Title
Hierarchical Integration of Heterogeneous Highly Structured Data: The Case of Functional Brain Imaging

Author
Li, Qian

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Hierarchical Integration of Heterogeneous Highly Structured Data: The Case of Functional Brain Imaging

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Qian Li

2018

ABSTRACT OF THE DISSERTATION

Hierarchical Integration of Heterogeneous
Highly Structured Data:
The Case of Functional Brain Imaging

by

Qian Li
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2018
Professor Donatello Telesca, Chair

Functional brain imaging technologies produce high dimensional data with structured dependency spanning along multiple dimensions. This dissertation focuses on the specific case of Electroencephalography (EEG), even though most methodological developments are applicable to other imaging modalities. The overarching goal is to provide methodological foundations to inferential problems involving population inference in the setting of cognitive experiments. Specifically, I address important challenges associated with the highly heterogenous nature of brain imaging measurements, by reframing complex inferential questions in the context of familiar analytical techniques involving regression, clustering, functional and longitudinal data analysis. These contributions focus on spatio-temporal modeling of EEG measurements, which characterize both intra- and inter-subjects variation within the contexts of neuronal synchronicity and differential band power dynamics. The methodological developments are used to provide analytical insight in several neuro-cognitive studies based on EEG data.

The dissertation of Qian Li is approved.

Shafali Spurling Jeste

Catherine Ann Sugar

Damla Şentürk

Donatello Telesca, Committee Chair

University of California, Los Angeles

2018

To my beloved parents

it has been a fun adventure

with you by my side

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I want to express my sincere gratitude to my advisor, Professor Donatello Telesca. You have been supportive more than academically since the first day I started working with you. I would like to thank you for the brilliant ideas you suggested in our research, and more importantly, for being a tremendous mentor to me along the journey. Your attention to quality of work has been my inspiration and will guide me on my career as a research scientist.

I would also like to thank my committee members, professor Damla Şentürk, professor Catherine A. Sugar and professor Shafali S. Jeste for letting me defend on such short notice and your terrific inputs to my projects and publications. Chapter Two is a version of [LSS16] and Chapter Three is in preparation for publication, both have received valuable inputs from all my committee members. Besides, Joel Frohlich and Charlotte DiStefano have been kindly providing their EEG data for my dissertation, and many thanks to you too. Aaron Scheffler, my awesome lab mate and friend, you have been a motivation to me with all your dedication and consistent efforts you put into research. It has been our constant discussion and feedback to each other that makes this journey enjoyable and fruitful.

I gratefully acknowledge the funding received toward my PhD study and this dissertation. The work in Chapter Two and Three was supported by National Institute of General Medical Sciences under the grant R01 GM111378-01A1. I am also grateful for a UCLA Dissertation Year Fellowship (2017-2018) that supported my final year of writing the thesis. My special thanks go out to Mr. Chuanhong Li and Mrs. Lina Li, for believing in my efforts and for the financial support I received through the Weige Foundation Scholarship.

A special thanks to my dear friends at UCLA because of whom I have enjoyed every single minute of the past few years.

Finally, I would like to thank my parents for everything. For Dad who has been a role model to me and your silent love. For Mom who has been supportive to me in all my pursuits. Thank you.

VITA

2008–2012 B.S. (Statistics), Peking University, Beijing, China.

2012–2014 M.S. (Biostatistics), UCLA, Los Angeles, California.

2013–2017 Teaching Assistant, Biostatistics Department, UCLA. Taught labs and discussions of BIOS110A, 110B, 100A, 100B for multiple quarters

2013–2017 Graduate Student Researcher, Biostatistics Department, UCLA.

2016–2017 Consultant, Institute for Digital Research and Education (IDRE), UCLA

PUBLICATIONS

Li, Q., Senturk, D., Sugar, C.A., Jeste, S., DiStefano, C., Frohlich, J. and Telesca, D., 2016. Inferring Brain Signals Synchronicity from a Sample of EEG Readings. (*to appear in JASA*)

Aziz, N., Detels, R., Quint, J.J., **Li, Q.**, Gjertson, D. and Butch, A.W., 2016. Stability of cytokines, chemokines and soluble activation markers in unprocessed blood stored under different conditions. *Cytokine*, 84, pp.17-24.

Li, Z., Tseng, C.H., **Li, Q.**, Deng, M.L., Wang, M. and Heber, D., 2014. Clinical efficacy of a medically supervised outpatient high-protein, low-calorie diet program is equivalent in prediabetic, diabetic and normoglycemic obese patients. *Nutrition & diabetes*, 4(2), p.e105.

CHAPTER 1

Brain Imaging Techniques and Statistical Methods for Integrative Inference

1.1 Introduction

Functional brain imaging (neuro-imaging) techniques have been widely investigated and discussed over the recent decades. They enable direct visualization of the brain, highlighting the underlying metabolic process, through typically non-invasive imaging protocols. For this reason, the technology is readily available for implementation and adaption to a variety of scientific investigation. Despite its attractiveness, however, the high dimensionality and the complex dependency intrinsic to the data, pose substantial challenges to both quantitative modeling and inferential techniques. Therefore, a tremendous amount of effort has been dedicated to these problems, and it is the main focus of my dissertation.

Another fundamental challenge from brain imaging techniques is that, measurements are inherently heterogeneous. This heterogeneity could be induced by trial-specific acquisition protocols, equipment setups [NCT12], but more commonly, idiosyncratic response from the measured participant [BDT15]. On the other hand, it is intuitively and statistically appealing to rely on readings from more than one individuals in order to highlight recurrent patterns and group characteristics in brain activities. As a consequence, integrating individual findings on a small to moderate sample of EEG measurements, that are intrinsically noisy and unpredictable, into statistically sound group level inference has been central to many methodological developments (see [HMS07] and [NA16]).

My dissertation focuses on the specific case of Electroencephalography (EEG), even

though most methodological developments are applicable to other imaging modalities, for example functional magnetic resonance imaging (fMRI). The overarching goal is to provide methodological foundations to inferential problems involving population inference in the setting of cognitive experiments. Specifically, I address important challenges associated with the highly heterogenous nature of brain imaging measurements, by reframing complex inferential questions in the context of familiar analytical techniques involving regression, clustering, functional and longitudinal data analysis. These contributions focus on spatio-temporal modeling of EEG data which characterizes both intra- and inter-subject variation within the contexts of neuronal synchronicity and differential band power dynamics. The ensuing methodological developments are used to provide analytical insight into several neuro-cognitive studies based on EEG data.

To lay out the background of related work, I review EEG techniques and analytical methods for EEG studies in Section 1.2.1 and 1.2.2; various approaches aimed at the estimation of power spectral densities are briefly discussed in Section 1.2.3; clustering methods as exploratory approaches are reviewed in Section 1.3. Group-level inference via hierarchical models in cluster analysis is discussed in Section 1.3.3, and mixed effects regression models in Section 1.4. Following a brief introduction, the application of these techniques in the context of EEG experiments, inference for neuronal synchronicity is discussed in Chapter 2, and inference for regionally referenced power bands dynamics is discussed in Chapter 3.

1.2 Electroencephalography (EEG) Data Analysis

Electroencephalography (EEG) captures the potential field and records the difference as electrical activity on human scalp, non-invasively through sensors that are essentially electrodes. This technique has been extensively used and studied for both clinical and research purposes since the first published work by Dr. Hans Berger in 1929 [Ber29]. More recently, EEG has gained increasing attention on account of multi-channel EEG recordings and fine-grained digitization, which in return has broadened its applications and benefitted scientific investigations that aim at: substrate of neuronal diseases, including Alzheimer's [HPS09],

epilepsy [ABS08] and Autism Spectrum Disorder (ASD) [BDW07]; functional and physiological investigation under the states of resting [BSB11] and attention [BCJ03]; and monitoring neurological injury, in the event of stroke [CBS76], trauma [TWG89] and coma [You00].

By comparison, EEG has significant advantages over the alternative neuroimaging modalities, for example functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), because of its high temporal resolution, allowing for stimuli to be registered precisely to their onset over the course of experiment. In response to the criticism of a lack of spatial resolution, to locate the electrophysiological source accurately, EEG has evolved from a single channel monitoring device to high-density geodesic net that captures the electrical fluctuations simultaneously at multiple locations, therefore gained in popularity for research purposes. This evolvement has dramatically enriched the volume of information at micro level, therefore precise inference became feasible longitudinally and spatially. On the other hand, it highlights the necessity of well engineered analytical metrics, known as quantitative EEG (qEEG) methods, and motivates the development of automatic and scalable procedures to analyze a massive amount of waveforms from EEG experiments. This section selectively reviews some relevant concepts and work. For a thorough review of the recent development in qEEG methodology, see [TT04].

1.2.1 EEG Measurements and Pre-processing

EEG non-invasively collects electrophysiological signals on the human scalp. More precisely, it measures the potential difference between a pair of electrodes: one placed at the point of interest, and the other typically placed on the top of head as a reference. Its measurements are expected to display the action potential induced by cortical neurons, that is either excitatory or inhibitory, corresponding to stimulus-based and resting-state characterization of physiological activities, in the targeted pyramidal layer of the cortex [MH03]. Fundamentally at the cellular level, neuronal activity causes electrical fluctuation that is projected to the surface of the scalp and further captured by EEG. Depending on the underlying process, it is possible to observe locally spiked signal as activation and remotely associated cooccurrence

as connectivity. Regardless of its applications, the observed EEG signal is hypothetically induced by the neuronal activity perpendicular to the cortical patch where the first electrode is placed upon, relative to its reference point.

In reality, however, other electrophysiological signals, besides the neuronal activity of interest, also appear in EEG recordings therefore need to be properly removed. These electrophysiological "noises", often induced by physical movements, muscle activities even heart beats, are often referred to as artifacts, and typically possess distinctive features in their waveforms. The pre-processing of EEG signals aims to identify and isolate artifacts by these features, and it has become common standard to all analytic EEG studies (see [JGK04]). In general, a typical pipeline of EEG preprocessing operates on the segmented recordings, by marking and replacing the suspicious channels at first, followed by the rejection of contaminated segments. The rejection procedure routinely starts with a manual rejection, which visually inspects for non-stereotyped artifacts of certain waveforms, then proceeds with an automated (algorithmic) rejection with combined *Independent Component Analysis* (ICA) and *Principal Component Analysis* (PCA). Ideally, the artifacts are separated out and rejected as non-signal components, without sacrificing the whole segment [DSM07].

On the other hand, preprocessing procedures complicate the structure of data. The first complication is that rejection leads to incompleteness, and significantly reduces the amount of available data and consequently the power of statistical inference. For example, when EEGs are recorded on a cohort of children diagnosed with Autism Spectrum Disorder (ASD), greater amount of rejections are noticed which can be partially attributed to inattention and frequent physical movements, relative to a control group of Typically Developing (TD) children. Besides, segments are normally re-referenced after segment-wise processing, therefore as a consequence, even adjacent segments are essentially discontinued over the course of the experimental time; this discontinuity impedes the application of statistical methods that operates on lengthy continuous time series. The effective unit admissible to statistical modeling has been shortened to segmented pieces that typically span 1 second in length.

In response, multiple measures have been taken to amplify the Signal-to-Noise Ratio

(SNR) of EEG recordings. Previous work can be roughly categorized into one among the following four:

- Elimination of noise sources;
- Rejection of noisy data;
- Removal of noise;
- Signal averaging.

Statistically speaking, all measures are often considered necessary, except for signal averaging since it often lead to loss of information with good intentions (see [HBM13]). Careful examination reveals that this compromise has abused the abundance of information along both longitudinal and spatial dimensions, where signals appear correlated instead of purely duplicated. This observation motivates a key objective of this dissertation, i.e. to structurally model multi-way dependent data.

1.2.2 Review of Quantitative Methods for EEG Data

Typical EEG studies extract quantitative features from: (a) time domain; (b) frequency domain; and (c) time-frequency representations. Most of methods in (a) and (b) fall into the *linear* category, as they hypothesizes stationarity of the EEG recordings and approach the problem linearly under a generalized regression framework. By contrast, (c) is more liberal in the sense that it aims at a dynamic representation of a potentially non-stationary process. In this context, the definition of time-varying spectra can capture more transient features, as well as longitudinal effects, and is often categorized as *non-linear* approach. A thorough review is provided by [TT04].

Frequency domain methods usually build upon power spectral analysis of each segment. The calculated power, in absolute or relative sense, is aggregated by five commonly used frequency bands: Delta (0-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz) and Gamma (30-50 Hz). Depending on the purpose of the study, absolute power is essentially

the raw estimate of power spectral density (PSD), whereas relative power is the normalized version that has the total variance adjusted in the units of segments.

Segmented recordings have a span of only a few seconds. Therefore, power estimates are highly volatile and potentially unreliable for use in influence. One practical way to stabilize the spectral estimate is to rely on averaging (smoothing). To be specific, the estimated powers are averaged over electrodes, within regions of interest (ROI), further over segments, with hypothetic stationarity spatially (i.e. within ROI) and temporally (i.e. across segments). This hypothesis is inappropriate in the sense that: (a) a scalp parcellation should be data-dependent rather than predefined by anatomy; (b) stationarity can be relaxed to locally such that a stable estimate of spectral power is available without compromising the spectral dynamics. Consequently, a Bayesian clustering method is motivated in Chapter 2 in accordance with these two observations.

Group-level inference can also improve on spectral differentiation, by relying on recordings from multiple subjects. Simple averaging scheme within group, however, overlook the heterogeneity of the measurements, especially in some pathological conditions, and often leads to attenuated or even inappropriate findings (see an example in [KJJ13]). Furthermore, inference at group level is still underdeveloped in the context of functional connectivity using EEG measurements. In this regard, we first introduce an integrative approach in Section 1.3.3, from a clustering perspective that is applicable to resting-state EEGs, then discuss Mixed Effects Models and their functional variants extendable to non-stationary time series from stimulus-based EEG experiments.

1.2.3 Review of Spectral Estimation

The spectral analysis of a time series can be conceptualized from two perspectives, strength of multiple periodic components and distribution of variance on the frequency domain. It can be shown that they are equivalent in principle, but often referred to as *periodogram* and *correlogram* in practice. More precisely, for a time series $x(t)$, $t = 1, \ldots, n$, the discrete

Fourier transform (DFT) is defined as

$$d(\omega_j) = n^{1/2} \sum_{t=1}^{n} x(t) \exp(-2\pi i \omega_j t) \tag{1.1}$$

for $j = 0, 1, \ldots, n-1$, where the frequencies $\omega_j = j/n$ are Fourier or fundamental frequencies.

The periodogram is defined as the squared modulus of the DFT,

$$\begin{aligned} I(\omega_j) = |d(\omega_j)|^2 &= n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} x(t)x(s) \exp(-2\pi i \omega_j (t-s)) \\ &= n^{-1} \sum_{h=-(n-1)}^{(n-1)} \sum_{t=1}^{n-|h|} [x(t+|h|) - \overline{x}(t)] [x(t) - \overline{x}(s)] \exp(-2\pi i \omega_j h) \qquad (1.2) \\ &= \sum_{h=-(n-1)}^{(n-1)} \widehat{\gamma}(h) \exp(-2\pi i \omega_j h) \qquad (1.3) \end{aligned}$$

where $\widehat{\gamma}(h)$ is the sample auto-covariance function (ACF),

$$\widehat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} [x(t+h) - \overline{x}(t)] [x(t) - \overline{x}(t)] \tag{1.4}$$

In (1.2), the mean $\overline{x}(t)$ is subtracted from $x(t)$ under the assumption that $x(t)$ is a zero-mean stationary time series. Then correlogram defined by (1.3) is essentially the DFT of the auto-covariance $\widehat{\gamma}(h)$.

Raw periodograms defined by either (1.1) or (1.3) are unbiased estimators of the spectral density $f(\omega)$. Their implementation is made efficient, by using the fast Fourier transform algorithm (FFT). The FFT has the most computational advantages when n is highly composite, for example $n = 2^p$ is a factor of 2 with order p. However, this estimator is inconsistent. As discussed by [SS10], $I(\omega_j)$ is the sum of squares of only two random variables for any sample size, which is problematic especially for lower frequencies. In practice, smoothed periodograms are preferred and set as default for common computation packages.

Smoothed Periodogram Estimator

Given a raw periodogram $I(\omega_j)$, the smoothed periodogram convolves with a kernel function h on the frequency domain,

$$\widehat{f}(\omega) = \sum_{k=-m}^{m} h_k I(\omega_j + k/n), \tag{1.5}$$

using weights $h_k > 0$ that satisfy $\sum_{-m}^{m} h_k = 1$ and spans over $2m+1$ values in the neighboring frequency of ω. Popular choices of h_k include: the Daniell kernel, corresponding to a moving average kernel of rectangular shape over $2m + 1$ points; the modified Daniell kernel which yields smoother estimates by assigning half weights at both end points, and it is often applied recursively by a convolution that leads to further smoothness in practice.

Lag Window Estimator

Alternatively, smoothness can also be achieved by operating on the ACF in (1.4). The so called lag window estimator is defined as,

$$\tilde{f}(\omega) = \sum_{|h|\leq r} w(h/r)\widehat{\gamma}(h) \exp(-2\pi i\omega h), \tag{1.6}$$

where the weight function $w(\cdot)$ satisfies: a) $w(0) = 1$; b) $|w(x)| \leq 1$ and $w(x) = 1$ for $|x| > 1$; and c) $w(x) = w(-x)$.

The smoothness is now controlled by both a lag window smoother, and truncation of the ACF with r lags at maximum. As suggested by [EOO15], the Parzen window works well for EEG applications, and the truncation parameter r can be selected based on cross-validation procedures (see [ORS01]).

Principal Components in the Frequency Domain

All univariate methods centered around the spectral density matrix $f(\omega)$ can be generalized to multivariate cases. These topics are rigorously covered in [SS10] (Chapter 7) and [Bri01] (Chapters 9 and 10) and only brief reviews are presented in this section.

For a pair of time series $x_j(t)$ and $x_k(t)$ defined on the same grid of time points, (1.4) can be generalized from an auto-covariance to a cross-covariance function $\widehat{\gamma}_{jk}(h)$. Therefore more generally, for a p-variate time series, $\boldsymbol{x}(t)$, one can define the $p \times p$ auto-covariance matrix,

$$\Gamma(h) = E(\boldsymbol{x}(t+h)\boldsymbol{x}(t)') - E(\boldsymbol{x}(t+h))E(\boldsymbol{x}(t)). \tag{1.7}$$

Therefore, similar to (1.6) with truncation for desirable smothness, the spectral density

matrix of the multivariate series $\boldsymbol{x}(t)$ is given by,

$$\widehat{F}(\omega) = \sum_{|h|\leq r} w(h/r)\widehat{\Gamma}(h)\exp(-2\pi i\omega h). \tag{1.8}$$

Notice that smoothing is performed universally across all frequency, such that $\widehat{F}(\omega)$ is guaranteed to be a complex, nonnegative-definite, Hermitian matrix. This is crucial for the principal component analysis. Similar to its classical definition, wee seek a linear combination of the multivariate time series at a given frequency ω, $y_t(\omega) = \boldsymbol{a}(\omega)'\boldsymbol{x}(t) = a_1(\omega)x_1(t) + \cdots + a_p(\omega)x_p(t)$, such that $\mathrm{var}(y_t(\omega))$ is maximized. Equivalently, we seek a complex vector $\boldsymbol{a}(\omega)$ such that

$$\max_{\boldsymbol{a}(\omega)\neq\boldsymbol{0}} \frac{\boldsymbol{a}(\omega)^*\widehat{F}(\omega)\boldsymbol{a}(\omega)}{\boldsymbol{a}(\omega)^*\boldsymbol{a}(\omega)} = \lambda_1(\omega),$$

if we denote $\{(\lambda_1(\omega), \boldsymbol{e}_1(\omega)), \ldots, (\lambda_p(\omega), \boldsymbol{e}_p(\omega))\}$ the eigenvalue and eigenvector pairs of $\widehat{F}(\omega)$, and $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_p(\omega) \geq 0$. Then at frequency ω, we consider the eigenvalue λ_1 as the first eigen-power that summarizes the multivariate time series $\boldsymbol{x}(t)$ as the leading principal component. Later on, we introduce power band calculation based on this characterization and how it connects with simple averaging commonly seen in practice in Chapter 3.

1.3 Clustering Methods

Clustering is well recognized as an exploratory method for data analysis. It usually aims at grouping similar objects together, and/or separating dissimilar objects apart. These two entangled objectives motivate two classes of methods, namely *bottom-up* and *top-down*. Top-down procedures are mostly segmentation-based methods, for example K-means, which iteratively finds the center and a partition based on the current guess of each other. On the other hand, bottom-up methods approach the problem reversely in an agglomerative manner, by which it recursively merges a pair of pre-existed clusters that are closest by certain distance metrics. The commonly used metrics, also known as linkage functions, include average, single and complete linkage. A comprehensive review of recent development in clustering method, please refer to [XT15].

In the case of functional brain imaging, clustering is scientifically appealing to conceptualize EEG electrode, or channels of EEG recordings, as objects, therefore a cluster identifies a collection of electrodes that hopefully locate a patch on scalp exhibiting certain degrees of similarity in its electrical activity. This idea has been investigated from a discriminative perspective by [Kra16] with known classes, and [CCP06] first suggested hierarchically cluster multi-variate time series using their spectral densities. As pointed out by both work, it is fundamental to first construct the similarity/dissimilarity measures which is discussed in Section 1.3.1. A hierarchical approach applicable to functional imaging data is reviewed in Section 1.3.2, followed by a model-based approach that integrates evidence for group inferences in Section 1.3.3.

1.3.1 Proximity Measures on EEG

For the ease of notation, we only consider a proximity measure defined between a pair of time series, $x(t)$ and $y(t)$. Supposedly a dissimilarity measure $d(x, y)$ is well defined for an arbitrary pair of electrodes, it can be readily extended to a $p \times p$ dissimilarity matrix $D(\cdot, \cdot)$ featuring $p(p-1)/2$ distinct pairwise distences.

Cross-Correlation

Cross-correlation between time series $x(t)$ and $y(t)$ at lag h, $\rho_{x,y}(h)$, is defined as,

$$\rho_{x,y}(h) = \frac{\mathrm{Cov}[x(t), y(t+h)]}{\sqrt{\mathrm{Var}[x(t)]\mathrm{Var}[y(t+h)]}}.$$

It summarizes the linear coupling relationship of two time series. When $h = 0$, a positive correlation close to 1 indicates two electrodes are simultaneously activated. Significant correlation with lag h can also be interpreted as delayed response inter-regionally. The numerator of cross-correlation is called cross-covariance function, which can be regarded as an extension to ACF. This also motivates a frequency-domain summary, called cross-coherence.

Cross-Coherence

Similar to the Fourier transform pair of ACF and spectral density, the cross-covariance function can be decomposed spectrally as cross-spectral density. Suppose that $f_{x,y}(\omega)$ is the DFT of $\mathrm{Cov}(x, y)$ at frequency ω, then coherence $\mathrm{Coh}_{x,y}(\omega)$ is defined as,

$$\mathrm{Coh}_{x,y}(\omega) = \frac{|f_{x,y}(\omega)|^2}{f_x(\omega) f_y(\omega)}.$$

Coherence measure implicitly accounts for lags between time series, and its squared value can be interpreted as the proportion of the power in one time series, which can be explained by a linear regression to the other time series [ZTS09]. It is worth noticing that the cross-coherence measure also relates to the spectral density matrix defined in (1.8), when normalized with diagonal elements being 1. Therefore the purpose of spectral PCA can be interpreted as a recombination of multivariate signals into zero coherence components, therefore it can be used as a summarizing metric referenced to local regions in Chapter 3.

Total Variation Distance

Assume that $f_x(\omega)$ and $f_y(\omega)$ are the power spectral densities for $x(t)$ and $y(t)$ respectively, and they are integrable (i.e. $\int f_x(\omega)\,\mathrm{d}\omega < \infty$, $\int f_y(\omega)\,\mathrm{d}\omega < \infty$). Then the *Total Variation Distance* (TVD) is defined between the normalized density of $f_x(\omega)$ and $f_y(\omega)$,

$$\mathrm{TVD}_{x,y} = 1 - \int \min\left\{ \frac{f_x(\omega)}{\int f_x(\omega')\,\mathrm{d}\omega'}, \frac{f_y(\omega)}{\int f_y(\omega'')\,\mathrm{d}\omega''} \right\} \mathrm{d}\omega.$$

TVD differs from coherence measure in the sense that it measures the affinity of the distributions of power purely over frequency domain. Recall that coherence is a function of frequency ω too, however, its definition is rooted on covariance function between two time series therefore regarded as linear predictability in the time domain. [EOO15] interprets this measure as spectral synchronicity, referring to the highly synchronized activity of the neurons that can be reflected by EEG signals. They also provide examples with negligible coherence, but their TVD is close to zero (highly synchronized).

1.3.2 Hierarchical Clustering

Based on the distance metric defined as TVD, [EOO15] suggested an agglomerative procedure called Hierarchical Merger Algorithm. Unlike typical agglomerative approaches with single-linkage or complete-linkage [XW05], they take advantage of the fact that the center spectra is representative of the cluster, which is also a valid spectral density. Therein, TVD can effectively link clusters by a recursive update of the distance matrix between the newly merged group and the rest.

To be more specific, the algorithm starts with p clusters of p electrodes by themselves, then TVD is calculated between every pair of spectra. The pair with lowest TVD is merged and replaced by their mean spectra. Eventually the algorithm iterates until all electrodes converge to a single cluster.

We find this approach enlightening for the reasons that, it summarizes spectral differentiation into a simple metric that is interpretable and scaled well (TVD $\in [0, 1]$), and it allows for a time-frequency representation that captures the dynamics of spectral evolution over time. However, this approach defines a partition tree deterministically without sufficient addresses on the uncertainty, also known as noise, in spectral estimation. At the same time, it doesn't integrate findings over repeated measured units, for example, segments even multiple individuals. To fully take into account these issues, a viable approach is introduced in Section 1.3.3 and a novel method that combines the benefits of both methods is finally explicated in Chapter 2.

1.3.3 Bayesian Consensus Clustering

The Bayesian Consensus Clustering (BCC), introduced by [LD13], aims at an integration of multi-source data for clustering. Two general approaches exist in this regard:

1. Separate clustering for each source, followed by a post-hoc integration.

2. Combining all data sources for a single "joint" clustering.

In a nutshell, BCC excels 1) in the sense that it captures inter-source associations that

promise a higher power for source-specific estimates and the integration is simultaneous which exploits the shared structure more efficiently. The approach 2) comes with a cost of sacrificing features that are specific to each data source, therefore considered suboptimal in practice. To illustrate the main idea of BCC, we first review model-based clustering via densities mixture, then revisit BCC as a hierarchical prior on the mixing probability.

Finite Mixture Models

To illustrate how BCC works, it is assumed that data X_n is given for N objects ($n = 1, \ldots, N$) with the goal of partitioning these objects into K clusters. We proceed with a model-based approach by assigning a probabilistic model $f(X_n|\theta)$ for X_n with parameter(s) θ. For example, f can be a Gaussian density with mean and variance, $\theta = (\mu, \sigma^2)$. Each X_n is independently drawn from a mixture of K Gaussian distributions. Let $C_n \in \{1, \ldots, K\}$ be the component label for X_n, then the probability that the object n belongs to cluster k:

$$\pi_k = \Pr(C_n = k).$$

Therefore, the generative model for X_n is

$$X_n \sim f(\cdot|\theta_k) \text{ with probability } \pi_k. \tag{1.9}$$

Alternatively, (1.9) can also be written as,

$$X_n \sim \sum_{k=1}^{K} \pi_k f(\cdot|\theta_k). \tag{1.10}$$

We prefer the annotation via data augmentation of latent group label parameter, C_n, since it is beneficial to express the shared structure hierarchically and connect findings at multiple levels.

Integrative Model

Now to extend the finite mixture model to M different sources of data $\mathbb{X}_1, \ldots, \mathbb{X}_M$, each of which is measured on a common set of N objects and the goal is still to cluster these objects

into K groups. Similarly, each individual data source can be characterized by a density $f_m(X_n|\theta_m)$ which has its own parameters θ_m.

