \mathrm{cov}(X(s), X(t))$. For two-way functional data, denote the individual realizations as $X_i(s,t)$, $i = 1,\ldots,n$, from an underlying random process $X(s,t) \in L^2(\mathcal{S}\times\mathcal{T})$, $s \in \mathcal{S} \subseteq \mathbb{R}^{d_1}$, $t \in \mathcal{T} \subseteq \mathbb{R}^{d_2}$. We assume the data have well defined mean $\mu(s,t)$ and covariance $C(s,t;u,v) = \mathrm{cov}(X(s,t), X(u,v))$. We use C to denote both the covariance operator and its kernel function. We assume the covariance is continuous, and $\mathcal{S}$ and $\mathcal{T}$ are compact.

One important application that relies on covariance structure estimation is functional principal component analysis (FPCA), and this is an area where there are extra consider-

ations that must be taken when working with two-way data as opposed to one-way data. Principal component analysis is a classical dimension reduction technique from multivariate analysis, which finds a small number of uncorrelated components along which the data vary the most, based on the eigen-decomposition of the covariance matrix. One-way FPCA is well developed (see Ramsay & Silverman (2005) for an introduction), and expands the data as a sum of projections onto the first few eigenfunctions of the covariance operator. The challenge of two-way FPCA is to obtain analogs of the eigenfunctions that allow one to separately analyze the effects of the spatial (s) and temporal (t) components.

One-way FPCA can be framed in terms of the Karhunen–Loève expansion, which decomposes the process as

$$X(t) = \mu(t) + \sum_{l=1}^{\infty} \xi_l h_l(t),$$

where the h_l are eigenfunctions of C, i.e., $\int_{\mathcal{T}} C(s,t)h_l(s)ds = \theta_l h_l(t)$ for eigenvalues $\theta_1 \geq \theta_2 \geq \ldots$. The ξ_l are the random projection scores of the process, i.e., $\xi_l = \int_{\mathcal{T}}(X(t) - \mu(t))h_l(t)dt$, which can be shown to be uncorrelated. If the above sum is truncated to the first K terms, then this approximation is optimal in that, of all expansions of $X(t) - \mu(t)$ with an orthonormal basis truncated to K terms, the mean squared L^2 norm of the difference between $X(t) - \mu(t)$ and the expansion using $h_1, \ldots, h_K$ will be the smallest. That is, for any set of orthonormal functions $h_1^*, \ldots, h_K^*$ and their corresponding projection scores $\xi_1^*, \ldots, \xi_K^*$, we have $E(\int_{\mathcal{T}}(X(t) - \mu(t) - \sum_{l=1}^{K} \xi_l^* h_l^*(t))^2 dt) \geq E(\int_{\mathcal{T}}(X(t) - \mu(t) - \sum_{l=1}^{K} \xi_l h_l(t))^2 dt)$. In other words, the first K eigenfunctions of C form the K-term representation of $X(t) - \mu(t)$ with the smallest unexplained variance.

Two-way data can be decomposed using the Karhunen–Loève (often abbreviated K-L) expansion in the same way, by writing

$$X(s,t) = \mu(s,t) + \sum_{l=1}^{\infty} \xi_l h_l(s,t),$$

where the h_l are now eigenfunctions of the covariance operator corresponding to the four-dimensional covariance $C(s,t;u,v)$. There are drawbacks to this method in the two-way case, however. For one, we wish to analyze the effects of the spatial and temporal components (s and t) separately. Additionally, there can be computational problems with using a nonparametric estimate of the full covariance, i.e., $C_n(s,t;u,v) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s,t) -$

$\bar{X}(s,t))(X_i(u,v) - \bar{X}(u,v))$, where $\bar{X}(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s,t)$. When each subject's data are recorded on a dense regular $p \times q$ grid, C_n can be obtained by stacking the data into $p \times q$ vectors and computing the $pq \times pq$ sample covariance. However, this is unfeasible for large p and q (which could be hundreds or thousands in practice), causing slow computing and storage problems.

We wish to find a suitable alternative that alleviates the difficulties associated with modeling the full covariance function, and also brings out the effects of the space-time interactions. Several ways of doing this are explored in Chen et al. (2017), the most easily interpretable of which is *product FPCA*,

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\psi_j(s)\phi_k(t), \tag{2.1}$$

where ψ_j and ϕ_k are eigenfunctions of the *marginal kernels*,

$$C_{\mathcal{S}}(s,u) = \int_{\mathcal{T}} C(s,t;u,t)dt \quad \text{and} \quad C_{\mathcal{T}}(t,v) = \int_{\mathcal{S}} C(s,t;s,v)ds, \tag{2.2}$$

respectively (we also refer to these as "marginal covariance functions"). The χ_{jk} are the *marginal projection scores*, defined as

$$\chi_{jk} = \int_{\mathcal{T}}\int_{\mathcal{S}} (X(s,t) - \mu(s,t))\psi_j(s)\phi_k(t)dsdt. \tag{2.3}$$

To be precise, we should first consider expanding $X(s,t)$ in terms of completed versions of the bases of marginal eigenfunctions, but since it can be shown that the scores χ_{jk} associated with the extra functions needed to complete the bases are 0, the expansion of $X(s,t)$ in Equation (2.1) holds.

Similar decompositions have been studied to develop PCA methods for multi-way matrix or tensor data (Ye, 2005; Zhang & Zhou, 2005; Lu et al., 2008; Hung et al., 2012). When strong separability is assumed, the product FPCA representation is the same as the K-L representation. However, when strong separability does not hold, we do not know whether the products of the marginal eigenfunctions are the optimal basis functions. Moreover, the marginal projection scores are not necessarily uncorrelated, and FPCA scores being uncorrelated is a property taken for granted in subsequent modeling such as functional

regression and functional additive models. Weak separability addresses these concerns, and gives insight into tensor methods for multi-way functional data.

Consider representing the process with a general basis of product functions. For orthonormal bases $\{f_j,\, j \geq 1\}$ in $L^2(\mathcal{S})$ and $\{g_k,\, k \geq 1\}$ in $L^2(\mathcal{T})$, the product functions $\{f_j(s)g_k(t),\, j \geq 1,\, k \geq 1\}$ form an orthonormal basis of $L^2(\mathcal{S} \times \mathcal{T})$, so we can have

$$X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \tilde{\chi}_{jk} f_j(s) g_k(t),$$

where $\tilde{\chi}_{jk} = \int_{\mathcal{T}}\int_{\mathcal{S}} (X(s,t) - \mu(s,t)) f_j(s) g_k(t) ds dt$.

Definition of weak separability: *$X(s,t)$ is weakly separable if there exist orthonormal bases $\{f_j,\, j \geq 1\}$ and $\{g_k,\, k \geq 1\}$ such that $\mathrm{cov}(\tilde{\chi}_{jk}, \tilde{\chi}_{j'k'}) = 0$ for $j \neq j'$ or $k \neq k'$, i.e., the scores $\{\tilde{\chi}_{jk},\, j \geq 1,\, k \geq 1\}$ are uncorrelated with each other.*

In the following, we develop important properties of weak separability, which allow it to be useful in many applications. Detailed proofs are given in Appendix A.

Lemma 1. *If X is weakly separable, the pair of bases $\{f_j,\, j \geq 1\}$ and $\{g_k,\, k \geq 1\}$ that satisfies weak separability is unique up to a sign, and $f_j(s) \equiv \psi_j(s)$ and $g_k(t) \equiv \phi_k(t)$, where $\psi_j(s)$ and $\phi_k(t)$ are the eigenfunctions of the marginal kernels $C_{\mathcal{S}}(s,u)$ and $C_{\mathcal{T}}(t,v)$ as defined in Equation (2.2). Moreover,*

$$C(s,t;u,v) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \eta_{jk}\psi_j(s)\psi_j(u)\phi_k(t)\phi_k(v), \tag{2.4}$$

where $\eta_{jk} = \mathrm{var}(\chi_{jk})$, and the convergence is absolute and uniform.

Lemma 1 shows that when the marginal projection scores are uncorrelated, the spatial and temporal bases are eigenfunctions of the marginal covariance functions. It also allows us to test the weak separability assumption by considering the covariance of the marginal projection scores (see Section 2.3). From Equation (2.4), we see that under weak separability the eigenfunctions of C are tensor products of the marginal eigenfunctions with eigenvalues η_{jk}. Thus, the product FPCA representation is the same as the K-L expansion in that the eigenfunctions $h_l(s,t)$ from the K-L expansion will each be some product $\psi_j(s)\phi_k(t)$. When weak separability does not hold, product FPCA can still be used as a dimension reduction approach, but as the product FPCA scores are correlated, one expects to have to use more terms in product FPCA than in conventional FPCA to explain the same amount of variance.

Lemma 2. *Strong separability defined as* $C(s,t;u,v) = aC_1(s,u)C_2(t,v)$ *with identifiability constraints* $\int_{\mathcal{S}} C_1(s,s)ds = 1$ *and* $\int_{\mathcal{T}} C_2(t,t)dt = 1$ *implies weak separability of* X*. And up to a constant scaling,* C_1 *and* C_2 *are the same as the marginal kernels.*

Lemma 2 shows that strong separability is a special case of weak separability. The following lemma shows that weak separability is much more flexible than strong separability:

Lemma 3. *Let* V *denote the array* $V = (\eta_{jk},\ j \geq 1,\ k \geq 1)$*. Strong separability is weak separability with an additional assumption that* $\text{rank}(V) = 1$*. Moreover, under strong separability* $V = a\Lambda\Gamma^T$*, where* $\Lambda = (\lambda_1, \lambda_2, \ldots)^T$ *and* $\Gamma = (\gamma_1, \gamma_2, \ldots)^T$ *are the eigenvalues of the marginal kernels, and* $a = 1/\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;s,t)dsdt$ *is a normalization constant.*

Define *nonnegative rank* as $\text{rank}_+(V) = \min\{\ell : V = V_1 + \ldots + V_\ell,\ V_i \geq 0,\ \text{rank}(V_i) = 1,\ \forall\, i\}$, where $V_i \geq 0$ means that V_i is entry-wise nonnegative. If $L = \text{rank}_+(V) < \infty$, then V can be decomposed as the nonnegative factorization $V = \sum_{l=1}^{L} a^l \Lambda^l (\Gamma^l)^T$, where $\Lambda^l = (\lambda_j^l)_{j\geq 1}$ and $\Gamma^l = (\gamma_k^l)_{k\geq 1}$ are all nonnegative for $l = 1, \ldots, L$. The constant a^l is for identifiability; for example, one can require $\sum_{j\geq 1} \lambda_j^l = 1$ and $\sum_{k\geq 1} \gamma_k^l = 1$.

Let $C_{\mathcal{S}}^l(s,u) = \sum_j \lambda_j^l \psi_j(s)\psi_j(u)$ and $C_{\mathcal{T}}^l(t,v) = \sum_k \gamma_k^l \phi_k(t)\phi_k(v)$. We can generalize Lemma 3 by saying when $L = \text{rank}_+(V) < \infty$ that we have *L-separability*, in which case we can write

$$C(s,t;u,v) = \sum_{l=1}^{L} a^l C_{\mathcal{S}}^l(s,u) C_{\mathcal{T}}^l(t,v). \tag{2.5}$$

Now the full covariance function is a sum of L strongly separable components. Strong separability corresponds to 1-separability. Note that the $C_{\mathcal{S}}^l(s,u)$ have common eigenfunctions for $l = 1, \ldots, L$, which are also eigenfunctions of the marginal kernel $C_{\mathcal{S}}(s,u)$. The case for $C_{\mathcal{T}}^l(t,v)$ is analogous.

Under strong separability, the value of $C(s,t;u,v)$ for given t and v is the same, regardless of the values of s and u, up to a constant, where the constant depends on s and u. By contrast, the L-separable decomposition above allows $C(s,t;u,v)$ for given t and v to be a weighted sum of L covariance structures $C_{\mathcal{T}}^l(t,v)$, where the weights depend on s and u. An analogous statement can be made about $C(s,t;u,v)$ for given s and u.

We can see that $\text{rank}(V) \leq \text{rank}_+(V)$. In the case that $\text{rank}(V) < \text{rank}_+(V)$, finding the nonnegative rank and computing the nonnegative factorization of V are challenging

problems (Lee & Seung, 2001; Donoho & Stodden, 2003; Arora et al., 2012; Dong et al., 2014). More details will be discussed in Chapter 4. In practice, we will truncate the product FPCA expansion to the first P eigenfunctions of $C_{\mathcal{S}}$ and the first K eigenfunctions of $C_{\mathcal{T}}$, i.e., we approximate the spatial effect with P components and the temporal effect with K components. The array $V_{P,K} = (\eta_{jk}, 1 \leq j \leq P, 1 \leq k \leq K)$ often satisfies $\text{rank}(V_{P,K}) = \text{rank}_+(V_{P,K}) = \min(P, K)$, in which case the decomposition of the covariance structure into several strongly separable covariances is relatively straightforward. In this case, assume without loss of generality that $P \leq K$, so that $\text{rank}_+(V_{P,K}) = P$. If we impose the condition that Λ^j is orthogonal to Λ^l for $1 \leq j < l \leq P$, then the decomposition of $V_{P,K}$ is unique (see Chapter 4), given by Λ^l being the lth column of the identity matrix I_P, and $a^l(\Gamma^l)^T$ being the lth row of $V_{P,K}$. Interestingly, with this orthogonality condition, we can also identify this decomposition of $V_{P,K}$ with the truncated product FPCA expansion, writing $X(s,t) \approx \mu(s,t) + \sum_{l=1}^{P} X_l(s,t)$, where $X_l(s,t) = \sum_{k=1}^{K} \chi_{lk}\psi_l(s)\phi_k(t)$. Under weak separability, each X_l has the strongly separable covariance structure $a^l C^l_{\mathcal{S}}(s,u)C^l_{\mathcal{T}}(t,v)$, and the X_l are uncorrelated with each other. Hence, our covariance decomposition lends itself to a simple interpretation, being the sum of covariances of uncorrelated processes. We apply this simple decomposition in the data analysis of Section 2.5.

In applications where one relies on the separable structure of the covariance for ease of computation and interpretation, for example in applications involving the inverse of the covariance, it is not clear whether and how one can modify the concept to work under the weak separability assumption (L additive separable terms). We defer this to future research.

2.3 TEST OF WEAK SEPARABILITY

By the definition of weak separability and Lemma 1, testing weak separability is equivalent to testing the covariance structure of the marginal projection scores, i.e., $H_0 : \text{cov}(\chi_{jk}, \chi_{j'k'}) = 0$ for $j \neq j'$ or $k \neq k'$. Assume we have a sample of smooth processes $X_i(s,t) \overset{\text{i.i.d.}}{\sim} X(s,t)$. For subject i, the marginal projection scores are $\chi_{i,jk} = \int_{\mathcal{T}} \int_{\mathcal{S}} (X_i(s,t) - \mu(s,t))\psi_j(s)\phi_k(t)dsdt$, where $\psi_j(s)$ and $\phi_k(t)$ are the eigenfunctions of the marginal covariances.

Hypothesis testing of a covariance structure is a classic problem in multivariate analysis. Suppose we have n i.i.d. copies of a p-variate random vector, from a distribution with mean μ and covariance matrix Σ, and we want to test the null hypothesis that Σ is diagonal, i.e., the p variables are uncorrelated. Under the traditional multivariate setting where p is fixed and does not increase with n, likelihood ratio methods can be used to test the diagonality of Σ (Anderson, 1984). Note that these methods require distributional assumptions. The high-dimensional problem has been studied in the context that $p/n \to \gamma \in (0, \infty)$ or even for p much larger than n (Ledoit & Wolf, 2002; Liu et al., 2008; Cai et al., 2011; Lan et al., 2015). Recently, Chang et al. (2017) applied the wild bootstrap procedure (Chernozhukov et al., 2017) for hypothesis testing of large covariance matrices with few distributional assumptions.

However, unlike in the traditional covariance testing problem, we do not observe the individual values $\chi_{i,jk}$. Instead they must be estimated from the sample curves $X_i(s,t)$, $i = 1, \ldots, n$, as

$$\hat{\chi}_{i,jk} = \int_{\mathcal{T}} \int_{\mathcal{S}} (X_i(s,t) - \bar{X}(s,t)) \hat{\psi}_j(s) \hat{\phi}_k(t) ds dt, \tag{2.6}$$

where $\bar{X}(s,t) = (1/n) \sum_{i=1}^n X_i(s,t)$ and $\hat{\psi}_j$ and $\hat{\phi}_k$ are eigenfunctions of the estimated marginal covariances, $\hat{C}_{\mathcal{S}}(s,u) = (1/n) \sum_{i=1}^n \int_{\mathcal{T}} (X_i(s,t) - \bar{X}(s,t))(X_i(u,t) - \bar{X}(u,t)) dt$, and $\hat{C}_{\mathcal{T}}(t,v) = (1/n) \sum_{i=1}^n \int_{\mathcal{S}} (X_i(s,t) - \bar{X}(s,t))(X_i(s,v) - \bar{X}(s,v)) ds$. In practice, if the data for each subject are observed on dense and regularly spaced grid points and recorded in matrices X_i, $i = 1, \ldots, n$, the above estimators can be simplified as $\hat{C}_{\mathcal{S}} = (1/n) \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$, and $\hat{C}_{\mathcal{T}} = (1/n) \sum_{i=1}^n (X_i - \bar{X})^T (X_i - \bar{X})$. In the case, for example, that the argument s has dimension greater than 1, the data cannot immediately be written as matrices, but as long as the observations are dense in $\mathcal{S}$ one can vectorize them along a certain ordering of s, compute the marginal covariances, and reorganize back accordingly. From $\hat{C}_{\mathcal{S}}$ and $\hat{C}_{\mathcal{T}}$, we obtain the estimated eigenfunctions $\hat{\phi}_j$ and $\hat{\psi}_k$ by standard eigen-decomposition methods for functional data (Chen et al., 2017), and we estimate the marginal projection scores by numerical approximation of the integrals. Note that we do not have to estimate the full covariance $C(s,t;u,v)$.

Although we can prove that the $\hat{\chi}_{i,jk}$ are $\sqrt{n}$-consistent estimators of the $\chi_{i,jk}$, most test statistics based on the $\hat{\chi}_{i,jk}$ have different null distributions from their counterparts using the $\chi_{i,jk}$, and this prevents us from directly using the testing procedures mentioned above.

In the following, we develop a test for weak separability based on the empirical correlations between the estimated scores $\hat{\chi}_{i,jk}$ and $\hat{\chi}_{i,j'k'}$ $(i = 1, \ldots, n;\ (j,k) \neq (j',k'))$. The proofs involve expansions of the differences between the estimated marginal eigenfunctions and their true values, i.e., $\hat{\psi}_j - \psi_j$ and $\hat{\phi}_k - \phi_k$, as well as multi-way tensor products with indices (j,k,j',k'). The asymptotic null distribution of the test statistic is found to be a χ^2-type mixture. No Gaussian assumption on X is imposed.

2.3.1 The test statistic and its properties

Some notation: Let H be a real separable Hilbert space, with inner product $\langle \cdot, \cdot \rangle$. Following standard definitions, we denote the space of bounded linear operators on H as $\mathcal{B}(H)$, the space of Hilbert–Schmidt operators on H as $\mathcal{B}_{HS}(H)$, and the space of trace-class operators on H as $\mathcal{B}_{Tr}(H)$. For any trace-class operator $T \in \mathcal{B}_{Tr}(H)$, we define the trace as $Tr(T) = \sum_{i \geq 1} \langle Te_i, e_i \rangle$, where $(e_i)_{i \geq 1}$ is an orthonormal basis of H. It is clear that this definition is independent of the choice of basis.

For H_1 and H_2 two real separable Hilbert spaces, we use $\otimes$ as the standard tensor product, i.e., for $x_1 \in H_1$ and $x_2 \in H_2$, $(x_1 \otimes x_2)$ is the operator from H_2 to H_1 defined by $(x_1 \otimes x_2)y = \langle x_2, y \rangle x_1$ for any $y \in H_2$. Let $H = H_1 \otimes H_2$ denote the tensor product Hilbert space, which contains all finite sums of $x_1 \otimes x_2$, with inner product $\langle x_1 \otimes x_2, y_1 \otimes y_2 \rangle = \langle x_1, y_1 \rangle \langle x_2, y_2 \rangle$, for $x_1, y_1 \in H_1$ and $x_2, y_2 \in H_2$. For $C_1 \in \mathcal{B}(H_1)$ and $C_2 \in \mathcal{B}(H_2)$, we let $C_1 \tilde{\otimes} C_2$ denote the unique bounded linear operator on $H_1 \otimes H_2$ satisfying

$$C_1 \tilde{\otimes} C_2 (x_1 \otimes x_2) = C_1 x_1 \otimes C_2 x_2, \quad \text{for all } x_1 \in H_1,\ x_2 \in H_2.$$

We consider a statistic based on the sample covariance of the estimated marginal projection scores:

$$T_n(j,k,j',k') = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\chi}_{i,jk} \hat{\chi}_{i,j'k'}, \quad \text{for } j \neq j' \text{ or } k \neq k'. \tag{2.7}$$

Also define $\mathcal{Z}_n = \sqrt{n}(C_n - C)$, where the sample covariance operator is defined as

$$C_n = (1/n) \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (X_i - \bar{X}).$$

The following two conditions are needed for the theorem below and the corollary following it:

Conditions:

Condition I: For some orthonormal basis $(e_j)_{j\geq 1}$ of $L^2(\mathcal{S}\times\mathcal{T})$, $\sum_j(\mathrm{E}(\langle X, e_j\rangle^4))^{1/4} < \infty$.

Condition II: For some integers P and K, we have $\min_{1\leq j\leq P}(\lambda_j - \lambda_{j+1}) > 0$ as well as $\min_{1\leq k\leq K}(\gamma_k - \gamma_{k+1}) > 0$.

Remark: According to Proposition 5 of Mas (2006), Condition I implies that $\mathcal{Z}_n$ converges to a Gaussian random element in $\mathcal{B}_{Tr}(L^2(\mathcal{S}\times\mathcal{T}))$.

We use similar notation and conditions as used by Aston et al. (2017). However, we note that to derive the asymptotic distribution of their test statistic for strong separability, they focus on deriving the asymptotic distribution of the difference between the sample covariance operator and its strongly separable approximation. Then by projecting onto the estimated marginal eigenfunctions, they check the requirement for strong separability that $\eta_{jk} = a\lambda_j\gamma_k$. They do not need further results on the estimation errors of the marginal eigenfunctions and marginal projection scores besides that they are consistent. By contrast, our proofs involve the expansion of $\hat{\psi}_j - \psi_j$ and $\hat{\phi}_k - \phi_k$, and four-way tensor products with indices (j, k, j', k'). This requirement also differentiates our proofs from those of Fremdt et al. (2013), in which a test of equality of covariances for two functional samples is presented.

Theorem 4. *Assume Conditions I and II hold, and that X is weakly separable. For $j, j' = 1, \ldots, P$ and $k, k' = 1, \ldots, K$ as defined in Condition II, we have*
(i) for $j \neq j'$ and $k \neq k'$,

$$T_n(j, k, j', k') = Tr\left(((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) + o_p(1),$$

(ii) for $j = j'$ and $k \neq k'$,

$$\begin{aligned} T_n(j, k, j, k') =& Tr\left(((\psi_j \otimes \psi_j)\tilde{\otimes}(\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) \\ &+ Tr\left((Id_1\tilde{\otimes}(\eta_{jk'}(\gamma_k - \gamma_{k'})^{-1}\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) \\ &+ Tr\left((Id_1\tilde{\otimes}(\eta_{jk}(\gamma_{k'} - \gamma_k)^{-1}\phi_{k'} \otimes \phi_k))\mathcal{Z}_n\right) + o_p(1), \end{aligned}$$

(iii) for $j \neq j'$ and $k = k'$,

$$\begin{aligned} T_n(j,k,j',k) =& Tr\left(((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_k))\mathcal{Z}_n\right) \\ &+ Tr\left(((\eta_{jk}(\lambda_{j'} - \lambda_j)^{-1}\psi_{j'} \otimes \psi_j)\tilde{\otimes} Id_2)\mathcal{Z}_n\right) \\ &+ Tr\left(((\eta_{j'k}(\lambda_j - \lambda_{j'})^{-1}\psi_j \otimes \psi_{j'})\tilde{\otimes} Id_2)\mathcal{Z}_n\right) + o_p(1), \end{aligned}$$

where Id_1 and Id_2 are identity operators on $L^2(\mathcal{S})$ and $L^2(\mathcal{T})$, respectively.

Remark: Since $\sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'})C\right)$ is zero under the null hypothesis, the first term is the same as $\sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'})C_n\right) = \frac{1}{\sqrt{n}}\sum_i \chi_{i,jk}\chi_{i,j'k'}$, i.e., the counterpart of T_n as if we had the true marginal projection scores. The second and third terms, if they exist, are the non-negligible estimation errors.

Corollary 5. *Assume Conditions I and II hold, and that X is weakly separable. For different sets of (j,k,j',k') satisfying $1 \leq j,j' \leq P$, $1 \leq k,k' \leq K$, and $(j,k) \neq (j',k')$, the $T_n(j,k,j',k')$ are asymptotically jointly Gaussian with mean zero and covariance structure Θ. The formula for Θ is given in the proof.*

2.3.2 Tests based on χ^2-type mixtures

Lemma 6. *For $j \neq j'$, $\sum_k \mathrm{E}(\chi_{jk}\chi_{j'k}) = 0$, and for $k \neq k'$, $\sum_j \mathrm{E}(\chi_{jk}\chi_{jk'}) = 0$. This also holds in the empirical version such that for $j \neq j'$, $\sum_k T_n(j,k,j',k) = 0$, and for $k \neq k'$, $\sum_j T_n(j,k,j,k') = 0$.*

The above lemma does not assume weak separability. Recall that principal component scores in traditional one-way FPCA are uncorrelated. This lemma is a generalized result for the marginal projection scores in the product FPCA representation.

Due to this linear relationship between the different terms of T_n, the asymptotic covariance Θ will be degenerate, and thus the statistic we consider is the sum of squares of T_n without normalizing by the estimated covariance. In practice, for suitably chosen P_n and K_n, we use the statistic defined as

$$S_n = \sum_{1 \leq j,j' \leq P_n;\, 1 \leq k,k' \leq K_n;\, (j,k)<(j',k')} (T_n(j,k,j',k'))^2, \tag{2.8}$$

where $(j,k) < (j',k')$ means $(j-1) * K_n + k < (j'-1) * K_n + k'$. This is used due to the symmetry $T_n(j,k,j',k') = T_n(j',k',j,k)$, so that we only include the "upper diagonal" part of the covariance among the marginal projection scores.

Take T_n to be a long vector of length $m = P_n K_n (P_n K_n - 1)/2$ created by stacking all of the $T_n(j,k,j',k')$, $1 \le j, j' \le P_n$, $1 \le k, k' \le K_n$, $(j,k) < (j',k')$. Then by Corollary 5, $T_n \sim N_m(0, \Theta)$ under H_0, where we now take Θ to be a covariance matrix. Define the spectral decomposition of Θ as $\Theta = UQU^T$, where Q is diagonal with diagonal entries $\sigma_1, \ldots, \sigma_m$, which are the eigenvalues of Θ ordered from largest to smallest, and $U = [u_1\ u_2\ \ldots\ u_m]$, where the u_i are orthonormal column vectors. Note that by Lemma 6, some of the σ_i are 0. Since $S_n = \|T_n\|^2 = \|U^T T_n\|^2$ and $U^T T_n \sim N_m(0, Q)$, we can write $S_n = \sum_{i=1}^m \sigma_i A_i$ where the A_i are i.i.d. χ^2_1, i.e., the null distribution of S_n is a weighted sum of χ^2 distributions, which we call a χ^2-type mixture.

The Welch–Satterthwaite approximation for a χ^2-type mixture (Zhang, 2013) approximates $S_n \sim \beta \chi^2_d$ and determines β and d from matching the first 2 cumulants (the mean and the variance). This results in $\beta = \text{var}(S_n)/(2\text{E}(S_n)) = Tr(\Theta^2)/Tr(\Theta)$ and $d = 2(\text{E}(S_n))^2/\text{var}(S_n) = (Tr(\Theta))^2/Tr(\Theta^2)$. By using a plug-in estimator of Θ, we can approximate the P-value for our test as an upper tail probability of $\beta\chi^2_d$. When the first (P_n, K_n) terms do not satisfy weak separability, we have $S_n \xrightarrow{p} \infty$, by noticing that for at least one set of (j,k,j',k'), the first term in Equation (A.1) (in the proof of Theorem 4) is on the order of $\sqrt{n}$.

The consistent selection of (P_n, K_n) for hypothesis testing is a challenging problem. The optimal choice of (P_n, K_n) needs to be defined according to the problem at hand and subsequent analysis of interest. Here we focus on the subspace where the subsequent product FPCA is going to be carried out. A criterion we will use to evaluate a given choice of P_n and K_n is the "fraction of variance explained" (FVE) by the first P_n and K_n components, defined as

$$\text{FVE}(P_n, K_n) = \frac{\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{P_n} \sum_{k=1}^{K_n} \hat{\chi}^2_{i,jk}}{\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \hat{\chi}^2_{i,jk}}. \tag{2.9}$$

This definition can be justified by noting its relation to the normalized mean squared L^2

loss of the truncated process $\tilde{X}(s,t) = \mu(s,t) + \sum_{j=1}^{P_n}\sum_{k=1}^{K_n}\chi_{jk}\psi_j(s)\phi_k(t)$. In particular,

$$\frac{\mathrm{E}(\|X-\tilde{X}\|^2)}{\mathrm{E}(\|X-\mu\|^2)} = 1 - \frac{\sum_{j=1}^{P_n}\sum_{k=1}^{K_n}\eta_{jk}}{\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\eta_{jk}}.$$

The latter term is approximated by our definition of FVE. The above equality only relies on the orthogonality of the eigenfunctions, not the weak separability assumptions. Thus, it still makes sense to consider this definition of FVE even when H_0 is not true.

We also define the "marginal FVEs" $\mathrm{FVE}_{\mathcal{S}}(P_n) = \sum_{j=1}^{P_n}\hat{\lambda}_j / \sum_{j=1}^{\infty}\hat{\lambda}_j$ and $\mathrm{FVE}_{\mathcal{T}}(K_n) = \sum_{k=1}^{K_n}\hat{\gamma}_k / \sum_{k=1}^{\infty}\hat{\gamma}_k$, where the $\hat{\lambda}_j$ are the eigenvalues of $\hat{C}_{\mathcal{S}}$ and the $\hat{\gamma}_k$ are the eigenvalues of $\hat{C}_{\mathcal{T}}$. In practice the infinite sums in the denominators of $\mathrm{FVE}(P_n, K_n)$, $\mathrm{FVE}_{\mathcal{S}}(P_n)$, and $\mathrm{FVE}_{\mathcal{T}}(K_n)$ will have to be replaced with the largest number of terms that can reasonably be considered nonzero.

Without assuming weak separability, we can derive that $\sum_j \mathrm{E}\chi_{jk}^2 = \gamma_k$, $\sum_k \mathrm{E}\chi_{jk}^2 = \lambda_j$ and $\sum_{j,k}\mathrm{E}\chi_{jk}^2 = \sum_j \lambda_j = \sum_k \gamma_k$ (to see, for example, that $\sum_k \mathrm{E}\chi_{jk}^2 = \lambda_j$, take $j = j'$ in the proof of Lemma 6). Hence, we have

$$\mathrm{FVE}(P_n, K_n) \gtrsim \mathrm{FVE}_{\mathcal{S}}(P_n) + \mathrm{FVE}_{\mathcal{T}}(K_n) - 1,$$

subject to estimation error. Therefore, we propose the following procedure to choose P_n and K_n: First choose P_n and K_n such that the marginal FVEs are at least 90%. If this choice results in $\mathrm{FVE}(P_n, K_n) \geq 90\%$, use these values of P_n and K_n. If not, use the values of P_n and K_n that have marginal FVEs at least 95%, in which case $\mathrm{FVE}(P_n, K_n)$ is expected to be above 90%.

2.3.3 Bootstrap approximation

As an alternative to the χ^2-type mixture approximation, we can consider a bootstrap approach to approximate the distribution of the test statistic. Theorem 4 provides theoretical support for the empirical and parametric bootstrap procedures of this section (Van Der Vaart & Wellner, 1996). Our simulations (see Section 2.4) show that the asymptotic approximation based on the χ^2-type mixture has very satisfactory performance and appears to be superior

to the bootstrap approximation. We still present the bootstrap procedure here since it is applicable in concept to similar tests where the asymptotic null distributions do not have closed form. Additionally, we have found the computational time of the empirical bootstrap can be lower than that of the χ^2-type mixture approximation as P_n and K_n become moderately large.

Empirical bootstrap: At each step, draw a random sample from the data $X_1, \ldots, X_n$ with replacement. Denote this sample as $X_1^*, \ldots, X_n^*$. Let

$$\hat{\chi}^*_{i,jk} = \int_{\mathcal{T}} \int_{\mathcal{S}} (X_i^*(s,t) - \bar{X}^*(s,t))\hat{\psi}_j^*(s)\hat{\phi}_k^*(t)dsdt, \tag{2.10}$$

where $\bar{X}^*$ is the sample mean of the X_i^*, and the $\hat{\psi}_j^*$ and $\hat{\phi}_k^*$ are the eigenfunctions of the estimated marginal covariances of the X_i^*. The signs of the $\hat{\psi}_j^*$ and $\hat{\phi}_k^*$ are chosen to minimize $\|\hat{\psi}_j^* - \hat{\psi}_j\|$ and $\|\hat{\phi}_k^* - \hat{\phi}_k\|$, respectively. Let

$$T_n^*(j,k,j',k') = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\chi}^*_{i,jk}\hat{\chi}^*_{i,j'k'}. \tag{2.11}$$

The empirical bootstrap test statistic is calculated as

$$S_n^* = \sum_{1\le j,j'\le P_n;\, 1\le k,k'\le K_n;\, (j,k)<(j',k')} (T_n^*(j,k,j',k') - T_n(j,k,j',k'))^2.$$

This procedure is repeated B times, and the P-value is approximated as the proportion of bootstrap test statistics S_n^* that are larger than the test statistic S_n.

Validity: Theorem 3.9.13 in Van Der Vaart & Wellner (1996) can be used to prove the validity of the bootstrap procedure, i.e., the conditional random laws (given data) of S_n^* are asymptotically consistent almost surely for estimating the laws of S_n, under the null hypothesis. By Theorem 4, we have that under the null hypothesis, T_n can be written as $\Phi'_P(\sqrt{n}(\mathcal{P}_n - P)) + o(1)$ and $T_n^* - T_n$ can be written as $\Phi'_P(\sqrt{n}(\mathcal{P}_n^* - \mathcal{P}_n)) + o(1)$, where Φ'_P is a linear continuous mapping that depends on the three different cases in Theorem 4. Thus, Theorem 3.9.13 applies.

Other than the above non-studentized empirical bootstrap based on S_n, we have also considered a bootstrap procedure based on a marginally studentized test statistic, in which we

divide each term of S_n by its corresponding estimated variance $\hat{\theta}(j,k,j',k')$ (which is the plug-in estimate of $\theta(j,k,j',k')$, the diagonal entry of Θ corresponding to the asymptotic variance of $T_n(j,k,j',k')$). However, we have found this procedure is much more time consuming, and requires substantially higher sample size to achieve high power, in comparison to the non-studentized empirical bootstrap method (see Section 2.6). This is not unexpected, since the form of $\theta(j,k,j',k')$ is very complicated and plug-in estimation adds extra variability. Therefore, we do not recommend the marginally studentized empirical bootstrap method.

Parametric bootstrap: While the empirical bootstrap procedure requires no distributional assumptions on X, the parametric bootstrap procedure assumes $X \sim F(\mu, C)$. We perform the parametric bootstrap procedure as follows: At each step, generate independent $\chi^*_{i,jk} \sim F(0, \frac{1}{n}\sum_{i'=1}^{n} \hat{\chi}^2_{i',jk})$ and then define $X^*_i(s,t) = \bar{X}(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi^*_{i,jk}\hat{\psi}_j(s)\hat{\phi}_k(t)$. In practice, the infinite sums will have to be replaced with the largest number of terms that can reasonably be considered nonzero. Note that, conditional on the data $X_1, \ldots, X_n$, the X^*_i have a weakly separable covariance structure. Calculate the $\hat{\chi}^*_{i,jk}$ and $T^*_n(j,k,j',k')$ as in Equation (2.10) and Equation (2.11), respectively. The parametric bootstrap test statistic is calculated as

$$S^*_n = \sum_{1\le j,j'\le P_n;\, 1\le k,k'\le K_n;\, (j,k)<(j',k')} (T^*_n(j,k,j',k'))^2.$$

This procedure is repeated B times, and the P-value is approximated as the proportion of bootstrap test statistics S^*_n that are larger than the test statistic S_n. It is common to take F to be normal when performing this procedure, and we shall do so in our simulations (see Section 2.6).

2.4 NUMERICAL STUDY

To numerically evaluate our test of weak separability for finite sample sizes, we perform the test on simulated data. We generate independent samples of data

$$X_i(s,t) = \sum_{j=1}^{P}\sum_{k=1}^{K} \chi_{i,jk}\psi_j(s)\phi_k(t) \ (i = 1, \ldots, n),$$

where the scores $\chi_{i,jk}$ are mean 0 random variables that we generate directly. We use $P = K = 8$ to evaluate the χ^2-type mixture and (non-studentized) empirical bootstrap procedures. For this setting, we do not consider the more time-intensive parametric bootstrap, which we defer to Section 2.6.

We let s and t take values from 0 to 1 on an evenly spaced grid of 20 points. We use $\psi_j(s) = -2^{1/2}\cos\{\pi(j+1)s\}$ for j odd and $\psi_j(s) = 2^{1/2}\sin(\pi j s)$ for j even, and we define the ϕ_k by taking the first 3 B-spline functions produced by Matlab's "spcol" function using order 4 with knots at 0, 0.5, and 1, combining these with the first 5 ψ_j, and orthonormalizing using Gram–Schmidt.

Let χ_i be the vector of $\chi_{i,jk}$ for $j = 1, \ldots, P; k = 1, \ldots, K$. We simulate each χ_i independently from either $N(0, \Sigma)$ or the multivariate t distribution. In the latter case, we first simulate a vector x of length PK from $N(0, \Sigma)$. One standard definition of a multivariate t vector is $x/(u/v)^{1/2}$, where u is a chi-squared random variable with v degrees of freedom that is independent of x. However, we use $x/\{u/(v-2)\}^{1/2}$ as our multivariate t vector so that its covariance matrix is Σ. We take $v = 6$ in our simulations.

The diagonal values of Σ, the covariance matrix of the χ_{jk}, are determined by the matrix $V = \{\mathrm{var}(\chi_{jk}),\ j = 1, \ldots, P; k = 1, \ldots, K\}$. We consider two choices for V, which we denote as V_1 and V_2. We choose V_1 and V_2 such that under H_0 (when all of the off-diagonal values of Σ are 0), V_1 corresponds to a strongly separable covariance structure, while V_2 corresponds to a weakly separable structure that is not strongly separable.

We choose V_1 and V_2 to both give $\lambda_j = \exp\{1.2(9-j)\}/\left\{\sum_{j'=1}^{8}\exp(1.2j')\right\}$ $(j = 1, \ldots, 8)$ and $\gamma_k = \exp\{1.6(9-k)\}/\left\{\sum_{k'=1}^{8}\exp(1.6k')\right\}$ $(k = 1, \ldots, 8)$ as the eigenvalues of the marginal covariances $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$, respectively. V_1 is defined as the rank 1 matrix computed from the outer product of the vectors of λ_j and γ_k, i.e.,

$$V_1 = \begin{bmatrix} 0.6989 \\ 0.2105 \\ 0.0634 \\ 0.0191 \\ 0.0058 \\ 0.0017 \\ 0.0005 \\ 0.0002 \end{bmatrix} \begin{bmatrix} 0.7981 & 0.1611 & 0.0325 & 0.0066 & 0.0013 & 0.0003 & 0.0001 & 0.0000 \end{bmatrix}$$

$$= 10^{-3} \times \begin{bmatrix} 557.7586 & 112.6095 & 22.7355 & 4.5902 & 0.9267 & 0.1871 & 0.0378 & 0.0076 \\ 167.9937 & 33.9173 & 6.8478 & 1.3825 & 0.2791 & 0.0564 & 0.0114 & 0.0023 \\ 50.5987 & 10.2157 & 2.0625 & 0.4164 & 0.0841 & 0.0170 & 0.0034 & 0.0007 \\ 15.2400 & 3.0769 & 0.6212 & 0.1254 & 0.0253 & 0.0051 & 0.0010 & 0.0002 \\ 4.5902 & 0.9267 & 0.1871 & 0.0378 & 0.0076 & 0.0015 & 0.0003 & 0.0001 \\ 1.3825 & 0.2791 & 0.0564 & 0.0114 & 0.0023 & 0.0005 & 0.0001 & 0.0000 \\ 0.4164 & 0.0841 & 0.0170 & 0.0034 & 0.0007 & 0.0001 & 0.0000 & 0.0000 \\ 0.1254 & 0.0253 & 0.0051 & 0.0010 & 0.0002 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}.$$

V_2 is a rank 2 matrix with first 2 rows multiples of each other and rows 3 through 8 multiples of each other:

$$V_2 = \begin{bmatrix} 0.6989 & 0 \\ 0.2105 & 0 \\ 0 & 0.0634 \\ 0 & 0.0191 \\ 0 & 0.0058 \\ 0 & 0.0017 \\ 0 & 0.0005 \\ 0 & 0.0002 \end{bmatrix} \begin{bmatrix} 0.8680 & 0.0955 & 0.0280 & 0.0072 & 0.0011 & 0.0002 & 0.0000 & 0.0000 \\ 0.0971 & 0.8194 & 0.0778 & 0.0003 & 0.0038 & 0.0011 & 0.0003 & 0.0001 \end{bmatrix}$$

$$= 10^{-3} \times \begin{bmatrix} 606.5983 & 66.7490 & 19.5788 & 5.0235 & 0.7557 & 0.1277 & 0.0193 & 0.0008 \\ 182.7039 & 20.1044 & 5.8970 & 1.5131 & 0.2276 & 0.0385 & 0.0058 & 0.0002 \\ 6.1565 & 51.9470 & 4.9350 & 0.0221 & 0.2398 & 0.0710 & 0.0202 & 0.0069 \\ 1.8543 & 15.6461 & 1.4864 & 0.0067 & 0.0722 & 0.0214 & 0.0061 & 0.0021 \\ 0.5585 & 4.7125 & 0.4477 & 0.0020 & 0.0218 & 0.0064 & 0.0018 & 0.0006 \\ 0.1682 & 1.4194 & 0.1348 & 0.0006 & 0.0066 & 0.0019 & 0.0006 & 0.0002 \\ 0.0507 & 0.4275 & 0.0406 & 0.0002 & 0.0020 & 0.0006 & 0.0002 & 0.0001 \\ 0.0153 & 0.1288 & 0.0122 & 0.0001 & 0.0006 & 0.0002 & 0.0001 & 0.0000 \end{bmatrix}.$$

To study power, for a given choice of V_1 or V_2, we take $\text{cov}(\chi_{i,12}, \chi_{i,21})$ to be the largest positive value such that Σ is positive definite, and we also consider half of this value. Alternatively, we let 3 off-diagonal terms, $\text{cov}(\chi_{i,12}, \chi_{i,21})$, $\text{cov}(\chi_{i,11}, \chi_{i,22})$, and $\text{cov}(\chi_{i,13}, \chi_{i,31})$, take their largest positive values such that Σ is positive definite.

For each of 200 trials, we simulate data $X_i(s,t)$ $(i = 1, \ldots, n)$ in the manner described above, estimate the marginal projection scores, calculate the test statistic, and obtain P-values from the test procedures described in Section 2.3, using $B = 1000$ for the bootstrap procedure. We show simulation results with (P_n, K_n) chosen by the FVE procedure described in Section 2.3.2, and this procedure ends up choosing $P_n = 3$ and $K_n = 2$ in most trials. We also consider using set values of $(P_n, K_n) = (2,2)$, $(3,3)$, or $(4,4)$ for all trials. Empirical rejection rates at the .05 significance level with $n = 50, 100, 500$ are shown in Tables 1 through 4.

We see that both the χ^2-type mixture approximation and the empirical bootstrap procedure are able to control the Type I error under all scenarios and achieve very good power

as n or the signal increase, with the χ^2-type mixture having slightly higher power than the empirical bootstrap for smaller n. Note that even when the chosen nonzero off-diagonal covariance terms are set to their maximum values, the other off-diagonal covariance terms of Σ are zero, and so the signal is moderate. The test procedures are slightly less powerful in the multivariate t case for smaller n, as the asymptotics likely come into play more quickly for the normal data. The rejection rates are in general stable across different choices of (P_n, K_n); although $(P_n, K_n) = (2, 2)$ seems to have higher power in some cases, the power stabilizes to a reasonable value for larger (P_n, K_n).

Table 1: Rejection rates for the χ^2-type mixture weak separability test procedure, using V_1 and choosing (P_n, K_n) with the fraction of variance explained procedure (FVE) or as $(2, 2)$, $(3, 3)$, or $(4, 4)$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
n = 50								
H_0	0.055	0.020	0.020	0.040	0.025	0.020	0.005	0.045
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.715	0.785	0.740	0.755	0.440	0.445	0.395	0.370
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	0.935	0.965	0.940	0.940
3 nonzero terms	1.000	1.000	1.000	1.000	0.960	0.990	0.990	0.965
n = 100								
H_0	0.035	0.075	0.045	0.050	0.025	0.040	0.020	0.010
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.985	0.985	0.985	0.985	0.800	0.810	0.785	0.710
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	1.000	0.990	0.990	0.990
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000
n = 500								
H_0	0.060	0.060	0.040	0.055	0.020	0.045	0.020	0.045
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	1.000	1.000	1.000	1.000	0.995	0.995	1.000	0.990
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

2.5 MORTALITY DATA APPLICATION

We apply our test of weak separability to a longitudinal mortality dataset, which has previously been discussed in Chen & Müller (2012). The data are obtained from the Human Mortality Database (www.mortality.org; Wilmoth et al. (2007)), and consist of period lifetables of different countries from around the world. These lifetables show mortality rates across age for a specific period of time, giving surfaces of the form $X_i(s,t)$, $i = 1, \ldots, n$, where $X_i(s,t)$ denotes the mortality rate of country i in calendar year t for subjects of age s. $n = 27$ countries were considered, and we assume the data from these countries to be independent. Each country in the dataset has mortality rates measured on the same equally spaced grid, where t takes integer values from 1960 to 2006, and s takes integer values from 60 to 100 (chosen since the interest is on death rates of older individuals). Mortality rates tend to increase with age and decrease with year.

The covariance structure $C(s,t;u,v)$ is of interest in studying changes in mortality over age and year, and also is essential in subsequent modeling and analysis such as FPCA. Note that C has dimension $41 \times 47 \times 41 \times 47$, resulting in about 3.7×10^6 entries. Direct nonparametric estimation and visualization of C may be challenging. Looking to represent C with separate components for age and year, we apply our test of weak separability using the procedures from Section 2.3. The values of P_n and K_n are selected to be $P_n = 2$ and $K_n = 4$ by the FVE procedure described in Section 2.3.2, which gives $\text{FVE}(P_n, K_n) = 0.9102$. We obtain a P-value of 0.0623 from the χ^2-type mixture approximation and a P-value of 0.1160 from the empirical bootstrap procedure with $B = 1000$. These results are borderline, and different choices of (P_n, K_n) result in similar P-values. The higher P-value for the empirical bootstrap could reflect the trend we saw in the simulations for the empirical bootstrap to be conservative in comparison to the χ^2-type mixture. For illustrative purposes, we will assume the data to be weakly separable.

We also apply the test of strong separability proposed by Aston et al. (2017) to this dataset via their R package "covsep" (Tavakoli, 2016). We use their empirical bootstrap test function with no studentization and $B = 1000$ to get a P-value of 0.011 for $P_n = 2$ and $K_n = 4$. Since this P-value is low, we can conclude that strong separability is rejected for

the dataset.

Following Equation (2.5), we can approximate the covariance structure $C(s,t;u,v)$ with two strongly separable terms,

$$\hat{C}(s,t;u,v) = a^1 \hat{C}^1_{\mathcal{S}}(s,u)\hat{C}^1_{\mathcal{T}}(t,v) + a^2 \hat{C}^2_{\mathcal{S}}(s,u)\hat{C}^2_{\mathcal{T}}(t,v).$$

Here, $a^1 = 0.2520 \times 10^{-3}$ and $a^2 = 0.0167 \times 10^{-3}$. The components of the two separable covariance structures are plotted in Figure 2. Here, $\hat{C}^l_{\mathcal{S}}(s,u)$ and $\hat{C}^l_{\mathcal{T}}(t,v)$ are the estimated versions of $C^l_{\mathcal{S}}(s,u)$ and $C^l_{\mathcal{T}}(t,v)$, respectively, calculated using the estimated versions of the marginal eigenfunctions and marginal projection scores. The covariance between mortality rates in years t and v is a weighted sum of two covariance structures, $\hat{C}^1_{\mathcal{T}}(t,v)$ and $\hat{C}^2_{\mathcal{T}}(t,v)$, which both show that the cross covariance decays as the two years t and v move further apart. The first component $\hat{C}^1_{\mathcal{T}}(t,v)$ indicates increased covariation around 1980-1990, possibly pointing to widespread societal changes during those years. The second component $\hat{C}^2_{\mathcal{T}}(t,v)$ shows increasing covariation in more recent years. $\hat{C}^1_{\mathcal{T}}(t,v)$ is weighted more heavily than $\hat{C}^2_{\mathcal{T}}(t,v)$ as determined mainly by a^1 and a^2, but we also observe from $\hat{C}^2_{\mathcal{S}}(s,u)$ that the weights on $\hat{C}^2_{\mathcal{T}}(t,v)$ increase for the oldest ages (post 90).

The covariance between mortality rates for two ages s and u is also a weighted sum of two covariance structures. The first component $\hat{C}^1_{\mathcal{S}}(s,u)$ captures the main trend of an increased covariance as age increases. The second component $\hat{C}^2_{\mathcal{S}}(s,u)$ characterizes a low negative cross covariance between the oldest (post 90) and less old mortalities. The first component $\hat{C}^1_{\mathcal{S}}(s,u)$ captures the main covariance structure over age, but the relative weights on $\hat{C}^2_{\mathcal{S}}(s,u)$ (as determined by the values of $\hat{C}^2_{\mathcal{T}}(t,v)$) increase in recent years (after 2000).

2.6 ADDITIONAL SIMULATIONS

In Section 2.3.3 we have mentioned a possible studentized version of the empirical bootstrap test statistic. Let $\theta(j,k,j',k')$ be a diagonal term of the asymptotic covariance Θ from

Corollary 5, so that $\theta(j,k,j',k')$ is the asymptotic variance of $T_n(j,k,j',k')$. We now consider a marginally studentized test statistic defined as

$$\tilde{S}_n = \sum_{1\le j,j'\le P_n,\, 1\le k,k'\le K_n,\, (j,k)<(j',k')} (T_n(j,k,j',k'))^2/\hat{\theta}(j,k,j',k'),$$

where $\hat{\theta}(j,k,j',k')$ is the plug-in estimate of $\theta(j,k,j',k')$, calculated using the procedure described in the proof of Corollary 5. Analogous to the procedure of Section 2.3.3, we approximate the distribution of $\tilde{S}_n$ using an empirical bootstrap procedure, with bootstrap test statistic defined as

$$\tilde{S}_n^* = \sum_{1\le j,j'\le P_n,\, 1\le k,k'\le K_n,\, (j,k)<(j',k')} (T_n^*(j,k,j',k') - T_n(j,k,j',k'))^2/\hat{\theta}^*(j,k,j',k'),$$

where $\hat{\theta}^*(j,k,j',k')$ is the version of $\hat{\theta}(j,k,j',k')$ calculated using the resampled data. We repeat several of the simulations from Section 2.4 using this marginally studentized empirical bootstrap method, and empirical rejection rates at the .05 significance level from 200 simulation runs are shown in Table 5 using V_2 (the results for V_1, not shown, were similar). Compared to the non-studentized empirical bootstrap method presented in Section 2.4, the marginally studentized empirical bootstrap requires substantially higher sample size to achieve high power, and is much more time consuming. Its inferior performance is not surprising, as the form of $\theta(j,k,j',k')$ is very complicated and plug-in estimation adds extra variability.

Although the performance of our weak separability testing procedures cannot be directly compared with existing strong separability testing procedures, we apply the strong separability test of Aston et al. (2017) to our simulations to illustrate some aspects of the simulation design. We perform their test using their R package "covsep" (Tavakoli, 2016). In Table 6 we include the results from 200 simulation runs for $n = 100$ and V_2 (H_0 corresponds to weak separability but not strong separability), using their asymptotic χ^2 test and bootstrap tests (with $B = 1000$). As expected, the strong separability tests reject with high power under H_0 and all cases of H_a. In Table 7, we show results for the case of V_1 (H_0 corresponds to strong separability), again using 200 trials. Here, their asymptotic χ^2 test, which relies on Gaussian assumptions, is invalid for the multivariate t setting, and requires large n ($n = 500$

in our simulations) to respect the .05 significance level under H_0 for the normal setting. Their empirical bootstrap procedures control the Type I error under H_0 when $n = 100$ for both normal and multivariate t data.

However, as shown in Table 7, the Aston et al. (2017) strong separability test does not achieve high power under the H_a cases with V_1. Under this scenario, the array $V = V_1$ does have rank one, but the covariance is not strongly separable since Σ has nonzero off-diagonal values (our Lemmas 2 and 3 show that strong separability is weak separability plus $\text{rank}(V) = 1$). The tests proposed by Aston et al. (2017) are based on the terms $T_N(r, s) = \sqrt{N}\left(\frac{1}{N}\sum_{k=1}^{N}\langle X_k - \bar{X}_N, \hat{v}_i \otimes \hat{u}_j\rangle^2 - \hat{\lambda}_r\hat{\gamma}_s\right)$ (defined in their Equation 2.4 and after), which in our understanding mainly characterize the difference between the empirical V and the outer product of the marginal eigenvalues $\Lambda\Gamma^T$ (Aston et al. (2017) normalize their marginal covariances so that a from our Lemma 3 is 1). Therefore, it is not unexpected that their test has very low power of detecting deviations from strong separability under this setting, where the deviation occurs in the off-diagonal terms of Σ.

To evaluate the parametric bootstrap weak separability testing procedure, we consider a setting that uses $P = K = 3$ instead of $P = K = 8$. We generate the marginal projection scores analogously to Section 2.4, defining the marginal eigenfunctions as $\psi_1(s) = -2^{1/2}\cos(2\pi s)$, $\psi_2(s) = 2^{1/2}\sin(2\pi s)$, and $\psi_3(s) = -2^{1/2}\cos(4\pi s)$, and setting the ϕ_k to be the first 3 B-spline functions produced by Matlab's "spcol" function using order 4 with knots at 0, 0.5, and 1. We consider for the variances of the scores an analog of V_2 defined as

$$V_3 = \begin{bmatrix} 1 & 0 \\ 0 & .6 \\ 0 & .4 \end{bmatrix} \begin{bmatrix} .6652 & .2447 & .0900 \\ .7856 & .1753 & .0391 \end{bmatrix} = \begin{bmatrix} 0.6652 & 0.2447 & 0.0900 \\ 0.4714 & 0.1052 & 0.0235 \\ 0.3142 & 0.0701 & 0.0156 \end{bmatrix}.$$

Under H_0, like V_2, V_3 corresponds to a weakly separable structure that is not strongly separable. Using V_3, the leading eigenvalues of the marginal covariances $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$ are (1.0000, 0.6000, 0.4000) and (1.4508, 0.4200, 0.1291). To study power in this setting, we take all off-diagonal values of Σ to be zero except $\text{cov}(\chi_{i,12}, \chi_{i,21}) = c$, where c is some constant. $c = 0$ indicates that H_0 is true, while $|c| > 0$ indicates that H_0 is violated, and H_0 is violated to a larger degree for larger $|c|$. We take c to be the largest positive value, rounded down to the nearest hundredth, such that Σ is positive definite. We also do simulations with c taken to be half of this value. Alternatively, we consider a covariance structure with

$\text{cov}(\chi_{i,11}, \chi_{i,12}) = b$, $\text{cov}(\chi_{i,21}, \chi_{i,22}) = -3b/4$, and $\text{cov}(\chi_{i,31}, \chi_{i,32}) = -b/4$, where b is some constant. Note that these covariance terms satisfy Lemma 6. We do simulations where b is chosen as the largest positive value, rounded down to the nearest hundredth, such that Σ is positive definite, as well as half of this value.

In Table 8, we show simulation results from 200 trials using $P = K = 3$ and V_3. (P_n, K_n) is chosen by the FVE procedure described in Section 2.3.2, and this procedure ends up choosing $P_n = 3$ and $K_n = 2$ in most trials. In performing the parametric bootstrap, using the notation of Section 2.3.3, we generate the resampled data by generating independent $\chi^*_{i,jk} \sim N(0, \frac{1}{n}\sum_{i'=1}^{n} \hat{\chi}^2_{i',jk})$, $j = 1, \ldots, 3$, $k = 1, \ldots, 3$, and then setting $X^*_i(s,t) = \bar{X}(s,t) + \sum_{j=1}^{3}\sum_{k=1}^{3} \chi^*_{i,jk}\hat{\psi}_j(s)\hat{\phi}_k(t)$. As in Section 2.4, the χ^2-type mixture and empirical bootstrap respect the null and achieve high power as n or the signal increase, though here the empirical bootstrap is more noticeably less powerful than the χ^2-type mixture for smaller n. The parametric bootstrap performs comparably to the χ^2-type mixture approximation for the normal data, but is invalid for multivariate t data. We observed similar results when we performed these simulations using a 3×3 analog of V_1 (not shown).

As an alternative simulation method, we generate the data $X_i(s,t)$ $(i = 1, \ldots, n)$ i.i.d. directly from a distribution with mean 0 and covariance structure C, defined as follows:

$$C(s,t;u,v) = \frac{1}{(t-v)^2+1}\exp\left(-\frac{(s-u)^2}{(t-v)^2+1}\right). \tag{2.12}$$

This covariance structure is taken from Example 1 in Gneiting (2002). It is a stationary covariance structure, meaning it depends only on the differences $s-u$ and $t-v$, and it is not strongly separable. Gneiting (2002) suggests covariance structures of this type to model space-time data, for example those pertaining to environmental factors such as wind speed. This covariance structure is also used by Aston et al. (2017) in their simulations as an example of a non-strongly separable covariance structure.

For 200 trials, we simulate data $X_i(s,t)$ $(i = 1, \ldots, n)$ from either multivariate normal or multivariate t with 6 degrees of freedom, using the above covariance structure. Table 9 shows the simulation results for the weak separability testing procedures. The rejection rates (excluding those of the parametric bootstrap procedure for multivariate t data, which as discussed above are invalid in this case) are near or below .05, suggesting that this covariance

structure, though not strongly separable, is weakly separable. Using the FVE-based rule of thumb described in Section 2.3.2, most trials in these simulations end up with $P_n = 2$ and $K_n = 2$.

We visualize a few slices of $C(s,t;u,v)$ from Equation (2.12) in Figure 3, and compare these to the weakly separable approximation

$$\hat{C}(s,t;u,v) = \sum_{j=1}^{2}\sum_{k=1}^{2}(\frac{1}{n}\sum_{i=1}^{n}\hat{\chi}_{i,jk}^2)\hat{\psi}_j(s)\hat{\psi}_j(u)\hat{\phi}_k(t)\hat{\phi}_k(v),$$

which is also plotted in Figure 3. Here, the estimated eigenfunctions and scores are obtained from a sample $X_i(s,t)$, $i = 1,\ldots,500$, that are generated i.i.d. normal with mean 0 and covariance structure C, with s and t taking values from 0 to 1 on an evenly spaced grid of 100 points. We see that $C(s,t;u,v)$ and $\hat{C}(s,t;u,v)$ are fairly similar, supporting the results of the test of weak separability.

Table 2: Rejection rates for the χ^2-type mixture weak separability test procedure, using V_2 and choosing (P_n, K_n) with the fraction of variance explained procedure (FVE) or as $(2,2)$, $(3,3)$, or $(4,4)$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
n = 50								
H_0	0.030	0.025	0.030	0.025	0.015	0.030	0.015	0.005
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	0.515	0.845	0.440	0.465	0.305	0.555	0.225	0.205
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	0.995	0.995	0.995	0.990	0.850	0.955	0.825	0.770
3 nonzero terms	1.000	1.000	1.000	1.000	0.965	0.985	0.950	0.970
n = 100								
H_0	0.045	0.055	0.035	0.040	0.010	0.050	0.040	0.020
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	0.920	0.990	0.930	0.920	0.625	0.900	0.605	0.500
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.990	1.000	0.965	0.955
3 nonzero terms	1.000	1.000	1.000	1.000	0.995	1.000	1.000	0.980
n = 500								
H_0	0.045	0.065	0.025	0.040	0.025	0.065	0.050	0.035
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	1.000	1.000	1.000	1.000	0.970	1.000	0.995	0.995
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.995
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 3: Rejection rates for the non-studentized empirical bootstrap weak separability test procedure, using V_1 and choosing (P_n, K_n) with the fraction of variance explained procedure (FVE) or as $(2,2)$, $(3,3)$, or $(4,4)$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
n = 50								
H_0	0.035	0.025	0.035	0.025	0.010	0.010	0.005	0.005
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.670	0.680	0.650	0.640	0.350	0.375	0.330	0.300
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	0.890	0.900	0.885	0.880
3 nonzero terms	1.000	1.000	1.000	1.000	0.920	0.920	0.920	0.915
n = 100								
H_0	0.070	0.060	0.060	0.060	0.015	0.010	0.015	0.015
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.990	0.995	0.985	0.985	0.735	0.785	0.700	0.680
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	0.970	0.970	0.965	0.960
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
n = 500								
H_0	0.065	0.050	0.060	0.060	0.055	0.050	0.055	0.055
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	1.000	1.000	1.000	1.000	0.995	0.995	0.995	0.995
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4: Rejection rates for the non-studentized empirical bootstrap weak separability test procedure, using V_2 and choosing (P_n, K_n) with the fraction of variance explained procedure (FVE) or as $(2,2)$, $(3,3)$, or $(4,4)$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
n = 50								
H_0	0.065	0.045	0.050	0.045	0.015	0.020	0.010	0.010
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	0.420	0.785	0.390	0.380	0.220	0.395	0.160	0.150
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	0.970	1.000	0.965	0.965	0.730	0.865	0.690	0.675
3 nonzero terms	1.000	1.000	1.000	1.000	0.895	0.925	0.885	0.885
n = 100								
H_0	0.025	0.010	0.020	0.020	0.035	0.035	0.030	0.020
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	0.950	1.000	0.955	0.955	0.585	0.855	0.575	0.515
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.980	0.995	0.980	0.975
3 nonzero terms	1.000	1.000	1.000	1.000	0.985	0.990	0.985	0.985
n = 500								
H_0	0.030	0.065	0.030	0.030	0.010	0.035	0.005	0.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

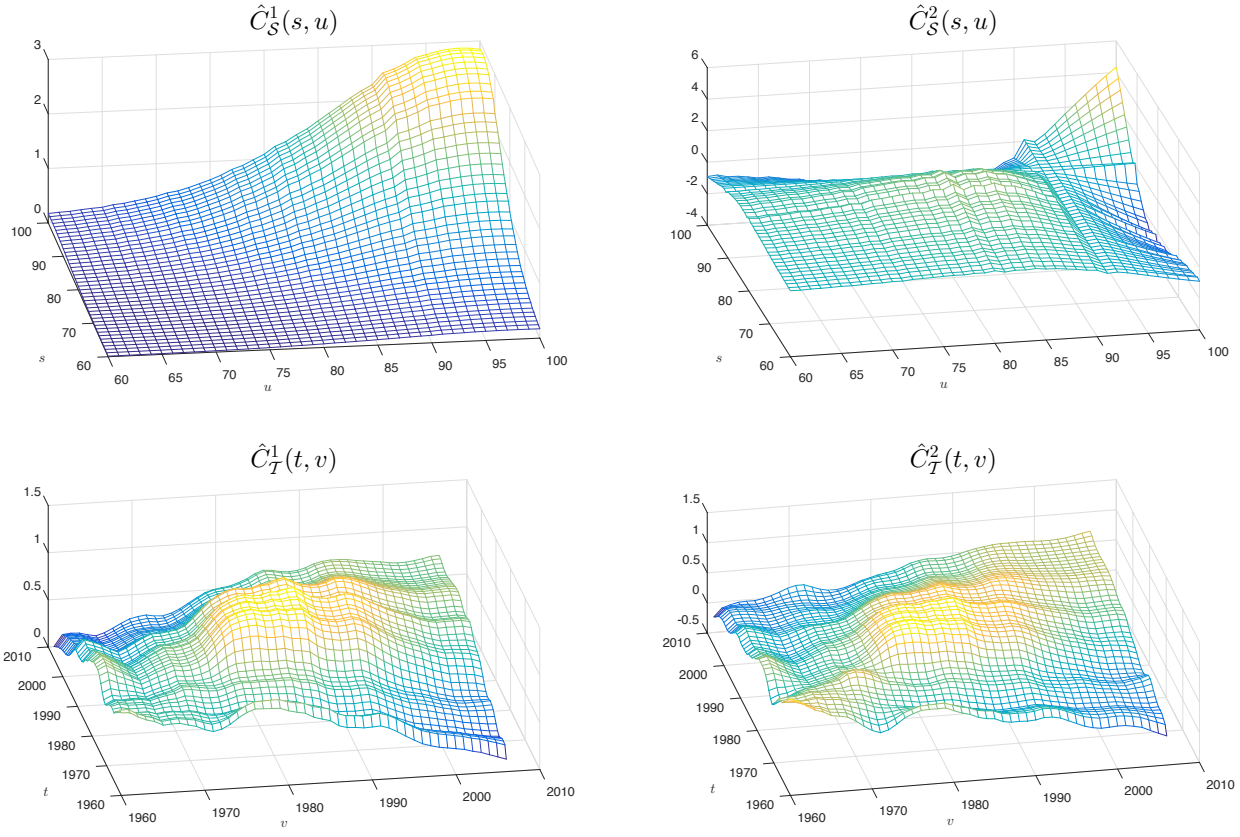

Figure 2: Plot of the components of the decomposition of $C(s,t;u,v)$ for the mortality data. To improve visibility, slight smoothing was done on $\hat{C}^1_{\mathcal{T}}$ and $\hat{C}^2_{\mathcal{T}}$.