We further assume that there exists clustering individually for each source $\mathbb{X}_m$, such that from a generative perspective, X_{mn}, $n = 1, \ldots, N$, are independently drawn from a K-component mixture distribution parameterized by $\theta_{m1}, \ldots, \theta_{mK}$. Let $L_{mn} \in \{1, \ldots, K\}$ denotes the source-specific label of X_{mn}, and C_n is held for a *consensus* label of object n. The source-specific clusterings $\mathbb{L}_m = (L_{m1}, \ldots, L_{mN})$ are anchored to a consensus clustering $\mathbb{C} = (C_1, \ldots, C_N)$ through:

$$\Pr(L_{mn} = k|C_n) = \nu(k, C_n, \alpha_m), \tag{1.11}$$

where ν has a simple form of

$$\nu(L_{mn}, C_n, \alpha_m) = \begin{cases} \alpha_m & \text{if } C_n = L_{mn} \\ \frac{1-\alpha_m}{K-1} & \text{otherwise} \end{cases} \tag{1.12}$$

We assume the mixing weights at consensus level, $\Pi = (\pi_1, \ldots, \pi_K)$, follows a Dirichlet distribution, then the source-specific mixing weights have a hierarchical representation:

$$\Pr(L_{mn} = k|\Pi) = \pi_k \alpha_m + (1 - \pi_k)\frac{1 - \alpha_m}{K - 1} \tag{1.13}$$

The fully conditional model reveals further insights on the mechanism as a hierarchically constructed mixture prior:

$$\Pr(L_{mn} = k|X_{mn}, C_n, \theta_{mk}) \propto \nu(k, C_n, \alpha_m) f_m(X_{mn}|\theta_{mk}) \tag{1.14}$$

$$\Pr(C_n = k|\mathbb{L}, \Pi, \alpha) \propto \pi_k \prod_{m=1}^{M} \nu(L_{mn}, k, \alpha_m). \tag{1.15}$$

To be specific, (1.14) indicates that a source-specific estimate is directly rooted to the likelihood of X_{mn}, that loosely adheres to the consensus cluster C_n, as a guidance, in proportion to the link function ν. At the consensual level, cluster label $\mathbb{C}$ integrates individual sources with an anchor of adherence of α_m if their labels coincide, in addition to a prior knowledge of the overall assignment through Π.

As for an application to EEG clustering, a population cluster $\mathbb{C}$ is loosely adhered to individual clusters $\mathbb{L}_m$ and we can further extend (1.13) to an extra level of segments, for a time-frequency representation within each subject. The link function inherits the formulation in (1.12) such that a dynamic cluster at time t is also loosely anchored to the individual cluster. We refer to this approach as Multilevel Integrative Clustering (MIC) in Chapter 2.

1.4 Mixed Effects Models

Mixed Effects Models were first introduced by [LW82] as a two-stage random-effects model. Their use was intended to handle unbalanced samples of serial measurements, that adjust for within- and between-individual variation explicitly. EEG recordings post to pre-processing are inherently unbalanced due to quality control measures, for example segment rejection. These structural assumption also coincide with well recognized characteristics of electrophysiological responses. In particular, within-subject variation of EEG is substantially due to the low signal-to-noise ratio, therefore can be attenuated by borrowing information over repeated measurements via structural assumptions. At the group level, heterogeneity is commonly observed on certain cohorts of individuals, Autistic spectrum disorder (ASD) for example, both etiologically and clinically, posing extra challenges in inferential problems that involve interpretations of EEG findings among these cohorts at group level.

Moreover, mixed effects models can handle both sequential observations and repeated measurements under a regression framework. The sequentially observed data points, often referred to as functional and longitudinal trends depending on applications, are subject to additional structural assumptions such as smoothness. In these scenarios, one may deal with single function observed randomly, or multiple functions that are correlated with each other. In both cases, we could regress observations on covariates, group, age and gender for example, as fixed effects, and encode appropriate assumptions on the covariance of the random effects. All of these are statistically appealing and supported with well established inferential techniques.

In this dissertation, I generalize the mixed effects model to the case of functional data

analysis (FDA). The proposed approach is applied in the study of regionally referenced EEG band power dynamics during the course of a cognitive experiment. To appropriately introduce and expand on the topic, linear mixed effects models are first reviewed in Section 1.4.1, followed by their functional variants in Section 1.4.2. Finally in Section 1.4.3, a latent factor regression model is introduced and later used to effectively approximate functional covariance via a reduced-rank structure.

1.4.1 Linear Mixed Effects Models

We start with a notation list similar to [LW82]. Let $\boldsymbol{\alpha}$ be a $p \times 1$ vector of fixed population parameters and $\boldsymbol{X}_i$ be the $n_i \times p$ design matrix relating $\boldsymbol{\alpha}$ to outcome $\boldsymbol{y}_i$. Let $\boldsymbol{b}_i$ denote a $k \times 1$ vector of individual random effects, that is related to $\boldsymbol{y}_i$ by a known design matrix $\boldsymbol{Z}_i$ of size $n_i \times k$. Then the outcome $\boldsymbol{y}_i$ is a $n_i \times 1$ vector which can be expressed as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{1.16}$$

where $\boldsymbol{e}_i$ is distributed as $\mathrm{N}(\boldsymbol{0}, \sigma^2 I_{n_i})$. The random effects $\boldsymbol{b}_i$ are distributed as $\mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, independently between subjects and of the noise term $\boldsymbol{e}_i$. $\boldsymbol{D}$ is a $k \times k$ positive-definite covariance matrix, that is common for all subjects.

Estimation and inference have been well studied and established by [LW82]. Both maximum likelihood (ML) and restricted maximum likelihood (REML) approaches are available in standard computational packages (see [BMB14]).

1.4.2 Functional Mixed Effects Models

Linear mixed effects models have been generalized for functional data analysis. As the major difference, the response variable $\boldsymbol{y}_i(t)$, at time t, is a snippet of repeated measure functionals that are assumed smooth in some functional space. To effectively bridge the gap, functional observations are projected onto pre-defined basis functions, for examples, polynomials, cubic splines, B-splines, wavelets, even functional Principal components which are considered as empirical choices. Due to a reduced and unified dimensionality of the functional domain,

the coefficient matrices are more tractable in practice. Here, a selective review of functional mixed models is provided in this section, however, the review paper by [RWC09] is highly recommended for a comprehensive outlook.

Functional Models with Smoothing Splines

[Guo02] considered a class of functional models in which smoothing splines are used to represent both fixed and random effects. It is regarded as a non-parametric extension to the linear mixed effects model, with smoothness guaranteed for both population-average and subject-specific curves.

To be specific, suppose that $y_i(t_{ij})$ $(i = 1, \ldots, n;\ j = 1, \ldots, m_i)$ is the response of the ith curve at time t_{ij} and can be modeled as

$$y_i(t_{ij}) = \boldsymbol{x}_i'\boldsymbol{\beta}(t_{ij}) + Z_i\boldsymbol{\alpha}_i(t_{ij}) + e_{ij}, \tag{1.17}$$

where $e_{ij} \sim \mathrm{N}(0, \sigma^2)$, $\boldsymbol{\beta}(t) = \{\beta_1(t), \ldots, \beta_p(t)\}^T$ is a $p \times 1$ vector of fixed functions, $\boldsymbol{\alpha}_i(t) = \{\alpha_{1i}(t), \ldots, \alpha_{qi}(t)\}^T$ is a $q \times 1$ vector of random functions that are modeled as realizations of Gaussian processes $\boldsymbol{A}(t) = \{a_1(t), \ldots, a_q(t)\}^T$ with zero mean, $X = \{\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}^T$ and $\boldsymbol{Z}_i$ are design matrices for fixed and random effects respectively.

In oder to achieve similar smoothness for both the population-average curve and subject-specific curve, $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}_i(t)$ are modeled in the same functional space, such that $\boldsymbol{\beta}(t)$ is considered as a single realization of a partially diffuse Gaussian process. Further details about estimation and inference are available in [Guo02].

In the context of EEG experiments, the curves can be conceptualized as longitudinal trends in response to a sequence of stimuli. For example, differential waveforms of EEGs are observed at the beginning and the end of a study, where a group of French infants were exposed to a stream of continuous speech that contains statistical cues for word segmentation [GSN10]. Moreover, statistically reconstructed functionals can dynamically reflect the functional response over time, for instance band-specific spectral power on experimental time, highlighting differential electrophysiological responses between groups.

Functional Models on Time-domain

Event-related potentials (ERPs) are routinely used in the study of electrophysiological responses to cognitive tasks. It time locks the EEG recording to the onset of stimuli, and requires a waveform analysis on time domain. [Dav09] described an application of mixed effects models to ERPs using a discrete wavelet transform (DWT). It essentially operates on the space of wavelet coefficients using a linear mixed effects model as (1.16), and transforms the sparse coefficient estimates of $\beta(t)$ and $\alpha_i(t)$ back to the time domain as fixed and random effects with the inverse DWT.

Functional Models on Frequency-domain

A mixed effects Cramér spectral representation is introduced by [KHG11]. Specifically, the spectrum is in a product form such that the resulting log-spectrum has a functional mixed effects representation where both the fixed and random effects are functions on the frequency domain. An iterative procedure is offered for the periodic smoothing spline estimation of the fixed effects, penalized estimation of the functional covariance of the random effects, and individual random effects prediction via the best linear unbiased predictor.

1.4.3 Sparse Bayesian Latent Factor Models

[BD11b] introduced a sparse modeling approach on high-dimensional covariance matrices, that can be easily generalized to flexibly handle the covariance of random effects under the settings of functional mixed effects models. The authors used a multiplicative gamma process shrinkage prior that automatically regulates factor loadings towards zero as column index increases. Essentially, their approach can efficiently find a unique solution to a low-rank decomposition of a $p \times p$ covariance matrix, $\Omega = \Lambda\Lambda^T + D$, where Λ is a $p \times k$ factor loading matrix with $k \ll p$ and a diagonal residual D.

In a generic form of a latent factor model,

$$\boldsymbol{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathrm{MN}(\mathbf{0}, D). \tag{1.18}$$

where the multivariate outcome $\boldsymbol{y}_i$ can be decomposed as k latent factors multiplying their loadings up to a residual that follows independent Normal distributions. To effectively induce automatic shrinkage on columns of the loading matrix Λ, the multiplicative gamma prior is defined as,

$$\lambda_{jk}|\phi_{jh}, \tau_h \sim \mathrm{N}(0, \phi_{jh}^{-1}\tau_h^{-1}), \quad \phi_{jh} \sim \mathrm{Ga}(\nu/2, \nu/2), \quad \tau_h = \prod_{l=1}^{h} \delta_l,$$
$$\delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_l \sim \mathrm{Ga}(a_2, 1), \quad l \geq 2, \quad \sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma).$$

The elements of factor loading matrix are handled individually, where each one has its own precision parameter $\phi_{..}$ and column-wise shrinkage parameter τ_h. To effectively regulate column as h increase, it is required that a_2>1 such that τ_h multiplicatively accumulates penalties by expectation.

This low-rank structure, also referred to as sparsity by the original authors, is practically appealing for large p, but also generalizable to include covariate effects. [MTN12a] extended the approach to functional and longitudinal data, and they explicitly encode covariate effects via the latent factor $\boldsymbol{\eta}_i$. It sacrificed the interpretability of covariate effects, in author's opinion, therefore a different generalization is proposed in this dissertation. In addition, we extend the framework to multivariate functional mixed model under the assumption of strongly separable and non-separable covariance structures. Further details regarding the proposed approach are elaborated in Chapter 3.

CHAPTER 2

Inferring Brain Signals Synchronicity from a Sample of EEG Readings

2.1 Introduction

Functional neuroimaging technologies, including MRI, PET, MEG, and EEG, aim to measure different aspects of brain function as they relate to specific mental processes. This chapter focuses on the analysis of Electroencephalography (EEG) data in the context of neuropsychology studies. EEG is a well-established noninvasive method for measuring spontaneous and event-related electrical activity across brain regions. The technology captures voltage fluctuation as signals, which reflect the distributed neuronal activities being projected on a cortical patch on which an EEG sensor is placed ([Tep02]). The general aim of an EEG study is often the identification of neural function and cognitive states. Diverse biomedical applications include epilepsy, sleep disorders, multiple sclerosis, brain tumors, lesions, schizophrenia, and mood disorders ([Tep02]).

Typical analyses in EEG studies focus primarily on inferring group differences in regions of interest. Such differences are assessed both in the frequency domain, by means of an amplified Signal-to-Noise Ratio (SNR) ([LKS03]), and, in the case of studies involving external stimuli, in the time domain, by means of averaging and smoothing over repeated applications of the stimuli ([HSJ15]).

Beyond differential activation of brain regions, mounting evidence is building a case for the deeper understanding of neural interactions ([DAS14], [CJY13]). In this setting, magnetic resonance imaging has become an established workhorse for the mapping and annotation of

the human connectome at the macro-scale. The key to the success of MRI technologies as a preferred measurement tool in functional connectivity studies lies in their ability to produce measurements at high spatial resolution. This ability comes, however, at the cost of low time resolution, and perhaps most importantly, at the cost of severe hardware limitations, intended as the need to rely on expensive and bulky MRI scanners, which make MRI studies hard to design in a logistically and financially feasible fashion.

On the other end, EEG is thought to provide reliable measurements of neuronal activity only for the brain cortical regions, with low spatial resolution and often low SNR. However, compared to other imaging techniques, EEG has the advantage of relying on less bulky hardware and is associated with robust and extremely non-invasive imaging protocols, making the technology readily available for implementation and adaptation to a variety of scientific investigations.

Recently , [EOO15] suggested exploiting EEG's excellent temporal resolution by defining the concept of spectral synchronicity. In particular, a pair of EEG signals are considered spectrally synchronized if they are both dominated by similar frequency oscillations. This idea formalizes the concept of coordinated neuronal activity and reflects recent empirical evidence, which suggests that differential patterns of coordinated neuronal activity may be associated with a range of neuropsychiatric and neurological processes, including memory formation ([FA11]) and mental disorders ([BDD08]).

From a statistical perspective, multi-subject studies of functional connectivity still pose substantial methodological challenges. Ideally, statistical inference should provide tools for the understanding of typical functional connectivity patterns, as well as quantification of familiar concepts like sample and population variability, and dependence on clinical phenotypes via regression. Even though some progress in the direction of population level inference has recently been made in the context of fMRI data ([NA15]; [SNM14]), typical analyses are still reliant on untenable assumptions of time-independence. The literature is, in fact, substantially silent on the subject of population level connectivity inference using EEG data. In this work, we aim to address this problem and introduce a simple and interpretable technique for the analysis of brain synchronicity from a sample of EEG readings. Our approach

relaxes the classical graphical modeling strategy into a simpler problem of clustering brain regions. As a consequence our analysis is perhaps coarser than what is usually done in the functional connectivity literature.

Our approach is based on the definition of cortical maps, identifying areas of synchronous neuronal activity specific to individual subjects and experimental epochs, intended as time intervals. Synchronized cortical regions are estimated via a mixture model of eigen-Laplacian vectors, obtained from appropriately constructed dissimilarity matrices. As the experiment evolves in time, subject and time-specific cerebral maps form a longitudinal ensemble. In this context, we posit that pooled information, within and between subjects, is amenable to statistical analysis via a hierarchical model involving mixture probabilities ([LD13]), which we call Multilevel Integrative Clustering (MIC). Our framework supports both the definition of coordinated neuronal activity via a mixture approach, and the formulation of probability statements describing inter-subject and intra-subject variability via the familiar toolset of hierarchical modeling.

This chapter is organized as follows. In Section 2.2 we describe a general framework for integrative clustering at the epoch, subject and population levels. In Section 2.3 we assess the operative characteristics of our proposed approach through experiments on engineered data. In Section 2.4 we apply the proposed framework to the analysis of a resting-state EEG study on typically developing (TD) children and children diagnosed with Autism Spectrum Disorder (ASD). We conclude with a critical discussion and potential extensions in Section 2.5.

2.2 Multilevel Integrative Clustering (MIC)

In the foregoing discussion we proceed to characterize coordinated neuronal activity via time-varying pairwise distances between the time series associated with a set of EEG sensors or electrodes. Our discussion builds on the approach of [EOO15], where synchronicity is defined in relation to pairwise similarities between the power spectral densities of electrode-level signals. In §2.2.1, we proceed by describing a data pre-processing step aimed at obtaining

stable time-varying estimates of the EEG spectral profiles. In §2.2.2, instead of directly operating on spectral densities, we model a set of related d-dimensional eigen-Laplacians via a multilevel model for clustering areas of synchronous neuronal activation. Inferential and computational details are discussed in §2.2.3 and §2.2.4.

In the following discussion we proceed to characterize coordinated neuronal activity via time-varying pairwise distances between the time series associated with a set of EEG sensors or electrodes. Our approach builds on [EOO15], who define synchronicity in relation to pairwise similarities between the power spectral densities of electrode-level signals. In §2.2.1, we describe a data meta-processing step aimed at obtaining stable time-varying estimates of the EEG spectral profiles. In §2.2.2, instead of directly operating on spectral densities, we model a set of related d-dimensional eigen-Laplacians via a multilevel model for clustering areas of synchronous neuronal activation. Inferential and computational details are discussed in §2.2.3 and §2.2.4.

2.2.1 From EEG Signals to Eigen-Laplacian Matrices

Let $\{Z_\tau,\ \tau = 0, \pm 1, \pm 2, \ldots\}$ be a zero mean, weakly stationary time series, with autocovariance $\mathcal{C}_Z(h) = E(Z_\tau, Z_{\tau+h})$, $(h = 0, \pm 1, \pm 2, \ldots)$. The second order properties of the series may be described by the spectral density function $\phi_Z(\omega)$ of Z_τ as in [Bri81], so that:

$$\phi_Z(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \mathcal{C}_Z(\tau) \exp(-i\tau\omega), \quad \omega \in [0, \pi].$$

Intuitively, $\phi_Z(\omega)$ may be interpreted as the variance contributed to the entire series by oscillations in a narrow frequency band around $\omega \in [0, \pi]$. The spectral analysis of neural signals is an important workhorse in EEG studies, as frequency bands are thought to be associated with specific cognitive, perceptive and cellular phenomena ([Tep02]).

EEG time-series signals are usually collected in relation to a geodesic net of p electrodes. Upon collection, raw signals are segmented into 1024ms time intervals for EEG preprocessing, which typically includes bandpass filtering, electrode and segments rejection, and artifacts inspection. Similar pipelines are common for EEG analysis, which can improve the SNR for spectral analysis ([BMK15]).

Let $i = 1, \ldots, n$ index n study subjects, $j = 1, \ldots, p$ index p EEG electrodes, and $s_{\ell i}$, $\ell_i = 1, \ldots, q_i$, index q_i-1024ms segments retained after data quality control. The filtered EEG data can be seen as an ensemble of time-series segments $Y_{ij}(s_{\ell_i})$, each composed of a number of measurements reflective of analog-to-digital sampling rates, typically 256/512Hz.

We are interested in the time-dynamics of neuronal synchronicity through a notion of time-varying spectral density via local stationarity ([FP95]; [RWS12]). In our formulation we fully acknowledge common pre-processing practices, which sees qualifying EEG segments being concatenated and re-referenced without time labelling. This practice typically leads to latent gaps in the post-processed series, providing a non-standard inferential framework for time-varying spectral estimation.

In order to obtain time-varying stable estimates of electrode-specific spectra, we operate on a combined set of γ adjacent segments $(s_{\ell_i}, \ldots, s_{(\ell_i+\gamma)})$, which we define as epochs. Furthermore, adjacent epochs smooth over the original time domain by overlapping over a $\delta \in (0, 1)$ fraction of segments. For each subject i, electrode j and epoch $t \in \{1, 2, \ldots, T_i\}$, we obtain estimates $\widehat{\phi}_{ij}(\omega, t)$ of the epoch-specific spectral density by averaging segment specific spectral density estimates obtained as in [ORS01]. The details of this procedure are reported in a supplementary document. Our approach stems from the idea introduced by [HSJ15] in the context of time-domain analyses. The use of overlapped sliding windows in the estimation of a time-dependent power spectral density mediates between the need for stable estimates and the potential for non-stationarity over the entire duration of the study. A study of inferential robustness to smoothing choices is reported in §2.3.

Following the approach by [EOO15], desynchronicity is measured by *total variation distance* (TVD) between a pair of normalized spectral densities estimated at each epoch, so that, for subject i, desynchronicity between electrode j and electrode k at epoch t is defined as:

$$d_{it}(\widehat{\phi}_{ij}, \widehat{\phi}_{ik}) = 1 - \int \min\{\widehat{\phi}_{ij}(\omega, t),\, \widehat{\phi}_{ik}(\omega, t)\}\, \mathrm{d}\omega.$$

For each subject and epoch, these pairwise distances produce a $p \times p$ dissimilarity matrix $D_i(t) = \left[d_{it}(\widehat{\phi}_{ij}, \widehat{\phi}_{ik})\right]$, summarizing information on differential synchronicity between the p

electrodes from different cortical regions.

Before clustering, each matrix is represented in the eigen-space spanned by the largest d eigenvectors of the graph-Laplacian associated with an affinity matrix $A_i(t) = 1 - D_i(t)$. More precisely, we take a *graph cuts* view of clustering and construct a normalized graph-Laplacian $\mathcal{G}_i(t) = diag\,[A_i(t)\mathbf{1}_p]^{1/2}\ A_i(t)\,diag\,[A_i(t)\mathbf{1}_p]^{1/2}$, representing a weighted undirected graph between EEG electrodes. In this setting, we follow ([NJW01]) and summarize the information in $\mathcal{G}_i(t)$ with its largest d eigenvectors $X_i(t) \in \mathbb{R}^{p\times d}$.

This strategy is intuitively motivated by the analysis of the isolated connected components "*ideal case*", in which $A_{jk}(t) > 0$ iff components j and k belong to the same cluster, and $A_{jk}(t) = 0$ otherwise. In this simplified setting, considering K clusters, the first K columns of $X_i(t)$ have non-zero elements corresponding to connected components in $A_i(t)$. Row-wise, $X_i(t)$ is piece-wise constant, suggesting K-means as a simple clustering rule to recover the connected components.

We work under the assumption that $\mathcal{G}_i(t)$ is a perturbation of the "*ideal case*" and in §2.2.2 we exploit this intuition to develop model-based clustering of electrodes at the epoch, subject and population level. Crucially, we avoid using a mixture model of spectral densities; instead model-based clustering of EEG signals over potentially non-convex manifolds is achieved using simpler location/scale-mixture models involving vectors in $\mathbb{R}^d$.

It is important to point out, that the measure of neuronal synchronicity, defined as spectral synchronicity, is indeed not essential for the application of multilevel integrative clustering. In particular, alternative means of quantifying similarity between time series, like coherence, cross correlation, partial correlation, etc. may be appropriate in specific investigations ([BS16]). Furthermore, if interest centers on specific band-power frequencies, discrepancies are easily defined over the appropriately truncated spectral densities.

2.2.2 Hierarchical Mixture Priors and Multilevel Inference

Let $X_{ij}(t) \in \mathbb{R}^d$, be a d-dimensional eigen-Laplacian vector associated with the EEG signal for subject i, $(i = 1, 2, \ldots, n)$; electrode j, $(j = 1, 2, \ldots, p)$; at epoch $t = 1, 2, \ldots, T_i$. In

practice, we observe subject-specific epochs t_{im_i}, $(m_i = 1, 2, \ldots, T_i)$. However, without loss of generality and for ease of notation, we maintain the lighter epoch indexing t throughout the chapter.

Within subject, at epoch t, we conceptualize synchronous patterns of cortical activity, by clustering electrodes according to the following mixture model. Denoting with $f\{\cdot \mid \cdot\}$ a generic density with respect to the Lebesgue measure on $\mathcal{B}\left(\mathbb{R}^d\right)$, we assume that each eigen-Laplacian vector $X_{ij}(t)$ is sampled from a K-components mixture distribution, indexed by parameters $\boldsymbol{\theta}_{ik}(t)$ and mixture probabilities $p_{ijk}(t) \in [0, 1]$, such that:

$$X_{ij}(t) \sim \sum_{k=1}^{K} p_{ijk}(t) f\{X_{ij}(t) \mid \boldsymbol{\theta}_{ik}(t)\}, \quad \sum_{k=1}^{K} p_{ijk}(t) = 1. \tag{2.1}$$

We find it convenient to re-express this sampling model with the equivalent hierarchical representation, mixing over cluster labels $L_{ij}(t) \in \{1, 2, \ldots, K\}$, s.t.:

$$\begin{aligned} X_{ij}(t) \mid L_{ij}(t) = k &\sim f\{X_{ij}(t) \mid \boldsymbol{\theta}_{ik}(t)\}, \\ \Pr\{L_{ij}(t) = k\} &= p_{ijk}(t). \end{aligned} \tag{2.2}$$

In this setting, echoing the clustering "*ideal case*" discussed in the previous section, we exploit the connection between K-means and Gaussian mixtures and represent the sampling density in (2.1) as a K-component location/scale mixture of Gaussian distributions. Specifically, let $\mu_{ik}(t) \in \mathbb{R}^d$ be a d-dimensional mean vector, and $\sigma^2_{ik}(t) > 0$ be a variance parameter. We assume:

$$f\{X_{ij}(t) \mid \boldsymbol{\theta}_{ik}(t)\} = N\{\boldsymbol{\mu}_{ik}(t), \sigma^2_{ik}(t) I_d\}. \tag{2.3}$$

Given the sampling model in (2.2), our proposed approach for the integration of information at the subject and population levels follows a conceptually simple strategy, building directly on the setting of multilevel modeling ([GH07]). Crucially, we maintain that mixture means and variances are independent across subjects and epochs, but posit that cluster configurations, conceptualizing synchronicity of brain regions, are likely to adhere to patterns of similarity within and between subjects.

We make this idea precise by specifying a hierarchical prior for the mixture probabilities, $p_{ijk}(t)$. This is achieved by defining conditionally exchangeable mixture configurations,

where epoch-level clusters $L_i(t)$ are obtained, *a priori*, as a stochastic perturbation of a time stable subject-level clustering, indexed by C_i. Similarly, subject level configurations, C_i, are obtained as a stochastic perturbation of a population-level cluster, indexed by S.

Let $C_{ij} \in \{1, 2, \ldots, K\}$ be the cluster label for electrode j at the level of subject i. Furthermore, let $\beta_i(t) \in [1/K, 1]$ be an adherence parameter, quantifying conformity between cluster assignments at epoch t and the subject-level label $C_i = (C_{i1}, \ldots, C_{ip})'$. We assume,

$$\Pr\{L_{ij}(t) = k \mid c_{ij}\} \equiv \nu_c\{k, c_{ij}, \beta_i(t)\} = \begin{cases} \beta_i(t) & \text{if } c_{ij} = k \\ \frac{1-\beta_i(t)}{K-1} & \text{otherwise} \end{cases}, \tag{2.4}$$

where the probability $\nu_c\{\cdot, \cdot, \cdot\}$ is defined implicitly. This prior defines a probabilistic anchor, relating epoch level patterns of synchronicity at the subject level via simple and interpretable parameters $\beta_i(t)$. The underlying assumption is that epoch-level patterns of synchronicity are allowed to vary dynamically with t, but that variation in cluster configurations is anchored at the subject-level by a consensus pattern C_i.

A similar anchoring strategy is pursued at the population level. Specifically, let $S_j \in \{1, 2, \ldots, K\}$ be a population level cluster label for electrode j, and $\alpha_i \in [1/K, 1]$ be an adherence parameter, quantifying conformity between cluster assignments for subject i and population level labels $S = (S_1, \ldots, S_p)'$. We assume,

$$\Pr(C_{ij} = k \mid s_j) \equiv \nu_s(k, s_j, \alpha_i) = \begin{cases} \alpha_i & \text{if } s_j = k \\ \frac{1-\alpha_i}{K-1} & \text{otherwise} \end{cases}, \tag{2.5}$$

where probability $\nu_s(\cdot, \cdot, \cdot)$ is defined implicitly. The model is completed by specifying population level prior proportions:

$$\Pr(S_j = k) = \pi_k, \quad (k = 1, 2, \ldots, K).$$

To build intuition about the nature of these priors, we note that, if $\alpha_i = 1$, we expect cluster assignments for subject i to match exactly the population-level labels with probability 1. In contrast, for α_i approaching the value $1/K$, electrode clustering configurations C_i, for subject i, are drawn independently of the population level labels S. Similar considerations apply to $\beta_i(t)$, as these paratmeters relate subject- and epoch-level cluster configurations.