Table 5: Rejection rates for the marginally studentized empirical bootstrap weak separability test procedure, using V_2 and choosing (P_n, K_n) with the FVE procedure or as $(2,2)$, $(3,3)$, or $(4,4)$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
$\mathbf{n = 100}$								
H_0	0.030	0.060	0.020	0.000	0.000	0.015	0.000	0.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	0.405	0.550	0.265	0.035	0.140	0.265	0.045	0.000
$\mathbf{n = 500}$								
H_0	0.040	0.065	0.040	0.010	0.025	0.025	0.020	0.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.970	0.980	0.935	0.675
$\mathbf{n = 1000}$								
H_0	0.070	0.060	0.060	0.030	0.040	0.030	0.020	0.010
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.970	0.970	0.970	0.930

Table 6: Rejection rates for the strong separability test procedures of Aston et al. (2017), using V_2 and $n = 100$, and choosing (P_n, K_n) with the FVE procedure or as $(2,2)$, $(3,3)$, or $(4,4)$. The test procedures include asymptotic χ^2, non-studentized empirical bootstrap ("non-studentized"), and marginally studentized empirical bootstrap ("marginal").

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
asymptotic χ^2								
H_0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3 nonzero terms	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
non-studentized								
H_0	1.000	1.000	1.000	1.000	1.000	0.985	1.000	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	1.000	1.000	1.000	1.000	0.990	0.985	0.990	0.990
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.975	0.950	0.975	0.975
3 nonzero terms	1.000	1.000	1.000	1.000	0.975	0.955	0.990	0.990
marginal								
H_0	1.000	1.000	1.000	1.000	0.995	0.970	0.995	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.055$	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000
$\text{cov}(\chi_{12}, \chi_{21}) = 0.11$	1.000	1.000	1.000	1.000	0.975	0.905	0.990	0.995
3 nonzero terms	1.000	1.000	1.000	1.000	0.975	0.915	0.990	1.000

Table 7: Rejection rates for the strong separability test procedures of Aston et al. (2017), using V_1 and choosing (P_n, K_n) with the FVE procedure or as $(2,2)$, $(3,3)$, or $(4,4)$. The test procedures include asymptotic χ^2, non-studentized empirical bootstrap ("non-studentized"), and marginally studentized empirical bootstrap ("marginal"). The asymptotic χ^2 procedure uses $n = 500$, while the bootstrap procedures use $n = 100$.

	Normal				Multivariate t			
	FVE	(2,2)	(3,3)	(4,4)	FVE	(2,2)	(3,3)	(4,4)
asymptotic χ^2								
H_0	0.050	0.035	0.050	0.040	0.290	0.285	0.380	0.560
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.045	0.040	0.050	0.040	0.320	0.295	0.415	0.610
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	0.080	0.085	0.095	0.090	0.290	0.230	0.360	0.545
3 nonzero terms	0.135	0.140	0.105	0.090	0.370	0.305	0.405	0.535
non-studentized								
H_0	0.055	0.055	0.060	0.060	0.055	0.060	0.055	0.055
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.070	0.070	0.070	0.070	0.060	0.060	0.060	0.060
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	0.065	0.065	0.065	0.065	0.055	0.055	0.055	0.055
3 nonzero terms	0.055	0.055	0.055	0.055	0.055	0.055	0.055	0.055
marginal								
H_0	0.045	0.050	0.060	0.040	0.035	0.040	0.060	0.035
$\text{cov}(\chi_{12}, \chi_{21}) = 0.065$	0.050	0.050	0.040	0.040	0.040	0.050	0.035	0.030
$\text{cov}(\chi_{12}, \chi_{21}) = 0.13$	0.055	0.060	0.055	0.045	0.045	0.050	0.050	0.050
3 nonzero terms	0.025	0.045	0.030	0.025	0.050	0.045	0.050	0.035

Table 8: Rejection rates for the weak separability test procedures using $P = K = 3$ and V_3. "χ^2" denotes the χ^2-type mixture approximation, "Emp" denotes the empirical bootstrap, and "Para" denotes the parametric bootstrap.

	Normal			Multivariate t		
	χ^2	Emp	Para	χ^2	Emp	Para
n = 50						
$c = 0$, $b = 0$	0.040	0.000	0.060	0.015	0.000	0.330
$c = .165$, $b = 0$	0.610	0.100	0.775	0.295	0.015	0.875
$c = .33$, $b = 0$	0.985	0.530	1.000	0.910	0.210	1.000
$c = 0$, $b = .145$	0.710	0.230	0.875	0.470	0.060	0.905
$c = 0$, $b = .29$	0.960	0.475	1.000	0.850	0.340	1.000
n = 100						
$c = 0$, $b = 0$	0.055	0.010	0.090	0.020	0.000	0.345
$c = .165$, $b = 0$	0.960	0.690	0.995	0.700	0.250	0.980
$c = .33$, $b = 0$	1.000	0.905	1.000	0.985	0.695	1.000
$c = 0$, $b = .145$	0.990	0.760	0.995	0.785	0.325	0.985
$c = 0$, $b = .29$	1.000	0.865	1.000	0.945	0.620	1.000
n = 500						
$c = 0$, $b = 0$	0.035	0.040	0.045	0.030	0.015	0.445
$c = .165$, $b = 0$	1.000	1.000	1.000	1.000	0.980	1.000
$c = .33$, $b = 0$	1.000	1.000	1.000	1.000	1.000	1.000
$c = 0$, $b = .145$	1.000	1.000	1.000	0.995	0.955	1.000
$c = 0$, $b = .29$	1.000	1.000	1.000	1.000	0.960	1.000

Table 9: Rejection rates for the weak separability test procedures using the covariance structure from Equation (2.12). "χ^2" denotes the χ^2-type mixture approximation, "Emp" denotes the empirical bootstrap, and "Para" denotes the parametric bootstrap.

Scenario	n=100			n=500		
	χ^2	Emp	Para	χ^2	Emp	Para
Normal	0.045	0.055	0.060	0.050	0.035	0.045
Multivariate t	0.020	0.050	0.215	0.025	0.035	0.210

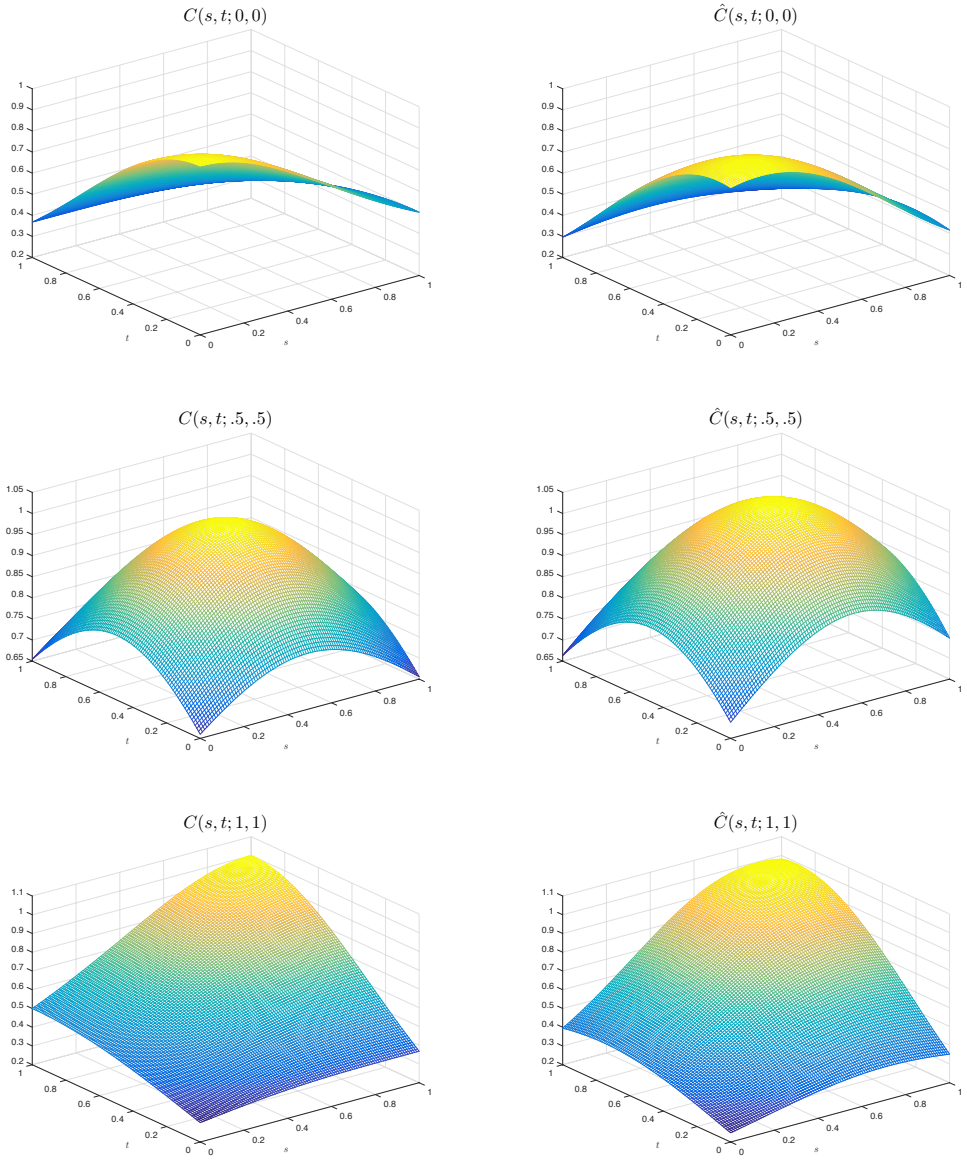

Figure 3: Plots of $C(s,t;u,v)$ and $\hat{C}(s,t;u,v)$ for fixed values of u and v, where C is the covariance structure from Equation (2.12).

3.0 BRAIN IMAGING DATA ANALYSIS

3.1 BACKGROUND

Brain imaging analysis is an area where functional data increasingly arise. In particular, there is an increasing interest in statistical modeling in studies using Functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) data (Lindquist et al., 2008; Chavez et al., 2010; Larson-Prior et al., 2013; Eloyan et al., 2014). An important goal in these studies is to analyze functional connectivity, which "is defined as the temporal dependency of neuronal activation patterns of anatomically separated brain regions" (Van Den Heuvel & Pol, 2010). We focus on MEG, which measures neuronal activity by recording magnetic fields generated within the brain. Due to the high temporal resolution and oscillatory nature of the MEG signal, MEG-based connectivity measures are calculated as functions of time and frequency, and thus can naturally be analyzed as two-way functional data. In this chapter we show how our methods can be applied to MEG data to analyze functional connectivity between two chosen regions of the brain.

We use MEG data collected through the Human Connectome Project (HCP), a study that has compiled a large amount of high quality multi-modal neural data (Van Essen et al., 2013). The project, which started in 2009, is led by groups from Washington University, University of Minnesota, and Oxford University. It has obtained fMRI and MEG scans for resting state and task experimental designs using 1200 human subjects, which include healthy adults with ages ranging from 22 to 35 years old. 100 subjects have MEG data, and these subjects are comprised of 50 monozygotic twin pairs (Larson-Prior et al., 2013), but in our analyses we do not use the pair info. Much of the data, both raw and processed to varying degrees, have been made freely accessible through the HCP's ConnectomeDB

database. Additionally, the HCP has continuously carried out analyses of the data, one of its aims being to generate a "parcellated connectome" of networks within the brain (Van Essen et al., 2013).

MEG uses a helmet of magnetometers around the subject's head to record magnetic fields generated within the brain. This gauges current density from within the brain that can represent neuronal activity. The temporal resolution of MEG is about 1 millisecond. Functional connectivity between spatially separated regions of the brain has been shown to change over time, and the high temporal resolution of MEG is useful in detecting coupling of distinct regions at various frequencies (Pizzella et al., 2014). However, because of the distance of the sensors from the brain (a few centimeters), the spatial resolution of MEG is no more than 10 millimeters at the surface of the brain (Larson-Prior et al., 2013), and source reconstruction methods are often used to try to recover the true signal originating at the cortical surface (see Section 3.2.2). We will focus on source-reconstructed data, but in Section 3.3.2 we will also consider analyses using only the raw signals from the sensors.

MEG data from the HCP is collected using a whole head MAGNES 3600 in a magnetically shielded room (Larson-Prior et al., 2013). Each scan uses around 250 magnetometer sensors, as well as around 20 reference sensors placed far from the head to estimate ambient noise levels. Different types of MEG studies have been done as part of the HCP, and we will focus on the motor and working memory tasks due to their relatively simple designs and numerous trials.

The MEG session in which motor task data is collected is split into several blocks, each with 10 trials. The subjects face a screen, and in each trial they are directed by an arrow on the screen to move a body part. Depending on the block, this body part is either the right hand, left hand, right foot, or left foot. In each trial of the working memory session, an image is displayed for 2 seconds, after which the subject has 0.5 seconds to indicate, with the push of a button, whether the image matches a "target" image that was shown earlier. In each trial of the "0-back" blocks, the subject must decide whether the displayed image matches the image that was shown at the beginning of the block. In each trial of the "2-back" blocks, the subject must decide whether the displayed image matches what was shown 2 images prior. More details about the tasks can be found in WU-Minn HCP (2017).

3.2 PROCESSING THE DATA

3.2.1 HCP preprocessed data

We start with the sensor-level preprocessed MEG data from the HCP's "tmegpreproc" pipeline, where the signal from each sensor has been separated into trials as described above. Faulty sensors have been removed, as well as trials/sensors with excessive noise or artifacts due to head motion (WU-Minn HCP, 2017). The sensor time series are demeaned and filtered through a band-pass filter of 1.3-150 Hz, and components of the signal resulting from artifacts such as cardiac pulsation or eye movement have been classified and removed using independent component analysis (Van Essen et al., 2013; WU-Minn HCP, 2017).

We download the data from the HCP Amazon S3 bucket using the "hcp" package of the R platform Neuroconductor (Muschelli, 2017). Credentials for access to the Amazon S3 bucket can be obtained from the HCP ConnectomeDB website (https://db.humanconnectome.org).

3.2.2 Source reconstruction

From the preprocessed MEG data, we derive two-way functional connectivity data using the Matlab package FieldTrip (Oostenveld et al., 2010). We are interested in calculating connectivity between different regions of interest (ROIs) of the brain. As the sensors are distant from the brain, using their signals to represent ROIs can lead to spurious connectivity measurements. This is due to the volume conduction/field spread problem, in which each sensor picks up the activity of several sources, as well as the common input problem, in which a common source provides input to a pair of signals that do not directly interact (Larson-Prior et al., 2013; Bastos & Schoffelen, 2015). With this in mind, we use source reconstruction to estimate the signals that we imagine directly arising at the locations of the ROIs. Source reconstruction is common in MEG analysis, but it is an inverse problem on which constraints must be placed to obtain a unique solution (Pizzella et al., 2014). The problem either models the sources as a few current dipoles corresponding to a few brain regions, or considers a grid of current dipoles across the brain. We will take the latter approach, generating signals for dipoles distributed across the cortical surface, and then

averaging the signals of all the dipoles within each ROI.

The source reconstruction method we use is Minimum Norm Estimation (MNE). This is a widely used method that is implemented in FieldTrip, but it should be noted that many competing source reconstruction methods exist, and each has strengths and drawbacks (Ou et al., 2009; Jensen & Hesse, 2010). We observe signals $x_i(t)$, $i = 1, \ldots, N$, where N is the number of sensors. We wish to determine the current magnitudes $y_j(t)$ of the dipoles, $j = 1, \ldots, M$, where M is the number of dipoles in a dense grid on the cortical surface, and M is usually much larger than N. We denote the signals of the sensors at time t as a vector $x = [x_1(t), \ldots, x_N(t)]^T$ and the current magnitudes as $y = [y_1(t), \ldots, y_M(t)]^T$. Given a model of the grid of dipoles and a volume conduction model of the head, Maxwell's equations in the quasistatic case can be used to derive the linear relationship $x = Ay$, known as the *forward model*, where A is a known $N \times M$ matrix (which is derived independently of x and does not depend on t) called the *lead field* (Dale & Sereno, 1993; Hämäläinen & Ilmoniemi, 1994; Jensen & Hesse, 2010). The forward model presents an underdetermined system of linear equations.

A more general version of the problem takes $x = Ay + n$, where n is a vector of independent measurement errors. y is taken to be a random vector, the covariance matrices of y and n are assumed to be known, and the solution is taken to be of the form $\hat{y} = Wx$ for some $M \times N$ matrix W (Dale & Sereno, 1993; Dale et al., 2000; Lin et al., 2004). FieldTrip uses the formulation of Lin et al. (2004), which, for estimates C and R of the covariance matrices of n and y, respectively, and a scaling parameter λ related to the signal-to-noise ratio, minimizes

$$\|C^{-1/2}(x - Ay)\|_2^2 + \lambda^2 \|R^{-1/2} y\|_2^2$$

over y, where $\|\cdot\|_2$ denotes the L_2 norm. This has the solution $W = RA^T(ARA^T + \lambda^2 C)^{-1}$. Note that in practice, the forward model takes the current at each dipole to be a vector quantity, and models the x, y, and z components of each current dipole separately, so M is actually 3 times the number of current dipoles. Thus, for each dipole we end up with 3 time series, and to get a single time series we project the current vector at each time point onto the orientation that was found to be the strongest over time.

For each subject, the HCP provides a volume conduction model of the head, and a source

model which defines positions of 8004 dipoles on the surface of the brain. The positions of the dipoles have been registered across subjects so that each dipole is comparable (WU-Minn HCP, 2017). FieldTrip provides an atlas file that assigns each of these dipoles to a region based on Glasser et al. (2016), and from this we choose which source reconstructed signals to average to generate a time series for a given ROI.

3.2.3 Time-frequency representation

MEG signals are inherently oscillatory, and synchronization at certain frequency ranges of the activity in different ROIs has been shown to be related to tasks performed by the brain (Pizzella et al., 2014). This synchronization is generally not captured by simple connectivity measures like Pearson correlation, as they ignore the temporal nature of the signals (Bastos & Schoffelen, 2015). To study how frequency-based coupling between ROIs changes over the course of a task, we calculate the time-frequency representations (TFRs) of their signals. The TFR of a signal is its representation at time t and frequency s as a complex number $A(s,t)e^{iB(s,t)}$, where $A(s,t)$ is the amplitude and $B(s,t)$ is the phase.

There are several ways to calculate the TFR, including Fourier decomposition, wavelet analysis, and Hilbert transformation, and these methods have been found to produce similar results (Le Van Quyen et al., 2001; Quiroga et al., 2002; Bastos & Schoffelen, 2015). We use the method of Morlet wavelets, which is implemented in FieldTrip's ft_freqanalysis function. For each pair of time and frequency points (s_0, t_0) of interest, we consider a window of time points around t_0 whose size depends on s_0, using a larger window for smaller s_0. We construct a complex Morlet's wavelet function (also known as a complex Gabor wavelet) as $w(s_0,t) = A\exp(-t^2/(2\sigma^2))\exp(i2\pi s_0 t)$, where $A = (\sigma\sqrt{\pi})^{-1/2}$ is a normalization constant and $\sigma = m/(2\pi s_0)$ determines how wide a time window around t_0 is considered, where m is a constant. Denoting our signal as $x(t)$, the TFR at (s_0, t_0) is calculated as the convolution of x with the wavelet, i.e., $\int_{-\infty}^{\infty} x(t)w(s_0, t_0 - t)dt$. Following Chavez et al. (2010), we choose $m = 7$, which is also FieldTrip's default value. This choice indicates the time window around t_0 includes approximately 7 cycles of the frequency s_0 (although the convolution is over infinity, the values of the wavelet function outside this window are seen as negligible).

For the motor data, the signal for each trial is recorded from -1.2 to 1.2 seconds in intervals of about 2 ms, where time 0 corresponds to the start of an Electromyographic (EMG) signal, which indicates motion of the body part using electrodes attached to the hands and feet. From plotting examples of the EMG signal, we find the motion usually lasts no longer than about .75 seconds. The signal at a time period shortly before time 0 is of interest, as it can represent brain activity when the subjects have received the movement cue but have not yet reacted to it. However, the trials are not disjoint, so the signal at times further before 0 overlaps with the signal from the end of the previous trial. Thus, in our analysis we consider times for each trial in the range of -.25 to .75 seconds. We only consider trials in which the right hand moved. In the preprocessed data, there are 61 subjects with motor data, and the subjects have an average of 75.38 of these trials.

For the working memory data, the signal for each trial is recorded from -1.5 to 2.5 seconds in intervals of about 2 ms, where time 0 is defined to be when the image is shown. The image is shown for 2 seconds before the subjects can respond by pushing the button, so we will consider times for each trial on the range of 0 to 2 s, since this is when subjects are using their working memory. We will consider two datasets, one using the 2-back trials, and another using the 0-back trials. In each, we will only use the trials where the subjects answered correctly. In the preprocessed data, there are 83 subjects with working memory data, and the average number of correct trials is 68.83 and 68.71 for the 2-back and 0-back designs, respectively.

MEG studies often group activity into 8 bands of frequencies, which are defined in Table 10, with task-based MEG usually modulating power within the theta through gamma bands (Larson-Prior et al., 2013). Because we calculate the time-frequency representation using wider time windows for lower frequencies, we are limited in how low of frequencies we can consider. Our preliminary results for power (see Section 3.2.4) show a lack of activity above 50 Hz in our data. With these considerations, we calculate the TFRs using frequencies from the alpha to gamma low bands, i.e., 8 to 50 Hz. We use a spacing of 1 Hz between frequency points, and .01 s between time points. Thus, the TFRs (and hence the connectivity functions of the following section) are calculated on 43×101 grids for the motor dataset and 43×201 grids for the working memory datasets.

Table 10: Frequency bands.

Frequency band	Frequency range (Hz)
Delta	1.5-4
Theta	4-8
Alpha	8-15
Beta low	15-26
Beta high	26-35
Gamma low	35-50
Gamma mid	50-76
Gamma high	76-120

3.2.4 Connectivity analysis

We construct two-way functional data as connectivity between the signals representing ROIs. Many functional connectivity measures are based on an analog of the cross-correlation function called the coherence (Bastos & Schoffelen, 2015). Given TFRs $A_{1,k}(s,t)e^{iB_{1,k}(s,t)}$ and $A_{2,k}(s,t)e^{iB_{2,k}(s,t)}$ for two signals recorded at trial k, $k = 1, \ldots, n_T$, the coherence is calculated as

$$\frac{(1/n_T)\sum_{k=1}^{n_T} A_{1,k}(s,t)A_{2,k}(s,t)e^{i(B_{1,k}(s,t)-B_{2,k}(s,t))}}{\sqrt{\{(1/n_T)\sum_{k=1}^{n_T} A_{1,k}^2(s,t)\}\{(1/n_T)\sum_{k=1}^{n_T} A_{2,k}^2(s,t)\}}}.$$

We consider the Phase Locking Value (PLV) (Lachaux et al., 1999), which disregards the amplitudes and considers only the magnitude of the average of the phase differences as unit vectors in the complex plane. It is defined as

$$\mathrm{PLV(s,t)} = (1/\mathrm{n_T})\left|\sum_{\mathrm{k=1}}^{\mathrm{n_T}} \mathrm{e}^{\mathrm{i(B_{1,k}(s,t)-B_{2,k}(s,t))}}\right|.$$

The PLV takes values from 0 to 1, with 1 indicating complete phase synchrony over trials and 0 indicating no phase synchrony. The PLV has gained popularity due to the belief that phase differences reveal more about functional connectivity than changes in amplitude (Lachaux et al., 1999; Aydore et al., 2013; Bastos & Schoffelen, 2015). A downside of the PLV is that mixing of the signals due to field spread can lead to artificially high connectivity

at 0 phase lag (Aydore et al., 2013). However, field spread should be less of an issue when source reconstruction is used. The imaginary part of the coherence has been proposed as an alternative connectivity measure that removes the 0 phase part of the signal (Bastos & Schoffelen, 2015), but it is harder to interpret due it taking on values between -1 and 1, where the sign depends on which signal is defined to be the first.

A conceptual illustration of what PLV measures can be seen in Figure 3 in Bastos & Schoffelen (2015). This figure shows 3 examples, denoted (A), (B), and (C), of 2 waveforms of frequency f over 4 trials, plotting their imaginary parts as well as the complex exponential of their phase difference, i.e., $e^{i(B_{1,k}(f,t)-B_{2,k}(f,t))}$. The latter is plotted as a vector in the complex plane. The waveforms can be thought of as TFRs with their amplitudes removed, both evaluated at frequency f. In (A), the 2 waveforms have 0 phase lag, i.e., $B_{1,k}(f,t) - B_{2,k}(f,t) = 0$, over all trials. In (B), the waveforms have a phase lag of $\pi/2$ radians, i.e., $B_{1,k}(f,t) - B_{2,k}(f,t) = \pi/2$, over all trials. In (C), the phase lags over the 4 trials are 0, $\pi/2$, π, and $3\pi/2$, respectively. (A) and (B) would both result in PLVs of 1 at frequency f, since in both examples $e^{i(B_{1,k}(f,t)-B_{2,k}(f,t))}$ is the same over all trials. (C) would result in a PLV of 0 at frequency f, since the vectors representing $e^{i(B_{1,k}(f,t)-B_{2,k}(f,t))}$ cancel out when summed over all trials.

Our connectivity analysis will focus on two regions that are spatially separated, likely activated during the task, and potentially functionally connected. For the motor data, these will be the left primary motor cortex (M1) and the right inferior parietal lobule (IPL). For the working memory data, we will use the left dorsolateral prefrontal cortex (DLPFC) and the left inferior parietal lobule (IPL). The DLPFC lies within the prefrontal cortex (PFC), and the IPL lies within the posterior parietal cortex (PPC). The M1, forming a ridge down the frontal lobe, is central in planning and enacting movement. Studies in humans and monkeys have shown the PPC to be functionally connected to the M1, and the IPL to be activated during motor tasks (Mattingley et al., 1998; Guye et al., 2003; Fogassi et al., 2005). We choose the left M1 because we use trials where the right hand moved, and we choose the right IPL because the left IPL is spatially close to the left M1. Working memory task studies in humans and monkeys have found the tasks activate the DLPFC (Friedman & Goldman-Rakic, 1994; Levy & Goldman-Rakic, 2000; Owen et al., 2005; Mars & Grol, 2007), as well

as the IPL (Friedman & Goldman-Rakic, 1994; LaBar et al., 1999; Owen et al., 2005). The locations of our ROIs on the cortical surface are plotted in Figure 4. There are 101 dipoles comprising the left M1, 299 comprising the right IPL, 325 comprising the left IPL, and 268 comprising the left DLPFC.

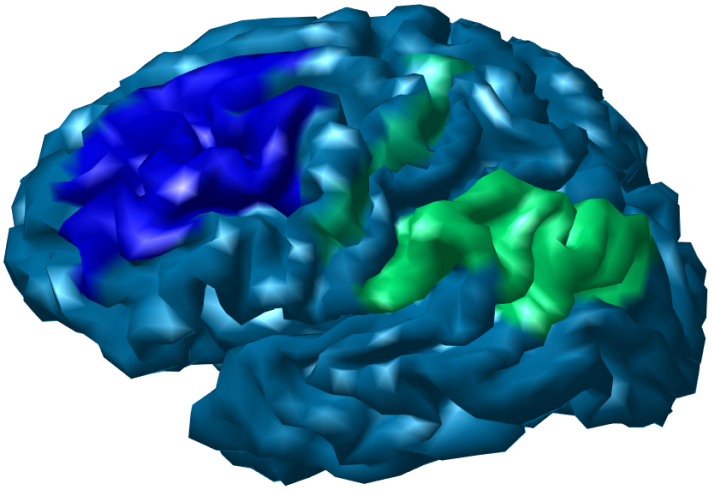

Figure 4: Plot of the positions of the ROIs in the left hemisphere. Dark green points represent the M1, dark blue points represent the DLPFC, and light green points represent the IPL.

In Figure 5, we plot the power of the source-reconstructed signal corresponding to each ROI, averaged over all trials and all subjects (including the motor subjects for the left M1 and right IPL, and the working memory subjects for the left DLPFC and left IPL). Power is defined as the squared amplitude of the TFR, and it gives an idea of the overall level of activity at different frequencies. These plots show that there is little activity as the frequency nears 50 Hz.

The PLV matrices are noisy, so we do smoothing on each subject's PLV values using a multivariate local linear regression estimator (Fan & Gijbels, 1996). This requires specifying some bandwidths bw_s and bw_t in the frequency and time directions, respectively. Denote

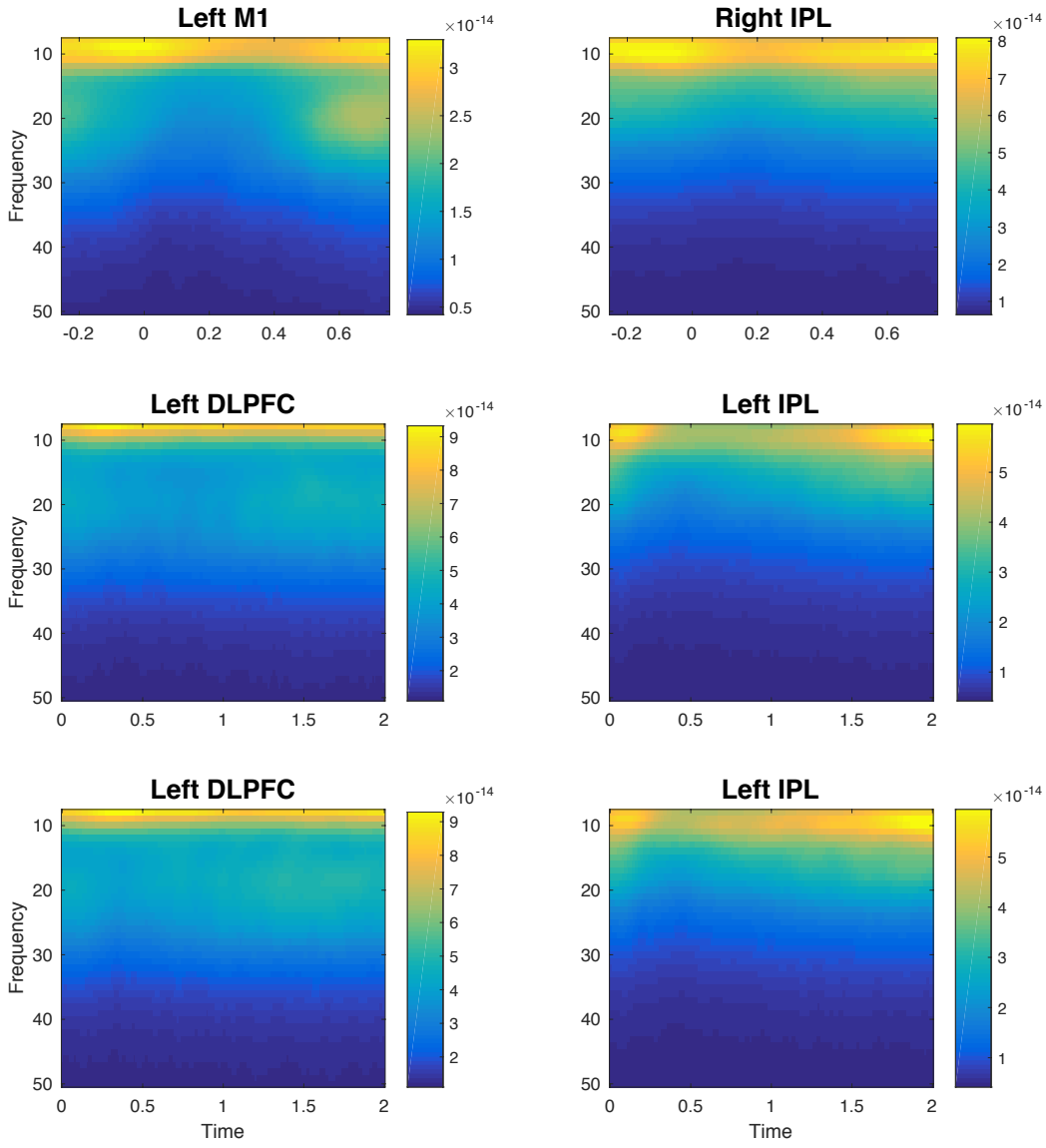

Figure 5: Plots of the power averaged over all trials and subjects. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively.

subject i's PLV as $X_i(s,t)$, and denote the grid points on which $X_i(s,t)$ is recorded as (s_j, t_k), $j \in \{1,\ldots,p\}$, $k \in \{1,\ldots,q\}$. For each grid point (s,t), the smoothed PLV value is calculated as the intercept $\hat{a}_0$ in the following weighted least squares problem:

$$(\hat{a}_0, \hat{a}_1, \hat{a}_2) = \text{argmin} \sum_{j=1}^{p} \sum_{k=1}^{q} [X_i(s_j,t_k) - a_0 - a_1(s_j - s) - a_2(t_k - t)]^2 K\left(\frac{s_j - s}{\text{bw}_\text{s}}\right) K\left(\frac{t_k - t}{\text{bw}_\text{t}}\right).$$

Here, $K(t)$ is a kernel function that is 0 for $|t| > 1$. We choose bw_s and bw_t as 15% of the lengths of their respective grids, rounded to the nearest Hz for s and nearest .01 s for t. This gives $\text{bw}_\text{s} = 6$ Hz, $\text{bw}_\text{t} = .15$ s for the motor data, and $\text{bw}_\text{t} = .30$ s for the working memory data. This choice was made by visual comparison, and an example of the PLVs for one subject using smoothing bandwidths of 10%, 15%, and 20% is shown in Figure 6.