This modeling strategy is loosely related to the idea of consensus clustering ([NC07]), as applied to the integration of multi-source data. Our specific formulation is a direct generalization to multilevel models of the approach taken by [LD13] to the integration of heterogeneous genomic data.

In our multilevel setting, the conditional posterior distribution for epoch-level cluster labels $L_{ij}(t)$ is easily defined as:

$$\begin{aligned} \Pr\{L_{ij}(t) = k \mid X_{ij}(t), c_{ij}, s_j, \boldsymbol{\theta}_i(t)\} &\propto f\{X_{ij}(t) \mid L_{ij}(t) = k, \boldsymbol{\theta}_i(t)\} \Pr\{L_{ij}(t) = k \mid c_{ij}\} \\ &= f\{X_{ij}(t) \mid \boldsymbol{\theta}_{ik}(t)\} \nu_c\{k, c_{ij}, \beta_i(t)\}. \end{aligned} \tag{2.6}$$

This form highlights how inference on $L_{ij}(t)$ integrates information from both data $X_i(t)$ at epoch t, and subject-level clustering C_i (assumed stable across epochs), through a weighting scheme proportional to the size of the adherence parameter $\beta_i(t)$.

At the subject-level, conditional posterior probabilities of cluster membership weigh epoch level configurations $L_i(t) = (L_{i1}(t), L_{i2}(t), \ldots, L_{ip}(t))'$ with population level configurations S, through adherences α_i as follows:

$$\begin{aligned} \Pr\{C_{ij} = k \mid \ell_{ij}(1), \ldots, \ell_{ij}(T_i), s_j\} &\propto \Pr\{\ell_{ij}(1), \ldots, \ell_{ij}(T_i) \mid C_{ij} = k\} \Pr\{C_{ij} = k \mid s_j\} \\ &= \textstyle\prod_{t=1}^{T_i} \nu_c\{\ell_{ij}(t), k, \boldsymbol{\beta}_i\} \, \nu_s(k, c_j, \alpha_i). \end{aligned} \tag{2.7}$$

Finally, at the population level, overall consensus labels S are determined according to the following conditional posterior probability:

$$\Pr(S_j = k \mid c_{1j}, \ldots, c_{nj}, \Pi, \boldsymbol{\alpha}) \propto \pi_k \prod_{i=1}^{n} \nu_s(k, c_{ij}, \alpha_i). \tag{2.8}$$

In summary, for each subject we infer a consensus cortical configuration C_i, combining epochs $L_i(t)$ trough coherence weights $\beta_i(t)$. Across subjects, group-level inference is conducted through a consensus configuration S, pooling subject-level configurations C_i through coherence parameters α_i. This stochastic structure allows for a very general conceptualization of dependence across epochs. We note that, in some cases, more structured priors could be warranted, including a fully exchangeable configuration, where $\beta_i(t) = \beta_i$, for all epochs; or the case of $\beta_i(t)$ being defined as a smooth function of the epoch index t. Similar restrictions may be applied to α_i, depending on inferential goals and experimental characteristics.

2.2.3 Posterior Inference

We discuss posterior inference for the model in § 2.2.2 on the basis of MCMC samples from the target distribution. Even though multilevel modeling of cluster labels is a somewhat non-standard approach in a hierarchical setting, conditionally conjugate analysis is indeed possible, resulting in significant simplifications in computation and inference.

Specifically, we consider a standard Dirichlet prior for population-level proportions, so that $\Pi = (\pi_1, \pi_2, \ldots, \pi_K)' \sim \text{Dirichlet}(\eta)$. Epoch-level means and variances, are chosen to be conjugate to the graph Laplacian likelihood in (2.3). Letting $\boldsymbol{\theta}_{ik}(t_{im}) = (\boldsymbol{\mu}_{ik}(t_{im})', \sigma_k^2(t_{im}))'$, we assume that $\boldsymbol{\theta}_{ik}(t_{im}) \sim N\Gamma^{-1}(\boldsymbol{\mu}_0, \lambda_0, \xi_{01}, \xi_{02})$. Finally, subject-level adherence parameters α_i and epoch-level adherence parameters $\beta_i(t_{im})$ are assigned truncated Beta priors, with left truncation at $1/K$, so that:

$$\alpha_i \sim \text{TBeta}(a_i, b_i, 1/K), \text{ and } \quad \beta_i(t) \sim \text{TBeta}(c_i, d_i, 1/K).$$

A justification for these truncated Beta priors may be obtained by considering the form of the marginal allocation probabilities at subject and epoch level. Given Π, subject-level allocation probabilities are expressed as:

$$p_{ik} = \Pr(C_{ij} = k \mid \pi_k) = \pi_k \alpha_i + (1 - \pi_k)\frac{1 - \alpha_i}{K - 1}.$$

Similarly, at the epoch level, we have:

$$\Pr\{L_{ij}(t) = k \mid \Pi\} = \sum_{c_{ij}} \Pr\{L_{ij}(t) = k \mid c_{ij}\} \Pr(c_{ij} \mid \Pi) = \beta_i(t) p_{ik} + (1 - p_{ik})\frac{1 - \beta_i(t)}{K - 1}.$$

At both levels, an adherence value of $1/K$ corresponds to allocation probabilities, which are independent of higher-level clustering realizations.

A Gibbs sampler targeting the posterior distribution is easily devised, by iterating through a transition sequence of full conditional posteriors. Specific details about the form of the conditional posterior densities are reported in a supplementary document.

At each level of the model, the posterior probability associated with set of clustering labels, for generality say $p(C \mid \mathbf{Y})$, and the corresponding MCMC samples, summa-

rize our knowledge about potential partitions of cortical regions into synchronously activated areas. Based on the information in this posterior, we may be interested in selecting a representative partition, say C^*. Following [Dah06], we avoid using the naïve maximum *a posteriori* (MAP) estimate and instead consider a point estimator based on least squares. More precisely, consider an MCMC sample of M $p-$dimensional label configurations, $\{C^{(r)} : r = 1, 2, \ldots, M\}$. For each sample, we define a $p \times p$ adjacency matrix $\mathcal{A}\left(C^{(r)}\right) = \left[\mathcal{A}\left(C^{(r)}\right)_{ij}\right] = \left[I(C_i^{(r)} = C_j^{(r)})\right]$. Let $\bar{\mathcal{A}}$ be an estimate of the posterior mean $E[A \mid Y]$. The least square estimate C^* is selected from posterior realizations which minimize the following Frobenius norm

$$C^* = \min_{C^{(r)}, r=1,\ldots,M} || \mathcal{A}\left(C^{(r)}\right) - \bar{\mathcal{A}} ||_2.$$

Uncertainty about clustering estimates can be obtained from the posterior distribution, locally by quantifying pairwise relative frequencies of synchronization or globally via the distribution of $\mathcal{D} = ||A^{(r)} - \bar{A}||_2$. Examining this quantity facilitates direct comparison between subject and population level clustering results, allowing for low dimensional assessment of cluster quality, population and individual-level variability.

Computation and inference for MIC is performed under the R environment. A readily compiled package is available from the corresponding author's GitHub page.

2.2.4 Number of Clusters and Identifiability

Posterior inference as described in §2.2.3 presumes a known number of clusters K and a known number of eigen-Laplacian components d. For given d, selection of the number of mixture components, K, may be based on information criteria. In our simulation studies we find that the Bayesian Information Criterion (BIC) ([Sch78]) tends to outperform more complicated indices. Our findings are in agreement with [SR10], who observed that BIC outperforms many other criteria including ICL, DIC, and AIC, especially in the case of Gaussian mixture models.

The choice of d is less trivial, even though, some theoretical results point to the inclusion of the first K eigenvectors as being sufficient in the task of separating K groups, ([NJW01]).

Guided by this general principle, we perform a joint search on the dimensionality of the eigen-Laplacian d, and the number of clusters K simultaneously. More precisely, within a specific dimension d, the optimal value of $(K \mid d)$ is determined by the maximal BIC. Starting from low dimensions, usually $K = d = 2$, we allow for up-transitions on dimensionality, when $K^* \mid d > d$. Stopping rules, aiming at achieving stable solutions around the equality of $d^* = K^*$ are determined heuristically. Details are reported in Algorithm 1. Crucially, we avoid complete enumeration over all (d, K) combinations, and propose a search strategy which is linear in the maximum number of clusters. Our empirical studies in §2.3 show good performance and fast convergence to well behaved solutions.

Algorithm 1 (d, K) Selection

```
Set d = 2, K = 2;
current_BIC = BIC(d, K);
while d ≤max_d do
    while BIC(d, K + 1) ≥ current_BIC do
        current_BIC = BIC(d, K + 1);
        K = K + 1;
    if d ≥ K then
        break;
    else
        d = K, K = K − 1;
        current_BIC = BIC(d, K);
return (d, K)
```

For given d and K, simulation based procedures, including MCMC, are usually prone to *label switching* ([CHR00]). In the setting of the model proposed in §2.2.2 the same phenomenon may occur both within and between data levels. An important aspect of simulation-based inference in multilevel clustering is, therefore, the enforcement of correspondence between component labels of epochs, subjects and population level clustering. Possible remedies include artificial identifiability constraints, relabeling procedures, and label invariant loss

functions ([JHS05]). Within the multilevel setting, we proceed with online class relabeling or alignment. More precisely, we operate within population and subject-level indexes to find permutations of labels that maximize adherence with the population level clustering. Specifically, all newly sampled labels are permuted to insure maximal alignment with the population indexes. If $\mathcal{A}_0$ is an adjacency matrix as defined in §2.2.3, representing the current state of the population level labels S, and $\mathcal{A}_q$ is an adjacency matrix representing the current state of any other level clustering, optimal alignments are obtained by maximizing $tr\left(\mathcal{A}_0'\mathcal{A}_q\right)$ over $k!$ possible permutations.

2.3 Monte Carlo Studies

To investigate the operating characteristics of the proposed framework, we simulate EEG signals with the desired oscillation features from a mixture of AR(2) processes. We seek to evaluate: (1) the sensitivity of MIC results to differing sliding window size, γ, (2) the accuracy of estimated quantities for varying group adherence, (3) the performance of the model selection strategy proposed in Algorithm 1, and (4) the behavior of population level clusters under varying signal to noise ratio (SNR) and varying sample size.

2.3.1 Simulation Setup for Spectrally Specified EEGs

We make an effort to tailor the simulation of engineered time series in a way that mimics a sample of EEG readings typically seen in practice. To this end, we note that EEGs are often expected to feature oscillation patterns at different frequency bands: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-50 Hz). Waveforms that are subdivided into bandwidths are thought to correspond to region-related activities on the cortex, both normally and pathologically.

Our strategy, aims to simulate this spectral distinguishability by allowing each spectrum to exhibit concentrated (peak-shaped) energy in at most two frequency bands. Given a family of spectra, EEG time-series are simulated from a linear mixture of second order auto-regressive AR(2) processes. Details about the data generating mechanism are reported in

a supplementary document. Furthermore, we represent potential non-stationarity by generating time-series as realizations from a piecewise stationary process, alternating randomly between two spectral configurations: a *main*-state (Fig 2.1(a)), and an *off*-state shown in Fig 2.1(b). The *main*-state has a time span $t_{\text{main},i} \sim \exp(\lambda)$, with $\lambda = .05$s, followed by the *off*-state which has a time span $t_{\text{off},i} \sim N(5, 1)$. Fig 2.1 (c) depicts this piecewise-stationarity for one electrode from the simulated samples. Cluster labels are generated as follows:

1. At the *population level*, we structure cluster labels S_j, $(j = 1, \ldots, p = 100)$ to partition 100 sensors into 4 balanced clusters.

2. Draw α from a Uniform$(0.5, 1)$ distribution. For each subject i, and $j = 1, \ldots, p = 100$; generate *subject level* labels $C_{ij} \in \{1, 2, 3, 4\}$ with probabilities $\Pr(C_{ij} = S_j) = \alpha$ and $\Pr(C_{ij} \neq S_j) = (1 - \alpha)/3$.

3. Given C_{ij}, generate piecewise stationary processes for 50 seconds, according to the *main*-state / *off*-state mechanism described previously.

Our Monte Carlo study is based on 100 datasets. Subject-level variation is induced semi-parametrically, via random reconfigurations of subject specific clusters, and random timing of the *main/off*-state segments. The number of subjects, electrodes and segments were chosen to mimic the sampling structure in our case study. Note that in this setting, knowledge of the timing of *main*-state, *off*-state would result in perfect agreement of cluster labels within subject. Our simulation is therefore engineered to detect specific sensitivity to alternative metapreprocessing strategies.

2.3.2 Operating Characteristics

In §2.2.1 we introduced a pre-processing step to smooth over the duration of the EEG recordings in order to obtain time-stable estimates of spectral densities. We start by assessing sensitivity of window size, $\gamma \in \{4, 6, 8, 10\}$, at a fixed $\delta = 0.5$ fraction of overlap between epochs. Algorithm 1 successfully selected the correct number of clusters ($K = 4$), in more than 76% of cases for all varying widow sizes, Fig 2.2(a).

Furthermore, we investigate the performance of MIC under varying degrees of subject-specific variability, by examining estimates of adherence between subject- and population-level clustering. Fig 2.2(b) depicts posterior medians $\widehat{\alpha}_i = E(\alpha_i \mid \mathbf{X})$ and their 90% credible intervals, based on the 5 and 95 percentiles, against the true α's. Posterior estimates are generally close to their true values, and over 99% of the credible intervals cover the true α's.

Clustering accuracy, defined as the percentage of correctly classified electrodes, is assessed both at the subject and population level, Fig 2.2(c). Estimated subject-level clusters tend to be recovered accurately ($> 98\%$), regardless of α values. As expected, accuracy in the recovery of population level patterns relies on the magnitude of subject-level adherence to the population, with accuracy approaching 100% as $\alpha \to 1$.

Finally, we investigate the relationship between subject-level and population-level clustering variance estimates as a function of adherence and meta-processing strategy, Fig 2.2(d). Our summaries focus on a measure of global variance $\mathcal{D}$, as defined in §2.2.3. More precisely, denoting the clustering variance by $\mathcal{D}_S$ at the population level, and by $\mathcal{D}_{C_i}$ at the level of subject i, we consider the average difference in clustering variance, defined as: $\Delta_{\mathcal{D}} = E(\mathcal{D}_S \mid \mathbf{X}) - \frac{1}{n}\sum_i E(\mathcal{D}_{C_i} \mid \mathbf{X})$. As the adherence simulation truth approaches a level of complete agreement ($\alpha \to 1$), the average difference in clustering variance $\Delta_{\mathcal{D}}$ converges to zero, indicating that average subject-level and population-level cluster variances reach similar magnitudes over strongly adherent clustering patterns, Fig 2.2(d).

A second set of simulation studies aims to assess the operative characteristics of the proposed method under different SNR and sample size settings. Specifically, we consider SNR $= 1, 5, 10$ and sample size $N = 10, 20, 40$. We assess performance of our method under stationarity and local stationarity. Details about the simulation procedure are reported in our supplementary materials document. Our experiments show that group-level inference is highly robust to SNR configurations. In both stationary and locally stationary settings, clustering accuracy increases with sample size, going from a minimum of 0.8 (N=10), to about 0.9 (N=40). Results are reported in Table 2.1.

From our experiments we conclude that estimation and clustering results tend to be

robust across a broad range of SNR and smoothing parameters. This feature is likely to be useful in many applications, where it is usually hard to develop meta-processing gold standards.

2.4 A Case Study on Resting State Brain Activity

Our study originates from an experiment aimed at understanding children's neurocognitive development. The study was carried out in the department of Psychiatry at UCLA and aims to cluster spectrally synchronized EEG signals recorded during *resting-state*. We provide technical background information about the study design and measurement structure in a web-based supplement. Here we investigate neuronal synchronicity in a group of typically developing (TD) children. We contrast group inference for the TD cohort against patterns of synchronicity in a cohort of children diagnosed with Autism Spectrum Disorder (ASD) in § 2.4.1. To our knowledge this is the first attempt at population level-inference for neuronal synchronicity in the setting of EEG studies.

2.4.1 MIC Analysis of TD and ASD Children

Autism Spectrum Disorder (ASD) describes a neurodevelopmental condition, characterized by social communication deficits, presence of repetitive behaviors, and/or restricted interest. Clinical presentation is highly variable, with heterogeneity in relation to medical conditions, behavioral challenges, and degree of intellectual impairments [PLB11]). Such behavioral and neurophysiological heterogeneity poses serious challenges to the study of the neurophysiological substrate. In this respect, resting-state EEG is a particularly advantageous, and therefore popular, brain imaging choice ([WBE13]).

Here we perform a comparative study between age-matched TD and ASD cohorts, under the framework of Multilevel Integrative Clustering (MIC). The study includes 9 participants (29-60 months of age) from the TD group, and 10 participants (27-99 months of age) from the ASD group. During the experiment, EEG was recorded at 250Hz using 129 channel geodesic nets with Ag/AgCl electrodes. Recordings took place while participants watched

videos of bubbles and other non-social images on a computer monitor for 2 to 6 minutes.

Starting with the TD cohort, our analysis follows the scheme detailed in § 2.2.1 and considers epochs composed of $\gamma = 6$ contiguous 1024ms segments, allowing for a $\delta = 0.5$ overlap between epochs. This choice was based on both substantive and empirical considerations. In particular, we consider a smoothing strategy that guarantees good average adherence. A sensitivity analysis to differential smoothing choices was carried out with respect to both the epoch length and the percent of overlap. While details are reported in a supplementary document, we observe fairly robust results, with only small changes in estimation and selection of the number of clusters, echoing our findings in the simulation setting.

An illustration of how the proposed method clusters electrodes in relation to their spectral features is provided in Figure 2.3. Here, we represent the epoch-level estimates of the spectral densities for each electrode, color-labeled by inferred cluster membership. For each subject, we report the epoch of highest coherence with subject-level clustering. This simple illustration shows how, pooling information at the level of cluster labels can be achieved without requiring the spectral structure of electrode-level time series to be aligned across subjects. We maintain that this feature is particularly appealing in resting-state neurocognitive settings, where complex and dynamic alignment issues may render extremely difficult any attempt at pooling EEG signals directly.

An informal comparison between TD and ASD groups is carried out in Figure 2.4. For both cohorts, we identify 5 spectrally synchronized areas, corresponding to the following cortical regions: frontal, left and right parieto-temporal, occipital, and peripheral, defined as a ring of outsidemost electrodes. At the population level, the least square estimates of cortical clusters are remarkably similar between the two cohorts, with the exception of an asymmetrical partition on the occipital and parieto-temporal regions, where the left parieto-temporal cluster seems to be leaning towards the left hemisphere for ASD, but towards the right hemisphere for TD.

Further, we examine local and global sources of cluster variability in both groups. At the electrode level, we report the entropy associated with posterior cluster label probabilities in

Figure 2.4: (1.b) for ASD and (2.b) for TD. Perfect partitions, e.g. an electrode assigned to cluster k with probability one, yields 0 entropy, whereas uniform assignment probabilities yield entropy equal to 1. We observe that the mid-, right-frontal and mid posterior regions are the most stable regions for both groups. Compared to the ASD group, the TD cohort exhibits more stable regions, for example, in the left-temporal (speech and language related), left-central, as well as some regions in the posterior and occipital areas of the cortex. The high entropy observed on the left-hemisphere among ASD children coincides with the abnormal left-hemispheric asymmetry findings in the literature on individuals with ASD ([SNT07], [BHI11]).

We gain more insight into the nature of variability of synchronized neuronal patterns by examining global sources of cluster variance at the subject-specific and population levels. In Figure 2.4: (1.c) for ASD and (2.c), we report subject and population level cluster assignments for both TD and ASD cohorts. For each subject we also report the posterior median coherence estimate. We note how ASD children exhibit higher clustering heterogeneity, with coherence estimates ranging from 0.63 to 0.82, compared to the TD cohort, with coherence estimates ranging from 0.70 to 0.81. A similar conclusion is noted in the higher entropy associated with ASD consensus estimates (1.b and 2.b). This observation echoes some of our previous findings in EEG studies of implicit-learning in ASD and TD children ([HSJ15]; [HST6a]; [HST6b]).

While formal covariate adjustments are outside the scope of this chapter, we attempted a post-hock analysis aimed at explaining subject-level cluster variability using subjects age, ASD vs. TD cohort indicators, and electrode-level band power estimates. Using cluster labels as a categorical outcome, we used random forests as a flexible tool to get a sense of variable importance in the classification of synchronous electrodes. We found the out-of-bag estimate of classification accuracy to be about 0.73, with subject's age explaining the largest mean decrease in accuracy and therefore being flagged as one of the most important predictors. None of the power bands had a specific predictive advantage in explaining subject-level cluster variability, confirming our intuition that, in the setting of resting state experiments, it may be inappropriate to pool subject-level spectral features directly, in order to infer

connectivity. A less stringent model, like the one proposed in this chapter, is therefore likely to be more robust in applications. More details for this analysis are included in the supplementary materials.

2.5 Discussion

This chapter proposes what to our knowledge is the first comprehensive statistical framework for population level inference of spectrally synchronized brain activity from a heterogeneous sample of EEG readings. A hierarchical model allows for the estimation of population level synchronicity patterns, with full consideration of intra- and inter-subjects variability. Crucially, information is borrowed at the latent level of cluster membership indicators. Dependent mixtures are based on a hierarchical Dirichlet prior, indexed by interpretable and informative parameters, which measure cluster adherence at all levels of the hierarchy.

Our approach melds non-parametric dimension reduction and fully model-based techniques through a graph-partitioning representation of clustering. This two-stage approach is likely to be useful in several experimental settings involving EEG measurements, where different scientific goals and different data meta-processing concerns may require substantial subject-matter input in the definition of similarity between cortical regions.

In our study we operate within the context of spectral synchronicity. It is however important to point out that alternative measures of neuronal affinity, for example partial correlation, coherence, and mutual information, are also amenable to MIC analysis. In this sense, the proposed framework is quite general and can be adapted to handle alternative neuroimaging data platforms, such as functional Magnetic Resonance Imaging (fMRI). This consideration also applies, with possible minor adjustments, to the integration of multiple imaging modalities. This flexibility traces back to the hierarchical prior, which relates cluster labels rather than cluster-specific parameters (location and scale for example), so that complex data alignment issues are resolved within a higher level of modeling abstraction. Clearly, technical preprocessing pipelines may differ substantially between and within modalities. Therefore, important analytic details should be thoughtfully engineered in practice.

	Stationary Setting			Locally Stationary Setting		
SNR	N=10	N=20	N=40	N=10	N=20	N=40
10	0.812	0.868	0.903	0.808	0.868	0.897
5	0.810	0.867	0.898	0.810	0.868	0.898
1	0.810	0.869	0.900	0.813	0.866	0.902

Table 2.1: Simulation study: Group-level clustering accuracy for varying sample size and signal to noise ratio.

Our simulation results in § 2.3.2, show that inference is robust to reasonable variants in the meta-processing strategies. In our experiments, simple information criteria like BIC tend do do well in the selection of the number of clusters K, when combined with a search over the number of eigen-Laplacians d. Our model, of course, offers a very simple representation of cluster variability within- and between-subjects. Therefore, modeling refinements are likely needed in applications where one can expect a strong dynamic evolution of synchronicity patterns; such as the setting of stimulus-based EEG studies.

Potentially useful extensions include a formal treatment of group comparison and covariate adjustments. In particular, predictors could, in principle, be introduced through cohesion functions as in [MQR11]. We note, however, that the multilevel and dynamic structure of cluster configurations may require significant efforts to extend available covariate adjustment strategies in clustering. Other options would include covariates through a regression on subject-level coherence parameters, which would perhaps lead to simpler and more interpretable models.

A user-friendly implementation of the proposed method is available online as an R package at: `https://github.com/Qian-Li/MIC2`.

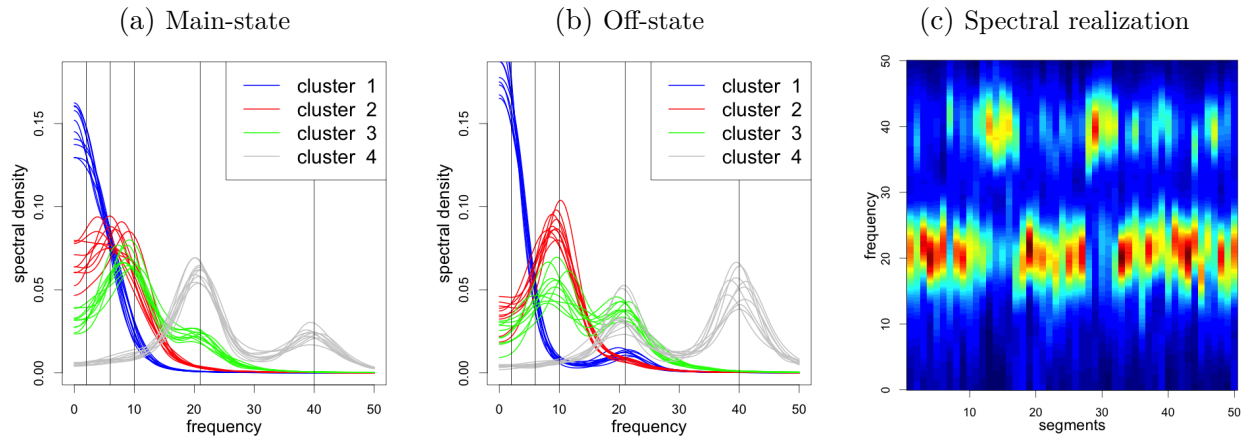

Figure 2.1: **Simulated spectral configurations**: (a) *main-state* spectral densities. (b) *off-state* spectral densities. (c) Segment-by-segment normalized power spectral densities for a piecewise stationary process simulated from cluster 4.

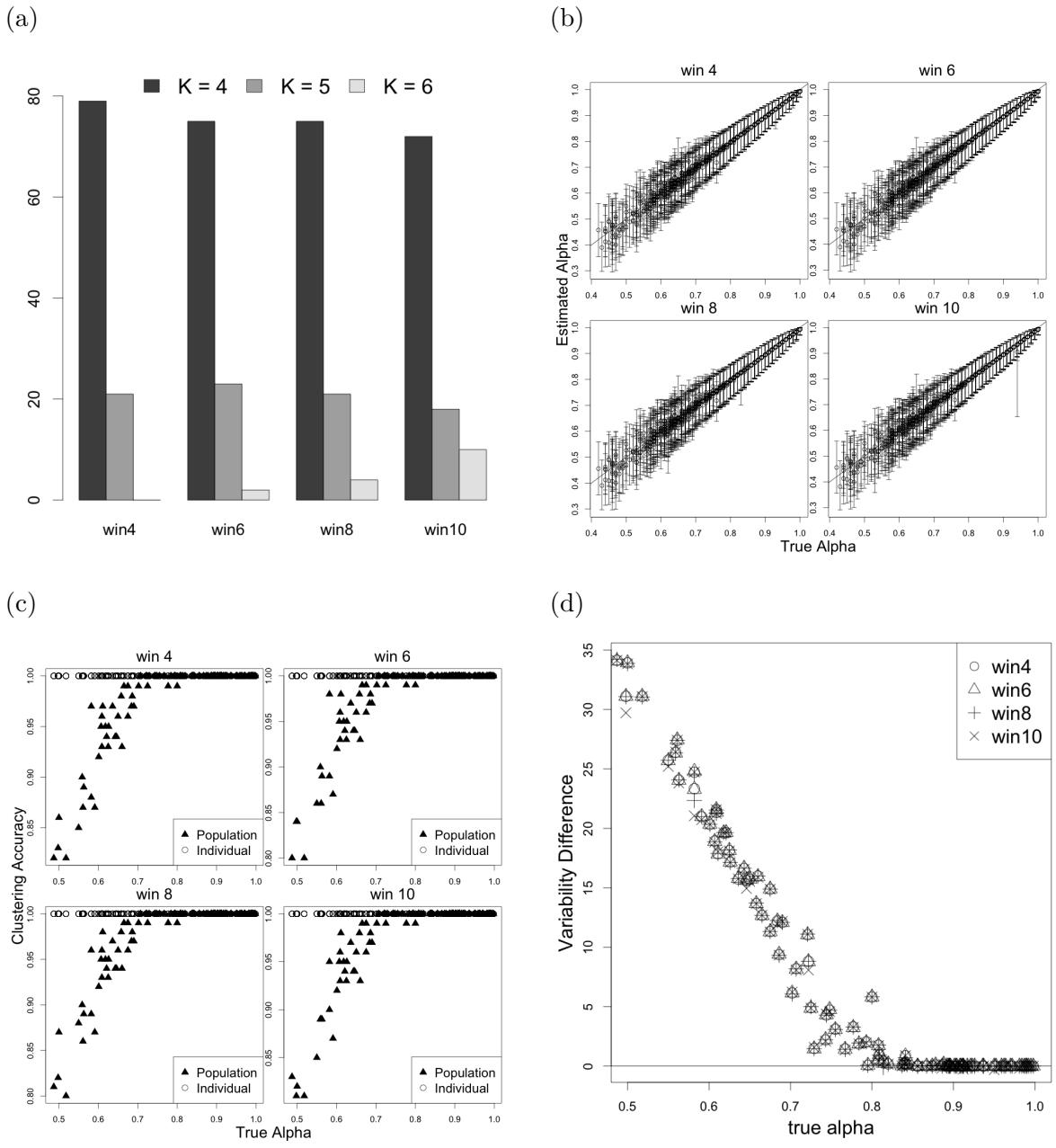

Figure 2.2: **Simulation results:** (a) Path-length for the search in Algorithm 1 for varying smoothing configurations in γ. (b) Estimated adherence parameters $\widehat{\alpha}$'s and 95% credible intervals against the data generating truth. (c) Clustering accuracy against generating α's at the subject- and population-level. (d) Average difference in clustering variance against true α's.