Figure 7 shows PLV matrices smoothed with a bandwidth of 15% for the 3 datasets (motor, 2-back, and 0-back), for each of 3 example subjects. The level of activity seems to vary between subjects, as seen by the second subject's relatively high values for all tasks. Figure 8 shows the averages of these PLV matrices over all subjects. The average motor PLV displays higher synchrony near the beginning of the movement (time 0) in the alpha and beta bands, and the individual subjects' plots also show higher values near time 0. The working memory data show some peaks in PLV at the lowest frequencies. However, for all the tasks the averages have small values overall, which indicates high variability between subjects, and points to the need to study covariance structure and eigen-decomposition.

3.3 WEAK SEPARABLE ANALYSIS AND PRODUCT FPCA

3.3.1 Source-level analysis

For the source-reconstructed datasets described above, we use product FPCA to find the components along which subjects' functional connectivity varies the most during the tasks. We justify this with weak separability, which provides the theoretical framework for the products of the marginal eigenfunctions to be our choice of basis functions. Using the FVE procedure described in Section 2.3.2, we choose the number of estimated marginal

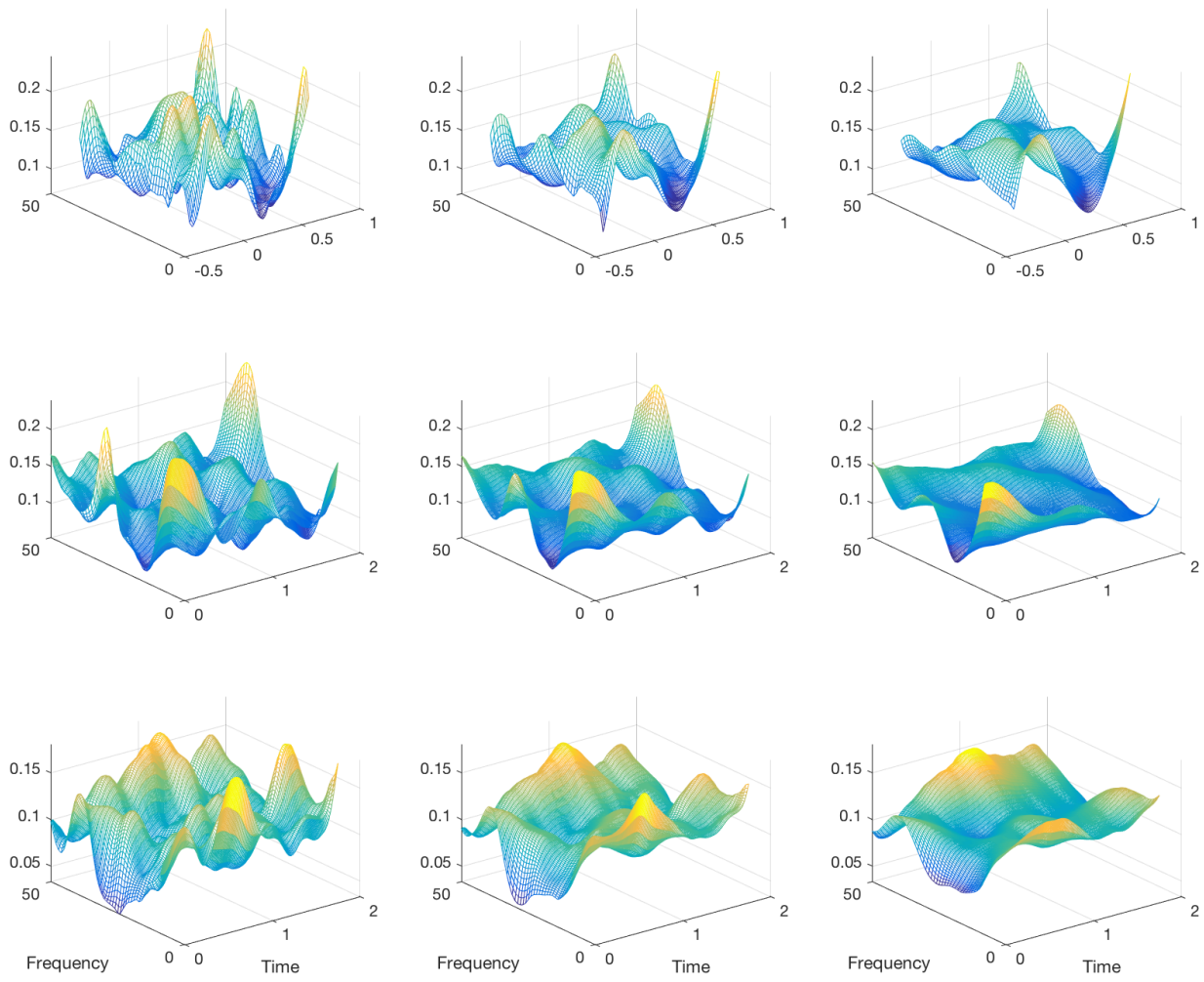

Figure 6: Plots of the source-level PLV for one subject using different levels of smoothing. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively. The columns from left to right show smoothing with bandwidths of 10%, 15%, and 20%, respectively.

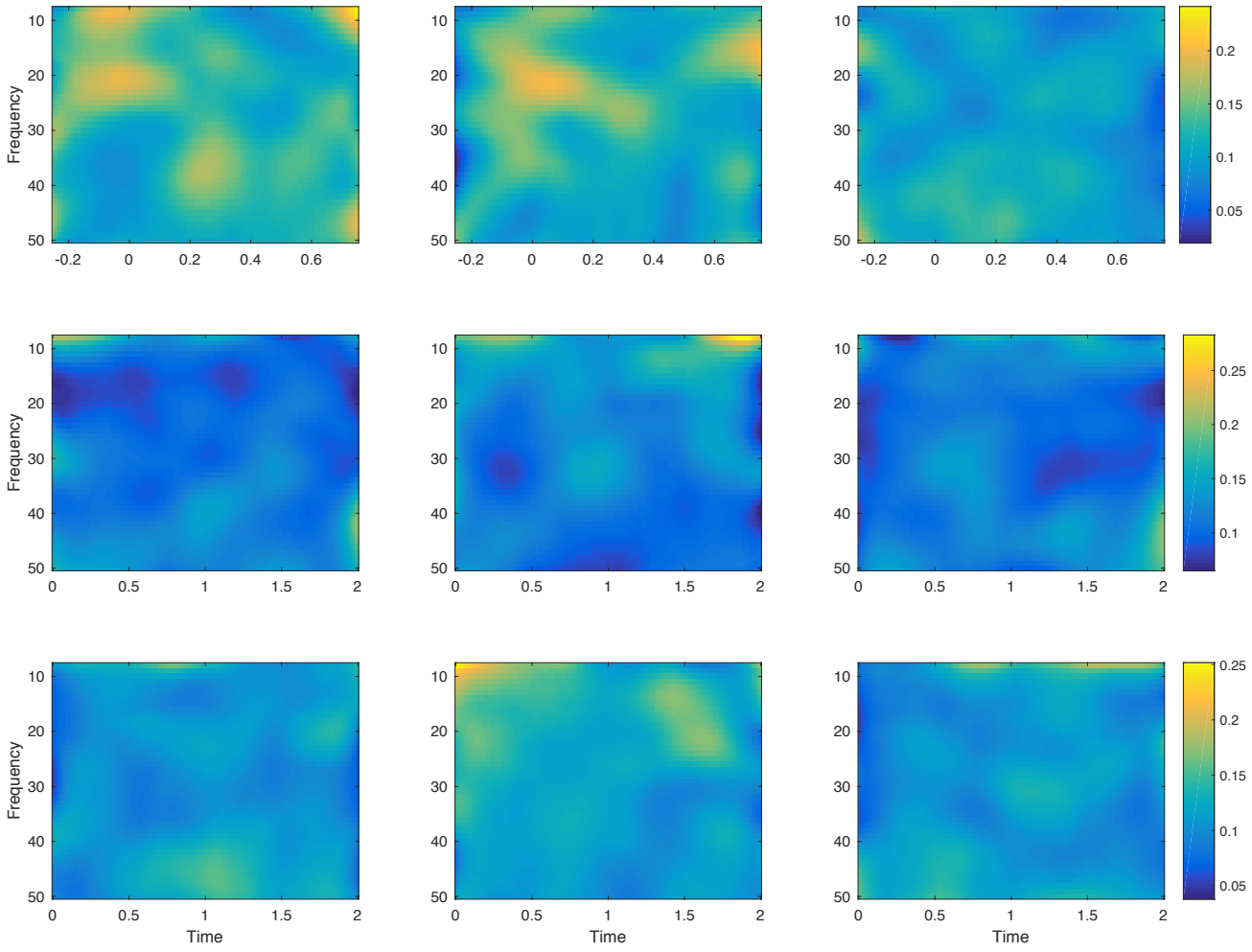

Figure 7: Plots of the source-level PLV using smoothing with a bandwidth of 15% for 3 subjects. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively. The columns from left to right correspond to the 3 subjects.

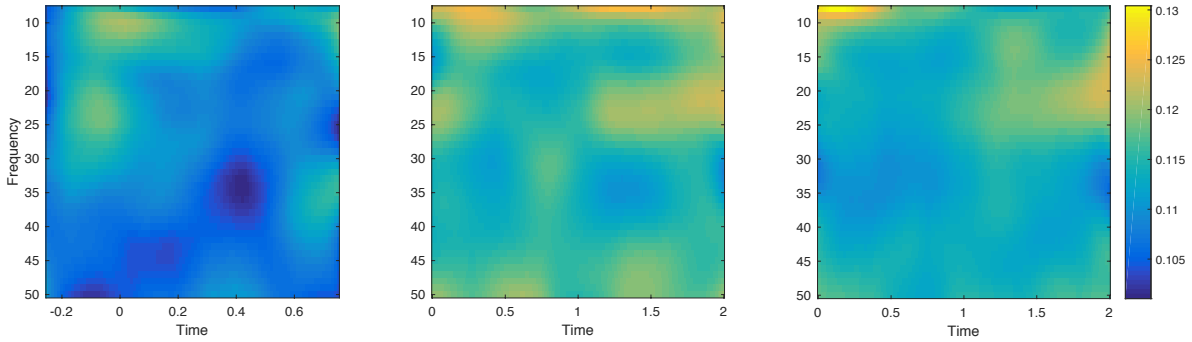

Figure 8: Plots of the average source-level PLV using smoothing with a bandwidth of 15%. The plots from left to right show the averages for the motor, 2-back, and 0-back data, respectively.

eigenfunctions to be $P_n = 7$ and $K_n = 7$ for the motor data, $P_n = 6$ and $K_n = 7$ for the 2-back data, and $P_n = 6$ and $K_n = 6$ for the 0-back data. Using these values, we apply the weak separability test using both the χ^2-type mixture approximation and empirical bootstrap (see Section 2.3). We also apply the strong separability test of Aston et al. (2017) via their R package "covsep" (Tavakoli, 2016). We use their empirical_bootstrap_test function with $B = 1000$ and no studentization, which can be viewed as the strong separable analog of our weak separable empirical bootstrap test. We also consider their asymptotic χ^2 test, though it should be noted that, unlike our χ^2-type mixture test for weak separability, this test is only valid for Gaussian data.

The P-values obtained for these tests are shown in Table 11. For all 3 datasets, the weak separability test does not reject the null hypothesis of weak separability. As was observed in the simulations and data example of Chapter 2, the empirical bootstrap procedure is conservative compared to the χ^2-type mixture approximation. The strong separability test rejects strong separability at the 5% significance level for all but the motor dataset.

Product FPCA, which decomposes the process as in Equation (2.1), is based on separable products $\psi_j(s)\phi_k(t)$ of the marginal eigenfunctions, where the $\psi_j(s)$ represent the frequency components and the $\phi_k(t)$ represent the time components. For the values of P_n and K_n we are

Table 11: P-values for the source-level datasets for the test of weak separability, as well as the test of strong separability from Aston et al. (2017). "Weak χ^2" denotes the weak separability test using the χ^2-type mixture approximation, "Weak Emp" denotes the weak separability test using the empirical bootstrap, "Strong χ^2" denotes the strong separability asymptotic χ^2 test with Gaussian assumptions, and "Strong Emp" denotes the strong separability test using the non-studentized empirical bootstrap.

Dataset	Weak χ^2	Weak Emp	Strong χ^2	Strong Emp
Motor	0.5293	0.926	1.198e-4	0.080
2-back	0.2050	0.637	8.634e-11	0.000
0-back	0.3228	0.575	1.409e-45	0.000

using, there will be many such products. We consider their relative importance by looking at their individual contributions to the FVE (see Section 2.3.2). Below, we show matrices $(\text{FVE})_{\text{j,k}}$ with entries $\hat{\eta}_{jk} / \sum_{j'} \sum_{k'} \hat{\eta}_{j'k'}$, $1 \leq j \leq P_n$, $1 \leq k \leq K_n$, where the $\hat{\eta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \hat{\chi}_{jk}^2$ are the empirical variances of the estimated marginal projection scores. We see that the 3 eigenfunction products that account for the most variance (in order) are $\hat{\psi}_1\hat{\phi}_1$, $\hat{\psi}_2\hat{\phi}_1$, and $\hat{\psi}_3\hat{\phi}_1$ for the motor and 2-back data, and $\hat{\psi}_1\hat{\phi}_1$, $\hat{\psi}_2\hat{\phi}_1$, and $\hat{\psi}_2\hat{\phi}_2$ for the 0-back data. Note that the variance explained by $\hat{\psi}_1\hat{\phi}_1$ is by far the largest, especially for the 0-back data, and the FVE values for the second product onward fall off gradually.

Motor data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.1453 & 0.0473 & 0.0311 & 0.0281 & 0.0238 & 0.0157 & 0.0092 \\ 0.0531 & 0.0325 & 0.0277 & 0.0256 & 0.0222 & 0.0162 & 0.0062 \\ 0.0516 & 0.0439 & 0.0242 & 0.0196 & 0.0153 & 0.0134 & 0.0048 \\ 0.0362 & 0.0196 & 0.0282 & 0.0185 & 0.0127 & 0.0077 & 0.0044 \\ 0.0255 & 0.0147 & 0.0203 & 0.0124 & 0.0082 & 0.0065 & 0.0039 \\ 0.0130 & 0.0082 & 0.0092 & 0.0052 & 0.0048 & 0.0033 & 0.0020 \\ 0.0070 & 0.0041 & 0.0059 & 0.0032 & 0.0035 & 0.0022 & 0.0017 \end{bmatrix}$$

2-back data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.2347 & 0.0278 & 0.0278 & 0.0236 & 0.0187 & 0.0119 & 0.0077 \\ 0.0669 & 0.0327 & 0.0324 & 0.0239 & 0.0173 & 0.0095 & 0.0068 \\ 0.0541 & 0.0387 & 0.0300 & 0.0166 & 0.0141 & 0.0096 & 0.0073 \\ 0.0262 & 0.0239 & 0.0175 & 0.0134 & 0.0093 & 0.0067 & 0.0058 \\ 0.0156 & 0.0161 & 0.0153 & 0.0097 & 0.0062 & 0.0040 & 0.0029 \\ 0.0098 & 0.0087 & 0.0072 & 0.0057 & 0.0052 & 0.0030 & 0.0021 \end{bmatrix}$$

0-back data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.4716 & 0.0257 & 0.0222 & 0.0184 & 0.0121 & 0.0063 \\ 0.0368 & 0.0331 & 0.0224 & 0.0155 & 0.0097 & 0.0064 \\ 0.0286 & 0.0237 & 0.0205 & 0.0163 & 0.0113 & 0.0070 \\ 0.0200 & 0.0156 & 0.0115 & 0.0115 & 0.0070 & 0.0037 \\ 0.0105 & 0.0083 & 0.0078 & 0.0074 & 0.0063 & 0.0028 \\ 0.0076 & 0.0064 & 0.0049 & 0.0040 & 0.0036 & 0.0025 \end{bmatrix}$$

The estimated marginal eigenvalues $\hat{\lambda}_j$ and $\hat{\gamma}_k$ are plotted in Figure 23 in Appendix B. These reflect the trends seen above, with the first eigenvalue dominating and the others falling off gradually. For the motor and 2-back data, there is also a slight drop between the second two $\hat{\lambda}_j$ and the rest, reflecting the fact that $\hat{\psi}_2\hat{\phi}_1$ and $\hat{\psi}_3\hat{\phi}_1$ are the products that explain the second and third highest amounts of variance, respectively.

The 3 product functions that explain the most variance for each the 3 datasets are plotted in Figure 12. The eigenfunctions that comprise these products are plotted for the motor, 2-back, and 0-back data in Figures 9, 10, and 11, respectively. For the motor dataset, these products capture modes of variation mainly around -.2 to .2 s, when the subject receives the cue to move to when they just start moving. This variation can be seen in $\hat{\phi}_1$, which peaks slightly after 0 s. $\hat{\psi}_1\hat{\phi}_1$ shows that, within this time range, subjects generally vary in synchrony from the alpha band to the beginning of the gamma low band (see Table 10), peaking within the beta band around 20-30 Hz. $\hat{\psi}_2\hat{\phi}_1$ shows a contrast between the beta low band and gamma low band. That is, subjects with higher χ_{21} values have lower synchrony in the beta low band and higher synchrony in the gamma low band. $\hat{\psi}_3\hat{\phi}_1$ shows a contrast between the alpha band (possibly lower), and the beta high band.

For the 2-back dataset, $\hat{\phi}_1$ shows a mode of variation at low and high time points, corresponding to when the subject is presented with the image and just before the subject has to choose if the image matches the target, respectively. $\hat{\psi}_1\hat{\phi}_1$ shows that, within these

two time ranges, subjects generally vary in their synchrony within the beta low band. $\hat{\psi}_2\hat{\phi}_1$ shows a contrast between the lowest frequencies and the gamma low band. $\hat{\psi}_3\hat{\phi}_1$ shows a mode of variation that is positive for the lowest and highest frequencies, and negative for a small range from around 20-25 Hz.

For the 0-back dataset, $\hat{\phi}_1$ is relatively flat, representing a mode of variation in subjects' overall activity across time, with slightly less importance in the range of time points from .5 to 1 s. Because $\hat{\psi}_1$ has a small peak within the beta low band, $\hat{\psi}_1\hat{\phi}_1$ shows a mode of variation within this range of frequencies, though less so from .5 to 1 s. $\hat{\psi}_2\hat{\phi}_1$ gives a mode of variation around the lowest frequencies across all time points, contrasted with frequencies in mid-beta and above. $\hat{\psi}_2\hat{\phi}_2$ contrasts the connectivity at the lowest frequencies between the first .5 seconds and the very end of the task, and also presents a less stark contrast between these time ranges for higher frequencies.

3.3.2 Sensor-level analysis

We compare our results from source analysis to a more naive method of generating connectivity between pairs of ROIs. In this method, for each ROI, we average the signals of a select few sensors found to be closest to the center of that ROI. Using Montreal Neurological Institute (MNI) coordinates, which are based on a standard template of the brain, we find the center coordinates for the DLPFC and IPL in Cohen et al. (2014), and the coordinates of the left M1 corresponding to the right hand in Landi et al. (2011).

In the HCP, sensor positions are given in BTI coordinates, in which each subject's coordinate system varies based on external landmarks of the head. For each subject, the HCP provides a homogenous transformation matrix from BTI to MNI coordinates, and we use this to find the MNI coordinates of the sensors. For each ROI within each subject, we take the sensor that is closest (in Euclidean distance based on the MNI coordinates) to the center of the ROI, and obtain its neighbors based on a template of MEG sensors using the FieldTrip function ft_prepare_neighbours. To represent each ROI, we use the sensors that are found to be closest (including the closest sensor and its neighbors) to that ROI in at least 80% of the subjects. This procedure results in 4, 5, 4, and 5 sensors representing the left M1, right IPL,

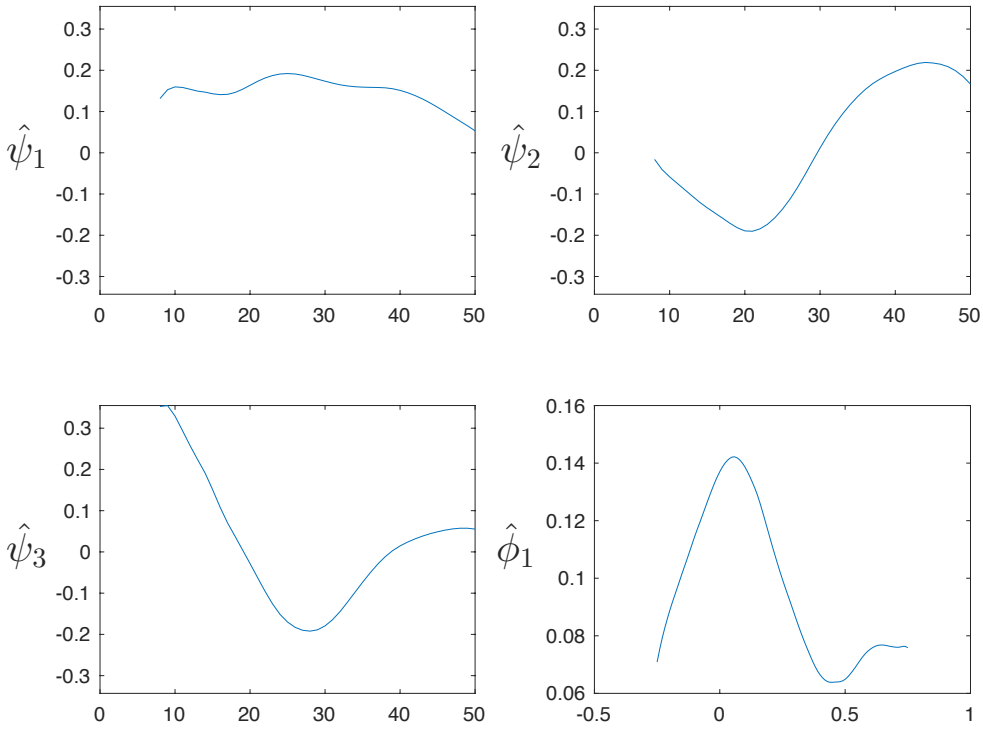

Figure 9: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the source-level motor data.

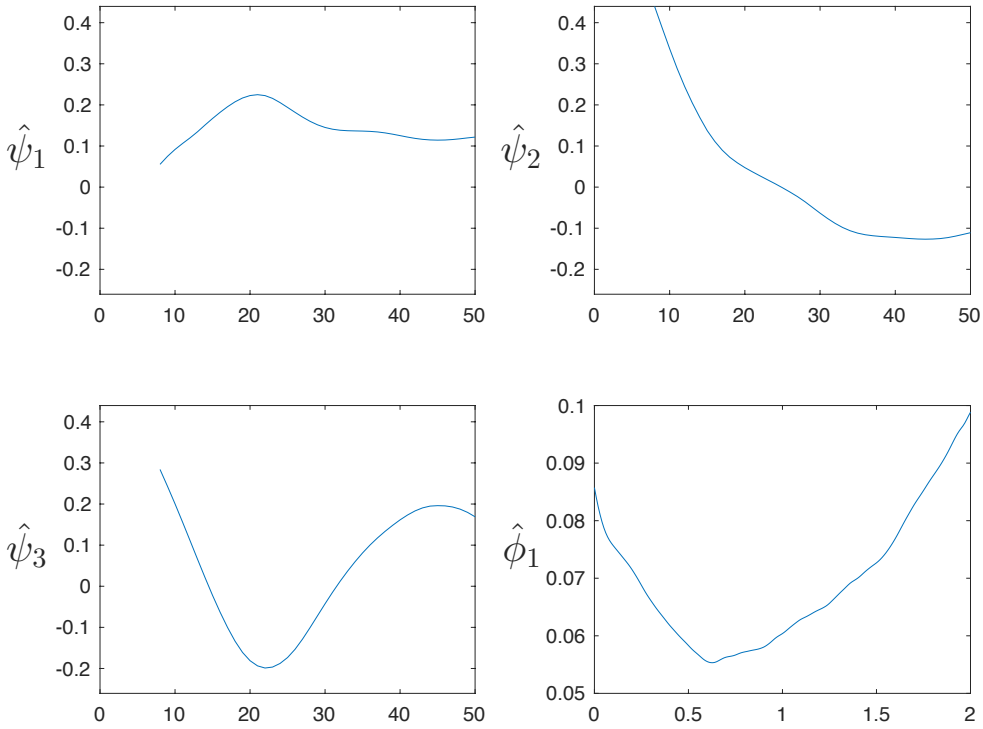

Figure 10: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the source-level 2-back data.

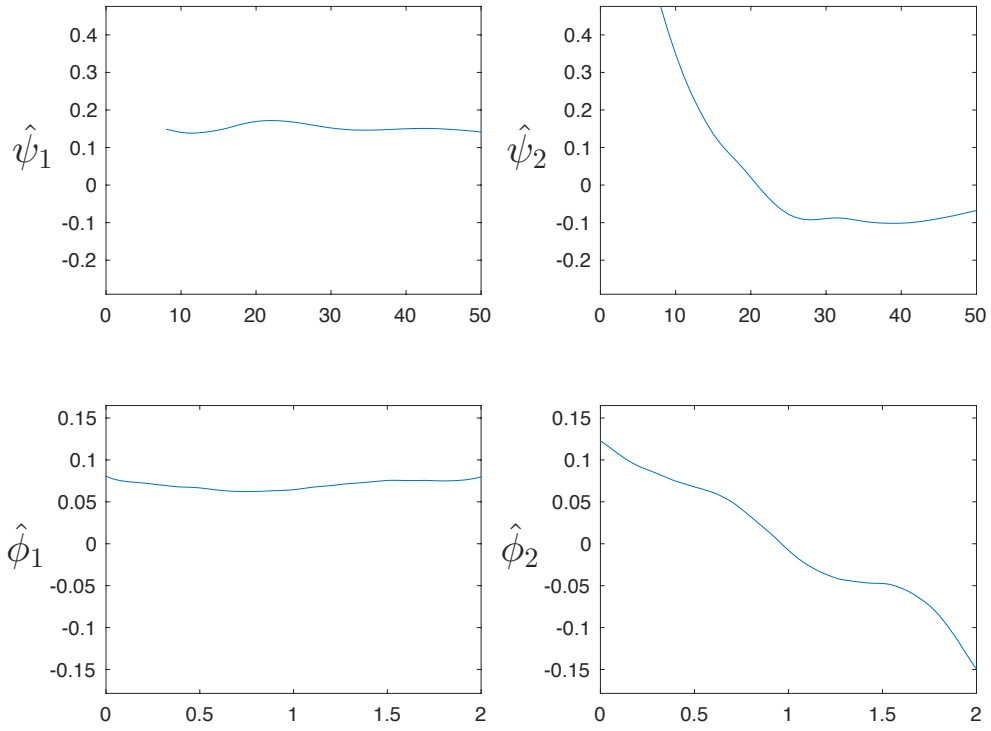

Figure 11: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the source-level 0-back data.

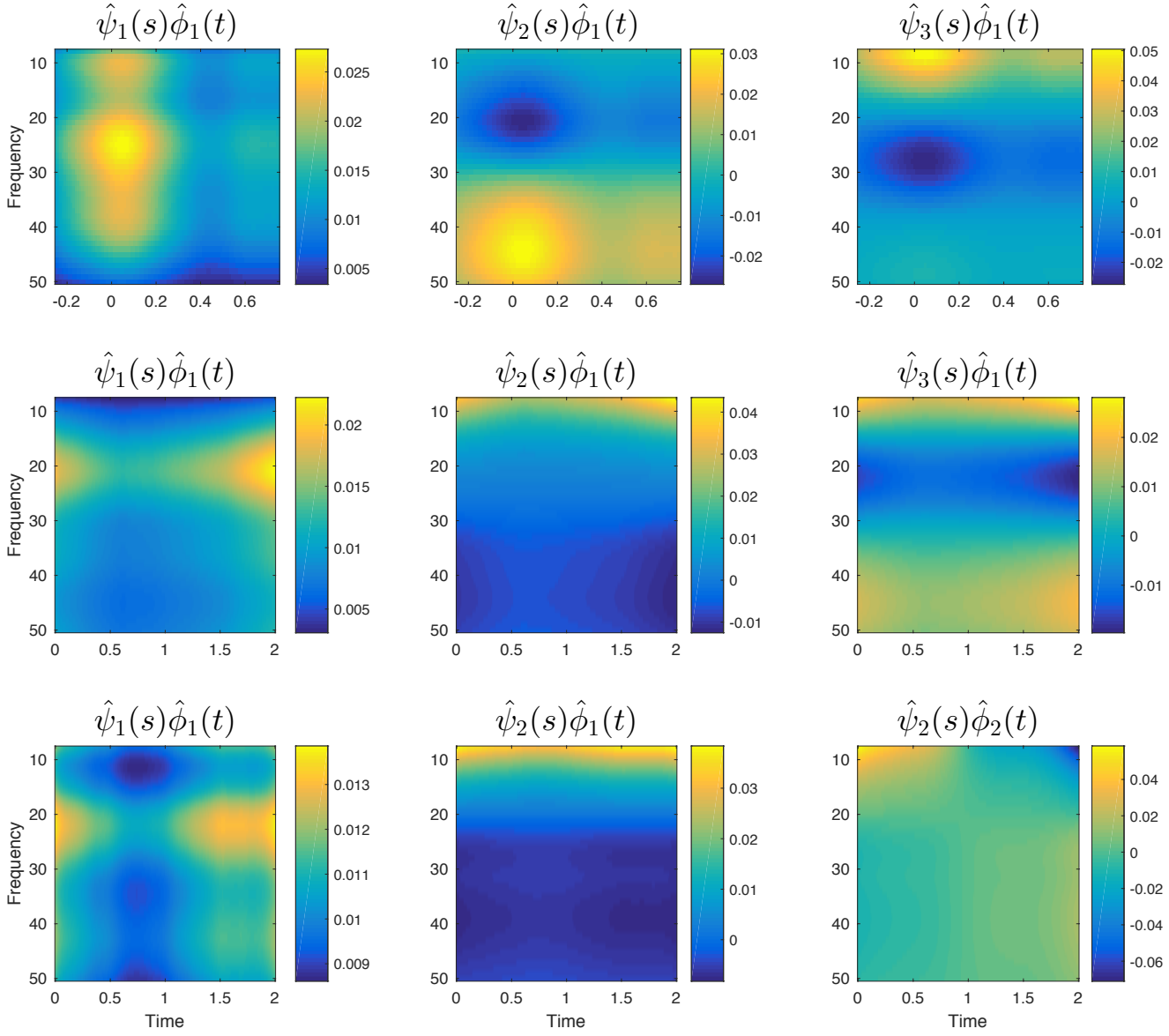

Figure 12: Plots of the products of the estimated eigenfunctions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ that explain the most variance (decreasing from left to right) for the source-level data. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively.

left IPL, and left DLPFC, respectively. For each subject, the signal we use for each ROI is the average of the signals of the group of sensors representing that ROI.

This method is less justified than source analysis because of the low spatial resolution of the sensor signals, as well as the possible contaminating effects of field spread. Additionally, unlike for source-level data, in which the source space is aligned across subjects so that the position of each dipole is comparable, the position of a given sensor here will not necessarily be at a comparable location around each subject's head. These issues could be especially worrisome when trying to define sensors to represent an ROI with thin area, such as the M1.

We proceed to calculate PLV from the sensor-level signals to derive motor, 2-back, and 0-back datasets, just as for the source-reconstructed data, and we again use smoothing with a 15% bandwidth. The means of these datasets are plotted in Figure 13. They look different than those of the source-level data, particularly for the motor data, in which the sensor-level mean has consistently higher synchrony at lower frequencies.

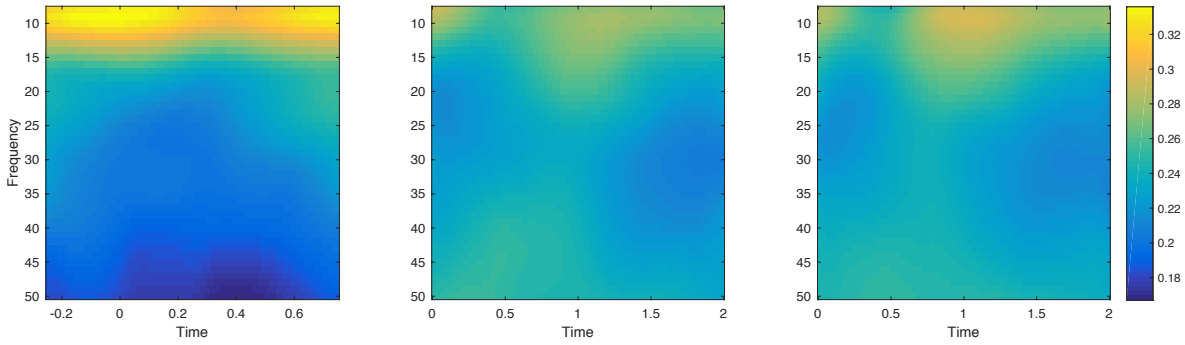

Figure 13: Plots of the average sensor-level PLV using smoothing with a bandwidth of 15%. The plots from left to right show the average for the motor, 2-back, and 0-back data, respectively.

Using the FVE procedure described in Section 2.3.2, we find $P_n = 5$ and $K_n = 5$ for all 3 sensor-level datasets. As shown in Table 12, weak separability seems to hold for all the datasets, while strong separability is rejected. Based on their contribution to the FVE, the three product functions in the product FPCA decomposition that explain the most variance are (in order) $\hat{\psi}_1\hat{\phi}_1$, $\hat{\psi}_2\hat{\phi}_1$, and $\hat{\psi}_3\hat{\phi}_1$ for all 3 datasets. Note that these are also the first 3

product functions for the motor and 2-back source-level data. The values of $(\text{FVE})_{\text{j,k}}$ are shown below, and the estimated eigenvalues are plotted in Figure 24 in Appendix B.

Table 12: P-values for the sensor-level datasets for the test of weak separability, as well as the test of strong separability from Aston et al. (2017). "Weak χ^2" denotes the weak separability test using the χ^2-type mixture approximation, "Weak Emp" denotes the weak separability test using the empirical bootstrap, "Strong χ^2" denotes the strong separability asymptotic χ^2 test with Gaussian assumptions, and "Strong Emp" denotes the strong separability test using the non-studentized empirical bootstrap.

Dataset	Weak χ^2	Weak Emp	Strong χ^2	Strong Emp
Motor	0.5044	0.688	1.385e-69	0.000
2-back	0.1126	0.231	7.816e-147	0.000
0-back	0.1831	0.358	1.458e-92	0.000

Sensor-level motor data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.4970 & 0.0364 & 0.0249 & 0.0177 & 0.0095 \\ 0.1436 & 0.0243 & 0.0223 & 0.0082 & 0.0059 \\ 0.0379 & 0.0106 & 0.0141 & 0.0075 & 0.0058 \\ 0.0219 & 0.0086 & 0.0055 & 0.0062 & 0.0039 \\ 0.0101 & 0.0052 & 0.0053 & 0.0033 & 0.0024 \end{bmatrix}$$

Sensor-level 2-back data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.4446 & 0.0374 & 0.0224 & 0.0079 & 0.0074 \\ 0.1849 & 0.0170 & 0.0180 & 0.0089 & 0.0049 \\ 0.0678 & 0.0136 & 0.0076 & 0.0061 & 0.0045 \\ 0.0399 & 0.0083 & 0.0067 & 0.0068 & 0.0035 \\ 0.0133 & 0.0051 & 0.0032 & 0.0024 & 0.0025 \end{bmatrix}$$

Sensor-level 0-back data:

$$(\text{FVE})_{\text{j,k}} = \begin{bmatrix} 0.4537 & 0.0437 & 0.0272 & 0.0130 & 0.0087 \\ 0.1656 & 0.0182 & 0.0155 & 0.0089 & 0.0053 \\ 0.0593 & 0.0123 & 0.0093 & 0.0062 & 0.0037 \\ 0.0371 & 0.0077 & 0.0075 & 0.0048 & 0.0031 \\ 0.0150 & 0.0053 & 0.0041 & 0.0033 & 0.0026 \end{bmatrix}$$

The 3 product functions that explain the most variance are plotted for the 3 sensor-level datasets in Figure 14, with the individual eigenfunctions plotted in Figures 25, 26, and 27 in Appendix B. These show fairly different patterns than the components of the source-level data. While the $\hat{\psi}_1\hat{\phi}_1$ show a mode of variation at lower and higher time points, the other products seem homogeneous across time. Unlike in the source-level data, the product functions of the 2-back and 0-back datasets look quite similar, and the values of $\hat{\psi}_2\hat{\phi}_1$ and $\hat{\psi}_3\hat{\phi}_1$ look quite similar across all the datasets. Generally, $\hat{\psi}_2\hat{\phi}_1$ represents a contrast between the alpha band and higher frequencies, and $\hat{\psi}_3\hat{\phi}_1$ represents a contrast between the beta low band and the other frequency ranges.

Acknowledgement: Data for this chapter were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

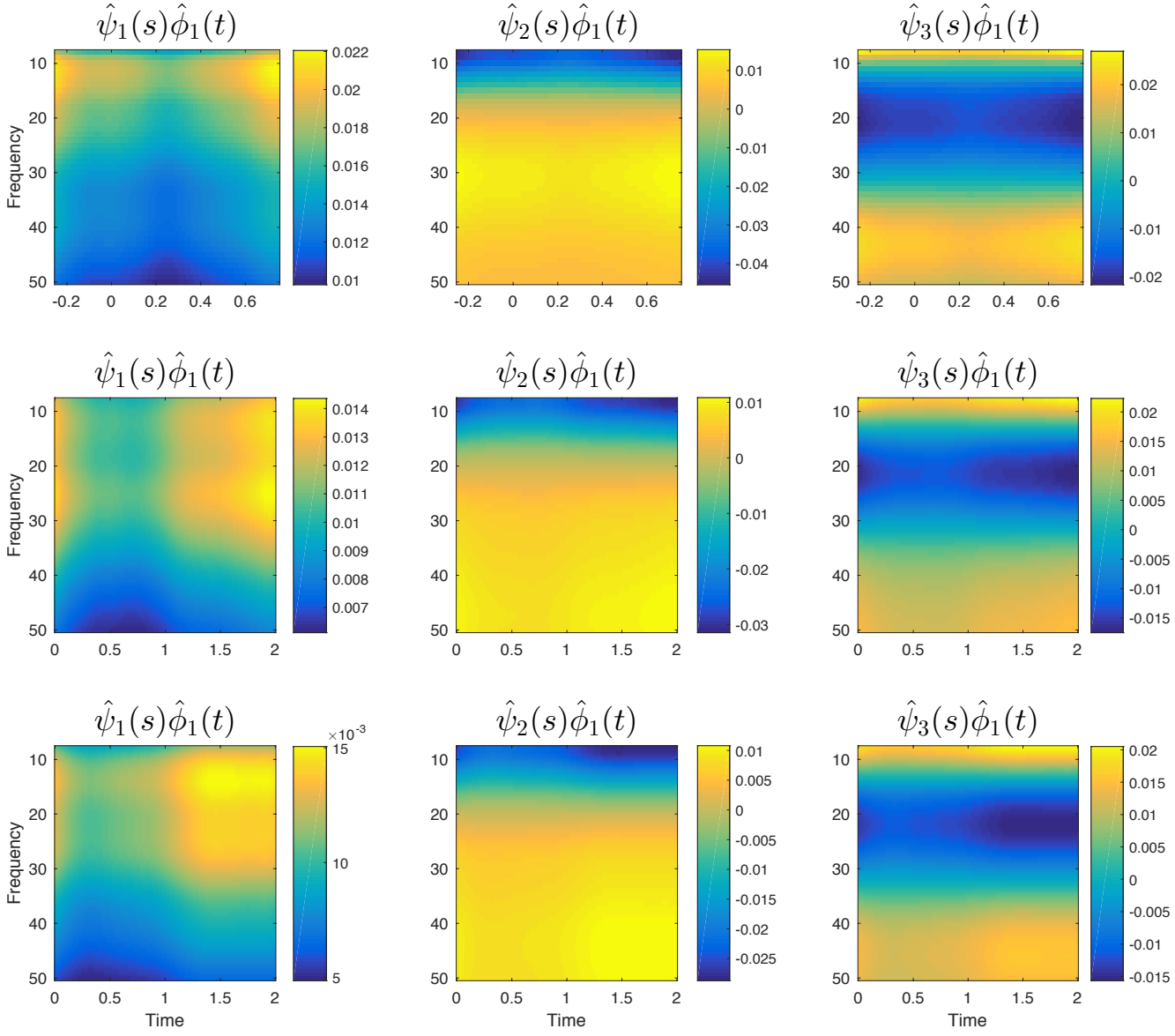

Figure 14: Plots of the products of the estimated eigenfunctions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ that explain the most variance (decreasing from left to right) for the sensor-level data. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively.

4.0 L-SEPARABILITY

4.1 MOTIVATION FOR L-SEPARABILITY AND RELATED APPROXIMATIONS OF THE COVARIANCE

In Section 2.2 we developed the concept of *weak separability* for two-way functional data. In general, when the data are weakly separable, the covariance structure C can be written as a weighted sum of infinitely many strongly separable components, as in Equation (2.4). In this chapter, as a means to more easily interpret the weakly separable covariance structure C, we study ways to represent C, either exactly or approximately, as the sum of only a few strongly separable components. Equation (2.5) shows a representation of C with L terms, where L is the nonnegative rank of V, the array of variances of the marginal projection scores. When L is finite, we say we have *L-separability*.

This approach still presents interpretational pitfalls, as L may be large, and the L-separable decomposition is not in general unique. In this chapter, we show that the problem of approximating the covariance structure with L or fewer terms is tied to the problem of nonnegative matrix factorization (NMF; Lee & Seung (2001)) on V. NMF is a difficult problem in that it is NP-hard (Arora et al., 2012; Dong et al., 2014), and has issues with identifiability (Donoho & Stodden, 2003; Gillis, 2012). We simplify the problem by imposing restrictions based on the *orthogonal NMF* of Ding et al. (2006). These allow for a global minimum to be computed, and we give an algorithm to do so that is feasible when the size of V is not too large. We also show that the orthogonal NMF restrictions allow for a more easily interpretable decomposition of C.

4.2 PROPERTIES OF L-SEPARABILITY

Using the notation of Section 2.2, assume we have a process $X(s,t)$ that is weakly separable, with covariance function $C(s,t;u,v)$. In this case we can represent $X(s,t)$ using product FPCA as in Equation (2.1), where the marginal projection scores χ_{jk} are uncorrelated. In the case that the nonnegative rank L of $V = (\text{var}(\chi_{jk}) : j \geq 1, k \geq 1)$ is finite, Equation (2.5) gives a representation of C as a sum of L strongly separable terms, where we set the spatial and temporal covariance structures comprising each term to have the eigenfunctions $\{\psi_j, j \geq 1\}$ and $\{\phi_k, k \geq 1\}$, which are the eigenfunctions of the marginal covariances $C_{\mathcal{S}}$ and $C_{\mathcal{T}}$, respectively. This choice is justified by the following lemma, where we denote the matrix Frobenius norm as $\|\cdot\|_F$, and the norm of a covariance structure as $\|C\|^2 = \int_{\mathcal{T}}\int_{\mathcal{S}}\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;u,v)^2 dsdtdudv$. Also, for an array F, we take $F \geq 0$ to mean F is entry-wise nonnegative, in which case we call F a *nonnegative array*.

Lemma 7. *Consider the problem of approximating $C(s,t;u,v)$ with a sum of d strongly separable covariance structures, i.e., $\min_{C^1,\ldots,C^d} \|C - \sum_{l=1}^{d} C^l\|$ such that $C^l(s,t;u,v) = C_1^l(s,u)C_2^l(t,v)$, where $d \leq L$ and the structures C_1^l and C_2^l are nonnegative definite. A solution to this problem sets $C_1^l(s,u) = \sum_j F_{j,l}\psi_j(s)\psi_j(u)$ and $C_2^l(t,v) = \sum_k G_{k,l}\phi_k(t)\phi_k(v)$, $l = 1,\ldots,d$, where $F_{j,l}$ is the (j,l)th entry of a nonnegative array F, $G_{k,l}$ is the (k,l)th entry of a nonnegative array G, and F and G are solutions to $\min_{F\geq 0, G\geq 0} \|V - FG^T\|_F$.*

This lemma implies that $C(s,t;u,v)$ cannot be decomposed exactly into a sum of fewer than L strongly separable terms, but this lemma also gives a method to approximate $C(s,t;u,v)$ using fewer than L strongly separable terms. For the rest of this chapter, in order to make finding solutions to $\min_{F\geq 0, G\geq 0} \|V - FG^T\|_F$ practical, we assume V is of finite dimension $P \times K$, where $P, K < \infty$. This can result from truncating the product FPCA expansion in Equation (2.1) to the first P and K terms, i.e., using the truncated process $X(s,t) = \mu(s,t) + \sum_{j=1}^{P}\sum_{k=1}^{K} \chi_{jk}\psi_j(s)\phi_k(t)$. When we are working with data and have to estimate the marginal eigenfunctions, marginal projection scores, and V, we are always limited to using a finite number of of terms, which we have denoted P_n and K_n (see Section 2.3.2).

In general, for a $P \times K$ nonnegative matrix V, the problem of finding a $P \times d$ nonnegative matrix F and a $K \times d$ nonnegative matrix G that solve

$$\min_{F \geq 0, G \geq 0} \|V - FG^T\|_F$$

is known as *nonnegative matrix factorization* (NMF; Lee & Seung (2001)). Equivalently, NMF seeks a sum of matrices $V_1 + \cdots + V_d$, where $\text{rank}(V_l) = 1$ and $V_l \geq 0$ for all $l \in \{1, \ldots, d\}$, that minimizes $\|V - (V_1 + \cdots + V_d)\|_F$. We take $1 \leq d \leq L$. The special case when $d = L$ is referred to as *exact NMF*, in which we can find a solution such that $V = FG^T$. Also, when $d = P$, setting $F = I_P$ and $G = V^T$ gives an exact NMF solution; and when $d = K$, setting $F = V$ and $G = I_K$ gives an exact NMF solution. From these observations, we note that $\text{rank}(V) \leq \text{rank}_+(V) \leq \min(P, K)$.

NMF, finding the nonnegative rank L, and exact NMF (even when the nonnegative rank is known), are all NP-hard problems (Arora et al., 2012; Dong et al., 2014). Algorithms that have been proposed for exact and non-exact NMF (Lee & Seung, 2001; Dong et al., 2014; Vandaele et al., 2015) are iterative and not guaranteed to converge to a global minimum. Even if we find solutions F and G, it is clear that there are infinitely many other choices F' and G' such that $F'G'^T = FG^T$, or more broadly, $\|V - F'G'^T\|_F = \|V - FG^T\|_F$. Donoho & Stodden (2003) and Gillis (2012) have attempted to make the NMF problem more well-posed by adding restrictions to the structures of F, G, and V, but these restrictions are unintuitive in our setting, and the sense in which their solutions are unique is complicated. To simplify the problem, in the following section we consider a constraint known as *orthogonal NMF* (Ding et al., 2006). We show that solutions to this problem have a natural interpretation in our setting, we introduce a new algorithm that finds a global minimum, and we prove uniqueness in the exact case.

4.3 ORTHOGONAL NMF

Orthogonal NMF, as defined in Ding et al. (2006), seeks a $P \times d$ nonnegative matrix F and a $K \times d$ nonnegative matrix G that solve

$$\min_{F \geq 0, G \geq 0} \left\| V - FG^T \right\|_F \ s.t.\ F^T F = I_d.$$

Equivalently, orthogonal NMF seeks a sum of matrices $V_1 + \cdots + V_d$; where $\text{rank}(V_l) = 1$ for all $l \in \{1, \ldots, d\}$, $V_l \geq 0$ for all $l \in \{1, \ldots, d\}$, and $V_l^T V_m = 0$ for all $l \neq m$; which minimizes $\|V - (V_1 + \cdots + V_d)\|_F$. We will consider $d \leq \min(P, K)$. When considering a d such that there is a solution with $V = FG^T$, we denote the problem as *exact orthogonal NMF*.

Condition A: We take the entries of V to be strictly positive.

This condition is natural since, in our context, the entries of V are variances of the marginal projection scores.

The following lemma shows that exact orthogonal NMF has a unique solution:

Lemma 8. *Under Condition A, when using the smallest d such that there exists a $P \times d$ nonnegative matrix F with $F^T F = I_d$ and a $K \times d$ nonnegative matrix G such that $V = FG^T$, F and G are unique up to a column-wise permutation.*

When we say the solution is unique up to a column-wise permutation, we mean that given F and G that solve the exact orthogonal NMF problem, for any $d \times d$ permutation matrix P, FP and GP will also give a solution. It is clear that $FG^T = (FP)(GP)^T$, and doing a column-wise permutation will not change the values of the corresponding rank-one matrices $V_1, \ldots, V_d$.

Remark: Proposition 1 of Ding et al. (2006) proves that an exact orthogonal NMF solution FG^T is unique up to a column-wise permutation among all solutions of the form $F'G'^T$ where $F' = FA$, $G' = GB$, and $AB^T = I_d$ (the authors show $A = B = P$ for P some permutation matrix). However, $F' = FA$ implies that F' has the same support (the same positions of nonzero entries) as F, while our Lemma 8 puts no additional assumptions on F' or G'. Additionally, Proposition 1 of Ding et al. (2006) implicitly assumes Condition A, since their proof takes each row of F to have exactly one nonzero element.

When $P \leq K$, there will always exist an exact orthogonal NMF solution with $d \leq P$, since we can set $F = I_P$ and $G = V^T$. We denote this as the "trivial decomposition." When $P \leq K$ and V has full rank, i.e., $\text{rank}(V) = P$, the trivial decomposition will be the unique exact orthogonal NMF solution. When $K < P$, we can get an analogous trivial decomposition by considering orthogonal NMF on V^T (see the end of this section).

An advantage of orthogonal NMF is that it allows for the corresponding decomposition of $C(s,t;u,v)$ to be interpreted as a sum of covariance structures of d uncorrelated processes $X_l(s,t)$, where $X(s,t) - \mu(s,t) = \sum_{l=1}^{d} X_l(s,t)$. Write an orthogonal NMF solution as $FG^T = \sum_{l=1}^{d} V_l$, where V_l is the nonnegative rank-one matrix obtained by taking the outer product of the lth columns of F and G. Note that F can have no more than one nonzero entry in each row, so we can define B_l to be the set of indices corresponding to the rows of F that have their nonzero element in column l. Under Condition A, the B_l will be nonempty, disjoint, and will have union $\{1, 2, \ldots, P\}$ (see the proof of Lemma 8). Set $X_l(s,t) = \sum_{j \in B_l} \sum_{k=1}^{K} \chi_{jk} \psi_j(s) \phi_k(t)$. From the product FPCA expansion $X(s,t) = \mu(s,t) + \sum_{j=1}^{P} \sum_{k=1}^{K} \chi_{jk} \psi_j(s) \phi_k(t)$, we see that $X(s,t) - \mu(s,t) = \sum_{l=1}^{d} X_l(s,t)$. Since under weak separability the χ_{jk} are uncorrelated, the $X_l(s,t)$ are uncorrelated. Let $C^l(s,t;u,v)$ denote the covariance structure of $X_l(s,t)$. Then

$$C^l(s,t;u,v) = \sum_{j \in B_l} \sum_{k=1}^{K} \eta_{jk} \psi_j(s) \psi_j(u) \phi_k(t) \phi_k(v),$$

and hence $C(s,t;u,v) = \sum_{l=1}^{d} C^l(s,t;u,v)$.

Setting $C^l_{\mathcal{S}}(s,u) = \sum_{j=1}^{P} F_{jl} \psi_j(s) \psi_j(u)$ and $C^l_{\mathcal{T}}(t,v) = \sum_{k=1}^{K} G_{kl} \phi_k(t) \phi_k(v)$ as in Lemma 7, we see $C^l(s,t;u,v) \approx C^l_{\mathcal{S}}(s,u) C^l_{\mathcal{T}}(t,v)$ (which holds exactly in the exact orthogonal NMF case), and hence the X_l are strongly separable in the exact orthogonal NMF case, and "approximately strongly separable" otherwise. Also, as in Section 2.2, we can write $C(s,t;u,v) \approx \sum_{l=1}^{d} a^l C^l_{\mathcal{S}}(s,u) C^l_{\mathcal{T}}(t,v)$, where the constant a^l allows for scaling of F and G to make $C^l_{\mathcal{S}}(s,u)$ and $C^l_{\mathcal{T}}(t,v)$ comparable; for example, we can scale F and G so that the entries in each of their columns add to 1, so that $Tr(C^l_{\mathcal{S}}) = Tr(C^l_{\mathcal{T}}) = 1$.

We also consider orthogonal NMF on V^T, which is equivalent to the variant of orthogonal NMF on V where we impose $G^T G = I_d$ instead of $F^T F = I_d$. This leads to an analogous

decomposition of $X(s,t)$ into terms $X'_l(s,t) = \sum_{j=1}^{P} \sum_{k \in B'_l} \chi_{jk} \psi_j(s) \phi_k(t)$, where B'_l is the set of indices corresponding to the rows of G that have their non-zero element in column l. This decomposition can be interpreted in the same manner as above, being a sum of uncorrelated and approximately strongly separable processes. For non-exact orthogonal NMF, there seems to be no way to know beforehand whether orthogonal NMF on V or orthogonal NMF on V^T will give a smaller error. However, we may be interested in one more than the other for interpretational reasons when working with a given dataset. Orthogonal NMF on V gives X_l (and hence C^l) that are composed of only a few distinct s components, while orthogonal NMF on V^T gives X'_l that are composed of only a few distinct t components, and one of these may be more desirable given the meanings of s and t in the context of the problem.

4.4 ALGORITHM FOR ORTHOGONAL NMF

We give an algorithm to find a global minimum of the orthogonal NMF problem for a given d. Our algorithm is not iterative, and only involves calculating singular value decompositions. We first consider orthogonal NMF for a fixed support of F, i.e., the problem where the positions of the nonzero entries of F are fixed. We show that this problem has a unique minimizing solution. Once we solve the problem for fixed support, we get a global minimum for orthogonal NMF by taking the minimizing solution over all possible supports of F. Note that we cannot guarantee that there is only one globally minimizing solution (except in the exact orthogonal NMF case, as described in Lemma 8), though we will have finitely many candidates that we can compare, subject to roundoff error.

Each support of F corresponds to a choice of sets B_l, $l = 1, \ldots, d$, where B_l is defined as in Section 4.3. Let $(V)_{B_l}$ be the $P \times K$ matrix with rows corresponding to B_l equal to those of V, and all other entries 0. The following lemma gives the unique solution to orthogonal NMF for a fixed support of F:

Lemma 9. *Under Condition A, the solution to orthogonal NMF with fixed support is unique and, for $l = 1, \ldots, d$, sets column l of F to be u_l and column l of G to be $\sigma_l v_l$, where σ_l is the largest singular value of $(V)_{B_l}$, and u_l and v_l are its corresponding left and right singular*

vectors, respectively.

To find the support of F that gives the orthogonal NMF solution with the smallest error, we only need to consider supports where the following are true:

(1) F has no columns of all zeros.

(2) F has no rows of all zeros.

(3) The columns of F are ordered according to their lowest-numbered row with a nonzero entry. That is, the first entry of the first column of F is nonzero; the second column of F has its jth entry nonzero, where j is the smallest integer such that the jth entry of the first column is 0; the third column of F has its kth entry nonzero, where k is the smallest integer such that the kth entries of the first two columns are both 0; and so on.

(1) follows from the fact that a solution where F has a column of zeros corresponds to a solution with a smaller value of d. (2) is implied by Condition A (see the proof of Lemma 8). (3) follows from the fact that column-wise permutations will give equivalent solutions.

To enumerate all the possible supports of F, we can do the following: First, enumerate all the possible numbers of nonzero entries in each column. For instance, if $P = 6$ and $d = 3$, we could have $(1, 1, 4)$ (1 nonzero entry in the first column, 1 in the second, and 4 in the third), $(1, 2, 3)$, $(1, 3, 2)$, etc. There will be $\binom{P-1}{d-1}$ of these choices. For a specific choice $(i_1, i_2, \ldots, i_d)$ of nonzero entries, the number of possible supports (following the 3 properties above) will be

$$\binom{P-1}{i_1-1}\binom{P-i_1-1}{i_2-1}\binom{P-i_1-i_2-1}{i_3-1}\cdots\binom{P-i_1-i_2-\cdots-i_{d-1}-1}{i_d-1}.$$

Table 13 shows the total number of supports that need to be searched for selected values of P and d. As functional data is usually well approximated by only a small number of components, and only small values of d are of interest, we expect this algorithm to be computationally feasible for the purposes of two-way functional data.

Note that Lemma 8 gives uniqueness of orthogonal NMF in the exact case, while Lemma 9 gives uniqueness of general (not necessarily exact) orthogonal NMF only for a fixed support of F. To see that we cannot remove the fixed support requirement in Lemma 9, consider the following examples where two different solutions (with different supports of F) can be the globally minimizing solutions and have the same error: Denote $V_{j\cdot}$ as the jth row of

Table 13: The number of supports of F that need to be considered for selected values of P and d.

	d=2	d=4	d=6	d=8	d=10	d=12	d=14
P=2	1						
P=4	7	1					
P=6	31	65	1				
P=8	127	1701	266	1			
P=10	511	34105	22827	750	1		
P=12	2047	611501	1323652	159027	1705	1	
P=14	8191	10391745	63436373	20912320	752752	3367	1

the $P \times K$ nonnegative matrix V. Consider the $d = 2$ case for a V that is 3×3 with $V_{3\cdot} = V_{1\cdot} + V_{2\cdot}$. By symmetry it is clear that an F with first column $[1\ 0\ 0]^T$ will give a minimizing value with the same error as an F with second column $[0\ 1\ 0]^T$. By inspection of output from our orthogonal NMF algorithm, we can find cases where these two values of F give the minimizing solution, for example when $V_{1\cdot} = [2\ 1\ 1]$ and $V_{2\cdot} = [1\ 1\ 2]$. We can find other counterexamples by setting $V_{3\cdot} = f(V_{1\cdot}, V_{2\cdot})$ for f some symmetric function.

We note that Ding et al. (2006) propose an algorithm for orthogonal NMF that is iterative, taking initial values F_0 and G_0 for F and G, respectively, and on each iteration using element-wise multiplicative update rules to obtain new values of F and G. They prove that their update rules converge to a local minimum, but they do not guarantee a global minimum. It can be shown that each update gives an F with the same support as F_0. Presumably, Ding et al. (2006) are interested in much larger P and K than we are, making searching over all supports impractical, and they expect the user to try many different initial values with different supports. However, our Lemma 9 gives a unique, explicit minimizing solution for orthogonal NMF with a fixed support, so we have improved on their method. Furthermore, for moderate P and K, we have described how to search over all supports to get an overall minimizing solution.

4.5 CHOOSING THE NUMBER OF TERMS IN THE ORTHOGONAL NMF DECOMPOSITION

In this section we discuss possible ways to choose the number of terms d to use in the orthogonal NMF decomposition. At the end of this section, we note the relationship between choosing d and choosing the number of clusters in clustering problems. This is an open problem, and likewise we do not present a definitive solution. One tactic is to, for each value of d, do a full search over all $B_1, \ldots, B_d$ to get the globally minimizing solution, and then compare the solutions for different d using some measurement of error, for instance the relative error $\|V - FG^T\|_F / \|V\|_F$. The relative error is guaranteed to decrease as d increases. The issue here is that there is no clear bound to put on the difference between consecutive error terms to decide whether the error is low enough. One approach could be to plot the error as a function of d and look for an "elbow."

A possibly more rigorous method for our setting would be to decide whether our d-term approximation is sufficient by testing whether each process $X_l(s,t)$, $l = 1, \ldots, d$, as defined in Section 4.3, is strongly separable. To test the null hypothesis that $X_l(s,t)$ is strongly separable, we cannot directly apply the test of strong separability from Aston et al. (2017), since we do not have the true values of $X_l(s,t)$ for each subject. We could apply their test to estimated versions of the $X_l(s,t)$, calculated as $\sum_{j \in B_l} \sum_{k=1}^{K} \hat{\chi}_{i,jk} \hat{\psi}_j(s) \hat{\phi}_k(t)$ for each subject i. However, a simpler testing procedure can be derived using the test statistic

$$T'_n(j,k) = \sqrt{n}(\hat{\eta}_{jk} - \sum_{j' \in B_l} \sum_{k'=1}^{K} \hat{\eta}_{jk'} \hat{\eta}_{j'k} / \sum_{j' \in B_l} \sum_{k'=1}^{K} \hat{\eta}_{j'k'}),$$

where $j \in B_l$ and $\hat{\eta}_{jk} = (1/n) \sum_{i=1}^{n} \hat{\chi}^2_{i,jk}$. This is related to the test statistic from Aston et al. (2017), which in our notation, when applied to the process $X(s,t)$, is $\sqrt{n}(\hat{\eta}_{jk} - \hat{\lambda}_j \hat{\gamma}_k / Tr\hat{(}C))$, where $\hat{\lambda}_j$ and $\hat{\gamma}_k$ are the eigenvalues of the estimated marginal covariances. We can derive the joint asymptotic null distribution of the $T'_n(j,k)$ over all pairs (j,k) by showing $\sqrt{n}(\hat{\eta}_{jk} - \frac{1}{n} \sum_{i=1}^{n} \chi^2_{i,jk}) = o_p(1)$ and then using the multivariate delta method to find the joint distribution of the $T'_n(j,k)$ under the null hypothesis that $X_l(s,t)$ is strongly separable, in which case $\eta_{jk} = \sum_{j' \in B_l} \sum_{k'=1}^{K} \eta_{jk'} \eta_{j'k} / \sum_{j' \in B_l} \sum_{k'=1}^{K} \eta_{j'k'}$ by Lemma 3. The asymptotic covariance of the $T'_n(j,k)$ will be degenerate, which can be seen by noting $\sum_{j \in B_l} T'_n(j,k) = 0$

and $\sum_{k=1}^{K} T'_n(j,k) = 0$. Thus, as for our test of weak separability, we compute the P-value of the test from a χ^2-type mixture approximation.

For each possible d, starting with $d = 1$, we could find the choice of $B_1, \ldots, B_d$ that minimizes the objective function of orthogonal NMF, and then decide to use this solution if the P-values for the corresponding X_l are above some cutoff.

We could also consider the following two greedy algorithms:

Forward: Starting with $d = 1$, 1) Apply the test of strong separability to all the X_l. 2) If, for all X_l, the test fails to reject, stop and use the current clustering. If not, consider all possible ways to split one of the sets B_l into two. 3) Choose the split that gives the smallest value of the objective function. d has increased by 1. Repeat.

Backward: Starting with $d = P$, 1) Apply the test of strong separability to all the X_l. 2) If the test rejects for some X_l, stop and use the previous clustering. If not, consider all possible combinations of two B_l. 3) Choose the combination that gives the smallest value of the objective function. d has decreased by 1. Repeat.

We finally note that choosing d in orthogonal NMF can be thought of as a clustering problem in which we seek to cluster the rows of V into d clusters. Denote $V_{j\cdot}$ as the jth row of V. It can be shown that under Condition A, orthogonal NMF is equivalent to minimizing

$$\sum_{l=1}^{d} \sum_{j \in B_l} \|V_{j\cdot}\|^2 \left(1 - \left(\frac{1}{\|V_{j\cdot}\|} V_{j\cdot} v_l\right)^2\right)$$

over d nonnegative unit vectors v_l of length K and d sets $B_1, \ldots, B_d$ that partition $\{1, \ldots, P\}$. As in Section 4.4, let $(V)_{B_l}$ be V with the rows not corresponding to B_l set to 0. For $B_1, \ldots, B_d$ fixed, it can be shown that the minimizing value for the above problem sets each v_l to be the first right singular vector of $(V)_{B_l}$, and using Lemma 9 one can show that this problem is equivalent to orthogonal NMF. From this result, we note the resemblance of orthogonal NMF to the problem of spherical clustering (Dhillon & Modha, 2001; Buchta et al., 2012) on the rows of V, which minimizes

$$\sum_{l=1}^{d} \sum_{j \in B_l} (1 - \cos(V_{j\cdot}, v_l)) = \sum_{l=1}^{d} \sum_{j \in B_l} (1 - \frac{1}{\|V_{j\cdot}\|} V_{j\cdot} v_l)$$

over d nonnegative unit vectors v_l of length K and d sets $B_1, \ldots, B_d$ that partition $\{1, \ldots, P\}$. We additionally note that orthogonal NMF on V^T can be thought of as an analogous clustering problem on the columns of V.