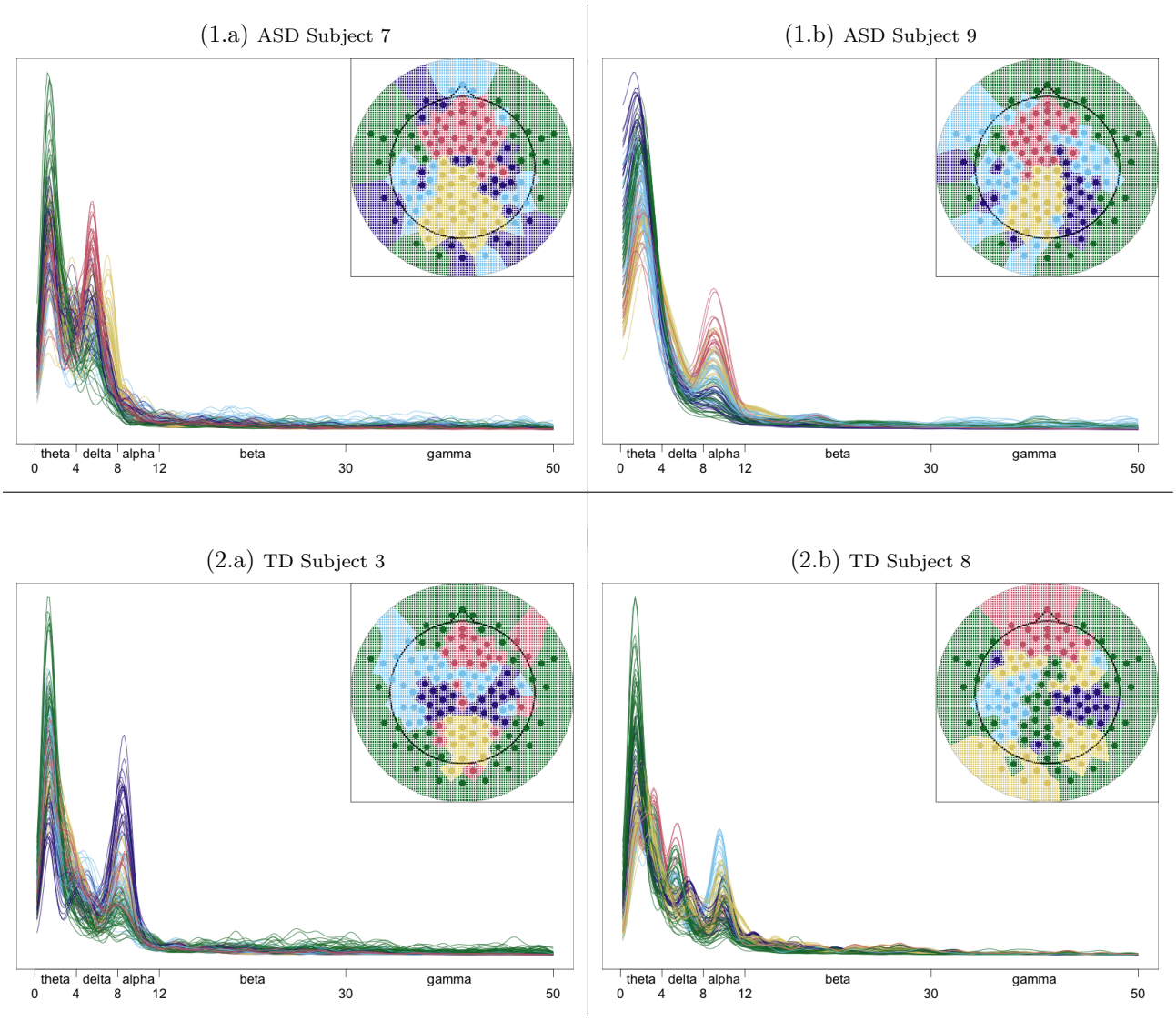

Figure 2.3: **Synchronicity and spectral features**: For each cohort, cluster configurations are depicted for two illustrative subjects. For each electrode, the estimated spectral density (normalized) is color coded by cluster membership. All plots refer to the epoch that is most coherent with subject-level clustering.

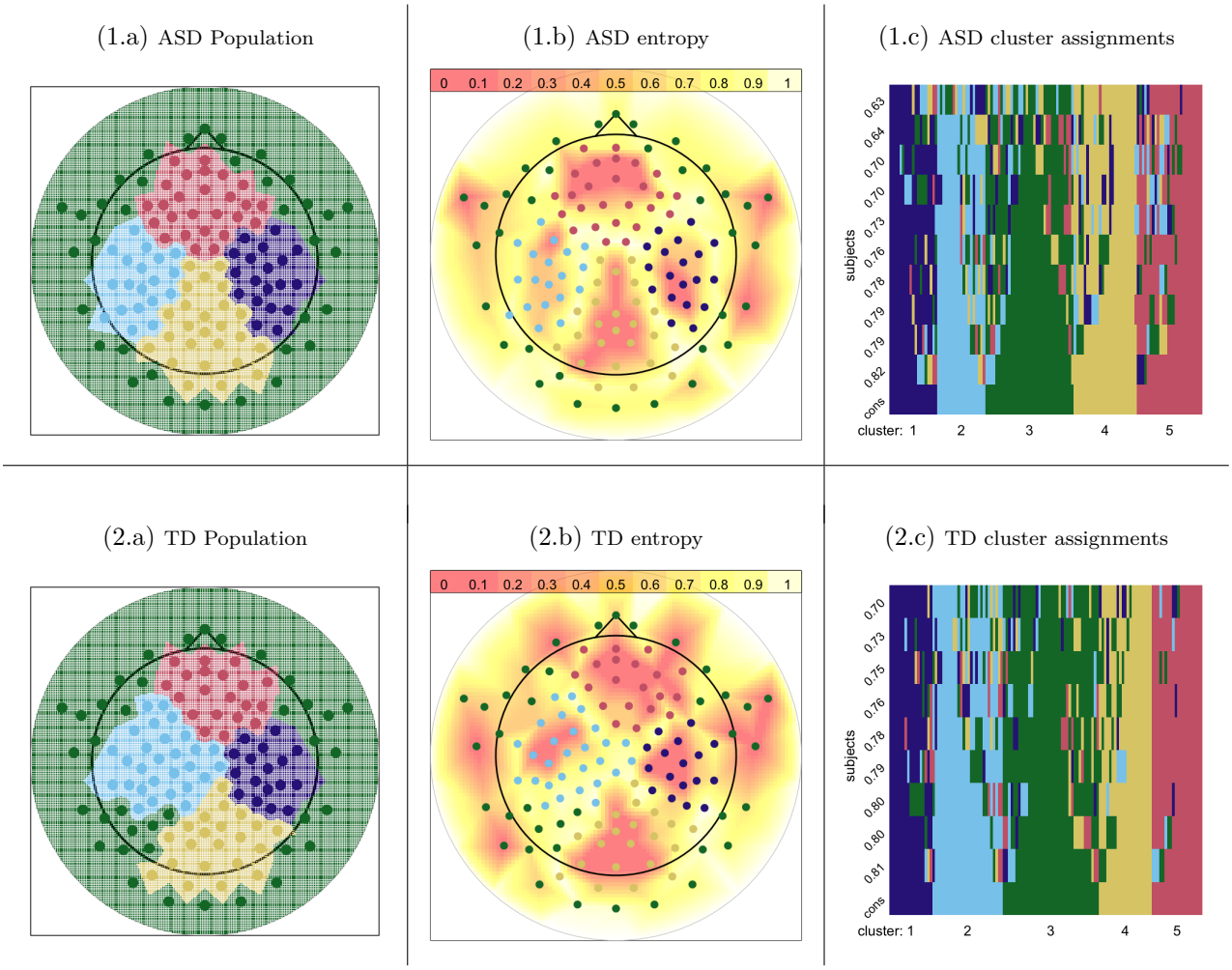

Figure 2.4: **Group contrasts, ASD (1) vs TD (2)**: (1.a) TD-cohort posterior least square synchronicity. (1.b) TD-cohort normalized posterior entropy. (1.c) TD-cohort subject- and population-level cluster assignments. (2.a) ASD-cohort posterior least square synchronicity. (2.b) ASD-cohort normalized posterior entropy. (2.c) ASD-cohort subject- and population-level cluster assignments. In the (c) panels, we report consensus labels as the last row. Subject-level labels are reported in each row, together with posterior median estimates of cluster adherence.

CHAPTER 3

Regionally Referenced Spectral Power Dynamics of EEG Signals: A Hierarchical Functional Modeling Approach

3.1 Introduction

The human brain is in central command to all neural activities. Either driven by individual neurons or their ensemble interactively, these activities see their involvement in every single aspect of the central nervous system, therefore affecting human behaviors like sleep, motor coordination, memory formation and perception ([BD04], [Kli99]), as well as pathological conditions, such as Parkinson's disease, Epilepsy and Autism ([BDD09], [US10], [MWG07]). To effectively observe it, one could study the macroscopic oscillations that were projected onto the scalp, in the form of electrophysiological signals, and record them by means of electroencephalography (EEG). Therein, EEG is widely recognized for research in investigating the *brain dynamics*, and for identifying causes and characterizing mechanism that give rise to the intrinsic oscillations in the appearance of electrical signals ([Bas04]), under various biomedical settings ([Tep02]).

Within the recent decades, EEG has seen more of its applications analytically and quantitatively by means of multi-channel recording on a geodesic net. This transition was both shaped by and benefitted from the shift in research focus, from localized observation to holistic assessment, when single-neuron neurophysiology faded in the shadow of quantitative EEG on the entire scalp. It has also revolutionized the research paradigm in the sense that the human brain is now considered an integrative system ([Bas12]), and meanwhile,

researchers have an extra dimension of the spectral features with fine grained resolution to explore on. Currently, typical multi-channel EEG research is geared with EEG-net, of up to 256 electrodes, that collects signals in milliseconds. Depending on the purpose of the study, event-related potentials (ERP) and event-related oscillations (ERO) are well investigated with focus on time- and frequency-domain respectively by means of ingeniously engineered features.

Despite the abundance of information provided, common practice still perform simple statistical tests for inference instead of characterizing brain activity by means of electrophysiological trajectories ([FF10], [JKR15]). Take our motivating example for instance, language acquisition, a critical cognitive process that relates to a number of developmental disorders, happens as early as 6 months old for infants, when they start to parse continuous stream of speech effectively and actively ([Kuh04]). Nevertheless, neither verbal instructions nor behavioral evaluations can be performed on toddlers, therefore EEG has greatly been appreciated under a lab environment to effectively and economically collect their responses. By means of an analysis on these measurements, research scientists hope to unravel the neural substates of this cognitive process. To do that, EEG oscillations are aggregated by frequency bands, namely Delta (0-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz) and Gamma (30-50 Hz), based on which extensive researches have been performed to simply compare the band metrics at multiple experimental stages. Findings include: 1) augmented Gamma power as relating to better performance in cognitive tasks and language acquisition ([GCB11] [FPM04]); 2) selective lateralization in frontal and temporal regions are observed during language-related tasks ([DOG74], [MCN93]), etc.

Even though desirable, full trajectory analysis of EEG band powers poses challenges in many folds. Most importantly among all, EEG recordings suffer from poor Signal-to-Noise Ratio (SNR), since by nature, all activities that result in electrophysiological responses are captured including neuronal activities on the scalp, muscle movements (myogenic potentials), eye movements and other artifacts. Worse off, to obtain an accurate estimate of power spectral density (PSD), which proceeds band aggregation, we need long pieces of electrical signals from a stationary process which are barely encountered in practice. On the analytical

side, it falls short on the flexibility of the modeling strategy to handle a complex covariance structure underneath. Traditional statistical methods do not scale up on densely collected functional observations, besides, we need more inferential tools for meaningful scientific findings in a rigorous way.

Recently, remarkable efforts have been made from the domain of Functional Data Analysis (FDA) by treating each electrophysiological trajectory as a function of interest. Under a regression framework, [Guo02] first introduced functional mixed effects models using smoothing splines, where both fixed and random effects are characterized in the same functional space. [CW11] maintained the flexibility of Guo's work, and further decomposed variation by intra- and inter-subjects components by their covariance structures. From a Bayesian perspective, [CRC07] extended the nonparametric modeling to the variance function as a variance process characterized by P-splines, therefore gave extra freedom in handling covariance structures along multi-dimensions. Additionally, a reduced rank structural assumption is utilized to approximate functional covariances more economically. Among all great works, we found both [ZHM10] and [MTN12a] are intriguing with a difference of using either leading principle components from a frequentist perspective, or shrinkage priors on the loading matrices of latent factors under a Bayesian framework, for an efficient approximation.

In this chapter, we present a Bayesian hierarchical functional model that decomposes functional observations flexibly in both the first and second moments: group mean and individual curves are estimated in the same functional space; multi-way functional covariances are flexibly handled with an automatic regularization of the redundant entries in a latent factor decomposition. Theoretically, the proposed framework is suitable for nonparametric estimation with any projection methods and it permits covariance structures that are non-separable between the functional and spatial dimensions. Our approach is fully Bayesian and accompanied with a rich set of inferential tools directly applicable to the Markov Chain Monte Carlo samples, all of which are highly optimized in a Rcpp package. We combine our modeling efforts together with a carefully engineered band power metric that references noisy spectral features to a robust regional summary, and illustrate the benefits of such innovations through an analysis of brain dynamics on a EEG dataset collected from a cognitive

experiment at UCLA. Our work conforms with and amplifies the existing findings in psychiatry with abundant and dynamic information on the learning process, and provide further insights to a more fundamental understanding of Autism from a statistical perspective.

The chapter is organized as follows. In Section 3.2, we lay out background information of the motivating example and a outline of the meta-preprocessing steps, including the Regionally smoothed Eigen band powers (Rs-EP), and its alternatives. The hierarchical functional model is introduced and elucidated under a fully Bayesian framework, in Section 3.3. We assess the performance of the proposed model on engineered datasets in Section 3.4, and report the results and findings by performing an analysis on the speech stream exposure data in Section 3.5. We conclude with final remarks and potential extension in Section 3.6.

3.2 Defining Regionally Referenced Spectral Power Dynamics

3.2.1 Research Background

Our motivating example of EEG recordings is collected by our collaborators at UCLA on 3 groups of young children over the course of a speech stream exposure task. One group of the children are typically developing (TD), and the other 2 groups are diagnosed with autism spectrum disorder (ASD) at different language level: verbal (v-ASD) and minimally verbal (mv-ASD). ASD is a developmental condition that affects individuals in two domains: communication and social interactions ([LRL00]). It also features late development of linguistic skills compared to typically developing (TD) children ([EMS11]). Despite the wealth of information uncovered about ASD, little is known about the 30% of the ASD population who remain minimally verbal. The main goal of this study is to compare and contrast EEG oscillation properties, i.e. band powers, between children in the mv-ASD, v-ASD and TD diagnostic groups, to infer their brain dynamics when processing linguistic inputs.

3.2.2 EEG Recordings from the Speech Stream Exposure Task

In our working example, 9 TD, 14 v-ASD and 19 mv-ASD children volunteered in the speech stream exposure task. Each participant listened to a stream of synthetic speech constructed with a collection of 12 different phonemes, of which certain pairs appeared deterministically but the others did not. In particular, as exemplified in Figure 3.1, two consecutive pairs that were deterministic in the order they appear consisted *words* such as **pa-bi-ku** and **da-ro-pi**, whereas the phoneme pairs in between *words* were indeterministic, therefore could not be expected, such as **ku-da** in block 1 and **ku-go** in block 2. All 4 *words* are pseudo-randomly permuted and repeated into blocks, to balance the frequency that indeterministic pairs appeared in the whole speech, then further repeated in the units of blocks. As hypothesized by our collaborators, the repeated co-occurrence of phoneme pairs facilitating the parsing of a continuous speech, or equivalently word segmentation, will provide substantial insights on this cognitive process, therefore hopefully, language related developmental disorders and delays frequently observed among autistic children. To capture underlying neuronal activities effectively, the experiment lasted about 144 seconds, during which EEG data were recorded for each participant at 128 electrodes and pre-processed in NetStation. The standardized pre-processing steps outlined in the Appendix B, resulted in segmented pieces of EEG recordings, each of 1 second in duration, and segments tagged as *inattention*, which flagged time windows when the participants were not complying with the requirements, were excluded for statistical analysis.

Previous studies have acknowledged differentiated neurophysiological responses across regions and oscillation frequencies, over the course of the experiment. Using functional magnetic resonance imaging (fMRI), which measures haemodynamic responses within brain, [MMD06] reported extensive activation in the temporal, frontal and parietal regions. In addition, converging evidences using multiple neuro-imaging techniques suggest that, in the left hemisphere, the posterior regions are involved in the storage and retrieval of phonological information, whereas the anterior regions are involved in segmentation, phonological assembling and word production ([Bur01], [BSB00], [ZE92]). Specific to EEG studies,

Gamma oscillations associated with information processing were noticed significantly increasing in power shortly after the onset of auditory stimulus in a segmentation task according to [KKP98]. Irregular activities, for example, slow increase of Delta and Theta band powers in the frontal and right temporal regions were detected among individuals with learning deficits, as contrasted against the control group ([APB07]). In this chapter, we investigate EEG oscillations by means of regionally referenced frequency-band power estimates, with explicit emphases on their dynamic changes in the previously recognized region, on three diagnostic groups and their contrasts.

3.2.3 Regionally Referenced Spectral Power Estimates

Multi-channel EEGs are highly structured and high-dimensional, however, considerable amounts of similarity exhibits in the neighboring electrodes and segments from both time and frequency domain. It has been a common practice to collapse ERP characteristics and EEG spectral powers over adjacent segments, to achieve an amplified Signal-to-Noise Ratio (SNR) on pre-processed signals. However, this strategy operates on the redundancy of segments available post to pre-processing. The rejection percentage ranging between 10% and 90% with an average of 50%, in our working example, forced our attention of dimensionality and noise reduction among neighboring electrodes. Specifically, electrodes are clustered into 11 anatomical regions of interests on the scalp, i.e., right/mid/left frontal (RF/MF/LF), right/left temporal (RT/LT), right/mid/left central (RC/MC/LC) and right/mid/left posterior (RP/MP/LP), therefore, power spectral density estimates are aggregated and referenced to predefined regions.

Within a specific region, we are interested in a zero mean p-dimensional stationary random process, $\boldsymbol{X}(t)$ $(t = 0, \pm 1, \ldots)$, which has a spectral representation:

$$\boldsymbol{X}(t) = \int_{-\pi}^{\pi} \boldsymbol{A}(\omega) \exp(\mathrm{i}\omega t)\, \mathrm{d}\boldsymbol{Z}(\omega),$$

where $\boldsymbol{A}(\omega)$ is a $p \times p$ Hermitian transfer function matrix and $\mathrm{d}\boldsymbol{Z}(\omega)$ is a zero mean and orthonormal random process. The spectral density matrix of $\boldsymbol{X}(t)$ is defined to be $\boldsymbol{I}(\omega) = \boldsymbol{A}(\omega)\overline{\boldsymbol{A}(\omega)}'$, where $\overline{\boldsymbol{A}}'$ is the transpose of the complex conjugate of $\boldsymbol{A}$.

Following [OH06], we define kernel smoothed spectral density matrices under the assumption of local stationarity on each segment. The smoothing bandwidth, that is allowed to vary by regions, is selected via a generalized cross validation (GCV) procedure (see Appendix B for details). Since we fix the smoothing bandwidth within each region, the corresponding smoothed spectral density matrix, $\tilde{\boldsymbol{I}}(\omega)$, stays Hermitian and non-negative definite, therefore all of its eigenvalues are non-negative real numbers. To aggregate power spectral densities regionally, one could define:

- *Regionally smoothed Average Power (Rs-AP)* is the arithmetic mean of p smoothed spectral power densities, or equivalently,

$$P_{\mathrm{AP}}(\omega) = \frac{\mathrm{tr}(\tilde{I}(\omega))}{\sum_{\omega} P_{AP}(\omega)}. \tag{3.1}$$

- *Regionally smoothed First Eigen-Power (Rs-EP$_1$)* is the first leading eigenvalue associated with the smoothed spectral density matrix normalized by the total Rs-EP$_1$ over all frequencies,

$$P_{\mathrm{EP}_1}(\omega) = \frac{\max_{|x|=1} x^T \tilde{I}(\omega) x}{\sum_{\omega} P_{\mathrm{EP}_1}(\omega)}. \tag{3.2}$$

- *Regionally smoothed First k Eigen-Powers (Rs-EP$_k$)* are the sum of k leading eigenvalues associated with the smoothed spectral density matrix normalized by its total power,

$$P_{\mathrm{EP}_k}(\omega) = \frac{\sum_{j=1}^{k} \max_{x_j \perp x_1 \ldots x_{j-1}} x_j^T \tilde{I}(\omega) x_j}{\sum_{\omega} P_{\mathrm{EP}_k}(\omega)}. \tag{3.3}$$

It can be shown that, Rs-EP$_k$ contains both Rs-EP$_1$ and Rs-AP as special cases when $k \to 1$ and $k \to p$, where k indicates the equivalent reduced dimensionality of the p-electrodes region. Only k incoherent components, which are linear combinations of the p original time series, are used to summarize the regional spectral feature at each given frequency ω. Similar to coherence measure defined directly on the spectral density matrix $\boldsymbol{I}(\omega)$, Rs-EP$_1$ equals Rs-AP when complete redundancy, therefore perfect cross-coherence, exists in the multivariate signals that are linearly transferrable. For this reason, we consider Rs-EP$_1$ an approximation of Rs-AP when the electrodes are highly coherent in region. Practically, it

excels at capturing locally coherent oscillations whereas attenuating incoherent oscillations in their spectral profile, and consequently, Rs-EP$_1$ is our choice of band power metrics per region for the case study.

3.3 Regionally Referenced Functional Hierarchical Model

3.3.1 Data Projection

We set to model spectral power dynamics over cortical regions, as brain activity adapts to stimuli over the course of an cognitive experiment. Responses to the external stimuli are regionally referenced and follow a multivariate stochastic process, by our assumptions, that is continuous over time. More precisely, the underlying process denoted by $\boldsymbol{Y}_i(t) = \{Y_{i1}(t), \ldots, Y_{ip}(t)\}'$, a p-dimensional vector of band-specific power metrics, highlights the trajectory of subject i, $(i = 1, 2, \ldots, n)$, at p cortical regions $(j = 1, 2, \ldots, p)$ over time $t \in (0, T)$. In practices, only a finite number of observations are collected discretely at $\boldsymbol{t}_i = \{t_{i1}, t_{i2}, \ldots, t_{im_i}\}$. Therefore, for ease of notation, we maintain a lighter indexing t, that varies on a common set of observed time points.

The ensuing assumptions admit a decomposition with a functional component and error terms of the observed band-powers,

$$\boldsymbol{y}_i(t) = \boldsymbol{f}_i(t) + \boldsymbol{\epsilon}_i(t), \quad \boldsymbol{\epsilon}_i(t) \sim N(0, \Sigma_\epsilon); \tag{3.4}$$

Residual covariance, $\Sigma_\epsilon = \text{diag}\{\epsilon_{1\epsilon}^2, \ldots, \epsilon_{p\epsilon}^2\}$, allows for independent heteroscedastic regional error that can be induced either by band-power processing or system noise. The functional component $\boldsymbol{f}_i(t)$ has its dependency encoded in two-way, namely regional and functional. With its regional dependency, it highlights the coordinated cortical activity instantaneously, whereas with the functional dependency, the underlying process is characterized smoothly with long and short term correlations.

To flexibly characterize the multivariate functional component, we project each of its element $f_{ij}(t)$, $j \in \{1, \ldots, p\}$ onto a functional space spanned by q cubic B-spline basis

$\{b_1(t), \ldots, b_q(t)\}$, such that,

$$f_{ij}(t) = \sum_{k=1}^{q} \theta_{ijk} b_k(t), \quad \text{or equivalently,} \quad \boldsymbol{f}_i(t) = \sum_{k=1}^{q} \boldsymbol{\theta}_{ik} b_k(t),$$

where $\boldsymbol{\theta}_{ik} = (\theta_{i1k}, \ldots, \theta_{ipk})^T$ is p-dimensional B-spline coefficients.

So far, we have projected band-power observations as well as the encoded functional dependency into a pre-defined functional space. And this first stage of our model is called *data projection*,

$$Y_i = \Theta_i B_i' + E_i, \tag{3.5}$$

with $\Theta_i = [\boldsymbol{\theta}_{i1}, \ldots, \boldsymbol{\theta}_{iq}]$ being a $p \times q$ coefficient matrix, $B_i = (b_k(t))_{tk}$ the $n_i \times q$ projection matrix and residuals $E_i \sim \mathcal{MN}(0, \Sigma_e, I_{n_i})$ follow a matrix normal distribution. Under this projection, both regional and functional dependency are preserved and encoded with unified and reduced dimensionality: $p \times q$.

3.3.2 Hierarchical Priors on Projected Coefficients

We proceed with hierarchical modeling of Θ_i in the projected functional space. Under a typical regression framework, one often collects subject-level covariates $\boldsymbol{x}_i$, characterizing study participants through a set of quantitative and/or qualitative variables. In this regard, the main objective of typical scientific investigations involving EEG imaging can be conceptualized as a characterization of the conditional expectation, $E\{\boldsymbol{y}_i(t) \mid \boldsymbol{x}_i\}$, or equivalently in the functional space $M(\boldsymbol{x}_i) \coloneqq E\{\Theta_i \mid \boldsymbol{x}_i\}$, while accounting for appropriate dependencies in the observed data.

Under a familiar mixed effects regression framework for Θ_i, one could construct both fixed and random effects in matrix notations,

$$\Theta_i = M(\boldsymbol{x}_i) + Z_i, \quad \text{vec}(Z_i) \sim \text{MN}(0, \Sigma_z); \tag{3.6}$$

where $M(\boldsymbol{x}_i)$ represents population-level projected dynamics associated with $\boldsymbol{x}$ as the fixed effect, and Z_i features subject-specific variations as random effect. The covariate-adjusted group mean trajectories are main quantities of inferential interests, and individual deviations should respect heterogeneity with tremendous flexibility.

Indeed, the model in (3.6) induces explicit assumptions about regional and functional dependence via Σ_z. For instance, when modeling multi-way functional covariances under the assumption of a *separable* structure, $\Sigma_z \propto \Sigma_p \otimes \Sigma_q$, it directly reflects at the observational level,

$$\text{cov}(y_{ij}(t), y_{ij'}(t')) = \Sigma_p(j, j') \cdot \boldsymbol{b}(t)^T \Sigma_q \boldsymbol{b}(t'),$$

that overall covariance is essentially an outer product between regional and functional covariances.

All centered around (3.6), we introduce three modeling approaches that correspond to hierarchical prior configurations on Θ_i. To be more specific, a) a Naive Bayesian(NB) prior inducing separable covariance consists of a constant smoothness penalty and a flexible regional covariance; b) a Strongly Separable(SS) prior originated from a Bayesian infinite factor model; c) a Non-Separable(NS) prior that operates on $vec(Z_i)$ in a similar approach as SS using latent factor characterization, however with minimal restrictions on the covariance separability.