5.0 CASE STUDY: PSYCHIATRIC DATA

5.1 EXPERIMENTAL DESIGN AND STRUCTURE OF THE DATA

We apply and extend the ideas of the previous chapters to MEG data collected by the University of Pittsburgh's Clinical Neurophysiology Research Laboratory, of which the members with whom we worked most closely were Brian Coffman (Post-Doctoral Researcher) and Dean Salisbury (Professor of Psychiatry). The data comes from task-based studies using psychologically normal control subjects, as well as patients diagnosed with schizophrenia, schizoaffective disorder, schizophreniform disorder, or psychotic disorder: not otherwise specified (hereafter, we refer to these simply as "schizophrenia"). During the experiment, each subject faces a screen that directs them to perform a simple attention-based task over many trials. The structure of each trial is illustrated in Figure 15. Each trial begins with the screen showing a cross for 500 ms, after which a cue appears for 500 ms. The cue consists of a circle of a given color (green in Figure 15). After the cue, the subject is shown 6 rings in a hexagonal arrangement, with three on each side of a central cross, only one of which (the "target") has a color matching that of the cue. Each ring is open on either its left or right side. The rings are shown for 500 ms, after which the subject must choose (with a button press) whether the target was open on its left or right side.

Two types of trials are considered, which we denote as "popout" and "flex." For popout, all of the rings besides the target are of the same color, while for flex the rings are all different colors. Flex is meant to be the more difficult design in that it requires greater attention. These tasks were designed to explore how connectivity differs between the control subjects and patients with schizophrenia among three ROIs, including the primary visual cortex (V1), the posterior parietal cortex (PPC), and the DLPFC. It is hypothesized that for the pairs

V1 vs. PPC and PPC vs. DLPFC, the connectivity in the patient group will be impaired compared to the control group, and this will be more prominent in the flex trials.

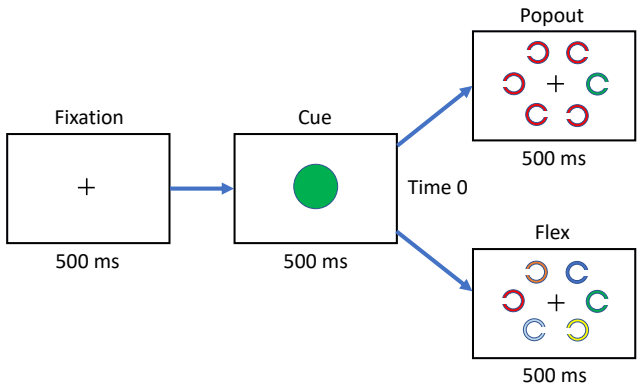

Figure 15: Illustration of the MEG trials.

There are 63 subjects in the data, 27 of whom are psychologically normal, and 36 of whom have schizophrenia. Source reconstruction was performed on the sensor-level MEG trials as in Section 3.2.2, and the PLV between the two pairs of ROIs was calculated on a grid of time and frequency points as in Section 3.2.4. Hence, the processed data consist of a PLV matrix for each combination of subject, design (popout or flex), ROI pair (from both sides of the brain), and visual field in which the target ring appeared (right or left). Based on preliminary results for overall levels of MEG activity, we focus our attention on trials that involve the left visual field, and we consider ROIs from the right hemisphere. Only the trials where the subjects answered correctly are used, and there is an average of 58.05 of these trials over all the subjects and the two designs. Each PLV matrix is recorded on 77 frequency points from 4 to 80 Hz (in increments of 1 Hz), and 176 time points from -200 to 500 ms (in increments of 4 ms), where time 0 corresponds to the switch from the cue to the 6 rings. As in Section 3.2.3, we truncate the PLV to 50 Hz based on visualization of the power. We smooth each subject's PLV using a 15% bandwidth as in Section 3.2.4.

From plotting the PLV for each individual subject, one subject with schizophrenia stands out as a possible outlier, as this subject has fairly uniformly high PLV across all frequencies for a large portion of the beginning of the time period. This subject's PLV matrices for the

two ROI pairs and designs are shown in Figure 16. In the analysis in the following sections, we remove this subject from the dataset.

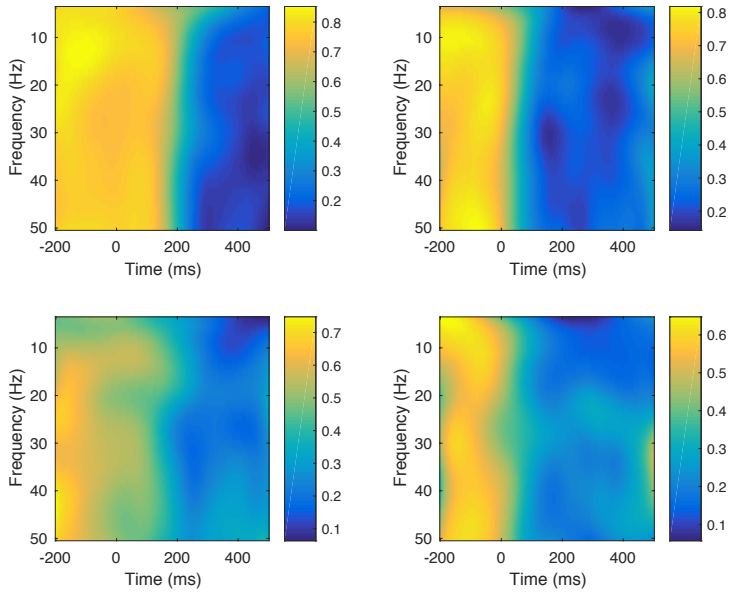

Figure 16: PLV for the outlier subject. The two columns correspond to flex (left) and popout (right), and the two rows correspond to V1 vs. PPC (top) and PPC vs. DLPFC (bottom).

5.2 CONFIDENCE BAND FOR DIFFERENCE IN MEANS

Figures 17 and 18 show, for the popout and flex data, respectively, side-by-side plots of the means for the two groups (those who have schizophrenia vs. those who are psychologically normal). Generally, it seems that the subjects without schizophrenia have slightly higher connectivity than the subjects with schizophrenia on average, though the average overall PLV is fairly low. To attempt to formally identify differences in mean PLV between the two subject groups, we calculate simultaneous confidence bands for the difference in mean functions of two independent samples of two-way functional data. In doing this, we do not assume weak separability, and we simply recast the method for functional data with one

input variable (Cao et al., 2012; Degras, 2017) in terms of two-way functional data.

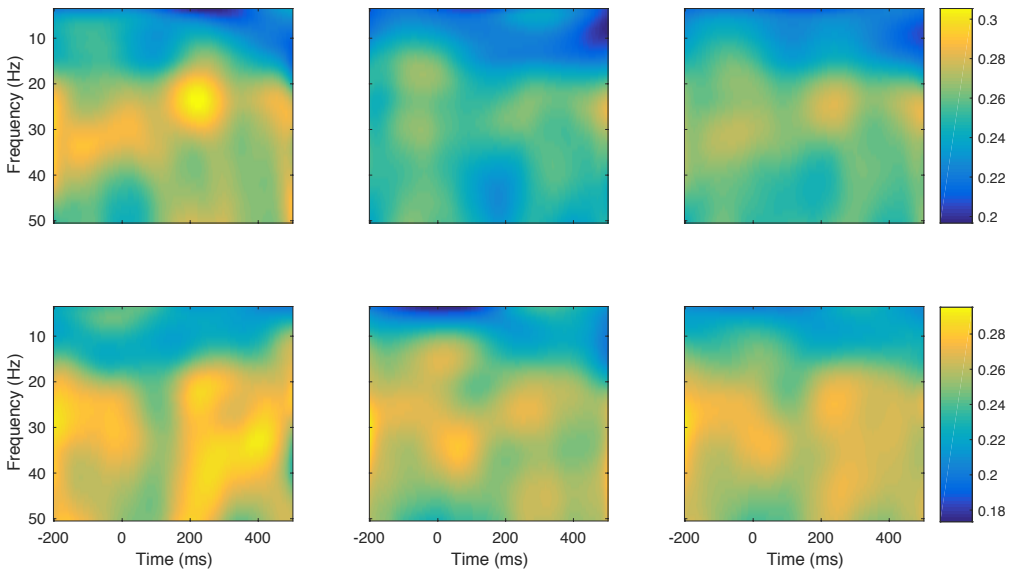

Figure 17: PLV for the popout data averaged over (from left to right) the subjects without schizophrenia, the subjects with schizophrenia, and all of the subjects. The rows correspond to V1 vs. PPC (top) and PPC vs. DLPFC (bottom).

For a given dataset (i.e., a given ROI pair and design), define $\mu_1(s,t)$ and $\mu_2(s,t)$ to be the mean PLV functions for the subjects with and without schizophrenia, respectively. Let $C_1(s,t;u,v)$ and $C_2(s,t;u,v)$ be their respective covariance functions, let $\hat{\mu}_1(s,t)$ and $\hat{\mu}_2(s,t)$ be their respective sample mean functions, and let n_1 and n_2 be their respective sample sizes. With regularity conditions on the observations, grid points, error, and smoothing, there is a functional central limit theorem $\sqrt{n_1}(\hat{\mu}_1 - \mu_1) \xrightarrow{d} \mathcal{G}(0, C_1)$ (and similarly for $\hat{\mu}_2$), where $\mathcal{G}$ denotes a two-way Gaussian process, defined as the joint distribution of a collection of random variables indexed by s,t such that the joint distribution of any finite collection of these variables is multivariate normal. We hence have $\hat{\mu}_1 - \hat{\mu}_2 \sim \mathcal{G}(\mu_1 - \mu_2, C_1/n_1 + C_2/n_2)$ approximately.

Define the standard deviation functions for the two groups as $\sigma_1(s,t) = C_1(s,t;s,t)^{1/2}$

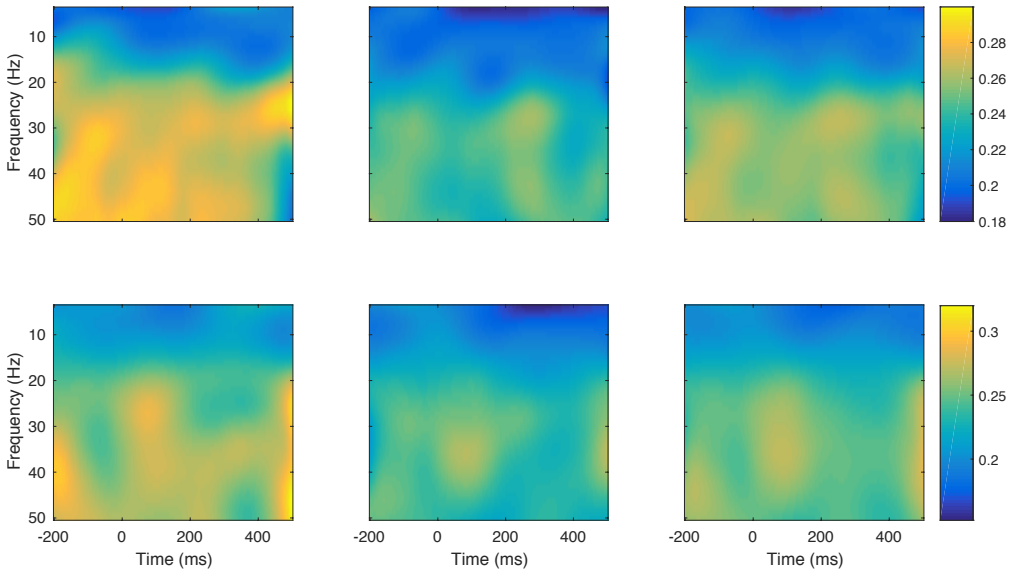

Figure 18: PLV for the flex data averaged over (from left to right) the subjects without schizophrenia, the subjects with schizophrenia, and all of the subjects. The rows correspond to V1 vs. PPC (top) and PPC vs. DLPFC (bottom).

and $\sigma_2(s,t) = C_2(s,t;s,t)^{1/2}$. If we denote the correlation function of $\hat{\mu}_1 - \hat{\mu}_2$ as

$$\rho(s,t;u,v) = \frac{C_1(s,t;u,v)/n_1 + C_2(s,t;u,v)/n_2}{\sqrt{(\sigma_1(s,t)^2/n_1 + \sigma_2(s,t)^2/n_2)(\sigma_1(u,v)^2/n_1 + \sigma_2(u,v)^2/n_2)}},$$

then we have the approximate distribution

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{G}(0,\rho). \tag{5.1}$$

Define the quantile $z_{\alpha,\rho}$ such that, for $Z \sim \mathcal{G}(0,\rho)$, $P(\sup_{s,t}|Z(s,t)| \leq z_{\alpha,\rho}) = 1-\alpha$. Also define $\hat{\sigma}_1(s,t)$, $\hat{\sigma}_2(s,t)$, and $\hat{\rho}(s,t;u,v)$ to be the estimated versions of the standard deviation and correlation functions calculated from the empirical covariances $\hat{C}_1(s,t;u,v)$ and $\hat{C}_2(s,t;u,v)$ (which we calculate simply by vectorizing the two-way data). Then an approximate $1-\alpha$ confidence band for $\mu_1 - \mu_2$ is

$$\hat{\mu}_1(s,t) - \hat{\mu}_2(s,t) \pm z_{\alpha,\hat{\rho}}\sqrt{\hat{\sigma}_1(s,t)^2/n_1 + \hat{\sigma}_2(s,t)^2/n_2}.$$

The quantile $z_{\alpha,\hat{\rho}}$ is estimated using the following process, which can be thought of as a parametric bootstrap of the standardized estimator in Equation (5.1): We consider the FPCA decomposition $\hat{C}_1(s,t;u,v)/n_1 + \hat{C}_2(s,t;u,v)/n_2 \approx \sum_{k=1}^{K}\theta_k\zeta_k(s,t)\zeta_k(u,v)$ for some integer K. If we define

$$h_k(s,t) = \sqrt{\theta_k}\zeta_k(s,t)/\sqrt{\hat{\sigma}_1(s,t)^2/n_1 + \hat{\sigma}_2(s,t)^2/n_2},$$

and let Z_k, $k = 1,\ldots,K$, be a sequence of i.i.d. $N(0,1)$ random variables, then $\sum_{k=1}^{K} Z_k h_k \sim \mathcal{G}(0,\hat{\rho})$ approximately. For B simulated sets of standard normal random variables Z_k, $k = 1,\ldots,K$, we calculate $\sup_{s,t}|\sum_{k=1}^{K} Z_k h_k(s,t)|$, and we estimate $z_{\alpha,\hat{\rho}}$ as the $1-\alpha$ sample quantile of this value. Additionally, we use $\sum_{k=1}^{K}\theta_k\zeta_k(s,t)^2$ to approximate $\hat{\sigma}_1(s,t)^2/n_1 + \hat{\sigma}_2(s,t)^2/n_2$. We use $B = 10000$ and $K = 20$. Using this moderate value of K can be seen as a "smoothing" procedure for the covariances, removing the "noise" terms that contribute less to the variability.

When we calculate 95% confidence bands for mean difference (schizophrenia minus no schizophrenia) for each ROI pair and design, we find that none of these confidence bands have their lower bounds rise above 0, and only the flex and PPC vs. DLPFC dataset has its upper bound fall below 0. The portion of this upper bound that is below 0 is plotted in

Figure 19. We see that it is only a small portion near the low frequency boundary where the subjects without schizophrenia seem to have a significantly higher mean PLV than the subjects with schizophrenia.

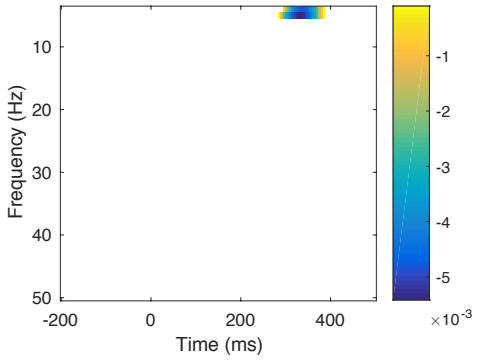

Figure 19: The portion of the upper bound of the 95% confidence band for difference in mean PLV (mean for schizophrenia minus mean for no schizophrenia) that falls below 0 for the flex and PPC vs. DLPFC data.

5.3 STRONG AND WEAK SEPARABILITY TESTS

Since there is little difference in the means of the two groups of subjects, we investigate their modes of variation. We justify the use of product FPCA by applying our test of weak separability. As before, we consider the sets of the subjects' PLV matrices for each combination of ROI pair and design to be separate datasets, and we apply the testing procedures to these 4 datasets separately. We use our χ^2-type mixture and non-studentized empirical bootstrap testing procedures, with the number of components P_n and K_n chosen using the FVE procedure described in Section 2.3.2, and the results are shown in Table 14. This table also shows the results of selected strong separability testing procedures from Aston et al. (2017), using the same values of P_n and K_n as for the weak separability procedures. For the empirical bootstrap procedures, we use $B = 1000$. We see that while the P-values for strong separability are low in all cases, the P-values for weak separability are moderately

high. Hence, in the following sections, we take weak separability to hold, and analyze the scores from product FPCA.

Table 14: For each dataset, the P_n and K_n from the FVE procedure, the weak separability P-values from the χ^2-type mixture ("Weak χ^2") and non-studentized empirical bootstrap ("Weak Emp") procedures, and the strong separability P-values from the Aston et al. (2017) asymptotic χ^2 ("Strong χ^2") and non-studentized empirical bootstrap ("Strong Emp") procedures. ROI pair 1 refers to V1 vs. PPC, and ROI pair 2 refers to PPC vs. DLPFC.

Design	ROI pair	P_n	K_n	Weak χ^2	Weak Emp	Strong χ^2	Strong Emp
Popout	1	5	6	0.3214	0.395	6.974e-22	0.011
	2	5	6	0.1488	0.242	1.292e-23	0.000
Flex	1	5	6	0.2097	0.354	8.272e-25	0.000
	2	5	6	0.0826	0.367	2.655e-12	0.000

Figure 20 shows the 3 products of estimated eigenfunctions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ that explain the most variance for the popout and V1 vs. PPC dataset. These products, which include $\hat{\psi}_1\hat{\phi}_1$, $\hat{\psi}_2\hat{\phi}_1$, and $\hat{\psi}_1\hat{\phi}_2$, contribute proportions of 0.4706, 0.0713, and 0.0643 to the FVE, respectively. Compared to $\hat{\psi}_1$ and $\hat{\psi}_2$, $\hat{\phi}_1$ is a fairly static function, so $\hat{\psi}_1\hat{\phi}_1$ and $\hat{\psi}_2\hat{\phi}_1$ mainly characterize variation based on frequency. $\hat{\psi}_1\hat{\phi}_1$ shows a mode of variation around the beta band (see Table 10). $\hat{\psi}_2\hat{\phi}_1$ shows a contrast between alpha and below, and beta high and above. $\hat{\psi}_1$ is fairly static in comparison to $\hat{\phi}_2$, so $\hat{\psi}_1\hat{\phi}_2$ shows a contrast between the time period before 0 (when the subjects are viewing the cue), and the time period after around 200 ms (which may correspond to when subjects have reacted to the change from the cue to the rings).

5.4 CLASSIFICATION OF SUBJECTS

To determine if the leading scores from product FPCA are associated with schizophrenia, we use these scores as features in binary classification models for schizophrenia. This strategy of

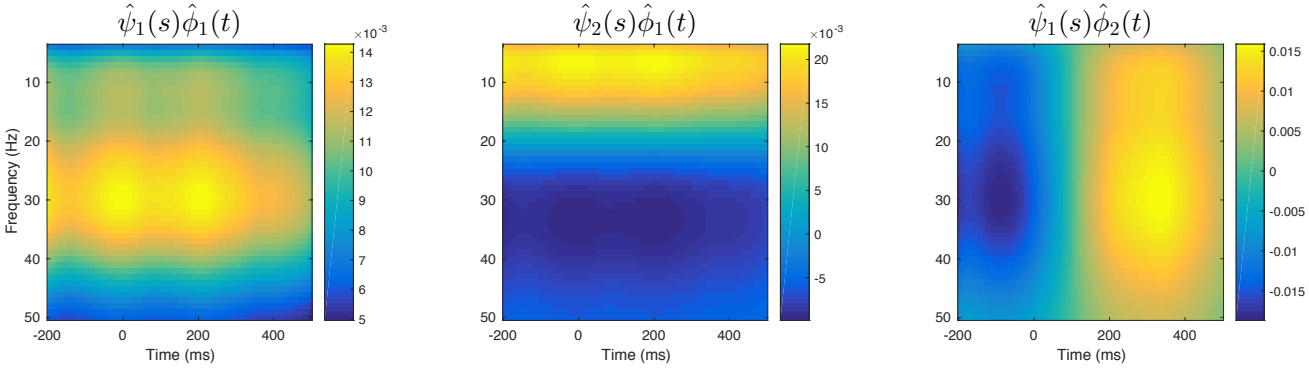

Figure 20: Plots of the products of estimated eigenfunctions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ that explain the most variance (decreasing from left to right) for the popout and V1 vs. PPC data.

using FPCA scores in classification models has been used, for example, in Müller (2005). For a given dataset (ROI pair and design), we train a classification model using the first 6 product FPCA scores for each subject, where the scores are calculated using the combined data from both groups (those with and without schizophrenia). We consider linear discriminant analysis (LDA), logistic regression, and k-nearest neighbors using the R functions lda (from the package MASS), glm, and knn (from the package CLASS), respectively.

Table 15 shows the misclassification rates from 10-fold cross-validation for each method and dataset. None of the methods perform much better than random assignment, the lowest misclassification rate being 0.412 for 1-nearest neighbor on the flex and V1 vs. PPC data.

5.5 COVARIANCE DECOMPOSITION BASED ON L-SEPARABILITY

Here, we use the orthogonal NMF methods from Section 4.3 to approximate the covariance as a sum of d strongly separable components, $C(s,t;u,v) \approx \sum_{l=1}^{d} a^l C^l_{\mathcal{S}}(s,u) C^l_{\mathcal{T}}(t,v)$, for some small d. Recall from Section 4.3 that applying orthogonal NMF to V, the variance matrix of the marginal projection scores, gives $C^l_{\mathcal{S}}(s,u)$ that are each composed of a few distinct frequency components $\psi_j(s)\psi_j(u)$, providing a more interpretable view of the covariation of

Table 15: Misclassification rates for schizophrenia using the first 6 scores from product FPCA. Here, k-NN 1 uses 1 nearest neighbor, k-NN 3 uses 3 nearest neighbors, etc. ROI pair 1 refers to V1 vs. PPC, and ROI pair 2 refers to PPC vs. DLPFC.

Design	ROI pair	LDA	Logistic regression	k-NN 1	k-NN 3	k-NN 5	k-NN 10
Popout	1	0.577	0.583	0.513	0.522	0.555	0.615
	2	0.518	0.525	0.512	0.575	0.632	0.572
Flex	1	0.467	0.472	0.412	0.415	0.450	0.472
	2	0.483	0.490	0.513	0.427	0.458	0.443

subjects' connectivity at different frequency ranges. Also recall that each term in the decomposition of C can be thought of as the covariance of an approximately strongly separable process X_l, $l = 1, \ldots, d$. For each design and ROI pair, we choose d by, starting with $d = 1$, finding an orthogonal NMF solution. We use this solution if the P-values from the strong separability test for the X_l (introduced in Section 4.5) are all above .05; otherwise, we repeat the procedure for the next integer value of d.

Denote $\hat{C}^l_{\mathcal{S}}(s,u) = \sum_{j=1}^{P_n} F_{jl}\hat{\psi}_j(s)\hat{\psi}_j(u)$ and $\hat{C}^l_{\mathcal{T}}(t,v) = \sum_{k=1}^{K_n} G_{kl}\hat{\phi}_k(t)\hat{\phi}_k(v)$ as the estimated versions of $C^l_{\mathcal{S}}(s,u)$ and $C^l_{\mathcal{T}}(t,v)$, where the matrices F and G are from orthogonal NMF applied to $\hat{V}$, the empirical version of V with dimension $P_n \times K_n$. Figure 21 plots each $\hat{C}^l_{\mathcal{S}}(s,u)$ and $\hat{C}^l_{\mathcal{T}}(t,v)$ for the popout and V1 vs. PPC data, for which we have chosen $d = 3$. Here, we scale F and G so that the entries in each of their columns add to 1, so that $Tr(\hat{C}^l_{\mathcal{S}}) = Tr(\hat{C}^l_{\mathcal{T}}) = 1$. In this case, we have $a^1 = 47.5996$, $a^2 = 10.2464$, and $a^3 = 15.1119$ in the decomposition $C(s,t;u,v) \approx \sum_{l=1}^{3} a^l \hat{C}^l_{\mathcal{S}}(s,u)\hat{C}^l_{\mathcal{T}}(t,v)$. The $\hat{C}^l_{\mathcal{T}}(t,v)$ all look fairly similar, which could be expected since they are linear combinations of all the $\hat{\phi}_k(t)\hat{\phi}_k(v)$, and they show fairly uniformly high covariance near the diagonal (with the exception of the low and high time boundaries). On the other hand, we get $\hat{C}^l_{\mathcal{S}}(s,u)$ with more distinct features, as $\hat{C}^1_{\mathcal{S}}(s,u) = \hat{\psi}_1(s)\hat{\psi}_1(u)$, $\hat{C}^2_{\mathcal{S}}(s,u) = \hat{\psi}_2(s)\hat{\psi}_2(u)$, and $\hat{C}^3_{\mathcal{S}}(s,u)$ is a linear combination of the remaining $\hat{\psi}_j(s)\hat{\psi}_j(u)$. $\hat{C}^1_{\mathcal{S}}(s,u)$ shows high positive covariance between connectivity at

frequencies within the beta low to gamma low bands. $\hat{C}^2_{\mathcal{S}}(s,u)$ shows high positive covariance within the theta to alpha bands, and negative covariance between this frequency range and beta high to gamma low. $\hat{C}^3_{\mathcal{S}}(s,u)$ shows high positive covariance among the lowest frequencies (which may be due to boundary effects), and negative covariance between these frequencies and slightly higher frequencies (within alpha to beta low).

5.6 PRODUCT FPCA WITH LOCALIZATION

In this section, we use the product FPCA setting to develop an interpretable and easily computable localization method for eigenfunctions of two-way functional data. For one-way functional data, localization modifies the eigenfunctions to have restricted support regions, i.e., limited intervals of the domain where they are nonzero, thus sacrificing variance explanation to allow for greater interpretability. We adapt the recent procedure for one-way localization developed in Chen & Lei (2015). This method, named localized functional principal component analysis (LFPCA), generates a sequence of orthogonal eigenfunctions that can have differing support regions. Using the data of this chapter, we apply LFPCA to the two marginal covariance functions separately to get localized marginal eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$, and product FPCA functions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ that are localized to rectangular support regions. In particular, the product functions $\hat{\psi}_j(s)\hat{\phi}_k(t)$ are 0 at a point (s,t) if and only if they are 0 at (s,t') for all $t' \in \mathcal{T}$, or 0 at (s',t) for all $s' \in \mathcal{S}$. We thus have a simple and interpretable framework for two-way localization that would not be guaranteed if we were to simply apply one-way localization or sparse PCA methods to the vectorized versions of the two-way data.

Here we give a brief description of LFPCA, further details of which can be found in Chen & Lei (2015): Essentially, for dense, regular one-way functional data, LFPCA adds a localization penalty to the eigenvalue problem for the discretized covariance matrix S (there is also the option for a smoothing penalty, but we do not use this). For a given localization parameter ρ_2, LFPCA modifies the eigenvalue problem on S by way of a convex relaxation involving the deflated Fantope to formulate the problem as a convex optimization.

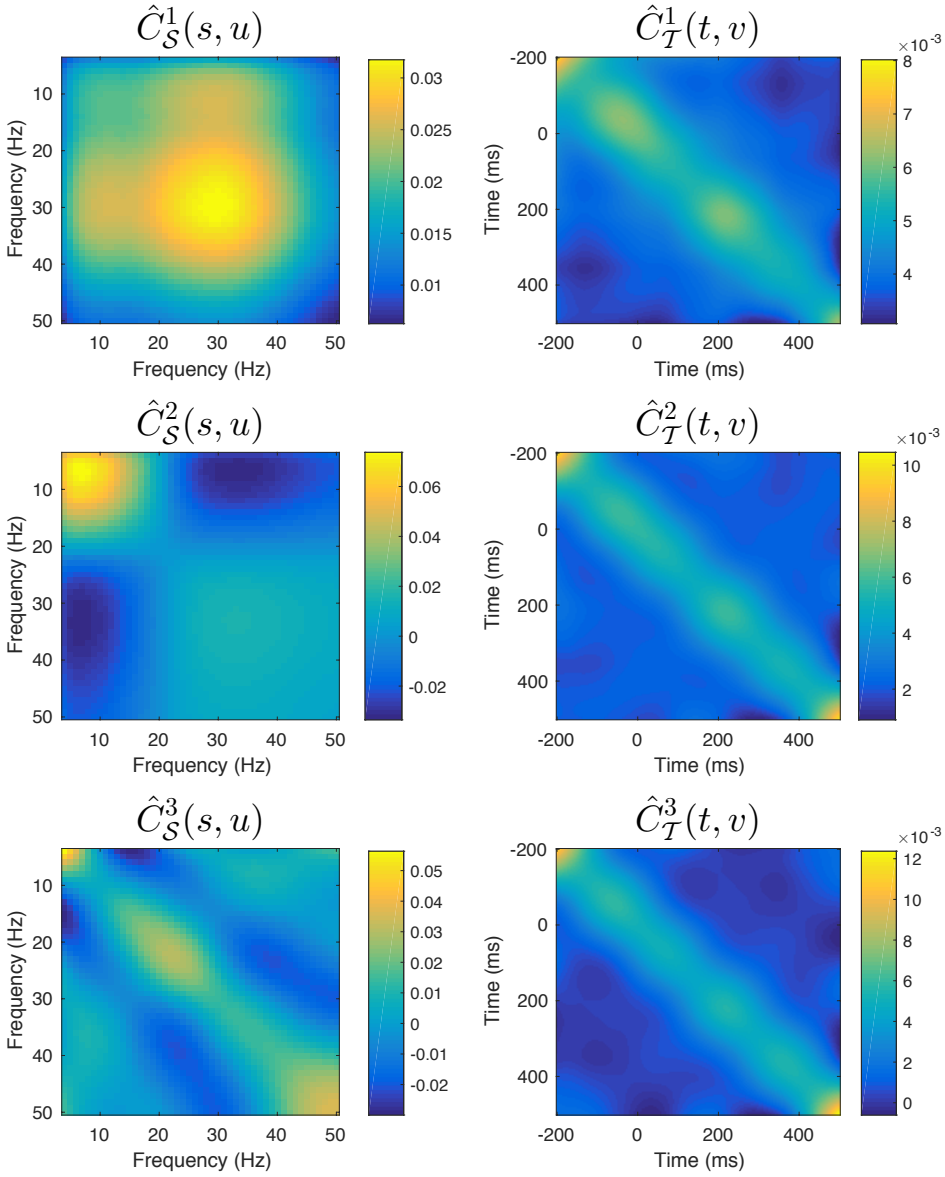

Figure 21: Estimated components of the covariance from orthogonal NMF on $\hat{V}$ using the popout and V1 vs. PPC data.

The eigenvectors v_j, which are discretized versions of the localized eigenfunctions, are estimated sequentially using an iterative algorithm based on the alternating direction method of multipliers (ADMM, Boyd et al. (2011)). The localization parameter ρ_2 can take values from 0 to 1, with higher values imposing more localization. LFPCA, as implemented in the accompanying MATLAB code for Chen & Lei (2015), allows for a different choice of localization parameter in estimating each eigenfunction. Let $\rho_{2,j}$ be the localization parameter used to obtain the jth eigenvector v_j. Chen & Lei (2015) give two methods of obtaining a suitable $\rho_{2,j}$. One method involves cross validation, and is shown to work well when the true eigenfunctions are localized. The other method, which we consider, chooses the most localized eigenfunctions that account for a fixed level of variance. That is, for some choice of $a \in [0,1]$, $\rho_{2,j}$ is chosen as the largest $\rho \in [0,1]$ such that $\frac{{v_j}^T(\rho)Sv_j(\rho)}{{v_j}^T(0)Sv_j(0)} \geq 1-a$, where $v_j(\rho)$ is the estimated value of v_j using ρ as the choice of localization parameter. Chen & Lei (2015) show that this method is most useful when the true eigenfunctions are not localized but we wish to gain more interpretable results while explaining a certain proportion of the variance. Note that higher values of a will give more localization.

Figure 22 shows localized versions of the 3 leading terms from product FPCA for the popout and V1 vs. PPC data. Compare these to the non-localized versions shown in Figure 20. Three levels of localization are considered, determined using $a = 0.1, 0.2, 0.3$. Most of the resulting products share the same basic features as their corresponding non-localized versions. Localization on $\hat{\psi}_1(s)\hat{\phi}_1(t)$ seems to mostly have the effect of setting the lowest and highest frequency values to be 0. Localization on $\hat{\psi}_2(s)\hat{\phi}_1(t)$ using $a = .1$ or $a = .2$ makes this product a contrast between theta to low alpha and beta high. For higher values of a, all the higher frequency values for this product are set to 0, which may be undesirable as this removes the interpretation of this product as a contrast. As a increases, $\hat{\psi}_1(s)\hat{\phi}_2(t)$ becomes a contrast between times nearer to 0 and times nearer to the end of the trial, focused more on frequencies from beta low to gamma low. We see localization has the effect of narrowing the focus to a smaller range of time and frequency values, which can provide a simpler interpretation of the components of product FPCA.