3.3.2.1 Naive Bayesian Prior

Our baseline model correspond to a Bayesian P-spline approach that is essentially a matrix normal prior on Θ_i, with strongly separable constant functional smoothness penalty and totally free regional covariance. To be precise,

$$Z_i \sim \mathcal{MN}(0, S, \Omega_0^{-1}), \qquad S^{-1} \sim W(\nu, S_0^{-1}/\nu). \tag{3.7}$$

In this parametrization, the flexibility totally hinges on regional dependency via the covariance matrix S, which indexes the column-wise covariance of the subject-level random effects Z_i. For completion, we consider a conjugate Inverse Wishart (IW) prior on S, and $E(S^{-1}) = S_0^{-1}$ a priori. In this setting, S_0 may be chosen to represent prior knowledge on the proximity of cortical regions through conditionally autoregressive representations ([GV03]).

On the other hand, the constant regularization matrix Ω_0 are constructed to achieve adaptive smoothing at the subject-level, by following a strategy similar to ([LB04]). At a

region j, the adaptiveness reflects on its region-specific regularization $D_j = s_{jj}^{-1}\Omega_0$ which also relieves the identifiability concern that commonly appears troublesome for multi-way covariance estimation. The choice of Ω_0 is based on an AR(1) process, and more details are documented in the Appendix B.

3.3.2.2 Strongly Separable Prior

We propose a two-way Bayesian latent factor model on Z_i which also coincide with the *strongly separable* covariance structure by [CL17]. First introduced by [BD11a] and further extended to functional data by [MTN12b], Bayesian latent factor model offers extensively flexibility in modeling structured covariance with minimal restrictions. Our work extends [MTN12b] to multivariate functional observations by introducing individual factor loading on both functional and regional dimensions, at the same time, preserves all merits to the original approach including sparsity of factors loading matrices and scalable sampling of factor loadings. More specifically, the two-way multiplicative shrinkage prior on Z_i is defined by a decomposition,

$$Z_i = \Upsilon H_i \Gamma' + R_i, \qquad \text{with} \quad R_i \sim \mathcal{MN}(0, \Sigma_p, \Sigma_q), \tag{3.8}$$

where $\Upsilon = \{\upsilon_{jr}\}$ and $\Gamma = \{\gamma_{ks}\}$ highlights regional and functional factor loadings respectively, that by their combination interacts with a matrix of latent fact scores, H_i, whose components are i.i.d $N(0,1)$. Furthermore, coefficient residuals R_i has separable covariances, Σ_p and Σ_q, both are diagonal with heteroscedastic variances. Similar to the multiplicative shrinkage prior (MGPS by [BD11a]), we complete with,

$$\begin{aligned}
\upsilon_{jr} \mid \phi_{jr}, \tau_r \sim \mathrm{N}(0, \phi_{jr}^{-1}\tau_r^{-1}), \quad \phi_{jr} \sim \mathrm{Ga}(\nu_1/2, \nu_1/2), \quad \tau_r = \prod_{u=1}^{r} \delta_u, \\
\delta_1 \sim \mathrm{Ga}(a_{11}, 1), \quad \delta_u \sim \mathrm{Ga}(a_{12}, 1), \quad \text{when } u > 1; \\
\gamma_{ks} \mid \psi_{ks}, \kappa_s \sim \mathrm{N}(0, \psi_{ks}^{-1}, \kappa_s^{-1}), \quad \psi_{ks} \sim \mathrm{Ga}(\nu_2/2, \nu_2/2), \quad \kappa_s = \prod_{v=1}^{s} \pi_v, \\
\pi_1 \sim \mathrm{Ga}(a_{21}, 1), \quad \pi_v \sim \mathrm{Ga}(a_{22}, 1), \quad \text{when } v > 1;
\end{aligned}$$

When the residuals R_i's are disregarded, the covariance of Z_i conditional on both factor loadings rewrites,

$$\text{cov}(Z_i \mid \Upsilon, \Gamma) = \Upsilon\Upsilon^T \otimes \Gamma\Gamma^T,$$

therefore strongly separable. It can also be regarded as a two-way generalization of [JHS00] with reduced-rank approximation to $\text{cov}(Z_i)$ by means of a latent factor decomposition. Therein, we call the proposed two-way MGPS prior a **strongly separable (SS) prior** on projected coefficients Θ_i.

3.3.2.3 Non-Separable Prior

To lift the constraints on separability, we propose an alternative to two-way MGPS by imposing the latent factor model on $vec(Z_i)$, which we refer to as block multiplicative shrinkage prior. The vectorization gives us freedom to exceed beyond linear characterization and decomposition of Σ_z, while scalability persists since the building blocks of the construction is bounded by regularizations of size $p \times p$ or $q \times q$. To be specific, the vectorized Z_i is decomposed in a similar way as [BD11a],

$$\text{vec}(Z_i) = \boldsymbol{\zeta}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{r}_i, \qquad \text{with} \quad r_i \sim \text{MN}(0, \Sigma_p \otimes \Sigma_q), \tag{3.9}$$

where the residuals stay the same, $\boldsymbol{r}_i = vec(R_i)$, the factor loading matrix Λ is $pq \times m$ and the latent factors $\boldsymbol{\eta}_i$ has an i.i.d standard normal similar to H_i. However, this prior differs from 3.8 by the hierarchical priors on Λ,

$$\boldsymbol{\lambda}_{(k)}^{(c)} \mid \Omega_k, \tau_c \sim \text{MN}(\mathbf{0}, \Omega_k^{-1}\tau_c^{-1}), \quad \Omega_k \sim \text{W}(\nu, W_0), \quad \tau_c = \prod_{l=1}^{c} \delta_l,$$
$$\delta_1 \sim \text{Ga}(a_1, 1), \quad \delta_l \sim \text{Ga}(a_2, 1), \quad \text{when } l > 1;$$

where $\boldsymbol{\lambda}_{(k)}^{(c)}$ indicates the k-th block, $k \in \{1, \ldots, q\}$, of the c-th column of Λ. As directly compared to 3.8, the shrinkage operates jointly on all regions via Ω_k which is the same across all latent factors. As a matter of fact, Σ_z can be represented regardless of the residuals $\boldsymbol{r}_i$,

$$\text{cov}(\boldsymbol{\zeta}_i \mid \Lambda) = \Sigma_z = \Lambda\Lambda^T,$$

therefore non-separable in essence. This block multiplicative shrinkage prior exploits on factor loading matrices and allows for factor-coefficient specific regularization by blocks. Considering the encoded assumption on the covariance structure, we call it a **non-separable prior** on projected coefficients Θ_i.

3.3.3 Linking the Priors

The Naive Bayesian prior distinguish from the other two with a constant regularization, as well as the lack of residual component R_i that can be treated as an extra level of adaptive smoothing. Even though it appears naive and simple, we find this approach surprisingly effective in reconstructing the fixed effects in simulation. By including this prior as the baseline approach, we aim to evaluate the effectiveness of the low-rank approximations, as well as the benefits of addressing non-separability on Σ_z.

At a second look, we can bridge the Strongly Separable (SS) and Non-Separable (NS) priors by revisiting the latent factor decomposition. Let us rewrite the loading matrices with their column vectors, namely $\Lambda = \{\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(m)}\}$, $\Upsilon = \{\boldsymbol{\upsilon}^{(1)}, \ldots, \boldsymbol{\upsilon}^{(k_1)}\}$ and $\Gamma = \{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(k_2)}\}$. Therefore,

$$\text{SS-prior:} \quad Z_i = \sum_{s=1}^{m} \eta_{is} \text{mat}(\boldsymbol{\lambda}^{(s)}),$$

$$\text{NS-prior:} \quad Z_i = \sum_{r=1}^{k_1} \sum_{s=1}^{k_2} \eta_{irs} \boldsymbol{\upsilon}^{(r)} \boldsymbol{\gamma}^{(s)'},$$

such that when $\text{mat}(\boldsymbol{\lambda}^{(\cdot)}) = \boldsymbol{\upsilon}^{(\cdot)} \boldsymbol{\gamma}^{(\cdot)'}$ and $m = k_1 * k_2$, the Non-Separable prior includes the Strongly Separable prior as a special case. Indeed, both priors induces a reduced-rank structure on $\text{cov}(Z_i)$, the NS prior escalates above the outer product of two-way marginal covariances, and potentially, recognize more sophisticated structures of Σ_z than its latent factor alternative.

3.3.4 Posterior Inference

Posterior inference through Markov Chain Monte Carlo sampling is implemented and realized through Gibbs Sampling for all three proposed priors. Detailed calculation of full conditional

distributions are reported in the Appendix B, for completeness. A highly optimized R package for data manipulation and inference is freely available on author's github page.

Once posterior samples are obtained for all parameters in the model, estimation and inference on functionals of interest is relatively straightforward. For the fixed effects, we follow [KKC10] to obtain simultaneous confidence bands on population-level trajectories, as well as the posterior means; for the random effects of individual functional estimates, only the posterior means are reported. Further details are reported in Web Appendix B.

3.4 Experiments on Engineered Data

This section describes and evaluates the finite sample performance of the proposed functional mixed-effects model, on Monte Carlo samples that are generated both strongly separable and non-separable. We engineer our simulated datasets with comparable structure as the working example, i.e. multivariate functional observations are collected on groups of individuals, where group trajectories are of major interests as first moment estimates and covariance of random effects highlight structured regional and functional dependency via the second moment estimators.

Following the data project in (3.5), we encode the (non)separability in B-spline coefficient matrix Θ_i. Specifically, let's denote Σ_{ss} and Σ_{ns} as $\text{cov}(vec(\Theta_i))$ separately for strongly separable and non-separable priors,

$$\Sigma_{ss} = S_0 \otimes \Omega_0^{-1},$$
$$\Sigma_{ns} = \Lambda\Lambda' + \Sigma_0.$$

Given the covariance of random effects, we sample Θ_i marginally for strongly separable prior, $\Theta_i \sim \text{MN}(\mathbf{0}, \Sigma_{ss})$, however conditionally for non-separable priors,

$$\boldsymbol{\eta_i} \sim \text{MN}(\mathbf{0}, I); \qquad \text{vec}(\Theta_i) = \Lambda\boldsymbol{\eta}_i + \boldsymbol{r}_i.$$

To reflect the regional and functional dependency via Λ, we follow the generative assumptions introduced in Section 3.3.2.3 given the block-wise covariance hyperprior W_0. Notice that

even the mean is fixed, the spline specific covariance still yield non-separability since it is separately sampled. To be consistent, we assume all pre-determined covariances, S_0, Ω_0^{-1} and W_0, follow AR(1) structures that have the same parameter on each dimension (regional or functional). As an example in Figure 3.2(c), the strongly separable covariance is more structured than its alternative with significant amount of dependency between nearby regions and B-spline coefficients encoded on adjacent knots.

The group mean functionals are selected from a pool of parametric curves, including constant, polynomials and trigonometric functions. Each group has its own regional mean trajectory and can be differentiated at each region (see Figure 3.2). To evaluate on the proposed approaches, we consider performance metrics of:

- average integrated mean square error (**aiMSE**):

$$\text{aiMSE}(\hat{\boldsymbol{f}}) = \text{avg}\left(\int_t (\hat{\boldsymbol{f}}(t) - \boldsymbol{f}(t))^2 \, \mathrm{d}t\right);$$

- average coverage of simultaneous confidence band (**a.cover**): per region, set to 1 if $\forall t$ the band covers $f_j(t)$;

- Frobenius norm between the estimated and true covariance of $\text{vec}(\Theta_i)$(**fCOV**):

$$\|\Sigma_\theta - \widehat{\Sigma}_\theta\|_f$$

Further details regarding the data generation schemes are available in the Appendix B, and all results presented can be reproduced with our R package and supplementary documents.

3.4.1 Synthetic Data: Strongly Separable and Non-Separable

We generate simulated samples for both strongly separable and non-separable covariance. For each dataset, 60 individuals evenly split into 3 groups have their functional observations collected over 6 regions. Since we simulate the projected B-spline coefficients, instead of functional observations, the degrees of freedom for B-spline bases is set to 12. The discretized observations are defined on a common grid $\in [0, 1]$, however, may randomly be discarded

with probability between 0 and 0.8. This missing percentage is fixed per dataset as p, and each simulated time points is assigned with binary label generated from $Bernoulli(p)$.

To perform a comprehensive assessment, we evaluate all three models under various scenarios. First, we consider the degrees of freedom for B-splines are unknown. Since the fCOV metric requires the estimation defined on the same space as the truth, we only report aiMSE and a.cover accordingly. Secondly, regional dependency formulated under an AR(1) process is examined at three levels, $\rho = 0.4$, $\rho = 0.6$ and $\rho = 0.8$. And finally, all three models are evaluated on each of the 500 simulated dataset, under every scenario previously mentioned, and we obtain the posterior mean estimates, 90% confidence bands on parameters of interests.

Results: The group mean trajectories are recovered consistently well by all three models, where non-separable condition poses further challenges than strongly separable condition, yielding almost doubled aiMSE overall. Indeed, each individual function is generated in a less constrained way such that estimation at the group level becomes essentially more difficult. In terms of average coverage, Naive Bayes approach is pervasively conservative and covers the true curves over 97% against the specified level of 90%. When the true degrees of freedom is unknown, all three models demonstrate sufficient robustness on both metrics, and under both simulation settings.

We highlight some enlightening findings in Table 3.1 and withhold the extensive report in the Appendix B. Under mild regional dependency setting, i.e. $\rho = 0.4$, both Strongly Separable (SS) prior and Non-Separable (NS) prior outperform the Naive Bayes (NB) prior by averaged coverage and covariance reconstruction. More interestingly, NS prior exceeds both SS and NB priors in covariance recovery when functionals are more spatially correlated, $\rho = 0.8$ in contrast to $\rho = 0.4$, and the bias metric inflated only 8% compared to 49% of SS prior and 61% of NB prior. Bearing in mind the stringent assumptions of functional covariance, which is proportional to a constant matrix, we imposed on NB prior, the comparison between NS and SS priors becomes the spotlight. For the non-separable simulations, NS and SS are mainly distinguished by the measure of fCOV, where NS has been consistent regardless of regional dependency (ρ) and outperforming SS. Different from the

separable simulations, SS prior estimates the covariance structure more accurately under higher regional correlation and a possible explanation is that more dependent structures impose less challenge under non-separable setting, i.e. more information can be borrowed locally, therefore some efficiency gain can be expected.

3.4.2 Synthetic Data: Sample size, SNR and Number of Latent Factors

A second batch of simulation studies aim to assess the operative characteristics of all three models under different Signal-to-Noise Ratios (SNRs), sample sizes and unknown number of latent factor settings. More specifically, we first examine under SNR= 0.2, 1.0 and 5.0 with a sample size of $N = 10, 20$ and 50 per group (3 in total). Both strongly separable and non-separable simulations are assessed with 100 repeated datasets per configuration. Secondly, we focus on the robustness to the number of latent factors when the true value is not given. Our experiments show that NS prior benefits the most with accruing sample size and SNRs, and its robustness and efficiency stands out under both simulation settings.

From our experiments we conclude that Non-Separable prior exceeds in the recovery of covariance estimation for random effects, and at the same time, yields satisfactory and robust estimation of the fixed effects trajectories under a broad range of settings. Therein, we recommend this implementation in practice, especially when separability cannot be hypothesized on an unknown structure of the random effects covariance.

3.5 Implicit Auditory Learning in a Study of Autism

Our case study follows the paradigm and metrics investigated in Section 3.2 to highlight the regionally referenced spectral power dynamics using our Hierarchical Functional Model. As a common practice, logarithm of power metrics are of major interest and we mainly emphasize on the **gamma** and **alpha** band oscillations for inference. We provide technical background and pre-processing pipelines in the Appendix B. Considering established robustness and efficiency from numerical experiments, we choose the non-separable model (NS-prior) that can flexibly handle both separable and non-separable conditions for inference. Under a

regression framework, our investigation provides a holistic view of regionally referenced brain activities under implicit learning research paradigm.

3.5.1 Group Mean Trajectory Analysis

Given all its manifestation at elder ages, for example social communication deficits, presence of repetitive behaviors and restricted interest, we want to discover what Autism means at a micro-level. Although the quantity of interest throughout this case study is the logarithm of Rs-EP$_1$ defined in Eq 3.2, our approach is readily generalizable to any self defined measures repeated observed at multiple locations on the scalp. The only restriction is that we require repeated observations are collected on the same grid of time points per individual, which practically makes sense.

It is worth mentioning that we still need to pick B-splines related parameters and the number of latent factors for the NS-prior. We empirically picked 12 B-spline with knots selected from quantiles of the distribution of time, considering a relatively small collection (150) of unique time stamps. However, authors do recognize more systematic even automatic schemes for optimizing the number and the position of the knots (see for example [Rup02] and [FS89]), therefore recommend in practice for a large span of functional time collected. As for the choice of the number of latent factors, we've assessed the robustness of our model to it in the simulation study and further on the case study. Both results agree that our model is capable of automatically regularize redundant factors and we chose 8 factors as model input therein.

To start with the alpha (8-12 Hz) power which is reportedly associated with cognitive activities [FB14]. By group means, TD and v-ASD are universally higher than mv-ASD across all 11 regions, even though none of the difference is significant by the 90% simultaneous confidence bands. The difference between TD and v-ASD is subtle and the trajectory is flat relative to experimental time among all three groups. In other words, the cognitive functionality, indistinguishable between TD and v-ASD, does not indicate the process of implicit learning and identifies the mv-ASD cohort directly at baseline, or reseting state. We

further looked into the asymmetry of alpha powers by calculating the difference between the left and right regions of frontal, temporal, central and posterior respectively. The asymmetry is significant over time at frontal regions, diminished at temporal regions and negligible at central and posterior regions.

Gamma (30-50 Hz) band activities have been long recognized as correlate to implicit learning process, therefore are further investigated under our research paradigm. As show in Figure 3.3(1.a & 1.b), TD group has significantly higher gamma band power than vASD and mvASD, especially at early stage of the experiment. Meanwhile, it also indicates a trend of decreasing in gamma band powers over time among TD children. Similar finding by [GM02] reported a decreased response when stimuli recur which they projected to be linked to a 'neural savings' mechanism within a cell assembly representing an object, i.e. a word in our case. This pattern is majorly discovered at temporal regions, both left and right, partially at frontal regions. Moreover, we observe significant frontal asymmetry throughout the study at gamma band too, where TD group exhibits a higher degree however gradually flattens over time. Although [RW14] has pointed out this left-dominant asymmetry as a discriminative feature between TD and ASD cohort, for the first time, we confirm it with region references by means of power dynamics over experimental time.

3.5.2 Effects of Age and Verbal-DQ

Further, we examine both the alpha and gamma band powers dynamics as effects of age and verbal-DQ. In addition to a small sized sample we acquired, the distribution of demographics among three groups is remarkably unbalanced. To be specific, the TD cohort is significantly older (94.7 ± 28.8, in months) with hider vDQ (120.6 ± 11.4), the v-ASD cohort are younger (67.1 ± 58) with medium vDQ (89.3 ± 22.7) and mv-ASD has widely distributed age (85.6 ± 24.0) however significantly lower by vDQ (23.6 ± 10.9). Due to the colinearity between group and vDQ, we include only age and vDQ as covariates below.

As shown in Figure 3.3(2.a & b), the effects of age and vDQ are flipped on alpha and gamma band powers. For the alpha band, higher power is correlated with higher vDQ scores

marginally, and significantly greater than 0 at two small time windows at both ends of the experiment with 90% confidence. For the other hand, we observe positive correlation between gamma band power and age especially in the middle and close to the end of the exposure. We tend to take extra caution making a scientific interpretation out of a relatively small sample, however, it casts some light on the comprehensive inferential toolkit we offer when handling functional observations under a bayesian regression framework.

3.6 Discussion

In this chapter, we propose a Bayesian hierarchical functional model that utilizes latent factor decomposition on B-spline coefficients for the analysis of EEG band power dynamics under the context of a cognitive experiment. We provide a high-performance computation package that implements three types of priors that can flexibly handle both separable and non-separable covariance structure of correlated functional observations. The proposed approach has satisfactory operating characteristics under extensive numerical experiments, in addition to its inferential competency from the posterior samples.

Our approach exploits on a latent factor structure with automatic regularization, that ensures identifiability and numerical stability, is the first attempt of hierarchical modeling on non-separable multi-way functional data to the best of author's knowledge. It is worth mentioning that this approach can be readily extended to various non-parametric and parametric projection methofds, for example polynomials, cubic splines, wavelets and empirical projections like fPCA. The latent factor structure is also easily scalable under both separable and non-separable conditions to high-dimensional tensor structures, with thoughtful engineering on the decomposition and prioritization of vectorization.

The case study of regionally referenced alpha and gamma band powers relies on a revisit to the eigen-power from spectral PCA, which is shown less prone to poor SNRs especially in highly coherent regions. However, it is important to point out the sample size sets the true bottleneck to our major findings above. The effects of age and vDQ are less interpretable in face of high correlation with group, therefore, a comprehensive study on a larger cohort

of individuals with balanced demographics is surely needed before any scientific claim being made. Given all the limitations to the available dataset, our contribution is more noticeable in the regards of methodological development and innovation. We also need to point out that we consider our approach under the category of functional-to-scaler regression, therefore all the covariates are expected to be time-invariant after all.

As shown in Figure 3.3(2.c), it is also possible to investigate the regional and functional correlation from the posterior samples and we obtain a point estimate of regional correlation matrix, from the gamma power analysis. Via a visual inspection, all 11 regions of interest can be aggregated into two groups: {LF, RF, LT, MC, RC} and {RC, RT, MF, LC, MP, LP, RP}, which intersects on Right Central (RC) region. The inverse of this correlation matrix is often interpreted as brain connectivity (for example see [HLS10]). To obtain this quantity, it is much easier under the separable condition, namely $\widehat{\Gamma}\widehat{\Gamma}'$, however more challenging for NS-prior. One possible remedy is an ad-hoc construction of this inverse covariance matrix under certain sparsity assumptions, but we believe it more interesting to encode both the sparse inverse constraint and low-rank structure simultaneously and analytically.

There are a couple of potential directions authors find illuminating for future reference. First of all, a comparative study of the eigen-power based on spectral PCA is desirable. Secondly, we believe modeling the full power spectral density (PSD) is totally feasible under the current framework instead of separate analysis on the band aggregated powers. The challenge is that our new observational units are functional surfaces of PSD over time that are naturally correlated across regions. We have seen significant efforts in this endeavor using functional PCA, and expect more from a bayesian perspective. Lastly, more inferential devices can be implemented to take the full advantage of the posterior samples of the functional observations. For example, to test on the hypothesis of time-invariant effect:

$$H_0: \quad a_j(t) = C_j \quad \longleftrightarrow \quad H_1: \quad a_j(t) \neq C_j$$

A user-friendly implementation of the proposed method is available online as an Rcpp package at: `https://github.com/Qian-Li/HFM`

Table 3.1: **Simulated results**

	$\rho = 0.4$			$\rho = 0.6$			$\rho = 0.8$		
Prior	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV
	Separable Simulation								
NB	0.1779	0.9757	23.1002	0.1773	0.9791	28.3420	0.1764	0.9813	37.2262
SS	0.1807	0.8981	7.7684	0.1806	0.8911	8.8431	0.1807	0.8819	11.5938
NS	0.1811	0.8958	10.5123	0.1809	0.8960	11.1077	0.1807	0.8948	11.3615
	Non-Separable Simulation								
NB	0.3309	0.9776	230.8164	0.3300	0.9788	253.0962	0.3280	0.9856	290.8440
SS	0.3331	0.8759	119.2790	0.3326	0.8775	116.2758	0.3309	0.8669	104.6439
NS	0.3297	0.8507	77.1105	0.3296	0.8519	76.4085	0.3286	0.8509	76.8683

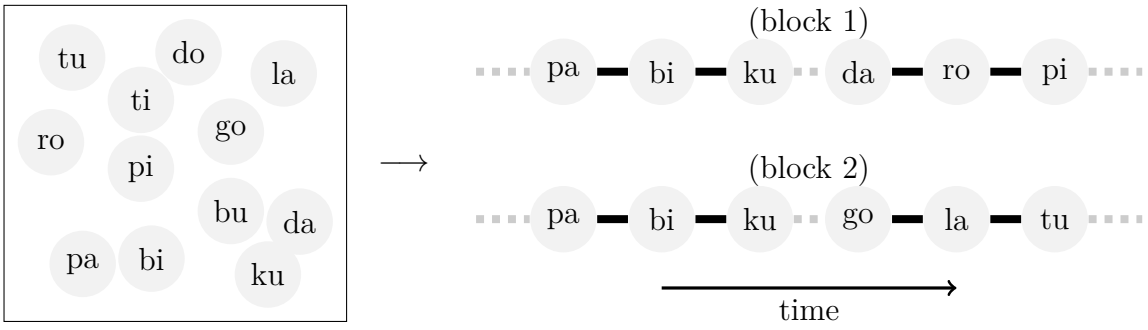

Figure 3.1: **Illustration of Word Segmentation Streams**: A Vocabulary consist of 12 syllables (phonemes) on the left; constructed tri-syllabic words are randomly concatenated into blocks on the right.

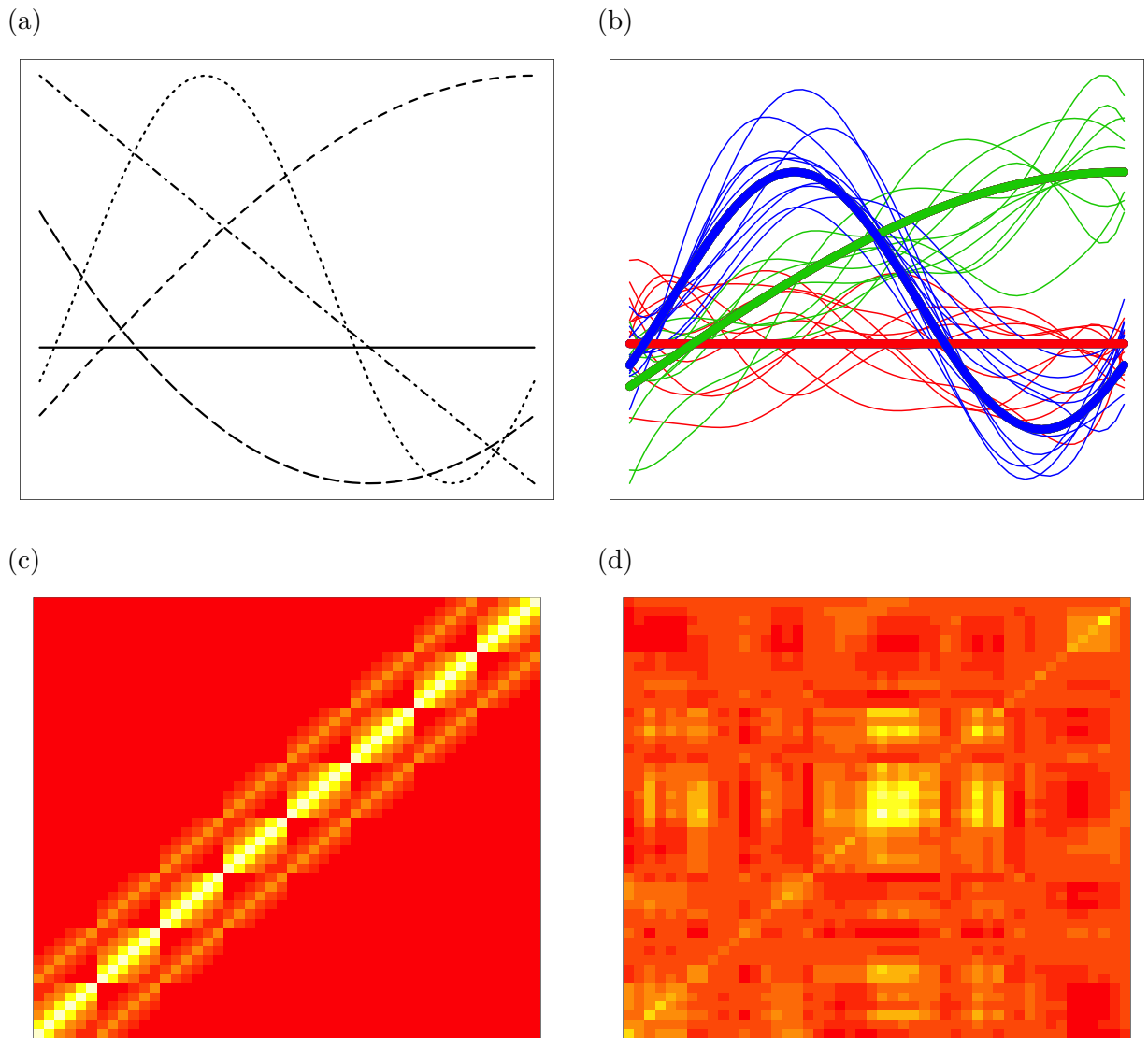

Figure 3.2: **Simulated curves**: a) Five group-regional means $\{\beta_0(t), \beta_1(t), \beta_2(t), \beta_3(t), \beta_4(t)\}$; b) Individual curves simulated from a multivariate Gaussian Process at one region; c) Simulated strongly separable covariance of Θ_i; d) Simulated non-separable covariance of Θ_i.

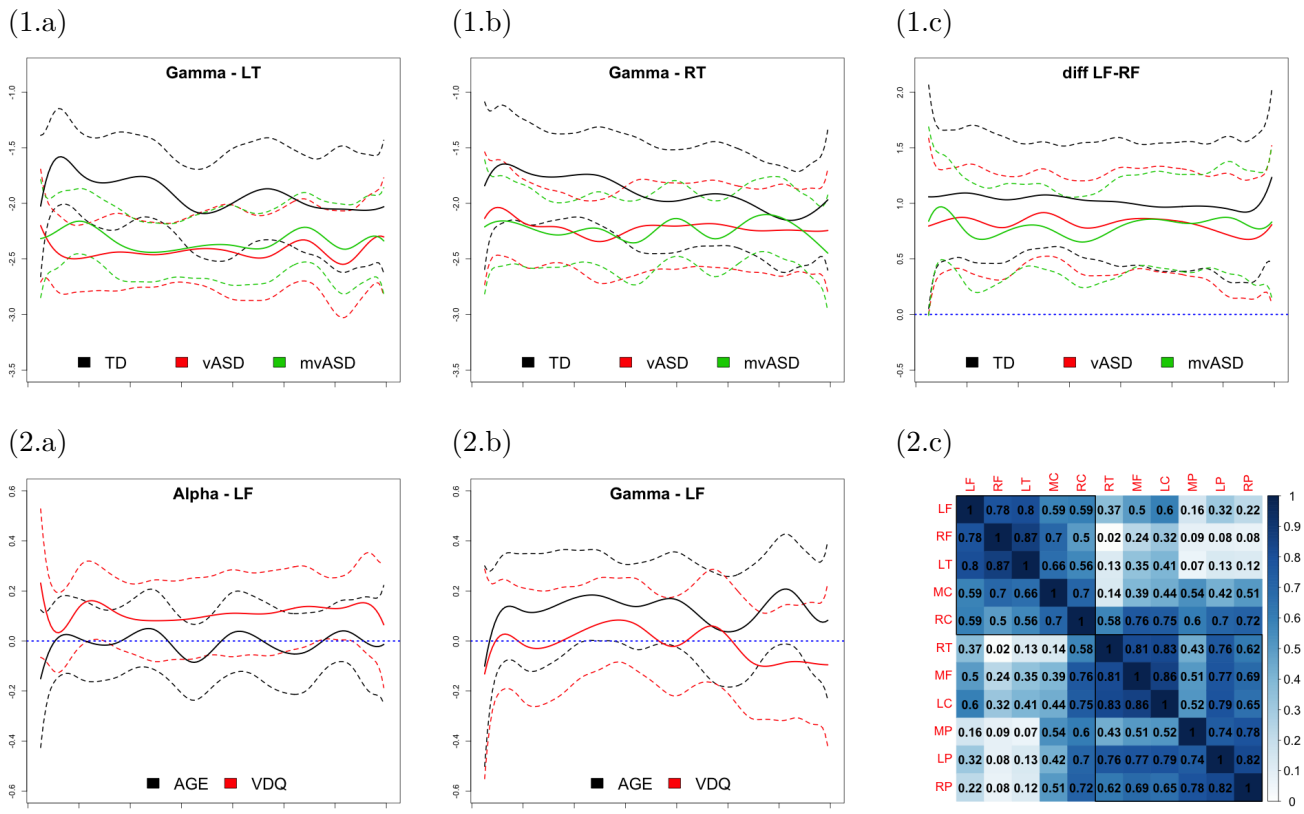

Figure 3.3: **Case Study Results**: Posterior group means and C.B. for Gamma band power at Left Temporal region (1.a), Right Temporal region (1.b) and contrasts between Left and Right Frontals (1.c); Posterior mean and C.B. for the marginal effects of AGE and Verbal-DQ at Left Frontal region: Alpha band powers (2.a) and Gamma band powers (2.b); Posterior mean of regional correlation of Gamma band power (2.c).

CHAPTER 4

Open Problems and Opportunities for Contributions

The contributions of this dissertation are likely to prove useful and provide a novel analytical view to the analysis of neuroimaging data collected in relation to neurocognitive experiments. In this Chapter, a discussion that has originated from the accomplished work of the previous chapters aims at the unrealized ideas and potential directions from where the author stands. Instead of peeks into the future, I consider all being reflections of the past on what could have been done from an angle of where more could have been seen.

4.1 Multimodal Neuroimaging

This dissertation is focusing on analytical approaches that are primarily engineered for EEG data. However, multimodal neuroimaging has gained increasing attention recently, which stands for a combination of imaging techniques (MRI, PET) and electrophysiological measurements (EEG, MEG). [LCL15] reviewed the recent advances in this endeavor specifically applications to neuropsychiatric disorders, however, very few contributions are seen from a statistical point of view. One key challenge that author noticed is that, multimodal techniques target distinct neuronal processes, therefore the measured responses are not fundamentally unifiable.

The integrative method from Chapter 2 has offered an alternative. In stead of seeking integration at observational level, it is plausible and viable to connect on some latent quantities and processes, for example clusters. Taking one step forward, the underlying process is mainly the target of interest, even though its representation per modality can be intractable. To tackle with this challenge, it requires a deep collaboration between research scientists and

statisticians, a sizable collection of multimodal data and a well-defined scientific problem of interest.

4.2 Joint Modeling of Exposure and Test Phases

The implicit learning paradigm in Chapter 3 has a two phases design: exposure and test phase. Participants get exposed to a continuous stream of artificially constructed linguistic input during the first phase, and receive "tests" of exposed and non-exposed words in the second phase. Methodologically speaking, analytics on phase one focus spectral features compared to waveform characteristics for the second phase. However, it is scientifically appealing to jointly model and interpret both phases simultaneously.

The statistical challenge in essence is a functional to functional regression problem. Taking the multivariate PSD dynamics from exposure phase as inputs, one could predict and/or interpret on ERP waveforms conditionally. It is strongly recommended to read [Mor15] for reference, where both functionals as predictors and outcomes are thoughtfully covered with great details.

4.3 Hierarchical and Latent Structures

Both hierarchical characterization and latent structure assumptions have seen their application in Chapter 2 and 3. As a matter of fact, it is just the tip of the iceberg how both techniques can benefit statistical modeling. Conceptually, hierarchical structure is a systematic simplification to a complicated hypothesis that often leads to numerical efficiency and direct interpretability. This dissertation has mainly dedicated to using this technique to utilize the structured dependency naturally observed in EEG data. More of its application are strongly encouraged, for example, author has also been working on a deep learning based multi-class image classification problem where hundreds of labels can follow a hierarchical structure directly encoded in the loss function during training time.

On the other hand, latent structure brings simplicity and sparsity from an alternative

perspective. The latent labels of clustering, also known as data augmentation, unifies and summarizes the exploratory findings at multiple levels, whereas the latent factor decomposition approximate a giant covariance matrix economically with a few factors. This highly organized and simplified conceptualization is signature to the creativity and originality of human been, therefore should be highlighted as valuable principles of statistical model building overall.

APPENDIX A

Supplementary files for Chapter 2

To company our main article in Chapter 2, more detailed formulations and computational procedures are described in this appendix. We structure this document as below: more technical details, including spectral estimation, Gibbs sampler and Information criteria are provided in Section A.1; detailed simulation of EEG signals along with extended results can be found in Section A.2; we conclude with EEG measurements and MCMC assessment of the case analysis in Section A.3

A.1 Technical Details

A.1.1 Estimation of the Spectrum

Since our procedure considers a Total Variation Distance between every pair of spectra, smoothness is desirable in spectral estimation. To have a continuous spectral density of $x(t)$, a truncated and smoothed correlogram is constructed by,

$$\tilde{f}(\omega) = \sum_{|h|\leq a} w(h/a)\widehat{\gamma}(h)\exp(-i2\pi\omega h),$$

where $\widehat{\gamma}(h)$ is the sample auto-covariance, defined as $1/(N-h)\sum_{h+1}^{N} x(t)\cdot x(t-h)$. The truncation with span a on the covariance is further smoothed by Parzen window, which is defined as,

$$w(u) = \begin{cases} 1-6|u|^2+6|u|^3, & \text{if } |u| < \frac{1}{2} \\ 2(1-|u|)^3, & \text{if } \frac{1}{2} \leq |u| \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

In practice, the span size a controls the smoothness of estimated spectrum. We used $a = N/2$ in both the numerical experiments and case study, which is relatively conservative for the purpose of smoothing periodogram. This is not problematic for our procedure, since cumulative evidence will be integrated, which can also be seen as smoothing, over epochs under the merits of hierarchical modeling. A more rigorous approach for choosing a can be found in [ORS01]

A.1.2 Gibbs Sampler

The estimation procedure is implemented under Bayesian framework, and a Gibbs sampler is designed targeting the posterior density. By default, the algorithm is initialized with a K partitioning using `pam` in R at each epoch, then iterates through conditional posterior distributions. At the m-th iteration, the sampling scheme is detailed as below:

- $\boldsymbol{\Theta}_i(t)|\mathbb{X}_i(t), \mathbb{L}_i(t) \sim f(\boldsymbol{\theta}_{ik}(t)|\mathbb{X}_i(t), \mathbb{L}_i(t) = k)$ for $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T_i\}$.

  Specifically, the posterior distribution for $\boldsymbol{\theta}_{ik}^{(m)}(t)$ is:

  $$\boldsymbol{\theta}_{ik}^{(m)}(t) \sim \text{NIW}(\boldsymbol{\mu}_{ik}^{(m)}(t), \boldsymbol{\lambda}_{ik}^{(m)}(t), {\sigma^2_{ik}}^{(m)}(t), \nu_{ik}^{(m)}(t))$$

  And these quantities will be updated using data specific to cluster k, based on $\mathbb{L}_i^{(m)}(t)$.

- $\mathbb{L}_i(t)|\mathbb{X}_i(t), \boldsymbol{\theta}_i(t), \boldsymbol{\beta}_i, \mathbb{C}_i \sim \prod_j \Pr(k|X_{ij}(t), C_{ij}, \boldsymbol{\theta}_{ij}(t), \gamma_i)$ for $j = 1, \ldots, p$ and $k = 1, \ldots, K$ where,

  $$\Pr(k|X_{ij}(t), C_{ij}, \boldsymbol{\theta}_{ik}(t), \boldsymbol{\beta}_i) \propto \nu_e(k, C_{ij}^{(m-1)}, \boldsymbol{\beta}_i^{(m-1)}) f(X_{ij}(t)|\boldsymbol{\theta}_{ik}^{(m)}(t))$$

  And $L_{ij}^{(m)}(t)$ will be sampled as $k \in \{1, \ldots, K\}$ with the probability proportional to the above expression.

- $\beta_i(t)|\mathbb{C}_i, \mathbb{L}_i(t) \sim \text{TBeta}(c_i + \tau_i, d_i + pT_i - \tau_i(t), 1/K)$ where $\tau_i(t)$ is the number of samples $(L_{ij}(t) = C_{ij})$ for $t \in \{1, \ldots, T_i\}$ and $j \in \{1, \ldots, p\}$. Also $\mathbb{C}_i = \mathbb{C}_i^{(m-1)}$ and $\mathbb{L}_i(t) = \mathbb{L}_i^{(m)}(t)$.

- $\mathbb{C}_i|\mathbb{L}_i(t), \mathbb{C}, \alpha_i, \boldsymbol{\beta_i}, \sim \prod_j \Pr(k|L_{ij}(t), \ldots, L_{ij}(T_i), C_j, \alpha_i, \boldsymbol{\beta_i})$, where

  $$\Pr(k|L_{ij}(1), \ldots, L_{ij}(T_i), C_j, \alpha_i, \boldsymbol{\beta_i}) \propto \left(\prod_{t=1}^{T_i} \nu_e(L_{ij}(t)^{(m)}, k, \beta_i^{(m)}(t)) \right) \nu_s(k, C_j^{(m-1)}, \alpha_i^{(m-1)}).$$

- $\alpha_i|\mathbb{C}_i, \mathbb{C} \sim \text{TBeta}(a_i + \psi_i, b_i + p - \psi_i, 1/K)$ where ψ_i is the number of electrodes ($C_{ij}^{(m)} = C_j^{(m-1)}$) for $j \in \{1, \ldots, p\}$.

- $C_j|\Pi, C_{1j}, \ldots, C_{nj}, \alpha_i \sim \Pr(k|C_{1j}, \ldots, C_{nj}, \Pi, \alpha_i)$, where

  $$\Pr(k|C_{1j}^{(m)}, \ldots, C_{nj}^{(m)}, \Pi^{(m-1)}, \alpha_i^{(m)}) \propto \pi_k^{(m-1)} \prod_{i=1}^{n} \nu_s(k, C_{ij}^{(m)}, \alpha_i^{(m)}).$$

- $\Pi|\mathbb{C} \sim \text{Dirichlet}(\boldsymbol{\eta} + \boldsymbol{\rho})$ where ρ_k is the number of samples clustered k in $\mathbb{C}^{(m)}$ and $\eta_k = 1$ is chosen non-informatively as a priori.

Markov chain Monte Carlo (MCMC) proceeds by iteratively sampling from the above conditional posterior distributions, and 1/5 of the iterations are used for burning-in after which parameter estimates and information criteria are calculated.

A.1.3 Model Assessment Criteria

We consider the model assessment measures at both epoch and population levels. Epoch levels fitting is favored by the number of clusters, K, that maximizes information criteria listed below, conditional on the eigen-Laplacian dimension d. Average adherence, $\widehat{\alpha}$, reflects population level fitting, which serves as valid measure across dimensionality.

- The BIC ([Sch78]) is defined as

  $$\text{BIC}(K) = 2 \log f(y|\widehat{\tau}, K) - d \log(n),$$

  where d is the number of free parameters.

  In our model:

  $$\text{BIC}(K) = \sum_i \sum_t^{T_i} 2 \log f(\mathbb{X}_i(t)|(\widehat{\boldsymbol{\theta}}_i(t), \widehat{\mathbb{L}}_i(t)), K) - d_f \log(P),$$

where d_f is the number of free parameters in the model,

$$d_f = 2G \cdot d \cdot N_{epoch} + N_{epoch} + n + K - 1,$$

and d is the dimensionality of the eigen-Laplacian data.

- The DIC ([SBC02]) uses *effective* model parameters p_d instead, which has a general form of,

$$\mathrm{DIC}(K) = -2 \log p(y|\widehat{\tau}, K) + 2p_d,$$

$$\text{where } p_d = \mathbb{E}_{\tau|y}(\log p(y|\widehat{\tau})) - \log p(y|\widehat{\tau}).$$

  [CFR06] suggested a form of DIC (as DIC_4 in their paper), which treats the labels $\mathbb{L}$ as missing data thus taking expectation with respect to them afterwards. The formal definition of DIC in our model is,

$$\begin{aligned}
\mathrm{DIC}(K) &= \sum_i \sum_t^{T_i} \{\mathbb{E}_{\mathbb{L}_i(t)l|\mathbb{X}_i(t)}[\mathrm{DIC}(\mathbb{X}_i(t), \mathbb{L}_i(t), K)]\} \\
&= \sum_i \sum_t^{T_i} \{-4\, \mathbb{E}_{\boldsymbol{\theta}, \mathbb{L}_i(t)|\mathbb{X}_i(t)}[\log f(\mathbb{X}_i(t), \mathbb{L}_i(t)|\boldsymbol{\theta}_i(t))] \\
&\quad + 2\, \mathbb{E}_{\mathbb{L}_i(t)|\mathbb{X}_i(t)}[\log f(\mathbb{X}_i(t), \mathbb{L}|\widehat{\boldsymbol{\theta}}_i(t))]\}
\end{aligned}$$

- The adjusted coherence ([LD13]) was suggested for choosing K,

$$\alpha_i^* = \frac{K \cdot \alpha_i - 1}{K - 1}; \quad \bar{\alpha}^* = \frac{K \cdot \bar{\alpha} - 1}{K - 1},$$

  By the definition of coherence parameter, $\alpha_i \in (1/K, 1)$. So adjustments were made such that α_i^* share the same range $(0, 1)$ regardless of K. In our model, α_i reflects the adherence between population and individual consensus, which complements the BIC and DIC for population inference. As pointed out by [LD13], this measure favors smaller K, which is more appropriate when a population pattern exists and being desired.

A.2 Extended Simulation Study

A.2.1 Simulating Band-Oscillating EEG Signals

EEG signal readings features by the oscillation at different frequency bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (30-50 Hz). Being motivated by [EOO15], a linear mixture of second order auto-regressive (AR(2)) models is used to simulate EEG readings. To be specific, an AR(2) process Z_t is defined to be

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \epsilon_t$$

where ϵ_t is a Gaussian white noise process. The oscillation properties of the procedure can be reflected by its characteristic polynomial, $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$. Its root, denoted as z_0^2 and z_0^2, has a polar representation that directly relates to the features in frequency domain,

$$|z_0^1| = |z_0^2| = M, \qquad arg(z_0) = \frac{2\pi\eta}{F_s},$$

where F_s is the sampling frequency; M is the amplitude or magnitude of the root ($M > 1$); and η is the frequency index. The spectrum of the AR(2) process with polynomial roots above will have peak frequency at η. The peak becomes broader as $M \to \infty$; it becomes narrower as $M \to 1^+$.

With a predetermined (F_s, η, M), EEG can be simulated as a sample of AR(2) process with the corresponding coefficients (ϕ_1, ϕ_2).

In our experiment, $M = 1.03$ and $F_s = 200$Hz are fixed. Let Z_t^m be the m-th AR(2) component ($m = 1, \ldots, 5$), oscillating at peak frequency $\eta = 2, 6, 10, 21, 40$ respectively. Thus the observed time series for electrode j is,

$$X_{jt} = \mathbf{e}_{L(j)}^T \cdot \mathbf{Z}_{jt} + \epsilon_{jt}, \qquad \mathbf{Z}_{jt} = (Z_{jt}^1, Z_{jt}^2, Z_{jt}^3, Z_{jt}^4, Z_{jt}^5)^T,$$

where $L(j) \in \{1, \ldots, K\}$ is the true group label for the electrode j and $\mathbf{e}_{L(j)}$ is the mixture weights associated with the AR(2) components. To be more specific, 4 groups of synchronized patterns ($K = 4$) are considered, and the corresponding weights are,

$$e_1 = (1, 2, 0, 0, 0)^T,\ e_2 = (0, 1, 2, 0, 0)^T,\ e_3 = (0, 0, 1, 1, 0)^T,\ e_4 = (0, 0, 0, 1, 1)^T$$

A.2.2 Simulating Non-stationary EEGs

To assess the robustness of proposed procedure to stationarity violations, we construct a piecewise stationary process that has randomly alternating *main* and *off* states. The non-stationarity is coded by two sets of AR(2) mixing weights, $\mathbf{e}_{L(j)}$ and $\mathbf{e}'_{L(j)}$ as *main* and *off* respectively:

$$
\begin{aligned}
e_1 = (1,2,0,0,0)^T &\longleftrightarrow e'_1 = (1,0,0,2,0)^T \\
e_2 = (0,1,2,0,0)^T &\longleftrightarrow e'_2 = (0,0,2,1,0)^T \\
e_3 = (0,0,1,1,0)^T &\longleftrightarrow e'_3 = (0,0,1,2,0)^T \\
e_4 = (0,0,0,1,1)^T &\longleftrightarrow e'_4 = (0,0,0,1,2)^T
\end{aligned}
$$

A.2.3 Extended Results

We examine our procedure against multiple window sizes at $\gamma = \{4, 6, 8, 10\}$ with fixed overlapping of 50%. Results are reported here on 100 repeated datasets, in terms of α estimatesand clustering accuracy against the known truth. Fig A.1 depicts the results for both cases. α estimates are stable over varying smoothing settings (a), and clustering accuracy behaves similar especially at individual level (b). Population level clusters are better established at $\gamma = 6$, which has fewer misclassified cases, but the overall trend as a function of true α holds for all settings.

A.3 Further Details of the Case Study

A.3.1 EEG measurements

The sample in our case study includes 9 participants (29-60 months of age) from the TD group, and 10 participants (27-99 months of age) from the ASD group. During the experiment, EEG was recorded at 250/512Hz using 129 channel geodesic nets with Ag/AgCl electrodes, while participants watched videos of soap bubbles and other non-social images on a computer monitor for 2 to 6 minutes.

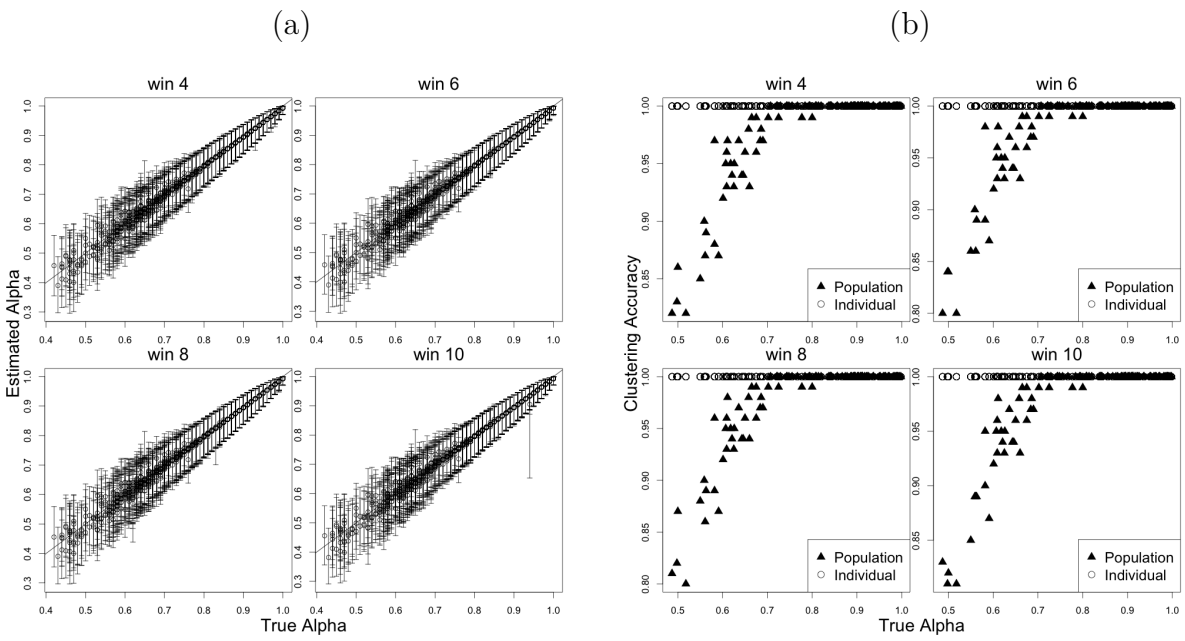

Figure A.1: Extended simulation results: (a) $\widehat{\alpha}$ and its 95% credible interval from MCMC samples, against the true α's. (b) Clustering accuracy at both individual and population level, as a function of the true α's.

The EEG recordings were bandpass filtered at 1-50 Hz using a finite impulse response (FIR) filter with EEGLAB toolbox ([DM04]). Recordings were then segmented into 1024 ms segments for preprocessing. Noisy or loose channels were spherically interpolated using EEGLAB, and EEG recording segments with > 11 interpolated channels were rejected. 4 eye channels were physically removed from the net before the recording session even began, and another reference channel was also excluded yielding a total of 124 channels to be clustered on. All remaining segments were manually inspected for non-stereotyped artifacts, e.g., electromyogram (EMG), and rejected based on qualitative inspection; then a combined principal component analysis (PCA) and independent component analysis (ICA) approach was used to eliminate stereotyped artifacts, e.g., ocular artifacts. All EEG recordings were re-referenced to the average prior to power calculations.

EEG was recorded for 165.1($\pm$74.3) seconds among ASD and 139.9($\pm$30.1) among TD. After the aforementioned preprocessing, 102.7($\pm$19.8, ASD) and 95.9($\pm$32.8, ASD) segments entered MIC analysis.

A.3.2 MCMC Mixing

We consider MCMC mixing by evaluating parameter and cluster estimates at a fixed configuration of smoothing ($\gamma = 6$ and $\delta = 3$) and model ($d = 5$ and $K = 5$). MCMC draws 20k and 40k samples respectively, and a comparison is presented in Table A.1 in terms of $\widehat{\alpha}$ and clustering results. To compare K-clusters partition, we used the `adjustedRandIndex` function of "`mclust`" package in R (see [FRS12], [FR02] and [HA85]).

	α C.I. 20K	α C.I. 40K	adjRandIndex
population	·	·	1.0000
subject 1	(0.7593, 0.9013)	(0.7300, 0.8992)	0.9086
subject 2	(0.7754, 0.9336)	(0.8144, 0.9573)	0.8616
subject 3	(0.6184, 0.7885)	(0.6123, 0.7828)	1.0000
subject 4	(0.7831, 0.9316)	(0.7930, 0.9290)	1.0000
subject 5	(0.5883, 0.8052)	(0.5979, 0.8106)	0.7987
subject 6	(0.7671, 0.9268)	(0.7233, 0.9177)	0.7446
subject 7	(0.7041, 0.8651)	(0.7039, 0.8697)	0.8959
subject 8	(0.6520, 0.8420)	(0.6042, 0.8106)	0.7447
subject 9	(0.6686, 0.8349)	(0.6583, 0.8230)	1.0000

Table A.1: MCMC mixing at 20k and 40k draws

The 95% credible intervals for α_i's are relatively close, and the subject clusters perfectly agree at subject 3,4 and 9. For the purpose of group inference, population estimate perfectly matches between 20k and 40k runs, which highlights the advantage of our proposed method when individual hard clusters are susceptible to missing segments and artifacts.

APPENDIX B

Supplementary files for Chapter 3

This document accompanies Chapter 3 and provides a detailed discussion of technical points. In particular, Section B.1 completes the research paradigm and the corresponding preprocessing procedures for EEG cleaning. In Section B.2, we give detailed formulation and calculation of the eigen powers, and discuss on an generalized cross validation procedure that automatically selects the smoothing bandwidth. Extensive technical details are provided in Section B.3, that contains prior and hyper-prior choices for our hierarchical functional models, step-by-step Gibbs samplers to carry out the posterior sampling and calculation of simultaneous confidence bands for inference. We conclude with an extended report of our simulation study, covering generative assumptions, settings for the numerical study, and extra findings under various scenarios in Section B.4. The implementation of our model is fulfilled in R and Rcpp, which is publicly available on author's Github (`https://github.com/Qian-Li/HFM`)

B.1 Research Paradigm and EEG Preprocessing

B.1.1 Extended Research Paradigm

The audio learning research paradigm hypothesized that participants were able to implicitly learn the boundary of artificially constructed words out of a continuous stream of phonemes. To be precise, there are 12 phonemes, as building blocks, constructing 4 *expected* words that would be repeatedly exposed during the task, thereby expected to be learned:

- pa-bi-ku
- da-ro-pi

- tu-da-ro
- go-la-tu

Exposure Phase: the exposure phase exposed participants with repetitions of "expected" words. Specifically, the above listed 4 words were pseudo-randomly permuted 3 times into a *block*, under two constraints: 1) each word was repeated 3 times in a block; 2) none of the words was repeated twice in a row. Afterwards, five individually constructed blocks were further repeated 3 times. As an example b1 = (w1 w2 w3 w4 w2 w1 w4 w3 w2 w4 w1 w3), b2 = (w2 w4 w1 w3 w4 w2 w1 w4 w3 w1 w2 w3), etc. Then overall (b1 b2 ... b5 b1 b2 ... b5 ...). At last, each *expected* word was repeated 45 times counting as 3(within block reps) $\times$ 5(blocks) $\times$ 3(blocks reps).