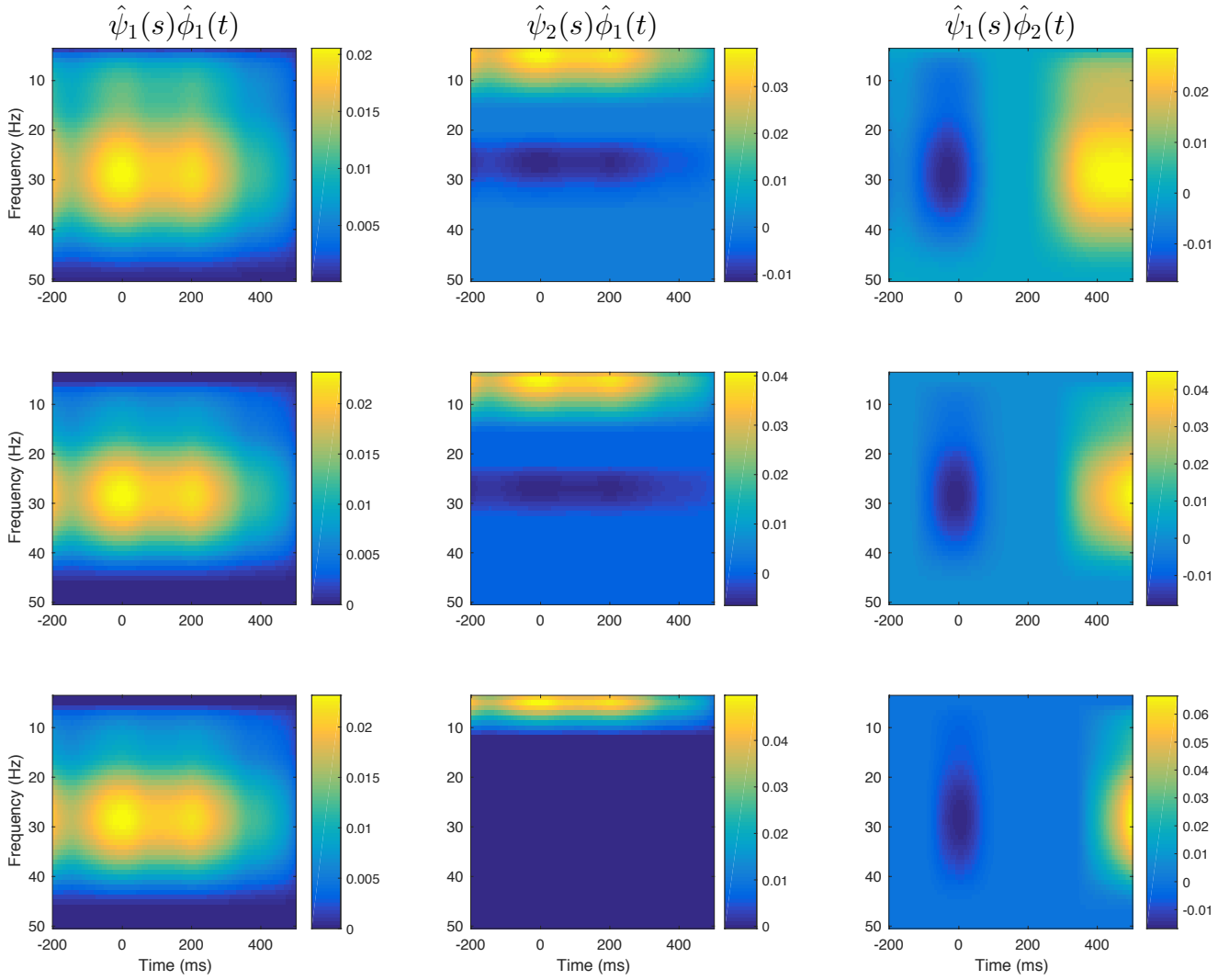

Figure 22: Leading products $\hat{\psi}_j(s)\hat{\phi}_k(t)$ for the popout and V1 vs. PPC data, localized using LFPCA with (from top row to bottom) $a = 0.1, 0.2, 0.3$.

5.7 DISCUSSION

In Sections 5.2 through 5.4, we were for the most part unable to differentiate the two groups of subjects (psychologically normal vs. schizophrenia) using the MEG connectivity data. The collaborators from the Clinical Neurophysiology Research Laboratory noted that both subject groups in both the popout and flex trials had success rates of over 95% in choosing whether the target opened on the left or right. Although flex was designed to be harder, it seems both designs may have been too easy and/or not different enough from each other, leading to a lack of difference in the observed brain connectivity. Additionally, the ROIs we have considered are fairly large, and the collaborators now hypothesize that the activity of interest during the tasks can be pinpointed to smaller subregions.

Based on our discussions with the collaborators, there are two possible issues in the processing of the data, and there are continuing efforts to investigate these problems, which are beyond the scope of this thesis. One potential issue is the lack of baseline correction. Baseline correction normalizes each subject's PLV values using data from the subject outside of the time period of the task, with the goal of capturing the change in PLV induced by the task and accounting for between-subject differences in overall PLV levels. This could possibly be implemented using data from the fixation period, but then an issue arises from the fact that the time during which the subject responds with the button press at the end of a given trial overlaps with the fixation period of the following trial. Another issue present in the processed data could be lack of correction for eye movement. Since the task involves focusing on stimuli that are assigned to a given hemifield, eye movements could give a misleading response in the brain.

APPENDIX A

PROOFS

Proof of Lemma 1

Let f_j $(j = 1, 2, \ldots)$ and g_k $(k = 1, 2, \ldots)$ be a pair of bases that satisfies weak separability. For $(j, k) \neq (j', k')$, we have $\langle Cf_j \otimes g_k, f_{j'} \otimes g_{k'} \rangle = E\left(\langle X - \mu, f_j \otimes g_k \rangle \langle X - \mu, f_{j'} \otimes g_{k'} \rangle\right) = E(\tilde{\chi}_{jk}\tilde{\chi}_{j'k'}) = 0$. The removal of the expectation from the inner product is allowed by the Fubini–Tonelli theorem, since $\mathcal{S}$ and $\mathcal{T}$ are compact and $E(|X(s,t) - \mu(s,t)||X(u,v) - \mu(u,v)|) \leq \sqrt{C(s,t;s,t)C(u,v;u,v)}$ for all $s, u \in \mathcal{S}$ and $t, v \in \mathcal{T}$ by the Cauchy–Schwarz inequality. Since the covariance operator C is diagonalized under the orthonormal basis $f_j \otimes g_k$ $(j = 1, 2, \ldots;\ k = 1, 2, \ldots)$, by Mercer's theorem,

$$C(s,t;u,v) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \eta_{jk} f_j(s) g_k(t) f_j(u) g_k(v),$$

where $\eta_{jk} = \langle Cf_j \otimes g_k, f_j \otimes g_k \rangle = \mathrm{var}\left(\langle X - \mu, f_j \otimes g_k \rangle\right)$, and the convergence is absolute and uniform.

The marginal kernel $C_{\mathcal{S}}(s,u)$ can then be written as

$$\begin{aligned} C_{\mathcal{S}}(s,u) &= \int_{\mathcal{T}} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \eta_{jk} f_j(s) g_k(t) f_j(u) g_k(t) dt \\ &= \sum_{j=1}^{\infty} \left(\sum_{k=1}^{\infty} \eta_{jk} \right) f_j(s) f_j(u). \end{aligned}$$

The exchange of the integral and sums is allowed by the Fubini–Tonelli theorem, by noticing that

$$\begin{aligned}
&\int_{\mathcal{T}} \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} |\eta_{jk} f_j(s) g_k(t) f_j(u) g_k(t)| dt \\
&\leq \int_{\mathcal{T}} \left\{ \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \eta_{jk} f_j^2(s) g_k^2(t) \right\}^{1/2} \left\{ \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \eta_{jk} f_j^2(u) g_k^2(t) \right\}^{1/2} dt \\
&= \int_{\mathcal{T}} C(s,t;s,t)^{1/2} C(u,t;u,t)^{1/2} dt \\
&\leq \int_{\mathcal{T}} \sup_{s,t} |C(s,t,s,t)| dt < \infty,
\end{aligned}$$

where we use the Cauchy–Schwarz inequality.

Thus, we see that the f_j are eigenfunctions of $C_{\mathcal{S}}$ with eigenvalues $\lambda_j = \sum_{k=1}^{\infty} \eta_{jk}$. An analogous computation shows that the g_k are eigenfunctions of $C_{\mathcal{T}}$ with eigenvalues $\gamma_k = \sum_{j=1}^{\infty} \eta_{jk}$.

Proof of Lemma 2

Under strong separability, we have $C(s,t;u,v) = aC_1(s,u)C_2(t,v)$. From the definition of $C_{\mathcal{S}}$ we have

$$C_{\mathcal{S}}(s,u) = \int_{\mathcal{T}} C(s,t;u,t) dt = aC_1(s,u) \int_{\mathcal{T}} C_2(t,t) dt = aC_1(s,u).$$

An analogous argument shows $C_{\mathcal{T}}(t,v) = aC_2(t,v)$. Note that $a = \int_{\mathcal{T}} \int_{\mathcal{S}} C(s,t;s,t) ds dt$. We can rewrite $C(s,t;u,v)$ as

$$C(s,t;u,v) = \frac{1}{a} C_{\mathcal{S}}(s,u) C_{\mathcal{T}}(t,v).$$

Therefore,

$$\mathrm{cov}(\chi_{jk}, \chi_{j'k'}) = \langle C\psi_j \otimes \phi_k, \psi_{j'} \otimes \phi_{k'} \rangle = \frac{1}{a} \langle C_{\mathcal{S}} \psi_j, \psi_{j'} \rangle \langle C_{\mathcal{T}} \phi_k, \phi_{k'} \rangle.$$

When $j \neq j'$ or $k \neq k'$, it is easy to see that $\mathrm{cov}(\chi_{jk}, \chi_{j'k'}) = 0$. Thus, we have weak separability.

Proof of Lemma 3

When V is of rank 1, V can be written $V = WZ^T$, where W and Z are column vectors with entries $(w_1, w_2, \ldots)$ and $(z_1, z_2, \ldots)$, respectively. Thus, $\eta_{jk} = w_j z_k$, and under weak separability, Equation (2.4) can be written

$$\begin{aligned} C(s,t;u,v) &= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} w_j z_k \psi_j(s)\psi_j(u)\phi_k(t)\phi_k(v) \\ &= \left\{\sum_{j=1}^{\infty} w_j \psi_j(s)\psi_j(u)\right\}\left\{\sum_{k=1}^{\infty} z_k \phi_k(t)\phi_k(v)\right\}. \end{aligned}$$

The above can be normalized to fit the definition of strong separability in Lemma 2.

Under strong separability, from the proof of Lemma 2 we have

$$C(s,t;u,v) = \frac{1}{\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;s,t)dsdt} C_{\mathcal{S}}(s,u)C_{\mathcal{T}}(t,v),$$

so $\eta_{jk} = \{1/\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;s,t)dsdt\}\langle C_{\mathcal{S}}\psi_j, \psi_j\rangle\langle C_{\mathcal{T}}\phi_k, \phi_k\rangle = \{1/\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;s,t)dsdt\}\lambda_j\gamma_k$, and then $V = \{1/\int_{\mathcal{T}}\int_{\mathcal{S}} C(s,t;s,t)dsdt\}\Lambda\Gamma^T$.

Proof of Theorem 4

We use the notation from Section 2.3.1. For H_1 and H_2 two real separable Hilbert spaces, we further define the partial trace with respect to H_1 as the unique bounded linear operator $Tr_1 : \mathcal{B}_{Tr}(H_1 \otimes H_2) \to \mathcal{B}_{Tr}(H_2)$ satisfying $Tr_1(C_1 \tilde{\otimes} C_2) = Tr(C_1)C_2$ for all $C_1 \in \mathcal{B}_{Tr}(H_1)$, $C_2 \in \mathcal{B}_{Tr}(H_2)$. The partial trace with respect to H_2 is defined symmetrically and denoted by Tr_2. With the notation of partial trace, we can see that $C_{\mathcal{T}} = Tr_1(C)$ and $C_{\mathcal{S}} = Tr_2(C)$. The estimated marginal covariance operators can also be written as $\hat{C}_{\mathcal{S}} = Tr_2(C_n)$ and $\hat{C}_{\mathcal{T}} = Tr_1(C_n)$. We use these equalities in proofs but not in computation. In practice, the estimated marginal covariances are calculated without having to calculate C_n.

From Condition I in Section 2.3.1 and the remark following it, $\mathcal{Z}_n = n^{1/2}(C_n - C)$ converges to a Gaussian random element in $\mathcal{B}_{Tr}\{L^2(\mathcal{S}\times\mathcal{T})\}$ with mean 0 and covariance structure $\Sigma_C = \mathrm{E}[\{(X-\mu)\otimes(X-\mu) - C\}\tilde{\otimes}\{(X-\mu)\otimes(X-\mu) - C\}]$.

For T_n as defined in Equation (2.7),

$$T_n(j,k,j',k') = \sqrt{n}\langle C_n(\hat{\psi}_j \otimes \hat{\phi}_k), \hat{\psi}_{j'} \otimes \hat{\phi}_{k'}\rangle = \sqrt{n}Tr((\hat{\psi}_j \otimes \hat{\psi}_{j'})\tilde{\otimes}(\hat{\phi}_k \otimes \hat{\phi}_{k'})C_n).$$

Using (5.1.8) in Hsing & Eubank (2015), we have

$$(\hat{\psi}_j - \psi_j) = \mathcal{M}_j(\hat{C}_\mathcal{S} - C_\mathcal{S})\psi_j + o_p(\hat{\psi}_j - \psi_j),$$

where $\mathcal{M}_j = \sum_{m\neq j}(\lambda_j - \lambda_m)^{-1}\psi_m \otimes \psi_m \in \mathcal{B}_{Tr}(\mathcal{S})$ and λ_j is the jth eigenvalue of $C_\mathcal{S}$. Analogously,

$$(\hat{\phi}_k - \phi_k) = \mathcal{M}'_k(\hat{C}_\mathcal{T} - C_\mathcal{T})\phi_k + o_p(\hat{\phi}_k - \phi_k),$$

where $\mathcal{M}'_k = \sum_{m\neq k}(\gamma_k - \gamma_m)^{-1}\phi_m \otimes \phi_m \in \mathcal{B}_{Tr}(\mathcal{T})$ and γ_k is the kth eigenvalue of $C_\mathcal{T}$. Here, Condition II is used to guarantee that $\mathcal{M}_j$ and $\mathcal{M}'_k$ exist for $j = 1, \ldots, P$ and $k = 1, \ldots, K$.

Using $\hat{C}_\mathcal{S} - C_\mathcal{S} = Tr_2(C_n - C)$ and $\hat{C}_\mathcal{T} - C_\mathcal{T} = Tr_1(C_n - C)$, we can write $T_n(j, k, j', k')$ as

$$\begin{aligned} T_n(j,k,j',k') =& \sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'})C\right) \\ &+ \sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'})(C_n - C)\right) \\ &+ \sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes (\mathcal{M}'_{k'}Tr_1(C_n - C)\phi_{k'}))C\right) \\ &+ \sqrt{n}Tr\left((\psi_j \otimes \psi_{j'})\tilde{\otimes}((\mathcal{M}'_kTr_1(C_n - C)\phi_k) \otimes \phi_{k'})C\right) \\ &+ \sqrt{n}Tr\left((\psi_j \otimes (\mathcal{M}_{j'}Tr_2(C_n - C)\psi_{j'}))\tilde{\otimes}(\phi_k \otimes \phi_{k'})C\right) \\ &+ \sqrt{n}Tr\left(((\mathcal{M}_jTr_2(C_n - C)\psi_j) \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'})C\right) \\ &+ o_p(1). \end{aligned} \tag{A.1}$$

Note that the first term in the above equation is zero under H_0, since under H_0 we have the representation $C(s,t,u,v) = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\eta_{jk}\psi_j(s)\psi_j(u)\phi_k(t)\phi_k(v)$, where $\eta_{jk} = \text{var}(\chi_{jk})$. Also, by Proposition C.1 in Aston et al. (2017), we have that $Tr(ATr_1(T)) = Tr((Id_1\tilde{\otimes}A)T)$, where Id_1 is an identity operator on $\mathcal{S}$, $A \in \mathcal{B}(\mathcal{T})$, and $T \in \mathcal{B}_{Tr}(\mathcal{S}\times\mathcal{T})$. An analogous identity holds for $Tr_2(T)$. Using these facts, we give a simplified form of $T_n(j,k,j',k')$ under H_0 for 3 cases:

Case (i): For $j \neq j'$ and $k \neq k'$,

$$T_n(j,k,j',k') = Tr\left(((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) + o_p(1).$$

Case (ii): For $j = j'$ and $k \neq k'$,

$$\begin{aligned} T_n(j,k,j',k') =& Tr\left(((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) \\ &+Tr\left((Id_1\tilde{\otimes}(\eta_{jk'}(\phi_k \otimes \phi_{k'})\mathcal{M}'_k))\mathcal{Z}_n\right) \\ &+Tr\left((Id_1\tilde{\otimes}(\eta_{jk}(\phi_{k'} \otimes \phi_k)\mathcal{M}'_{k'}))\mathcal{Z}_n\right) + o_p(1). \end{aligned}$$

Case (iii): For $j \neq j'$ and $k = k'$,

$$\begin{aligned} T_n(j,k,j',k') =& Tr\left(((\psi_j \otimes \psi_{j'})\tilde{\otimes}(\phi_k \otimes \phi_{k'}))\mathcal{Z}_n\right) \\ & Tr\left(((\eta_{jk}(\psi_{j'} \otimes \psi_j)\mathcal{M}_{j'})\tilde{\otimes}Id_2)\mathcal{Z}_n\right) \\ &+Tr\left(((\eta_{j'k}(\psi_j \otimes \psi_{j'})\mathcal{M}_j)\tilde{\otimes}Id_2)\mathcal{Z}_n\right) + o_p(1). \end{aligned}$$

In each of the above cases, two or more of the terms in Equation (A.1) end up being zero due to the orthogonality of the eigenfunctions. The latter 2 cases can be simplified to get the result in the statement of the theorem by noting that $\eta_{jk'}(\phi_k \otimes \phi_{k'})\mathcal{M}'_k = \eta_{jk'}(\gamma_k - \gamma_{k'})^{-1}\phi_k \otimes \phi_{k'}$ and $\eta_{jk}(\psi_{j'} \otimes \psi_j)\mathcal{M}_{j'} = \eta_{jk}(\lambda_{j'} - \lambda_j)^{-1}\psi_{j'} \otimes \psi_j$.

Proof of Corollary 5

From Theorem 4, we can see that all the terms of $T_n(j,k,j',k')$ can be written in the form $Tr(A_1\tilde{\otimes}A_2\mathcal{Z}_n)$ for some $A_1 \in \mathcal{B}(\mathcal{S})$ and $A_2 \in \mathcal{B}(\mathcal{T})$. Since $\mathcal{Z}_n$ converges to a Gaussian random element and $Tr(A_1\tilde{\otimes}A_2\mathcal{Z}_n)$ is a continuous linear mapping, the $T_n(j,k,j',k')$ are asymptotically jointly Gaussian for different sets of (j,k,j',k'). Let Θ be the covariance structure of the asymptotic joint distribution of the $T_n(j,k,j',k')$, and define $\mathcal{Z}$ to be a Gaussian random element with the limiting distribution of $\mathcal{Z}_n$. By the continuous mapping theorem, Θ can be calculated from terms of the form

$$\mathrm{E}(Tr(A_1\tilde{\otimes}A_2\mathcal{Z})Tr(B_1\tilde{\otimes}B_2\mathcal{Z})) = Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right), \tag{A.2}$$

where Σ_C is defined as in the proof of Theorem 4.

Recall the K-L expansion of the process $X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\chi_{jk}\psi_j(s)\phi_k(t)$. We define $u_{ij} = \psi_i \otimes \psi_j \in \mathcal{B}_{HS}(\mathcal{S})$, $v_{ij} = \phi_i \otimes \phi_j \in \mathcal{B}_{HS}(\mathcal{T})$, $\beta_{i,i',j,j',k,k',l,l'} = \mathrm{E}(\chi_{ii'}\chi_{jj'}\chi_{kk'}\chi_{ll'})$ and $\eta_{ii'} = \mathrm{E}(\chi^2_{ii'})$. With weak separability, we have

$$
\begin{aligned}
&Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right)\\
&= \sum_{i,i',j,j',k,k',l,l'} \beta_{i,i',j,j',k,k',l,l'} Tr[A_1u_{ij}]Tr[A_2v_{i'j'}]Tr[B_1u_{kl}]Tr[B_2v_{k'l'}]\\
&\quad - \sum_{i,i',j,j'} \eta_{ii'}\eta_{jj'} Tr[A_1u_{ii}]Tr[B_1u_{jj}]Tr[A_2v_{i'i'}]Tr[B_2v_{j'j'}].
\end{aligned}
$$

Each of the trace terms in the above equation can be evaluated using $Tr[Id_1u_{ij}] = I(i=j)$, $Tr[Id_2v_{i'j'}] = I(i'=j')$, $Tr[(\psi_{j_1}\otimes\psi_{j_1'})u_{ij}] = I(i=j_1)I(j=j_1')$, and $Tr[(\phi_{k_1}\otimes\phi_{k_1'})v_{i'j'}] = I(i'=k_1)I(j'=k_1')$. From these identities and the possible forms of A_1, A_2, B_1, and B_2 given in Theorem 4, it follows that the second sum is always 0. The first sum can be simplified by considering 9 cases, as follows:

Case (1): For $A_1 = a_1\psi_{j_1}\otimes\psi_{j_1'}$, $A_2 = a_2\phi_{k_1}\otimes\phi_{k_1'}$, $B_1 = b_1\psi_{j_2}\otimes\psi_{j_2'}$, $B_2 = b_2\phi_{k_2}\otimes\phi_{k_2'}$,

$$
Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1a_2b_1b_2\beta_{j_1,k_1,j_1',k_1',j_2,k_2,j_2',k_2'}.
$$

Case (2): For $A_1 = Id_1$, $A_2 = a_2\phi_{k_1}\otimes\phi_{k_1'}$, $B_1 = b_1\psi_{j_2}\otimes\psi_{j_2'}$, $B_2 = b_2\phi_{k_2}\otimes\phi_{k_2'}$,

$$
Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_2b_1b_2\sum_{i=1}^{\infty}\beta_{i,k_1,i,k_1',j_2,k_2,j_2',k_2'}.
$$

Case (3): For $A_1 = a_1\psi_{j_1}\otimes\psi_{j_1'}$, $A_2 = Id_2$, $B_1 = b_1\psi_{j_2}\otimes\psi_{j_2'}$, $B_2 = b_2\phi_{k_2}\otimes\phi_{k_2'}$,

$$
Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1b_1b_2\sum_{i'=1}^{\infty}\beta_{j_1,i',j_1',i',j_2,k_2,j_2',k_2'}.
$$

Case (4): For $A_1 = a_1\psi_{j_1}\otimes\psi_{j_1'}$, $A_2 = a_2\phi_{k_1}\otimes\phi_{k_1'}$, $B_1 = Id_1$, $B_2 = b_2\phi_{k_2}\otimes\phi_{k_2'}$,

$$
Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1a_2b_2\sum_{k=1}^{\infty}\beta_{j_1,k_1,j_1',k_1',k,k_2,k,k_2'}.
$$

Case (5): For $A_1 = a_1\psi_{j_1}\otimes\psi_{j_1'}$, $A_2 = a_2\phi_{k_1}\otimes\phi_{k_1'}$, $B_1 = b_1\psi_{j_2}\otimes\psi_{j_2'}$, $B_2 = Id_2$,

$$
Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1a_2b_1\sum_{k'=1}^{\infty}\beta_{j_1,k_1,j_1',k_1',j_2,k',j_2',k'}.
$$

Case (6): For $A_1 = Id_1$, $A_2 = a_2\phi_{k_1} \otimes \phi_{k_1'}$, $B_1 = Id_1$, $B_2 = b_2\phi_{k_2} \otimes \phi_{k_2'}$,

$$Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_2b_2\sum_{i=1}^{\infty}\sum_{k=1}^{\infty}\beta_{i,k_1,i,k_1',k,k_2,k,k_2'}.$$

Case (7): For $A_1 = Id_1$, $A_2 = a_2\phi_{k_1} \otimes \phi_{k_1'}$, $B_1 = b_1\psi_{j_2} \otimes \psi_{j_2'}$, $B_2 = Id_2$,

$$Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_2b_1\sum_{i=1}^{\infty}\sum_{k'=1}^{\infty}\beta_{i,k_1,i,k_1',j_2,k',j_2',k'}.$$

Case (8): For $A_1 = a_1\psi_{j_1} \otimes \psi_{j_1'}$, $A_2 = Id_2$, $B_1 = Id_1$, $B_2 = b_2\phi_{k_2} \otimes \phi_{k_2'}$,

$$Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1b_2\sum_{i'=1}^{\infty}\sum_{k=1}^{\infty}\beta_{j_1,i',j_1',i',k,k_2,k,k_2'}.$$

Case (9): For $A_1 = a_1\psi_{j_1} \otimes \psi_{j_1'}$, $A_2 = Id_2$, $B_1 = b_1\psi_{j_2} \otimes \psi_{j_2'}$, $B_2 = Id_2$,

$$Tr\left((A_1\tilde{\otimes}A_2)\widetilde{\bigotimes}(B_1\tilde{\otimes}B_2)\Sigma_C\right) = a_1b_1\sum_{i'=1}^{\infty}\sum_{k'=1}^{\infty}\beta_{j_1,i',j_1',i',j_2,k',j_2',k'}.$$

In the above, a_1, a_2, b_1, and b_2 are scalar constants. Using the above, all the terms in Θ can be obtained from straightforward but tedious calculations.

To illustrate the calculation of $\Theta(j,k,j',k',l,m,l',m')$, the term in Θ corresponding to the asymptotic covariance of $T_n(j,k,j',k')$ and $T_n(l,m,l',m')$, we consider as an example the case where $j \neq j'$, $k \neq k'$, $l \neq l'$, and $m \neq m'$. Here,

$$\begin{aligned}
&\Theta(j,k,j',k',l,m,l',m') \\
&\overset{\text{by Thm. 4 (i)}}{=} \mathrm{E}\left(Tr\left[\{(\psi_j\otimes\psi_{j'})\tilde{\otimes}(\phi_k\otimes\phi_{k'})\}\mathcal{Z}\right]Tr\left[\{(\psi_l\otimes\psi_{l'})\tilde{\otimes}(\phi_m\otimes\phi_{m'})\}\mathcal{Z}\right]\right) \\
&\overset{\text{by Eq. (A.2)}}{=} Tr\left[\{(\psi_j\otimes\psi_{j'})\tilde{\otimes}(\phi_k\otimes\phi_{k'})\}\widetilde{\bigotimes}\{(\psi_l\otimes\psi_{l'})\tilde{\otimes}(\phi_m\otimes\phi_{m'})\}\Sigma_C\right] \\
&\overset{\text{by Case (1)}}{=} \beta_{j,k,j',k',l,m,l',m'} = \mathrm{E}(\chi_{jk}\chi_{j'k'}\chi_{lm}\chi_{l'm'}),
\end{aligned}$$

where we have used $A_1 = \psi_j \otimes \psi_{j'}$, $A_2 = \phi_k \otimes \phi_{k'}$, $B_1 = \psi_l \otimes \psi_{l'}$, and $B_2 = \phi_m \otimes \phi_{m'}$.

Proof of Lemma 6

Let $X_N^*(s,t) = \mu(s,t) + \sum_{j=1}^{N}\sum_{k=1}^{N}\chi_{jk}\psi_j(s)\phi_k(t)$, and let C_N^* denote the covariance structure of X_N^*. Thus,

$$C_N^*(s,t;u,v) = \sum_{j=1}^{N}\sum_{j'=1}^{N}\sum_{k=1}^{N}\sum_{k'=1}^{N}\mathrm{cov}(\chi_{jk},\chi_{j'k'})\psi_j(s)\psi_{j'}(u)\phi_k(t)\phi_{k'}(v).$$

It is easy to show that C_N^* converges to C in Hilbert–Schmidt norm. Let $C_{\mathcal{S},N} = Tr_2(C_N^*)$, which converges to $C_{\mathcal{S}}$ because Tr_2 is continuous and linear. We know that $\langle C_{\mathcal{S}}\psi_j, \psi_{j'}\rangle = 0$ for $j \neq j'$, and therefore $\lim_N \langle C_{\mathcal{S},N}\psi_j, \psi_{j'}\rangle = 0$. By definition,

$$\begin{aligned}
&\langle C_{\mathcal{S},N}\psi_j, \psi_{j'}\rangle = \int_{\mathcal{S}}\int_{\mathcal{S}} \left\{ \int_{\mathcal{T}} C_N^*(s,t;u,t)dt \right\} \psi_j(s)\psi_{j'}(u)dsdu \\
=&\int_{\mathcal{S}}\int_{\mathcal{S}}\int_{\mathcal{T}} \sum_{l=1}^{N}\sum_{l'=1}^{N}\sum_{k=1}^{N}\sum_{k'=1}^{N} \mathrm{cov}(\chi_{lk}, \chi_{l'k'})\psi_l(s)\psi_{l'}(u)\phi_k(t)\phi_{k'}(t)\psi_j(s)\psi_{j'}(u)dtdsdu \\
=&\sum_{k=1}^{N} \mathrm{cov}(\chi_{jk}, \chi_{j'k}).
\end{aligned}$$

Therefore, $\lim_N \sum_{k=1}^{N} \mathrm{cov}(\chi_{jk}, \chi_{j'k}) = 0$, i.e., $\sum_{k=1}^{\infty} \mathrm{cov}(\chi_{jk}, \chi_{j'k}) = 0$ for $j \neq j'$.

The same argument holds for the empirical version as follows: Let

$$\hat{C}_N^*(s,t;u,v) = \sum_{j=1}^{N}\sum_{j'=1}^{N}\sum_{k=1}^{N}\sum_{k'=1}^{N} \frac{1}{\sqrt{n}} T_n(j,k,j',k')\hat{\psi}_j(s)\hat{\psi}_{j'}(u)\hat{\phi}_k(t)\hat{\phi}_{k'}(v).$$

We can write the empirical covariance as

$$\begin{aligned}
C_n(s,t;u,v) =& \frac{1}{n}\sum_{i=1}^{n}(X_i(s,t) - \bar{X}(s,t))(X_i(u,v) - \bar{X}(u,v)) \\
=& \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\infty}\sum_{j'=1}^{\infty}\sum_{k=1}^{\infty}\sum_{k'=1}^{\infty} \hat{\chi}_{i,jk}\hat{\chi}_{i,j'k'}\hat{\psi}_j(s)\hat{\psi}_{j'}(u)\hat{\phi}_k(t)\hat{\phi}_{k'}(v) \\
=& \sum_{j=1}^{\infty}\sum_{j'=1}^{\infty}\sum_{k=1}^{\infty}\sum_{k'=1}^{\infty} \frac{1}{\sqrt{n}} T_n(j,k,j',k')\hat{\psi}_j(s)\hat{\psi}_{j'}(u)\hat{\phi}_k(t)\hat{\phi}_{k'}(v),
\end{aligned}$$

and it is clear that $\hat{C}_N^*$ converges (with respect to N) to C_n in Hilbert–Schmidt norm. Let $\hat{C}_{\mathcal{S},N} = Tr_2(\hat{C}_N^*)$, which converges to $\hat{C}_{\mathcal{S}}$ because Tr_2 is continuous and linear. We know that $\langle \hat{C}_{\mathcal{S}}\hat{\psi}_j, \hat{\psi}_{j'}\rangle = 0$ for $j \neq j'$, and therefore $\lim_N \langle \hat{C}_{\mathcal{S},N}\hat{\psi}_j, \hat{\psi}_{j'}\rangle = 0$. By definition,

$$\begin{aligned}
&\langle \hat{C}_{\mathcal{S},N}\hat{\psi}_j, \hat{\psi}_{j'}\rangle = \int_{\mathcal{S}}\int_{\mathcal{S}} \left\{ \int_{\mathcal{T}} \hat{C}_N^*(s,t;u,t)dt \right\} \hat{\psi}_j(s)\hat{\psi}_{j'}(u)dsdu \\
=&\int_{\mathcal{S}}\int_{\mathcal{S}}\int_{\mathcal{T}} \sum_{l=1}^{N}\sum_{l'=1}^{N}\sum_{k=1}^{N}\sum_{k'=1}^{N} \frac{1}{\sqrt{n}} T_n(l,k,l',k')\hat{\psi}_l(s)\hat{\psi}_{l'}(u)\hat{\phi}_k(t)\hat{\phi}_{k'}(t)\hat{\psi}_j(s)\hat{\psi}_{j'}(u)dtdsdu \\
=&\sum_{k=1}^{N} \frac{1}{\sqrt{n}} T_n(j,k,j',k).
\end{aligned}$$

Therefore, $\lim_N \sum_{k=1}^{N} T_n(j,k,j',k) = 0$, i.e., $\sum_{k=1}^{\infty} T_n(j,k,j',k) = 0$ for $j \neq j'$.

Analogous calculations can be done for $k \neq k'$ to show that $\sum_{j=1}^{\infty} \text{cov}(\chi_{jk}, \chi_{jk'}) = 0$ and $\sum_{j=1}^{\infty} T_n(j,k,j,k') = 0$.