Timing: all *expected* words were recorded then played continuously in a stream, and further time-locked and tagged during the recording (however not available in our sample). With an average duration of 803 milliseconds per word, the total length of exposure phase is:

$$803 \times 4 \times 45 = 144540 \text{ ms} = 144.54 \text{ s}$$

Test Phase (Not used): the test phase consisted of words that were expected to be implicitly learned and part-words, that had also appeared in the stream as tri-phenome combination however less frequently, during the exposure phase. As an example, a part-word was constructed by the last phoneme of an expected word together with the first two phonemes of the following word. During the stream where "pa-bi-ku" was followed by "go-la-tu", "ku-go-la" was a part-word.

A total of 96 trials were tested in this phase, where a mixture of 4(expected words)$\times$12(reps)= 48 case trials and 12(part-words)$\times$4(reps)= 48 control trials had been randomly scrambled. A 500-700 ms inter-trial interval was inserted to definitely inform word boundaries, therefore the total length of this phase was roughly:

$$803 \times 96 + 600 \times 96 = 134688 \text{ ms} = 134.68 \text{ s}$$

The test phase data was time-locked to trails, and was further segmented and separated as "words" and "part-words" datasets. Within each trial, the event-related potentials (ERPs) were registered to -100ms relative to the onset of stimuli and the averaged amplitude within the 400-600ms time window was of major interest for interpretation.

B.1.2 EEG Preprocessing Pipeline

The exposure phase data was preprocessed in a similar manner to resting-state EEG. A pipelined procedure was followed:

$$\text{Filtering} \rightarrow \text{Segmentation} \rightarrow \text{initial BCR} \rightarrow \text{Auto Reject}$$
$$\rightarrow \text{Manu Reject} \rightarrow \text{Final BCR} \rightarrow \text{re-Reference}$$

- Segmentation: the exposure phase is word-locked according to the protocol, but still segmented into 1s in the given dataset.
- Initial Bad Channel Replacement (BCR): mark and replace **channels** that contains EMG, high impedances or being unstable and erratic throughout the recording.
- Automatic Rejection: use "Auto Artifact Detection" toolbox builtin in NetStation.
- Manual Artifacts Detection: mark segments bad if they contain eye blinks, saccades, EMG/muscle or clusters of bad channels (based on initial BCR).
- Final BCR: mark all channels bad again that were marked bad during initial review, also mark all eye channels, and replace.
- Re-Reference to Average: standard re-referencing

Eventually the segments marked "good" are saved and output, with the rejection log manually recorded in a separate file.

The exposure phase has a total of 153 segments, with 77.53 good segments left after preprocessing on average (MIN= 14, MAX= 146). All 129 channels (including 4 eye channels and 1 ref channel) recording EEG at 512Hz are available for 45 subjects. Discrepancy was noticed between data distribution and manual log. Further summaries are produced for all subjects (45) on their demographical variables.

IDs	17	437*	439*	451*	3002	3015	3020
Group	TD	vASD	mvASD	vASD	mvASD	mvASD	mvASD
Segments in Data	65	84	102	134	70	17	40
Segments in Log	67	153	69	94	72	18	39

Table B.1: Mismatches between manual log and actual data

Group	N	M/F	Age in m.	vDQ	nvDQ
TD	9	6/3	94.67(28.82)	120.56(11.40)	115.22(15.19)
vASD	16	9/7	67.05(5.84)	89.33(22.69)	95.62(19.77)
mvASD	20	17/3	85.65(24.00)	23.64(10.87)	43.09(13.60)
Total	45	32/13	80.73(22.99)	68.37(43.18)	76.95(34.63)

Table B.2: Demographical Variables, with Mean and (SD)

Individuals, 437, 439 and 451 were excluded from the case study due to a significant amount of mismatches between the data distribution and their processing log, therefore it leads to our sample size of 42 persons. It is also worth mentioning that age and verbal-DQ are highly correlated with the group definition of TD, vASD and mvASD. The separability of verbal-DQ is reasonable since that directly relates to the clinical definition of ASD. We consider the unbalanced distribution of age to be problematic especially the vASD cohort are significantly younger than TD children, and one might suspect that their deficits in verbal-DQ are purely due to the effects of age.

B.2 Spectral Principle Component Analysis

The principle component analysis on spectral domain is extensively covered by [SS10] in Chapter 7. This section mainly focuses on the technical details of implementing the method by [OH06] and a generalized cross-validation approach for automatic band selection. Author has implemented this procedure in R, however, all work can be easily translated and migrated to other statistical packages for computation. For simplicity, we refer to this approach as spectral PCA in the details that follow.

B.2.1 Calculation of Spectral PCA

The spectral PCA decomposes the multi-channel signals into incoherent components, and provides a lower dimensional representation of the regionally characteristic spectrum. In our application, only the first eigenvalues are considered at various Fourier frequencies. Due to the reported inconsistency of raw spectral estimates and a relatively short signal length (1s), smoothing is highly recommended to effectively reduce variance of the estimated spectrum. In this regard, each univaraite time series is normalized ($var = 1$) such that PSDs are comparable over frequencies, followed by an GCV-based procedure that automatically determines the optimal bandwidth of modified Daniell's kernel at each frequency.

Detrend: each channel of EEG signals is normalized per segment. A standard normalization is followed, by which we first remove a linear trend as well as the intercept from the univariate process, then rescale the residuals to unit variance as detrended signal.

Spectral Matrix: For a p-variate time series of length M, denoted as $\mathbf{X} = \{X_1, \ldots, X_p\}$, collecting signals in p channels, its Fourier coefficients at frequency ω is defined:

$$\mathbf{d}(\omega) = \frac{1}{\sqrt{M}} \sum_t \mathbf{X}(t) \exp(-i2\pi\omega t). \tag{B.1}$$

Consequently, the spectral matrix ($p \times p$), which has real diagonal entries and complex off-diagonal entries, is,

$$I(\omega) = \mathbf{d}(\omega)\overline{\mathbf{d}}(\omega)'. \tag{B.2}$$

Its implementation in R depends on the `mvspec` function of `astsa` package. De-trending can also be performed by this function with the corresponding option. An alternative implementation can utilize a basic-R function, `mvfft`, even though extra attentions are necessary regarding the normalization.

B.2.2 Optimal bandwidth via GCV

To smooth the spectral matrices over varying Fourier frequencies, we use the modified Daniell's kernel with a bandwidth of w ($span = 2w + 1$). The optimal bandwidth is determined by a generalized cross-validation (GCV) procedure, see [OH06]. Specifically for a given bandwidth w, we first estimate the corresponding spectral density matrix $\tilde{I_w}(\omega)$. The GCV error is defined individually on each univariate process and we summarize by summing up to a total error: $GCV(w) = \sum_{j=1}^{p} GCV_j(w)$. For each channel,

$$GCV_j(w) = \frac{2}{df(w)^2} \sum_{k=0}^{M/2-1} q_k \left[\frac{I_{jj}(\omega_k)}{\widehat{I}_{jj}(\omega_k, w)} - \log \frac{I_{jj}(\omega_k)}{\widehat{I}_{jj}(\omega_k, w)} - 1 \right], \tag{B.3}$$

where $df(w) = 1 - \text{trace}(S(w))/(M/2+1)$ is the degree of freedom, $q_k = 1/2$ when $k = 0, M/2$ and 1 elsewhere.

Eventually, the optimal bandwidth, $w^* = \min_w GCV(w)$, minimizes the combined GCV error across all frequencies. Therefore the smoothed spectral matrix is guaranteed Hermitian.

As for implementation, the optimal value is determined on a grid of preselected values whichever eventually yields the minimal loss. With the optimal bandwidth, we extract the first eigenvalues as the Regionally smoothed Eigen-power (Rs-EP$_1$) by means of eigen-decomposition on $\tilde{I}_{w^*}(\omega)$.

B.3 Technical Details

This section lays out the technical details for implementing all three priors on the coefficient matrix Θ_i, which were proposed in our hierarchical functional models. We refer to Naive Bayesian, Strongly Separable and Non-Separable priors and NB, SS and NS respectively.

B.3.1 Choice of Ω_0

To choose an appropriate hyper prior for the inverse Wishart distribution, especially the prior mean Ω_0, we borrow structural assumption that has been well established in the spatial statistics literature. Precisely,

$$\Omega_0 = (D_W - \rho W), \quad \text{with } \rho \in (1/\lambda_{(1)}, 1/\lambda_{(p)}),$$

and $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(p)}$ are the ordered eigenvalues of $D_W^{-1/2} W D_W^{-1/2}$, $D_W = \text{diag}(w_{1+}, \ldots, w_{p+})$, W is the a pre-defined adjacency matrix. Therefore Ω_0 is guaranteed to be positive definite. Hyper-parameter ρ can be empirically estimated to simplify the problem, and it is highly recommended to use a non-informative choice.

B.3.2 Gibbs Sampler

To unify the use of notations, we denote matrices as A, vectors as $\boldsymbol{a}$, c-th column of matrix A as $A^{(c)}$, the b-th block of matrix A as $A_{(b)}$. Under the representation from the main article, we complete the model with,

$$\begin{aligned}
NS: \quad & \boldsymbol{\lambda}_{(k)}^{(c)} | \Omega_k \tau_c \sim \text{MN}(\mathbf{0}, \Omega_k^{-1} \tau_c^{-1}), \quad \Omega_k \sim \text{W}(\nu, W_0), \quad \tau_c = \prod_{l=1}^{c} \delta_l, \quad k \in \{1, \ldots, q\} \\
& \delta_1 \sim \text{Ga}(a_1, 1), \quad \delta_l \sim \text{Ga}(a_2, 1), \quad \sigma_{\cdot}^{-2} \sim \text{Ga}(a_0, b_0); \\
SS: \quad & \upsilon_{jr} | \phi_{jr}, \tau_r \sim \text{N}(0, \phi_{jr}^{-1} \tau_r^{-1}), \quad \phi_{jr} \sim \text{Gamma}(\nu_1/2, \nu_1/2), \quad \tau_r = \prod_{u=1}^{r} \delta_u, \\
& j \in \{1, \ldots, p\}, \quad r \in \{1, \ldots, k_1\} \\
& \gamma_{ks} | \psi_{ks} \kappa_s \sim \text{N}(0, \psi_{ks}^{-1} \kappa_s^{-1}), \quad \psi_{ks} \sim \text{Gamma}(\nu_2/2.\nu_2/2). \quad \kappa_s = \prod_{v=1}^{s} \pi_v, \\
& k \in \{1, \ldots, q\}, \quad s \in \{1, \ldots, k_2\} \\
& \delta_1 \sim \text{Ga}(a_{11}, 1), \quad \delta_u \sim \text{Ga}(a_{12}, 1), \\
& \pi_1 \sim \text{Ga}(a_{21}, 1), \quad \pi_v \sim \text{Ga}(a_{22}, 1), \sigma_{\cdot}^{-2} \sim \text{Ga}(a_0, b_0). \\
\text{Mean:} \quad & \text{vec}(M_i)^T = \boldsymbol{x}_i^T B, \quad \text{where } B | \Sigma_r \sim \mathcal{MN}(0, g(X'X)^{-1}, \Sigma_r),
\end{aligned}$$

where $\Sigma_r = \text{cov}(\Lambda \boldsymbol{\eta}_i + Z_i) = \Lambda\Lambda' + \Sigma_q \otimes \Sigma_p$, g is a constant set to be large.

Gibbs Sampler using NS-prior

Step 1 Full Gibbs update of Λ, including $\Omega_\cdot$'s and $\delta_\cdot$'s. For $k \in \{1,\ldots,q\}$, $j \in \{1,\ldots,p\}$, the diagonal precision matrixes $\Sigma_p^{-1} = \text{diag}\{\sigma_j^{-2}\}$, $\Sigma_q^{-1} = \text{daig}\{\sigma_k^{-2}\}$

- For $k \in \{1,\ldots,q\}$,
  - $\Pr(\Lambda_{(k)}|\Omega,\delta,\ldots)$ by block: the vectorized version of which has a multivariate Gaussian posterior distribution,

$$\pi(\text{vec}(\Lambda_{(k)})|-) \sim \text{MN}(\sigma_k^{-2} S^{-1}(H' \otimes \Sigma_p^{-1})\text{vec}(\Theta^{(k)} - M^{(k)}), S^{-1})$$

    where $S = T \otimes \Omega_k + \sigma_k^{-2} H'H \otimes \Sigma_p^{-1}$, and $T = \text{diag}(\tau_1,\ldots,\tau_m)$.
  - $\Pr(\Omega_k|\Lambda_{(k)}, T)$

$$\pi(\Omega_k|-) \propto \text{W}(\nu + m, W_0 + \Lambda_{(k)} T \Lambda'_{(k)}),$$

- $\Pr(T|\Lambda,\Omega)$, and $\tau_c = \prod_{l=1}^{c} \delta_l$. To effectively sample δ's, we need to distinguish the baseline precision δ_1 and add-on penalties δ_l's where $l > 1$:

$$\begin{aligned}
\pi(\delta_1|-) &\sim \text{Gamma}(a1 + \frac{pqm}{2}, 1 + \frac{1}{2}\sum_{c=1}^{m} \tau_c^{(1)} \sum_{k=1}^{q} \boldsymbol{\lambda}_{(k)}^{(c)\,T} \Omega_k \boldsymbol{\lambda}_{(k)}^{(c)}), \\
&\sim \text{Gamma}(a1 + \frac{pqm}{2}, 1 + \frac{1}{2}\sum_{k=1}^{q} \text{tr}(T^{(1)} \cdot \Lambda'_{(k)} \Omega_k \Lambda_{(k)}), \\
\pi(\delta_l|-) &\sim \text{Gamma}(a2 + \frac{pq(m-l+1)}{2}, 1 + \frac{1}{2}\sum_{c=l}^{m} \tau_c^{(l)} \sum_{k=1}^{q} \boldsymbol{\lambda}_{(k)}^{(c)\,T} \Omega_k \boldsymbol{\lambda}_{(k)}^{(c)})), \\
&\sim \text{Gamma}(a2 + \frac{pq(m-l+1)}{2}, 1 + \frac{1}{2}\sum_{k=1}^{q} \text{tr}(T^{(l)} \Lambda_{(k)}^{(l-)\prime} \Omega_k \Lambda_{(k)}^{(l-)})), \\
&\quad \text{for } l \geq 2,
\end{aligned}$$

where $\tau_c^{(l)} = \prod_{t=1, t\neq c}^{l} \delta_t$ for $l = 1,\ldots,m$ and $T^{(l)} = \text{diag}(\tau_\cdot^{(l)})$

Step 2 Update Σ_p and Σ_q for $j \in \{1,\ldots,p\}$ and $k \in \{1,\ldots,q\}$,

$$\sigma_j^{-2}|\Sigma_q,- \propto \text{Gamma}(a_0 + \frac{nq}{2}, b_0 + \frac{\sum_{i=1}^{n}(\Theta_i - M_i - \text{mat}(\Lambda\boldsymbol{\eta}_i))_{j\cdot} \Sigma_q^{-1} (\Theta_i - M_i - \text{mat}(\Lambda\boldsymbol{\eta}_i))_{j\cdot}^T}{2}),$$
$$\sigma_k^{-2}|\Sigma_p,- \propto \text{Gamma}(a_0 + \frac{np}{2}, b_0 + \frac{\sum_{i=1}^{n}(\Theta_i - M_i - \text{mat}(\Lambda\boldsymbol{\eta}_i))_{\cdot k}^T \Sigma_p^{-1} (\Theta_i - M_i - \text{mat}(\Lambda\boldsymbol{\eta}_i))_{\cdot k}^T}{2}).$$

Step 3 Update the variance of system error $\epsilon_i(t)$, σ_e^2

$$\sigma_e^{-2}|- \propto \text{Gamma}(a_e + \frac{\sum_i n_i}{2}, b_e + \frac{\sum_{i=1}^{n} \text{vec}(Y_i - \Theta_i)'\text{vec}(Y_i - \Theta_i)}{2}).$$

Step 4 Update the mean function M_i or $\boldsymbol{\mu}_i$ (B)

$$\boldsymbol{\mu}_i = \text{vec}(M_i) = B' \cdot \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i$ is a $d \times 1$ vector, that contains both discrete and continuous covariates and B is the coefficient matrix of size $d \times pq$ with p-blocks of spline coefficients corresponding to d covariates. In our working example, we are explicitly interested in group-specific means and the effects from age and verbal-DQ. The coefficients are therefore modeled flexibly,

$$\boldsymbol{\theta}_i^* = \text{vec}(\Theta_i) = B' \cdot \boldsymbol{x}_i + \boldsymbol{r}_i, \qquad \text{where } \boldsymbol{r}_i \sim \text{MN}(\mathbf{0}, \Sigma_r),$$

In a matrix form, $\Theta = XB + R$, let $\Sigma_r = \Lambda\Lambda + \Sigma_q \otimes \Sigma_p$ is conditioned on, therefore the full conditional posterior for B is,

$$\pi(B|\Sigma_r, -) \sim \mathcal{MN}(B_n, \frac{g}{g+1}(X'X)^{-1}, \Sigma_r),$$

and $B_n = \frac{g}{g+1}(X'X)^{-1}(X'\Theta)$.

Step 5 Update the latent variable $\boldsymbol{\eta}_i$ by marginalizing out Θ_i and back projection $B_i(B_i'B_i)^{-1}$

$$Y_i = M_i B_i' + \text{mat}(\Lambda\boldsymbol{\eta}_i)B_i' + Z_i B_i' + E_i,$$

which is equivalent to,

$$Y_i^* = \text{mat}(\Lambda\boldsymbol{\eta}_i)B_i' + Z_i B_i' + E_i,$$

and Y_i^* being the centered observations $Y_i^* = Y_i - M_i B_i'$. Multiply both sides by $B_i(B_i'B_i)^{-1}$ and vectorize,

$$\boldsymbol{y}_i^{**} = \text{vec}(Y_i^* B_i(B_i'B_i)^{-1}) = \Lambda\boldsymbol{\eta}_i + \text{vec}(Z_i) + \text{vec}(E_i B_i(B_i'B_i)^{-1}),$$

where $\boldsymbol{y}_i^{**} \sim \text{MN}(\Lambda\boldsymbol{\eta}_i, \Sigma_q \otimes \Sigma_p + \sigma_e^2(B_i'B_i)^{-1} \otimes I_p)$ has fixed covariance, denoted as Σ^*

Therefore, sample $\boldsymbol{\eta}_i$ from its full conditional posterior,

$$\pi(\boldsymbol{\eta}_i|-) \sim \text{MN}((\Lambda'\Sigma^{*-1}\Lambda + I_k)^{-1}(\boldsymbol{\eta}_0 + \Lambda'\Sigma^{*-1}\boldsymbol{y}_i^{**}), (\Lambda'\Sigma^{*-1}\Lambda + I_k)^{-1})$$

Step 6 Update the coefficient matrices Θ_i, or $\boldsymbol{\theta}_i$ (must vectorize!) We have the data level:

$$Y_i = \Theta_i B_i' + E_i, \quad \text{where } E_i \sim \mathcal{MN}(0, I_p, \sigma_e^2 I_{n_i})$$

And the projection level, which is the prior in this sampler:

$$\text{vec}(\Theta_i) \sim \text{MN}(\text{vec}(M_i) + \Lambda\eta_i, \Sigma_q \otimes \Sigma_p)$$

Therefore, the full conditional posterior of $\text{vec}(\Theta_i')$ is,

$$\boldsymbol{\theta}_i^* = \text{vec}(\Theta_i') = \text{MN}(S^{-1}(\text{vec}(\sigma_e^{-2}B_i'Y_i') + (\Sigma_p^{-1} \otimes \Sigma_q^{-1}\text{vec}(M_i' + \text{mat}(\Lambda\eta_i)')), S^{-1}).$$

where $S = (\sigma_e^{-2}I_p \otimes (B_i'B_i) + \Sigma_p^{-1} \otimes \Sigma_q^{-1})$

Gibbs Sampler using SS-prior

Step 1 Update Υ and Γ: sample υ_{jr}, γ_{ks}, ϕ_{jr}, ψ_{ks}, δ's and π's.

- Sample the j-th row of Υ, for $j = 1, \ldots, p$:

$$\pi(\Upsilon_{j\cdot}|-) \sim \text{MN}(S_j^{-1}\sigma_j^{-2}\sum_i (H_i\Gamma'\Sigma_q^{-1}\Theta_{j\cdot}), S_j^{-1}),$$

  with $S_j = D_j^{-1} + \sigma_j^{-2}\sum_i (H_i\Gamma'\Sigma_q^{-1}\Gamma H_i')$ and $D_j^{-1} = \text{diag}(\phi_{j1}\tau_1, \ldots, \phi_{jk_1}\tau_{k_1})$.

- Sample ϕ_{jr}:

$$\pi(\phi_{jr}|-) \sim \text{Gamma}(\frac{\nu_1 + 1}{2}, \frac{\nu_1}{2} + \frac{\tau_r\upsilon_{jr}^2}{2})$$

- Sample δ_u for $u = 1$ and $u \geq 1$

$$\pi(\delta_1|-) \sim \text{Gamma}(a_{11} + \frac{pk_1}{2}, 1 + \frac{1}{2}\sum_{c=1}^{k_1}\tau_c^{(1)}\sum_{j=1}^{p}\phi_{jc}^2)$$
$$\pi(\delta_u|-) \sim \text{Gamma}(a_{12} + \frac{p(k_1-u+1)}{2} + 1 + \frac{1}{2}\sum_{c=u}^{k_1}\tau_c^{(u)}\sum_{j=1}^{p}\phi_{jc}^2)$$

with definitions similar to (a).

– Sample the k-th row of Γ, for $k = 1, \ldots, q$:

$$\pi(\Gamma_{k\cdot}|-) \sim \text{MN}(S_k^{-1}\sigma_k^{-2}\sum_i H_i'\Upsilon'\Sigma_p^{-1}\Theta_{\cdot k}, S_k^{-1}),$$

with $S_k = D_k^{-1} + \sigma_k^{-2}H_i'\Upsilon'\Sigma_p^{-1}\Upsilon H_i$ and $D_k^{-1} = \text{diag}(\psi_{k1}\kappa_1, \ldots, \psi_{kk_2}\kappa_{k_2})$.

– Sample ψ_{ks}:

$$\pi(\psi_{ks}|-) \sim \text{Gamma}(\frac{\nu_2+1}{2}, \frac{\nu_2}{2} + \frac{\kappa_s\gamma_{ks}^2}{2})$$

– Sample π_v for $v = 1$ and $v \geq 1$

$$\pi(\pi_1|-) \sim \text{Gamma}(a_{21} + \frac{qk_2}{2}, 1 + \frac{1}{2}\sum_{c=1}^{k_2}\kappa_c^{(1)}\sum_{k=1}^{q}\psi_{kc}^2)$$
$$\pi(\pi_u|-) \sim \text{Gamma}(a_{22} + \frac{q(k_2-u+1)}{2} + \frac{1}{2}\sum_{c=u}^{m}\kappa_c^{(u)}\sum_{k=1}^{q}\psi_{kc}^2)$$

with definitions similar to (a).

Step 2 Update Σ_p and Σ_q the same as (NS) except for plugging in

$$\text{mat}(\Lambda\boldsymbol{\eta}_i) = \Upsilon H_i\Gamma'$$

Step 3 Update the variance of system error $\epsilon_i(t)$, σ_e^2 the same as (NS)

Step 4 Update the mean function M_i or $\boldsymbol{\mu}_i$ the same as (NS)

Step 5 Update the latent variable $\boldsymbol{\eta}_i$ by marginalizing out Θ_i

We adopt a different strategy, but still borrow notations from (a):

$$\begin{aligned} Y_i^* &= \Upsilon H_i \Gamma' B_i' + A_i \\ &= \sum_{r,s} \eta_{irs} \upsilon^{(r)} \gamma^{(s)'} B_i' + A_i \\ \text{vec}(Y_i^*) &= \mathbf{1}_{k_1 k_2} \text{vec}(\upsilon^{(r)} \gamma^{(s)'} B_i')' \boldsymbol{\eta}_i + \text{vec}(A_i) \\ \boldsymbol{y}_i^* &= B_i^* \boldsymbol{\eta}_i + \boldsymbol{\alpha}_i \end{aligned}$$

And we have closed form updating density for $\boldsymbol{\eta}_i$

$$\pi(\boldsymbol{\eta}_i|-) \sim \text{MN}((B_i^{*'} \Sigma^{-1} B_i^*)^{-1} (\boldsymbol{\eta}_0 + B_i^{*'} \Sigma^{-1} \boldsymbol{y}_i^*), (B_i^{*'} \Sigma^{-1} B_i^*)^{-1})$$

Step 6 Update the Θ_i the same as (a)

Gibbs Sampler using NB-prior

Gibbs Sampler using Matrix normal priors are trivial, therefore not listed in details. To make all implementations fairly comparable, author borrowed the steps 3,4 and 6 from the NS and SS priors. This implementation is an example of hierarchical centering in Bayesian analysis.

B.3.3 Simultaneous Confidence Bands

To construct a simultaneous confidence band, we followed [KKC10] that slightly modified an approach from [CRC07]. First of all, functional estimates are recorded the same time as posterior means are updated online during the MCMC sampling. Suppose that $\widehat{f}$ is the posterior mean to the estimated functional, whose standard deviations are updated as well per functional observation, $\sqrt{\widehat{var}(\widehat{f}(t))}$. Under a normality assumption on the posterior samples, one could derive the $(1-\alpha)$ quantile c_b of,

$$\max_t \left| \frac{f^{(i)}(t) - \widehat{f}(t)}{\sqrt{\widehat{var}(\widehat{f}(t))}} \right|, \quad i = 1, \ldots, N,$$

where n is the total number of posterior samples. Therefore we can derive a hyperrectangular band at each time point,

$$\left[\widehat{f}(t) - c_b \cdot \sqrt{\widehat{var}(\widehat{f}(t))}, \widehat{f}(t) + c_b \cdot \sqrt{\widehat{var}(\widehat{f}(t))}\ \right] \tag{B.4}$$

B.4 Extended Simulation

This section expand on the numerical study reported in the main article.

B.4.1 Simulation Settings

Our simulation considered three groups of interest, each consists of 20 individuals, and their signals are simulated over 6 regions on a grid of 100 observational points evenly distributed by default. For each individual, their functional observations deviate randomly from its corresponding region-specific group means, which are selected among a predetermined collection of parametric curves:

$$\beta_0(t) = 1; \quad \beta_1(t) = 5 \cdot \sin(t * \pi/60); \quad \beta_2(t) = 2 - 3 \cdot \cos((t-25) * \pi/15)$$
$$\beta_3(t) = 5 - 0.2t; \quad \beta_4(t) = -1 + (20-t)^2/100,$$

The data generation process follows the generative assumptions listed in the main article. Let's denote the individual random effects projected onto the B-spline functional space as a coefficient matrix Θ_i^*, we define:

$$\text{Separable}: \quad \Theta_i^* \sim \mathcal{MN}(\mathbf{0}, S_0, \Omega_o^{-1})$$
$$\text{Non-Separable}: \quad \Theta_i^* = \text{mat}(\Lambda \boldsymbol{\eta}_i), \quad \textit{where} \quad \boldsymbol{\eta}_i \sim \text{MN}(0, I)$$

In fairness, we further included perturbations to Θ_i^*, $R_i \sim \mathcal{MN}(0, \Sigma_p, \Sigma_q)$, where Σ_p and Σ_q are diagonal matrices with its entries sampled from a inverse gamma distribution. Such noise in coefficient matrix is not considered the same as system errors, since they are no longer independent any more after B-spline projection to data level.

B.4.2 More Simulation Results

Extensive numerical experiments are performed on all three proposed priors to our hierarchical functional models. To organize our work, we group them by: 1) Separable and Non-separable simulation with incomplete functional observations; 2) Separable and Non-separable simulation with varying sample size and Signal-to-Noise Ratios (SNRs); 3) Sensitivity to unknown number of latent factors. Similarly as the above sections, we use the abbreviations of NB, SS and NS for Naive Bayesian, Strongly Separable and Non-Separable priors in our model.

Incomplete functional observations

In this case, 500 repeatedly sampled dataset were engineered with a functional missing percentage uniformly sampled between 0 and 0.8. We also investigate when the true number of B-spline basis is unknown, therefore to evaluate the robustness of three priors in this setting. Finally, we repeat this scheme on both non-separable and separable generative process, each with 500 engineered dataset.

Results:

For separable simulation, all three priors recovered the group mean curves very accurately. NB slightly outperform SS and NS, however the difference is very marginal. When the true degree of freedom is unknown for B-spline basis, our approaches are very robust to both over- and under-specified basis functions. On the simultaneous confidence bands side, NS and SS consistently covered the ground truth at the specified level of 90%, whereas NB yielded conservative results that had a coverage over 96%. We also varied the regional dependency parameter, $\rho = \{0.4, 0.6, 0.8\}$ during the generative process. NS and SS recovered both the group mean and the random effects covariance, $\text{cov}(\Theta_i)$ accurately with good confidence band coverage. NB is shy on the coverage and covariance part even though it conforms with the generating process. Between NS and SS, SS has yielded better covariance estimates especially when the regions are less dependent, however its advantage faded as more dependency were

introduced where more benefits were picked up by the low-rank structure of NS.

Compared with the Separable simulation, the Non-Separable simulation posed more challenges. Most of the findings have been reported by our main article, therefore we skip extra interpretations. Please refer to Table B.3 for more results.

Table B.3: Incomplete Functional Observations: Number of B-Spline basis Unkown

	df.BS = 8		df.BS = 12*		df.BS = 20	
Prior	aiMSE	a.cover	aiMSE	a.cover	aiMSE	a.cover
	Separable Simulation					
NB	0.1723	0.9887	0.1773	0.9791	0.1850	0.9667
SS	0.1734	0.8971	0.1806	0.8911	0.1957	0.8952
NS	0.1735	0.9029	0.1809	0.8960	0.1960	0.9034
	Non-Separable Simulation					
NB	0.3094	0.9844	0.3300	0.9788	0.3481	0.9709
SS	0.3115	0.8954	0.3326	0.8775	0.3586	0.8698
NS	0.3095	0.8736	0.3296	0.8519	0.3529	0.8565

Sample Size and SNRs

To extend on the finite sample operating characteristics, all three priors are tested on varying sample size and SNRs. We vary the sample size per group, $n = \{10, 20, 50\}$, and maintain the same size between groups. SNR is tested under $\{0.2, 1.0, 5.0\}$ on all three sample size configurations. Consequently, we have 9 combinations of sample size and SNR settings, within each case 120 datasets were simulated under both separable and non-separable assumption for evaluation.

Results:

Under the separable setting, NB outperformed NS and SS in group mean recovery especially on challenging conditions (small sample size and low SNR), however, it fell short in all other aspects. NS and SS are very comparable under all scenarios, and both of them benefit

by all metrics as sample size and SNR increased. On the other hand, the non-separable simulation stilled posed more challenges and we begin to notice a divergence in performance among three priors. Since NB has been easily outperformed by its alternatives, we mainly focus on the SS and NS under this scenario. The efficiency gain was obvious for NS in terms of covariance estimates, where 16% gained noticed at SNR= 0.2, 23% at 1.0 and 38% at 5.0, however barely for SS. By the group mean recovery, NS and SS yielded almost the same performance and gain as more information became available. And finally, we did notice better confidence band coverage from especially and strong SNR.

Table B.4: Sample Size and SNRs: Separable Simulations

	$n = 10$			$n = 20$			$n = 50$		
	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV
	SNR = 0.2								
NB	0.3300	0.9454	16.6004	0.2357	0.9407	16.0672	0.1543	0.9204	15.9142
SS	0.3836	0.9227	16.5972	0.2722	0.9199	15.6221	0.1748	0.9000	14.5899
NS	0.3862	0.9431	17.5774	0.2722	0.9380	16.2057	0.1738	0.9185	15.0789
	SNR = 1.0								
NB	0.2644	0.9703	26.2468	0.1862	0.9717	26.1408	0.1185	0.9689	26.4185
SS	0.2726	0.9057	10.3331	0.1920	0.8956	9.4470	0.1220	0.8865	8.7190
NS	0.2731	0.9119	13.4917	0.1922	0.9047	11.9307	0.1221	0.8870	9.9558
	SNR = 5.0								
NB	0.2404	0.9844	29.6935	0.1680	0.9861	29.4679	0.1073	0.9853	29.3421
SS	0.2412	0.8775	9.1318	0.1686	0.8809	8.2208	0.1077	0.8742	7.5274
NS	0.2413	0.8885	12.9454	0.1687	0.8826	10.6144	0.1077	0.8683	8.5531

Table B.5: Sample Size and SNRs: Non-Separable Simulations

	$n = 10$			$n = 20$			$n = 50$		
	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV	aiMSE	a.cover	fCOV
	SNR = 0.2								
NB	0.5733	0.9296	147.5943	0.4142	0.9236	146.6023	0.2754	0.9111	147.9268
SS	0.7015	0.9134	128.9884	0.4923	0.9208	125.9013	0.3150	0.9023	123.1696
NS	0.7049	0.9241	121.6705	0.4926	0.9269	113.1884	0.3144	0.9125	102.6813
	SNR = 1.0								
NB	0.5089	0.9765	231.0302	0.3544	0.9837	213.8458	0.2295	0.9775	215.7354
SS	0.5266	0.8779	121.1370	0.3642	0.8999	119.1692	0.2352	0.8817	117.8870
NS	0.5242	0.8664	96.5781	0.3620	0.8879	86.3803	0.2340	0.8669	74.5817
	SNR = 5.0								
NB	0.4601	0.9966	313.1928	0.3216	0.9987	314.2198	0.2096	0.9983	310.7135
SS	0.4590	0.8721	114.0363	0.3205	0.8944	111.4064	0.2092	0.8767	110.3721
NS	0.4572	0.7980	76.0761	0.3192	0.8304	63.3445	0.2086	0.8165	47.6799

REFERENCES

[ABS08] Kezban Aslan, Hacer Bozdemir, Cenk Şahin, Seyfettin Noyan Oğulata, and Rızvan Erol. "A radial basis function neural network model for classification of epilepsy using EEG signals." *Journal of medical systems*, **32**(5):403–408, 2008.

[APB07] Martijn Arns, Sylvia Peters, Rien Breteler, and Ludo Verhoeven. "Different brain activation patterns in dyslexic children: evidence from EEG power and coherence patterns for the double-deficit theory of dyslexia." *Journal of integrative neuroscience*, **6**(01):175–190, 2007.

[Bas04] Erol Başar. *Memory and brain dynamics: Oscillations integrating attention, perception, learning, and memory*. CRC, 2004.

[Bas12] Erol Başar. *Brain Function and Oscillations: Volume II: Integrative Brain Function. Neurophysiology and Cognitive Processes*. Springer Science & Business Media, 2012.

[BCJ03] Robert J Barry, Adam R Clarke, and Stuart J Johnstone. "A review of electrophysiology in attention-deficit/hyperactivity disorder: I. Qualitative and quantitative electroencephalography." *Clinical neurophysiology*, **114**(2):171–183, 2003.

[BD04] György Buzsáki and Andreas Draguhn. "Neuronal oscillations in cortical networks." *science*, **304**(5679):1926–1929, 2004.

[BD11a] A. Bhattacharya and D. B. Dunson. "Sparse Bayesian infinite factor models." *Biometrika*, pp. 291–306, 2011.

[BD11b] Anirban Bhattacharya and David B Dunson. "Sparse Bayesian infinite factor models." *Biometrika*, pp. 291–306, 2011.

[BDD08] SJ Broyd, C Demanuele, S Debener, K Helps, CJ James, and EJ Sonuga-Barke. "Default-mode brain dysfunction in mental disorders: a systematic review." *Nature Biobehavioral Review*, **33**(3):279–296, 2008.

[BDD09] Samantha J Broyd, Charmaine Demanuele, Stefan Debener, Suzannah K Helps, Christopher J James, and Edmund JS Sonuga-Barke. "Default-mode brain dysfunction in mental disorders: a systematic review." *Neuroscience & biobehavioral reviews*, **33**(3):279–296, 2009.

[BDT15] Lisa Byrge, Julien Dubois, J Michael Tyszka, Ralph Adolphs, and Daniel P Kennedy. "Idiosyncratic brain activation patterns are associated with poor social comprehension in autism." *The Journal of Neuroscience*, **35**(14):5837–5850, 2015.

[BDW07] R Bernier, G Dawson, S Webb, and M Murias. "EEG mu rhythm and imitation impairments in individuals with autism spectrum disorder." *Brain and cognition*, **64**(3):228–237, 2007.

[Ber29] Hans Berger. "Über das elektrenkephalogramm des menschen." *European Archives of Psychiatry and Clinical Neuroscience*, **87**(1):527–570, 1929.

[BHI11] Courtney P Burnette, Heather A Henderson, Anne Pradella Inge, Nicole E Zahka, Caley B Schwartz, and Peter C Mundy. "Anterior EEG asymmetry and the modifier model of autism." *Journal of autism and developmental disorders*, **41**(8):1113–1124, 2011.

[BMB14] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting linear mixed-effects models using lme4." *arXiv preprint arXiv:1406.5823*, 2014.

[BMK15] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A Robbins. "The PREP pipeline: standardized preprocessing for large-scale EEG analysis." *Frontiers in Neuroinformatics*, **9**:16, 2015.

[Bri81] D R Brillinger. *Time series: data analysis and theory*. Holden-Day Inc. Oakland, California, 1981.

[Bri01] David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 2001.

[BS16] A M Bastos and JM Schoffelen. "A Tutorial Review of functional connectivity analysis methods and their interpretation pitfalls." *Frontiers in System Neuroscience*, **9**:175, 2016.

[BSB00] Martha W Burton, Steven L Small, and Sheila E Blumstein. "The role of segmentation in phonological processing: an fMRI investigation." *Journal of cognitive neuroscience*, **12**(4):679–690, 2000.

[BSB11] Maria Boersma, Dirk JA Smit, Henrica de Bie, G Caroline M Van Baal, Dorret I Boomsma, Eco JC de Geus, Henriette A Delemarre-van de Waal, and Cornelis J Stam. "Network analysis of resting state EEG in the developing young brain: structure comes with maturation." *Human brain mapping*, **32**(3):413–425, 2011.

[Bur01] Martha W Burton. "The role of inferior frontal cortex in phonological processing." *Cognitive Science*, **25**(5):695–709, 2001.

[CBS76] Bernard Allan Cohen, Enrique J Bravo-Fernandez, and Anthony Sances. "Quantification of computer analyzed serial EEGs from stroke patients." *Electroencephalography and clinical neurophysiology*, **41**(4):379–386, 1976.

[CCP06] Jorge Caiado, Nuno Crato, and Daniel Peña. "A periodogram-based metric for time series classification." *Computational Statistics & Data Analysis*, **50**(10):2668–2684, 2006.

[CFR06] Gilles Celeux, Florence Forbes, Christian P Robert, D Michael Titterington, et al. "Deviance information criteria for missing data models." *Bayesian analysis*, **1**(4):651–673, 2006.

[CHR00] Gilles Celeux, Merrilee Hurn, and Christian P Robert. "Computational and inferential difficulties with mixture posterior distributions." *Journal of the American Statistical Association*, **95**(451):957–970, 2000.

[CJY13] R C Craddock, S Jbadi, C G Yan, J T Vogelstein, F X Castellanos, A Di Martino, C Kelly, K Heberlein, S Colcombe, and M P Milham. "Imaging human connectomes at the macroscale." *Nature Methods*, **10**(6):524–539, 2013.

[CL17] Kehui Chen and Brian Lynch. "Weak Separablility for Two-way Functional Data: Concept and Test." *arXiv preprint arXiv:1703.10210*, 2017.

[CRC07] Ciprian M Crainiceanu, David Ruppert, Raymond J Carroll, Adarsh Joshi, and Billy Goodner. "Spatially adaptive Bayesian penalized splines with heteroscedastic errors." *Journal of Computational and Graphical Statistics*, **16**(2):265–288, 2007.

[CW11] Huaihou Chen and Yuanjia Wang. "A penalized spline approach to functional mixed effects model analysis." *Biometrics*, **67**(3):861–870, 2011.

[Dah06] D.B Dahl. "Model-based clustering for expression data via a Dirichlet process mixture model." In Marina Vannucci, Kim-Anh Do, and Peter Müller, editors, *Bayesian inference for gene expression and proteomics*, pp. 201–215. Cambridge: Cambridge University Press, 2006.

[DAS14] A Di Martino, Fair D A, T D Satterthwaite, F X Castellanos, M E Thomason, R C Craddock, B Luna, B L Leventhal, Zuo X N, and M P Milham. "Unraveling the miswired connectome: a developmental perspective." *Neuron*, **17**(83):1335 – 1353, 2014.

[Dav09] Doug J Davidson. "Functional mixed-effect models for electrophysiological responses." *Neurophysiology*, **41**(1):71–79, 2009.

[DM04] Arnaud Delorme and Scott Makeig. "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis." *Journal of neuroscience methods*, **134**(1):9–21, 2004.

[DOG74] Joseph C Doyle, Robert Ornstein, and David Galin. "Lateral specialization of cognitive mode: II. EEG frequency analysis." *Psychophysiology*, **11**(5):567–578, 1974.

[DSM07] Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis." *Neuroimage*, **34**(4):1443–1449, 2007.

[EMS11] Inge-Marie Eigsti, Ashley B de Marchena, Jillian M Schuh, and Elizabeth Kelley. "Language acquisition in autism spectrum disorders: A developmental review." *Research in Autism Spectrum Disorders*, **5**(2):681–691, 2011.

[EOO15] Carolina Euan, Hernando Ombao, and Joaquin Ortega. "Spectral Synchronicity in Brain Signals." *arXiv preprint arXiv:1507.05018*, 2015.

[FA11] J Fell and N Axamcher. "The role of phase synchronization in memory processes." *Nature Reviews Neuroscience*, **12**(2):105–118, 2011.

[FB14] Andreas Fink and Mathias Benedek. "EEG alpha power and creative ideation." *Neuroscience & Biobehavioral Reviews*, **44**:111–123, 2014.

[FF10] Al A Fingelkurts and An A Fingelkurts. "Short-term EEG spectral pattern as a single event in EEG phenomenology." *The open neuroimaging journal*, **4**(1), 2010.

[FP95] Gernot Florian and Gert Pfurtscheller. "Dynamic spectral analysis of event-related EEG data." *Electroencephalography and clinical neurophysiology*, **95**(5):393–396, 1995.

[FPM04] SP Fitzgibbon, KJ Pope, Lorraine Mackenzie, CR Clark, and JO Willoughby. "Cognitive tasks augment gamma EEG power." *Clinical Neurophysiology*, **115**(8):1802–1809, 2004.

[FR02] Chris Fraley and Adrian E Raftery. "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American statistical Association*, **97**(458):611–631, 2002.

[FRS12] Chris Fraley, AE Raftery, and L Scrucca. "Normal mixture modeling for model-based clustering, classification, and density estimation." *Department of Statistics, University of Washington*, **23**:2012, 2012.

[FS89] Jerome H Friedman and Bernard W Silverman. "Flexible parsimonious smoothing and additive modeling." *Technometrics*, **31**(1):3–21, 1989.

[GCB11] Zhenkun Gou, Naseem Choudhury, and April A Benasich. "Resting frontal gamma power at 16, 24 and 36 months predicts individual differences in language and cognition at 4 and 5 years." *Behavioural brain research*, **220**(2):263–270, 2011.

[GH07] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press , New York, 2007.

[GM02] Thomas Gruber and Matthias M Müller. "Effects of picture repetition on induced gamma band responses, evoked potentials, and phase synchrony in the human EEG." *Cognitive Brain Research*, **13**(3):377–392, 2002.

[GSN10] Louise Goyet, Scania de Schonen, and Thierry Nazzi. "Words and syllables in fluent speech segmentation by French-learning infants: An ERP study." *Brain Research*, **1332**:75–89, 2010.

[Guo02] Wensheng Guo. "Functional mixed effects models." *Biometrics*, **58**(1):121–128, 2002.

[GV03] Alan E Gelfand and Penelope Vounatsou. "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, **4**(1):11–15, 2003.

[HA85] Lawrence Hubert and Phipps Arabie. "Comparing partitions." *Journal of classification*, **2**(1):193–218, 1985.

[HBM13] Stefanie Andrea Hutka, Gavin M Bidelman, and Sylvain Moreno. "Brain signal variability as a window into the bidirectionality between music and language processing: Moving from a linear to a nonlinear model." *Frontiers in psychology*, **4**:984, 2013.

[HLS10] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, Alzheimer's Disease NeuroImaging Initiative, et al. "Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation." *NeuroImage*, **50**(3):935–949, 2010.

[HMS07] RN Henson, J Mattout, Krish Devi Singh, GR Barnes, A Hillebrand, and K Friston. "Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields." *Neuroimage*, **38**(3):422–438, 2007.

[HPS09] Willem de Haan, Yolande AL Pijnenburg, Rob LM Strijers, Yolande van der Made, Wiesje M van der Flier, Philip Scheltens, and Cornelis J Stam. "Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory." *BMC neuroscience*, **10**(1):1, 2009.

[HSJ15] K Hasenstab, C Sugar, S Jeste, D Telesca, K McEvoy, and D Senturk. "Identifying longitudinal trends within EEG experiments." *Biometrics*, **71**(1090 - 1100), 2015.

[HST6a] K Hasenstab, C A Sugar, D Telesca, S Jeste, and D Şentürk. "Robust functional clustering of ERP data with application to a study of implicit learning in autism." *Biostatistics*, **3**:484–498, 2016a.

[HST6b] K Hasenstab, A Scheffler, D Telesca, C A Sugar, C DiStefano, and D Şentürk. "A multi-dimensional functional principal components analysis of EEG data." Technical report, UCLA-Biostatistics, 2016b.

[JGK04] Carrie A Joyce, Irina F Gorodnitsky, and Marta Kutas. "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation." *Psychophysiology*, **41**(2):313–325, 2004.

[JHS00] GM James, TJ Hastie, and CA Sugar. "Principal component models for sparse functional data." *Biometrika*, **87**(3):587–602, 2000.

[JHS05] Ajay Jasra, CC Holmes, and DA Stephens. "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling." *Statistical Science*, pp. 50–67, 2005.

[JKR15] Lutz Jäncke, Jürg Kühnis, Lars Rogenmoser, and Stefan Elmer. "Time course of EEG oscillations during repeated listening of a well-known aria." *Frontiers in human neuroscience*, **9**:401, 2015.

[KHG11] Robert T Krafty, Martica Hall, and Wensheng Guo. "Functional mixed effects spectral analysis." *Biometrika*, **98**(3):583–598, 2011.

[KJJ13] Valesca Kooijman, Caroline Junge, Elizabeth K Johnson, Peter Hagoort, and Anne Cutler. "Predictive brain signals of linguistic development." 2013.

[KKC10] Tatyana Krivobokova, Thomas Kneib, and Gerda Claeskens. "Simultaneous confidence bands for penalized spline estimators." *Journal of the American Statistical Association*, **105**(490):852–863, 2010.

[KKP98] Christina M Krause, Pirjo Korpilahti, Bodil Pörn, Joakim Jäntti, and Heikki A Lang. "Automatic auditory word perception as measured by 40 Hz EEG responses." *Electroencephalography and clinical neurophysiology*, **107**(2):84–87, 1998.

[Kli99] Wolfgang Klimesch. "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis." *Brain research reviews*, **29**(2):169–195, 1999.

[Kra16] Robert T Krafty. "Discriminant Analysis of Time Series in the Presence of Within-Group Spectral Variability." *Journal of time series analysis*, **37**(4):435–450, 2016.

[Kuh04] Patricia K Kuhl. "Early language acquisition: cracking the speech code." *Nature reviews neuroscience*, **5**(11):831–843, 2004.

[LB04] Stefan Lang and Andreas Brezger. "Bayesian P-splines." *Journal of computational and graphical statistics*, **13**(1):183–212, 2004.

[LCL15] Sidong Liu, Weidong Cai, Siqi Liu, Fan Zhang, Michael Fulham, Dagan Feng, Sonia Pujol, and Ron Kikinis. "Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders." *Brain informatics*, **2**(3):167–180, 2015.

[LD13] E F Lock and D B Dunson. "Bayesian consensus clustering." *Bioinformatics*, **29**(20):2610–2616, 2013.

[LKS03] H Laufs, K Krakow, P Sterzer, E Eger, A Beyerle, A Salek-Haddadi, and Kleinschmidt A. "Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest." *Proceedings of the national academy of sciences of the United States of America*, **100**(19):11053–11058, 2003.

[LRL00] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. "The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism." *Journal of autism and developmental disorders*, **30**(3):205–223, 2000.

[LSS16] Qian Li, Damla Senturk, Catherine A Sugar, Shanali Jeste, Charlotte DiStefano, Joel Frohlich, and Donatello Telesca. "Inferring Brain Signals Synchronicity from a Sample of EEG Readings." *arXiv preprint arXiv:1609.09532*, 2016.

[LW82] Nan M Laird and James H Ware. "Random-effects models for longitudinal data." *Biometrics*, pp. 963–974, 1982.

[MCN93] Debra L Mills, Sharon A Coffey-Corina, and Helen J Neville. "Language acquisition and cerebral specialization in 20-month-old infants." *Journal of Cognitive Neuroscience*, **5**(3):317–334, 1993.

[MH03] Karl E Misulis and Thomas Channing Head. *Essentials of clinical neurophysiology*, volume 1. Garland Science, 2003.

[MMD06] Kristin McNealy, John C Mazziotta, and Mirella Dapretto. "Cracking the language code: Neural mechanisms underlying speech parsing." *The Journal of neuroscience*, **26**(29):7629–7639, 2006.

[Mor15] Jeffrey S Morris. "Functional regression." *Annual Review of Statistics and Its Application*, **2**:321–359, 2015.

[MQR11] P Müller, F Quintana, and G L Rosner. "A Product Partition Model with regression on covariates." *Journal of Computational and Graphical Statistics*, **20**(1):260–278, 2011.

[MTN12a] Silvia Montagna, Surya T Tokdar, Brian Neelon, and David B Dunson. "Bayesian latent factor regression for functional and longitudinal data." *Biometrics*, **68**(4):1064–1073, 2012.

[MTN12b] Silvia Montagna, Surya T. Tokdar, Brian Neelon, and David B. Dunson. "Bayesian Latent Factor Regression for Functional and Longitudinal Data." *Biometrics*, **68**(4):1064–1073, 2012.

[MWG07] Michael Murias, Sara J Webb, Jessica Greenson, and Geraldine Dawson. "Resting state cortical connectivity reflected in EEG coherence in individuals with autism." *Biological psychiatry*, **62**(3):270–273, 2007.

[NA15] M Narayan and G I Allen. "Population Inference for Node Level Differences in Functional Connectivity." In *IEEE International Workshop on Pattern Recognition in Neuroimaging*, 2015.

[NA16] Manjari Narayan and Genevera I Allen. "Mixed Effects Models for Resampled Network Statistics Improves Statistical Power to Find Differences in Multi-Subject Functional Connectivity." *Frontiers in neuroscience*, **10**, 2016.

[NC07] Nam Nguyen and Rich Caruana. "Consensus clusterings." In IEE Computer Society, editor, *In Proceedings of the 7th IEEE International Conference on Data Mining*, pp. 607–612, 2007.

[NCT12] Scott Neu, Karen Crawford, and Arthur W Toga. "Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories." *Frontiers in neuroinformatics*, **6**:8, 2012.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. "On Spectral Clustering: Analysis and an Algorithm." In *Advances in Neural Information Processing Systems*, pp. 849–856. MIT Press, 2001.

[OH06] Hernando Ombao and Moon-ho Ringo Ho. "Time-dependent frequency domain principal components analysis of multichannel non-stationary signals." *Computational statistics & data analysis*, **50**(9):2339–2360, 2006.

[ORS01] H C Ombao, J A Raz, R L Strawderman, and Von Sachs R. "A simple generalized cross-validation method of span selection for periodogram smoothing." *Biometrika*, **88**(4):1186–1192, 2001.

[PLB11] Jeremy R Parr, Ann Le Couteur, Gillian Baird, Michael Rutter, Andrew Pickles, Eric Fombonne, Anthony J Bailey, International Molecular Genetic Study of Autism Consortium, et al. "Early developmental regression in autism spectrum disorder: Evidence from an international multiplex sample." *Journal of autism and developmental disorders*, **41**(3):332–340, 2011.

[Rup02] David Ruppert. "Selecting the number of knots for penalized splines." *Journal of computational and graphical statistics*, **11**(4):735–757, 2002.

[RW14] Donald C Rojas and Lisa B Wilson. "γ-band abnormalities as markers of autism spectrum disorders." *Biomarkers in medicine*, **8**(3):353–368, 2014.

[RWC09] David Ruppert, Matt P Wand, and Raymond J Carroll. "Semiparametric regression during 2003–2007." *Electronic journal of statistics*, **3**:1193, 2009.

[RWS12] Ori Rosen, Sally Wood, and David S Stoffer. "AdaptSPEC: Adaptive spectral estimation for nonstationary time series." *Journal of the American Statistical Association*, **107**(500):1575–1589, 2012.

[SBC02] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4):583–639, 2002.

[Sch78] Gideon Schwarz et al. "Estimating the dimension of a model." *The annals of statistics*, **6**(2):461–464, 1978.

[SNM14] A Shou, H adn Eloyan, M B Nebel, A Mejia, J J Pekar, S Mostofsky, B Caffo, M A Lindquist, and C Crainiceanu. "Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI." *Neuroimage*, **15**(102):938–944, 2014.

[SNT07] Tatiana A Stroganova, Gudrun Nygren, Marina M Tsetlin, Irina N Posikera, Christopher Gillberg, Mikael Elam, and Elena V Orekhova. "Abnormal EEG lateralization in boys with autism." *Clinical Neurophysiology*, **118**(8):1842–1854, 2007.

[SR10] Russell J Steele and Adrian E Raftery. "Performance of Bayesian model selection criteria for Gaussian mixture models." *Frontiers of statistical decision making and bayesian analysis*, pp. 113–130, 2010.

[SS10] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, 2010.

[Tep02] M Teplan. "Fundamentals of EEG Measurements." *Measurement Science Review*, **2**(2), 2002.

[TT04] Nitish V Thakor and Shanbao Tong. "Advances in quantitative electroencephalogram analysis methods." *Annu. Rev. Biomed. Eng.*, **6**:453–495, 2004.

[TWG89] Robert W Thatcher, RA Walker, I Gerson, and FH Geisler. "EEG discriminant analyses of mild head trauma." *Electroencephalography and clinical neurophysiology*, **73**(2):94–106, 1989.

[US10] Peter J Uhlhaas and Wolf Singer. "Abnormal neural oscillations and synchrony in schizophrenia." *Nature reviews neuroscience*, **11**(2):100–113, 2010.

[WBE13] Jun Wang, Jamie Barstein, Lauren E Ethridge, Matthew W Mosconi, Yukari Takarae, and John A Sweeney. "Resting state EEG abnormalities in autism spectrum disorders." *Journal of neurodevelopmental disorders*, **5**(1):1, 2013.

[XT15] Dongkuan Xu and Yingjie Tian. "A comprehensive survey of clustering algorithms." *Annals of Data Science*, **2**(2):165–193, 2015.

[XW05] Rui Xu and Donald Wunsch. "Survey of clustering algorithms." *IEEE Transactions on neural networks*, **16**(3):645–678, 2005.

[You00] G Bryan Young. "The EEG in coma." *Journal of Clinical Neurophysiology*, **17**(5):473–485, 2000.

[ZE92] Robert J Zatorre, Alan C Evans, et al. "Lateralization of phonetic and pitch discrimination in speech processing." *Science*, **256**(5058):846, 1992.

[ZHM10] Lan Zhou, Jianhua Z Huang, Josue G Martinez, Arnab Maity, Veerabhadran Baladandayuthapani, and Raymond J Carroll. "Reduced rank mixed effects models for spatially correlated hierarchical functional data." *Journal of the American Statistical Association*, **105**(489):390–400, 2010.

[ZTS09] Dongli Zhou, Wesley K Thompson, and Greg Siegle. "MATLAB toolbox for functional connectivity." *Neuroimage*, **47**(4):1590–1607, 2009.