Proof of Lemma 7

Expand the C_1^l and C_2^l terms using the marginal eigenfunctions. That is, write

$$C_1^l(s,u) = \sum_j \sum_{j'} \alpha_{jj'}^l \psi_j(s)\psi_{j'}(u)$$

and

$$C_2^l(t,v) = \sum_k \sum_{k'} \beta_{kk'}^l \phi_k(t)\phi_{k'}(v).$$

Under weak separability we can then write

$$\begin{aligned}
\|C - \sum_{l=1}^{d} C^l\|^2 = \int\int\int\int [&\sum_j \sum_k \eta_{jk}\psi_j(s)\psi_j(u)\phi_k(t)\phi_k(v) \\
&- \sum_{l=1}^{d} \sum_j \sum_k \sum_{j'} \sum_{k'} \alpha_{jj'}^l \beta_{kk'}^l \psi_j(s)\psi_{j'}(u)\phi_k(t)\phi_{k'}(v)]^2 ds dt du dv \\
&= \sum_j \sum_k \eta_{jk}^2 + \sum_j \sum_k \sum_{j'} \sum_{k'} \left(\sum_{l=1}^{d} \alpha_{jj'}^l \beta_{kk'}^l \right)^2 - 2 \sum_j \sum_k \eta_{jk} \sum_{l=1}^{d} \alpha_{jj}^l \beta_{kk}^l \\
&= \sum_j \sum_k \left(\eta_{jk} - \sum_{l=1}^{d} \alpha_{jj}^l \beta_{kk}^l \right)^2 + \sum_{j \neq j'} \sum_{k \neq k'} \left(\sum_{l=1}^{d} \alpha_{jj'}^l \beta_{kk'}^l \right)^2 .
\end{aligned}$$

Note that for each l, the arrays of $\alpha_{jj'}^l$ and $\beta_{kk'}^l$ must be nonnegative definite. Hence, the α_{jj}^l and β_{kk}^l must be nonnegative, and it is clear that the minimizing value of $\|C - \sum_{l=1}^{d} C^l\|$ can be attained by setting $\alpha_{jj'}^l = 0$ for $j \neq j'$ and $\beta_{kk'}^l = 0$ for $k \neq k'$. Then $\|C - \sum_{l=1}^{d} C^l\|^2 = \|V - FG^T\|_F^2$, where F is a nonnegative array with (j,l)th entry α_{jj}^l and G is a nonnegative array with (k,l)th entry β_{kk}^l.

Note that if the sets $(\psi_j)_{j>1}$ or $(\phi_k)_{k>1}$ do not form complete bases, we have to consider expanding C_1^l or C_2^l using completions of them. However, for $\psi_{j'}$ one of the functions added to complete the set $(\psi_j)_{j>1}$, we will have $\eta_{j'k} = 0$ for all k, so $\alpha_{j'j'}^l = 0$. Likewise, we will have $\beta_{k'k'}^l = 0$ for $\phi_{k'}$ added to complete the set $(\phi_k)_{k>1}$, so the result does not change.

Proof of Lemma 8

Some notation: For a matrix A, denote A_{jk} as its (j,k)th entry, $A_{j\cdot}$ as its jth row, and $A_{\cdot k}$ as its kth column.

Consider an exact orthogonal NMF solution $V = V_1 + \cdots + V_d = FG^T$, where $F \geq 0$ is $P \times d$, $G \geq 0$ is $K \times d$, and $F^T F = I_d$. By Condition A we are guaranteed, for orthogonal NMF in general, that each row of F will have exactly one nonzero entry. To see this, note that when $F_{j\cdot} = 0$, for all $l \in \{1, \ldots, d\}$ row j of V_l will be all zeros. For some V_l, we could replace this row of zeros with another row of V_l multiplied by a scalar small enough to make each of its entries smaller than the corresponding entries of $V_{j\cdot}$. Doing this would give a smaller error $\|V - (V_1 + \cdots + V_d)\|_F$, and would be equivalent to changing one of the entries of $F_{j\cdot}$ to be nonzero.

Hence, we define π_j to be the column number of the nonzero entry in $F_{j\cdot}$. Using the notation of Section 4.3, $B_l = \{j : \pi_j = l\}$. Note that the B_l will be disjoint and have union $\{1, 2, \ldots, P\}$. Each B_l will be nonempty, since otherwise $F_{\cdot l}$ would be all zeros, which would correspond to a solution with a smaller d.

For any $j \in \{1, \ldots, P\}$, an exact orthogonal NMF solution will have $V_{j\cdot} = F_{j\pi_j} G^T_{\cdot\pi_j}$. Define $a_j = \|G_{\cdot\pi_j}\| F_{j\pi_j}$, and for all $l \in \{1, \ldots, d\}$, define $v_l = \frac{G_{\cdot l}}{\|G_{\cdot l}\|}$, where $\|\cdot\|$ is the vector L_2 norm. Then for any $j \in \{1, \ldots, P\}$, $V_{j\cdot} = a_j v^T_{\pi_j}$. That is, there exist unit vectors $v_1, v_2, \ldots, v_d$ of length K such that, for all $j \in \{1, \ldots, P\}$, $V_{j\cdot}$ is a scalar multiple of v_{π_j}.

Note that, when we use the smallest possible d that gives an exact orthogonal NMF solution, the unit vectors $v_1, v_2, \ldots, v_d$ are distinct. To see this, note that if we had $v_l = v_{l'}$ for some $l \neq l'$, then $G_{\cdot l'}$ would be a scalar multiple of $G_{\cdot l}$. Thus, we could move the nonzero entries of $F_{\cdot l'}$ (scaled by $\|G_{\cdot l'}\|/\|G_{\cdot l}\|$) into their corresponding rows of $F_{\cdot l}$, remove $F_{\cdot l'}$ and $G_{\cdot l'}$, and rescale to get an exact orthogonal NMF solution with a smaller d.

Suppose we had another solution $V = V'_1 + \cdots + V'_d = F'(G')^T$, where $F' \geq 0$ is $P \times d$, $G' \geq 0$ is $K \times d$, and $F'^T F' = I_d$. Define π'_j to be the column number of the nonzero entry in $F'_{j\cdot}$. Since for any $j \in \{1, \ldots, P\}$, $V_{j\cdot} = F'_{j\pi'_j}(G'_{\cdot\pi'_j})^T$, the columns of G' must be scalar multiples of the unit vectors $v_1, v_2, \ldots, v_d$. Permute the columns of F' and G' so that for all $l \in \{1, \ldots, d\}$, $G'_{\cdot l}$ is a scalar multiple of v_l. Then for all $j \in \{1, \ldots, P\}$, $V_{j\cdot}$ is a scalar multiple of both v_{π_j} and $v_{\pi'_j}$, and since the v_l are distinct this implies $\pi_j = \pi'_j$. That is, F and F' have the same supports. But by Lemma 9 (which does not depend on the results of this lemma), the orthogonal NMF solution for a fixed support is unique, so $V'_1 + \cdots + V'_d$ is the same solution as $V_1 + \cdots + V_d$ (that is, the V'_l can be reordered such that $V'_l = V_l$ for all

$l \in \{1, \ldots, d\}$). Equivalently, F and G are unique up to a column-wise permutation.

Proof of Lemma 9

Define $(V)_{B_l}$ as in Section 4.4. Define V_l to be the rank-one matrix obtained by taking the outer product of the lth columns of F and G, i.e., $V_l = F_{\cdot l} G_{\cdot l}^T$. For a fixed support, i.e., fixed sets B_l, $l = 1, \ldots, d$, V_l is restricted to having entries of 0 in all rows but those corresponding to B_l. Since $(V)_{B_l}$ also only has nonzero entries in rows corresponding to B_l, we can write

$$\|V - (V_1 + \cdots + V_d)\|_F^2 = \|\sum_{l=1}^{d}((V)_{B_l} - V_l)\|_F^2 = \sum_{l=1}^{d}\|(V)_{B_l} - V_l\|_F^2.$$

Define $(V)_{B_l}^-$ and V_l^- to be $(V)_{B_l}$ and V_l, respectively, but with their rows of zeros removed. Then minimizing the above sum is equivalent to minimizing $\|(V)_{B_l}^- - V_l^-\|_F^2$ for each l. The optimal rank-one approximation of $(V)_{B_l}^-$ (in terms of minimizing the Frobenius norm) is given by the Eckart-Young-Mirsky theorem to be $\sigma_l u_l^- v_l^T$, where σ_l is the largest singular value of $(V)_{B_l}^-$, and u_l^- and v_l are its corresponding left and right singular vectors, respectively. By Condition A, $(V)_{B_l}^-$ only has positive entries, so by Perron's theorem, u_l^- and v_l have only positive entries, σ_l is positive, and the minimizing value $\sigma_l u_l^- v_l^T$ is unique (Lax, 2007). Thus, the unique solution to orthogonal NMF with a fixed support is given by setting $V_l = \sigma_l u_l v_l^T$ for each l, where u_l is the vector of length P that has the entries of u_l^- in the rows corresponding to B_l, and zeros elsewhere. It is clear that this solution satisfies the orthogonality requirement $V_l^T V_m = 0$ for all $l \neq m$, and it is clear that σ_l, u_l, and v_l are the largest singular value of $(V)_{B_l}$ and its corresponding left and right singular vectors, respectively. To satisfy $\|F_{\cdot l}\|^2 = 1$ for each l, we must have $F_{\cdot l} = u_l$, and $G_{\cdot l} = \sigma_l v_l$.

APPENDIX B

ADDITIONAL FIGURES

Here we include additional figures referenced in Chapter 3.

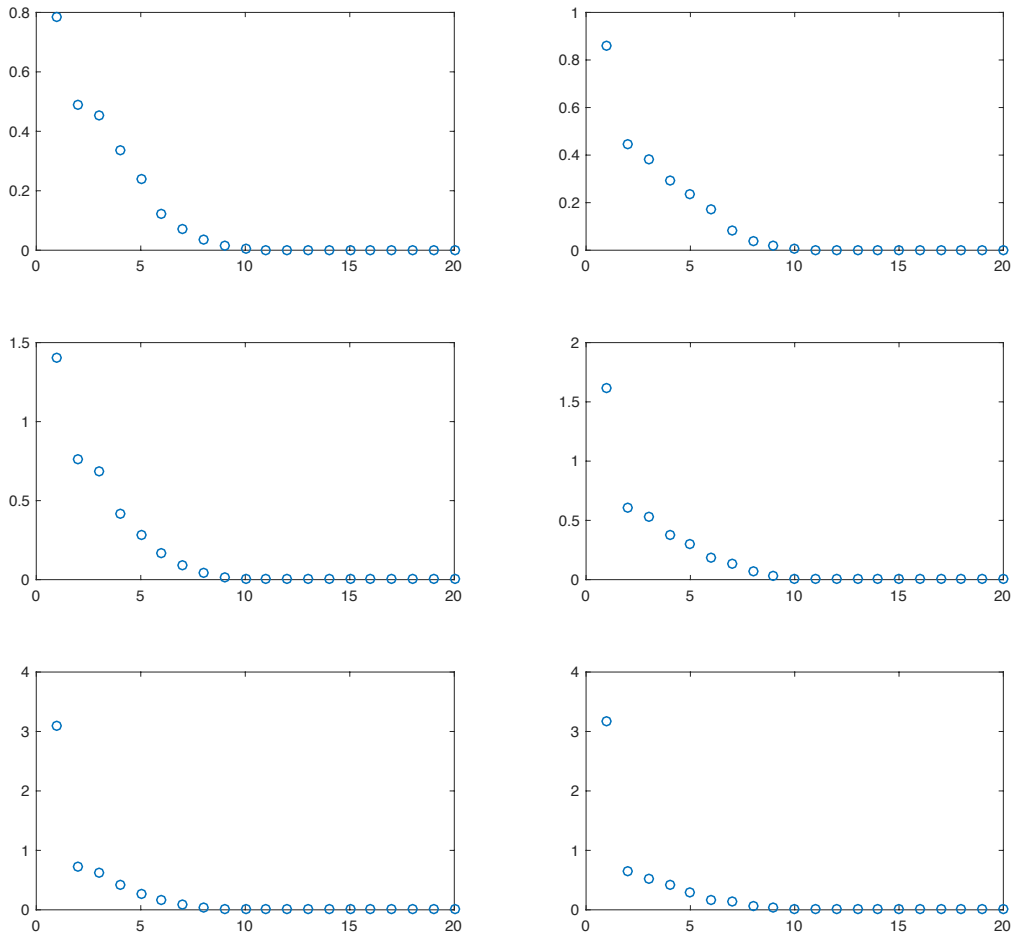

Figure 23: Plots of the first 20 estimated marginal eigenvalues $\hat{\lambda}_j$ (left column) and $\hat{\gamma}_k$ (right column) for the source-level datasets. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively.

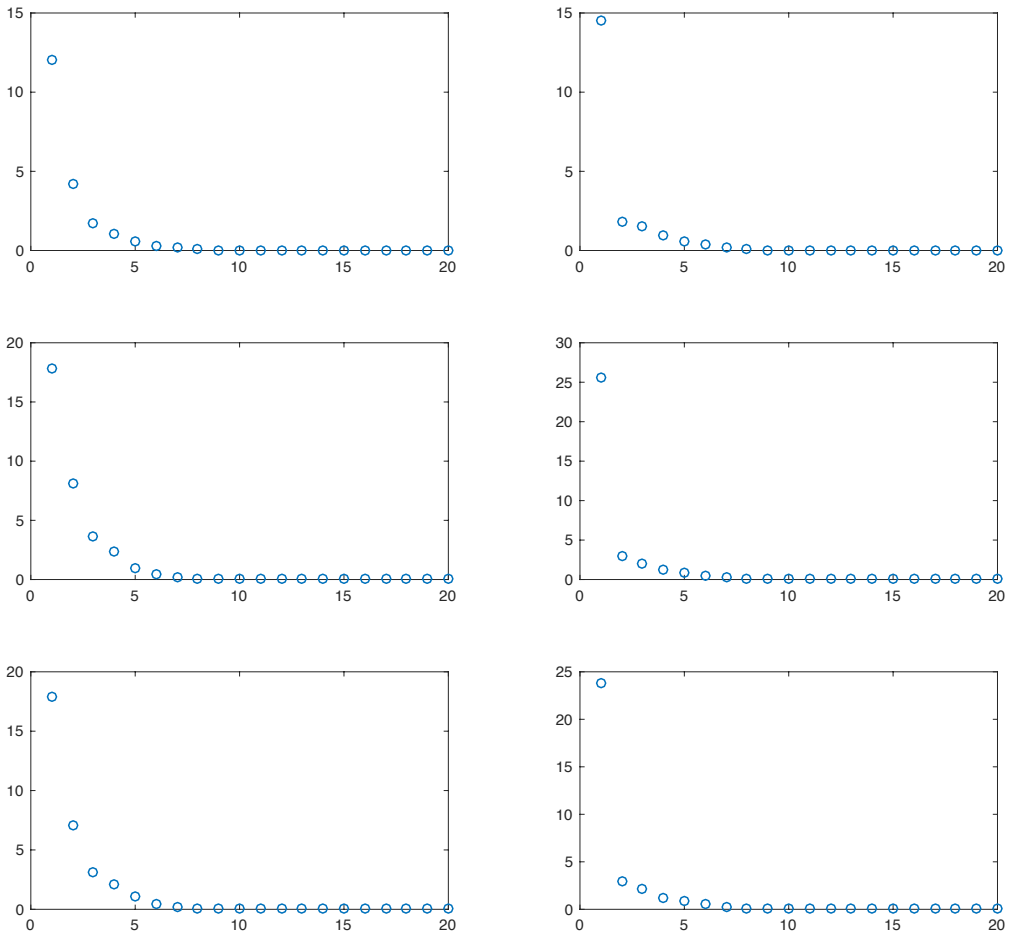

Figure 24: Plots of the first 20 estimated marginal eigenvalues $\hat{\lambda}_j$ (left column) and $\hat{\gamma}_k$ (right column) for the sensor-level datasets. The rows from top to bottom correspond to the motor, 2-back, and 0-back data, respectively.

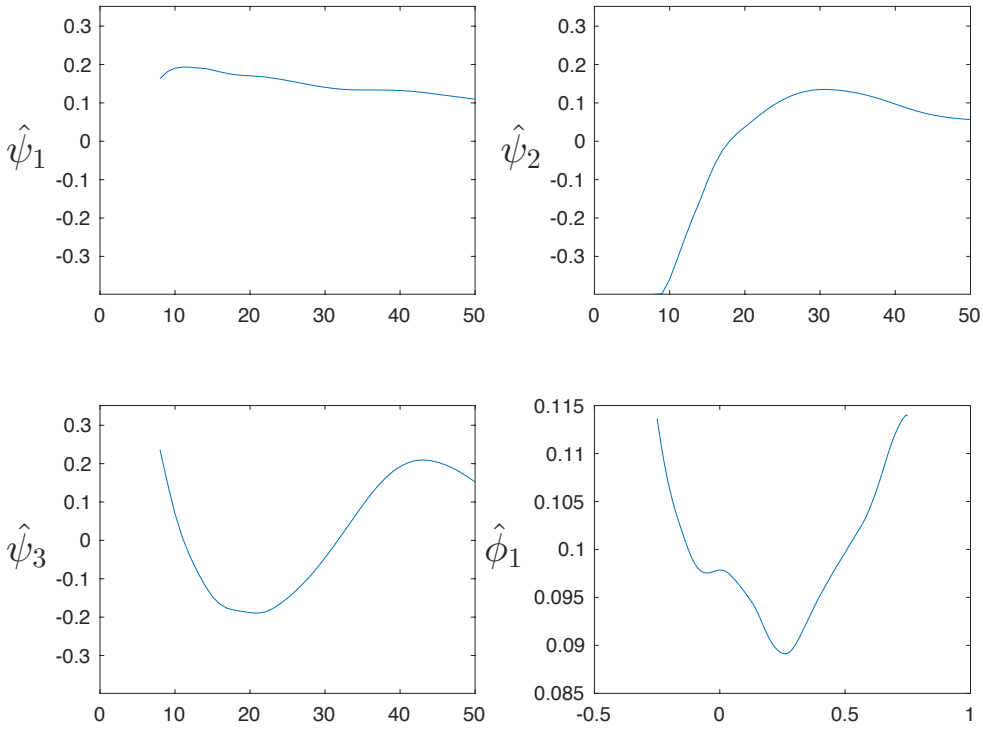

Figure 25: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the sensor-level motor data.

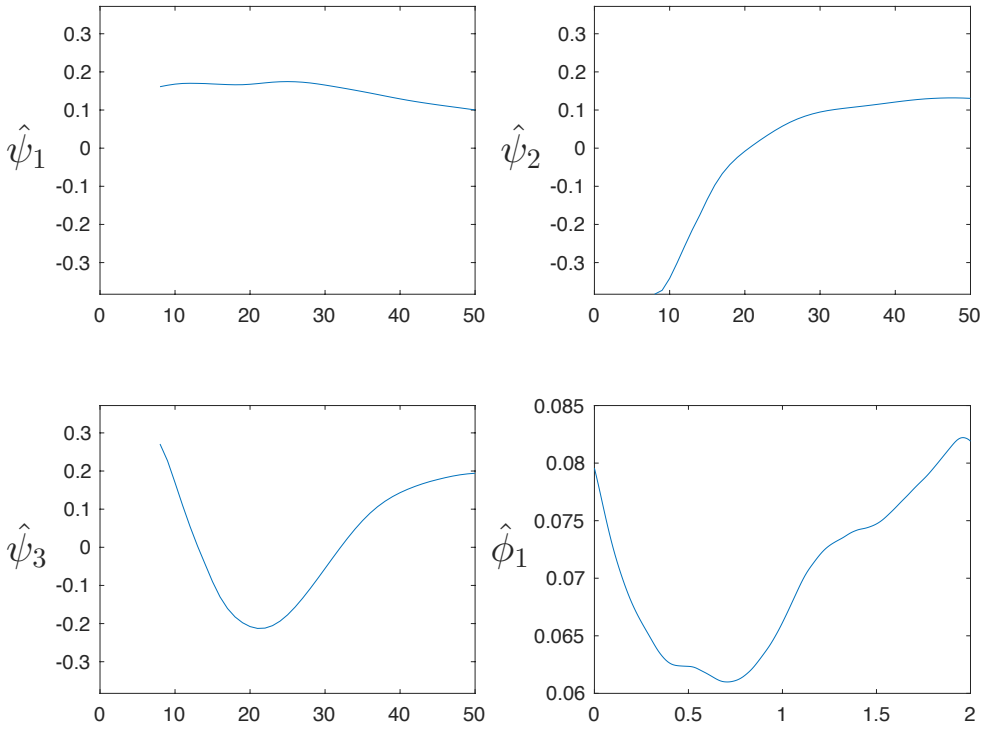

Figure 26: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the sensor-level 2-back data.

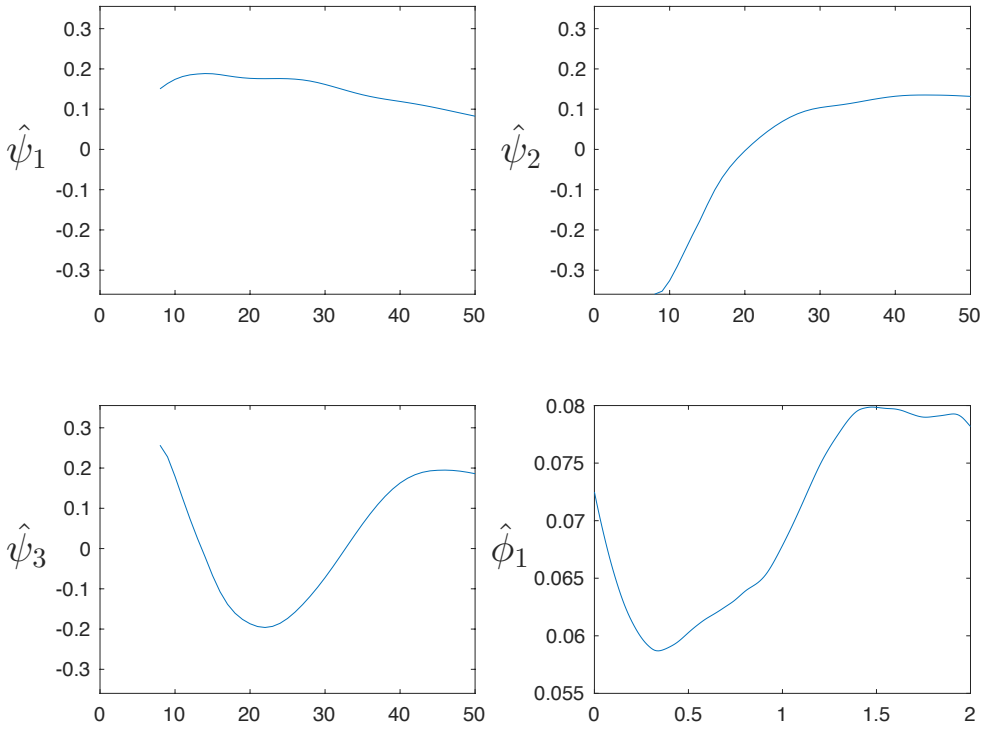

Figure 27: Plots of the estimated eigenfunctions $\hat{\psi}_j(s)$ and $\hat{\phi}_k(t)$ whose products explain the most variance for the sensor-level 0-back data.

BIBLIOGRAPHY

Allen, G. I., Grosenick, L., & Taylor, J. (2014). A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, *109*(505), 145–159.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Wiley New York, 2nd ed.

Arora, S., Ge, R., Kannan, R., & Moitra, A. (2012). Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, (pp. 145–162). ACM.

Aston, J. A., Pigoli, D., Tavakoli, S., et al. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, *45*(4), 1431–1461.

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

Bastos, A. M., & Schoffelen, J.-M. (2015). A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, *9*.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Buchta, C., Kober, M., Feinerer, I., & Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, *50*(10), 1–22.

Cai, T. T., Jiang, T., et al. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics*, *39*(3), 1496–1525.

Cao, G., Yang, L., & Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of nonparametric statistics*, *24*(2), 359–377.

Chang, J., Zhou, W., Zhou, W.-X., & Wang, L. (2017). Comparing large covariance matrices under weak conditions on the dependence structure and its application to gene clustering. *Biometrics*, *73*(1), 31–41.

Chavez, M., Valencia, M., Latora, V., & Martinerie, J. (2010). Complex networks: new trends for the analysis of brain connectivity. *International Journal of Bifurcation and Chaos*, *20*(06), 1677–1686.

Chen, K., Delicado, P., & Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(1), 177–196.

Chen, K., & Lei, J. (2015). Localized functional principal component analysis. *Journal of the American Statistical Association*, *110*, 1266–1275.

Chen, K., & Müller, H.-G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association*, *107*(500), 1599–1609.

Chen, K., Zhang, X., Petersen, A., & Müller, H.-G. (2015). Quantifying infinite-dimensional data: Functional data analysis in action. *Statistics in Biosciences*, (pp. 1–23).

Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2017). Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, *45*(4), 2309–2352.

Cohen, J. R., Gallen, C. L., Jacobs, E. G., Lee, T. G., & D'Esposito, M. (2014). Quantifying the reconfiguration of intrinsic networks during working memory. *PloS one*, *9*(9), e106636.

Constantinou, P., Kokoszka, P., & Reimherr, M. (2017). Testing separability of space-time functional processes. *Biometrika*, *104*(2), 425–437.

Cressie, N., & Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, *94*(448), 1330–1339.

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67.

Dale, A. M., & Sereno, M. I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. *Journal of cognitive neuroscience*, *5*(2), 162–176.

Degras, D. (2017). Simultaneous confidence bands for the mean of functional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *9*(3).

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*(1), 143–175.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The annals of applied statistics*, *3*(1), 458.

Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 126–135). ACM.

Dong, B., Lin, M. M., & Chu, M. T. (2014). Nonnegative rank factorization - a heuristic approach via rank reduction. *Numerical Algorithms*, *65*(2), 251–274.

Donoho, D., & Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems (NIPS)*.

Eloyan, A., Li, S., Muschelli, J., Pekar, J. J., Mostofsky, S. H., & Caffo, B. S. (2014). Analytic programming with fMRI data: A quick-start guide for statisticians using R. *PloS one*, *9*(2), e89470.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, *308*(5722), 662–667.

Fremdt, S., Steinebach, J. G., Horváth, L., & Kokoszka, P. (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, *40*(1), 138–152.

Friedman, H. R., & Goldman-Rakic, P. S. (1994). Coactivation of prefrontal cortex and inferior parietal cortex in working memory tasks revealed by 2DG functional mapping in the rhesus monkey. *Journal of Neuroscience*, *14*(5), 2775–2788.

Fuentes, M. (2006). Testing for separability of spatial–temporal covariance functions. *Journal of statistical planning and inference*, *136*(2), 447–466.

Gillis, N. (2012). Sparse and unique nonnegative matrix factorization through data preprocessing. *The Journal of Machine Learning Research*, *13*(1), 3349–3386.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, *97*(458), 590–600.

Greven, S., Crainiceanu, C., Caffo, B., & Reich, D. (2011). Longitudinal functional principal component analysis. *Recent Advances in Functional Data Analysis and Related Topics*, (pp. 149–154).

Gromenko, O., Kokoszka, P., Zhu, L., & Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *The Annals of Applied Statistics*, (pp. 669–696).

Guye, M., Parker, G. J., Symms, M., Boulby, P., Wheeler-Kingshott, C. A., Salek-Haddadi, A., Barker, G. J., & Duncan, J. S. (2003). Combined functional MRI and tractography to demonstrate the connectivity of the human primary motor cortex in vivo. *Neuroimage*, *19*(4), 1349–1360.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical and biological engineering and computing*, *32*(1), 35–42.

Hoff, P. D., et al. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, *6*(2), 179–196.

Hsing, T., & Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

Huang, J. Z., Shen, H., & Buja, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, *104*(488), 1609–1620.

Hung, H., Wu, P.-S., Tu, I., Huang, S.-Y., et al. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, *99*(3), 569–583.

Hyndman, R. J., & Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, *38*(3), 199–211.

Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942–4956.

Jensen, O., & Hesse, C. (2010). Estimating distributed representations of evoked responses and oscillatory brain activity. *MEG Introd. Methods Hansen PC Kringelbach ML Salmelin R Eds*, (pp. 156–185).

LaBar, K. S., Gitelman, D. R., Parrish, T. B., & Mesulam, M.-M. (1999). Neuroanatomic overlap of working memory and spatial attention networks: a functional MRI comparison within subjects. *Neuroimage*, *10*(6), 695–704.

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

Lan, W., Luo, R., Tsai, C.-L., Wang, H., & Yang, Y. (2015). Testing the diagonality of a large covariance matrix in a regression setting. *Journal of Business & Economic Statistics*, *33*(1), 76–86.

Landi, S. M., Baguear, F., & Della-Maggiore, V. (2011). One week of motor adaptation induces structural changes in primary motor cortex that predict long-term memory one year later. *Journal of Neuroscience*, *31*(33), 11808–11813.

Larson-Prior, L. J., Oostenveld, R., Della Penna, S., Michalareas, G., Prior, F., Babajani-Feremi, A., Schoffelen, J.-M., Marzetti, L., de Pasquale, F., Di Pompeo, F., et al. (2013). Adding dynamics to the Human Connectome Project with MEG. *Neuroimage*, *80*, 190–201.

Lax, P. (2007). *Linear Algebra and Its Applications*. No. v. 10 in Linear algebra and its applications. Wiley.

Le Van Quyen, M., Foucher, J., Lachaux, J.-P., Rodriguez, E., Lutz, A., Martinerie, J., & Varela, F. J. (2001). Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *Journal of neuroscience methods*, *111*(2), 83–98.

Ledoit, O., & Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *The Annals of Statistics*, *30*(4), 1081–1102.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, (pp. 556–562).

Levy, R., & Goldman-Rakic, P. S. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. In *Executive control and the frontal lobe: Current issues*, (pp. 23–32). Springer.

Lin, F.-H., Witzel, T., Hämäläinen, M. S., Dale, A. M., Belliveau, J. W., & Stufflebeam, S. M. (2004). Spectral spatiotemporal imaging of cortical oscillations and interactions in the human brain. *Neuroimage*, *23*(2), 582–595.

Lindquist, M. A., et al. (2008). The statistical analysis of fMRI data. *Statistical Science*, *23*(4), 439–464.

Liu, W.-D., Lin, Z., Shao, Q.-M., et al. (2008). The asymptotic distribution and Berry–Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability*, *18*(6), 2337–2366.

Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *Neural Networks, IEEE Transactions on*, *19*(1), 18–39.

Lu, N., & Zimmerman, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters*, *73*(4), 449–457.

Mars, R. B., & Grol, M. J. (2007). Dorsolateral prefrontal cortex, working memory, and prospective coding for action. *The Journal of neuroscience*, *27*(8), 1801–1802.

Mas, A. (2006). A sufficient condition for the CLT in the space of nuclear operators—Application to covariance of random functions. *Statistics & probability letters*, *76*(14), 1503–1509.

Mattingley, J. B., Husain, M., Rorden, C., Kennard, C., & Driver, J. (1998). Motor role of human inferior parietal lobe revealed in unilateral neglect patients. *Nature*, *392*(6672), 179–182.

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 179–199.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, *32*, 223–240.

Muschelli, J. (2017). *hcp: Human Connectome Project*. R package version 0.5. URL https://db.humanconnectome.org

Nerini, D., Monestiez, P., & Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis*, *101*(2), 409–418.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, *2011*.

Ou, W., Hämäläinen, M. S., & Golland, P. (2009). A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, *44*(3), 932–946.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, *25*(1), 46–59.

Pizzella, V., Marzetti, L., Della Penna, S., de Pasquale, F., Zappasodi, F., & Romani, G. L. (2014). Magnetoencephalography in the study of brain dynamics. *Functional neurology*, *29*(4), 241.

Quiroga, R. Q., Kraskov, A., Kreuz, T., & Grassberger, P. (2002). Performance of different synchronization measures in real data: a case study on electroencephalographic signals. *Physical Review E*, *65*(4), 041903.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer, second ed.

Srivastava, M. S., von Rosen, T., & von Rosen, D. (2009). Estimation and testing in general multivariate linear models with Kronecker product covariance structure. *Sankhyā: The Indian Journal of Statistics, Series A*, *71*, 137–163.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*, *100*(469), 310–321.

Tavakoli, S. (2016). *covsep: Tests for determining if the covariance structure of 2-dimensional data is separable*. R package version 1.0.0.
URL https://CRAN.R-project.org/package=covsep

Van Den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534.

Van Der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. Springer Verlag.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, *80*, 62–79.

Vandaele, A., Gillis, N., Glineur, F., & Tuyttens, D. (2015). Heuristics for exact nonnegative matrix factorization. *Journal of Global Optimization*, (pp. 1–32).

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., & Vachon, P. (2007). Methods protocol for the human mortality database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality.org [version 31/05/2007]*, *9*, 10–11.

WU-Minn HCP (2017). *1200 subjects data release reference manual*.
URL https://www.humanconnectome.org

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, *61*(1-3), 167–191.

Zhang, D., & Zhou, Z.-H. (2005). (2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, *69*(1), 224–231.

Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC Press.