Machine Learning Methods for Statistical Analysis and Representation Learning on Neuroimaging Data

by

FAN YANG

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2022

To God and my beloved everyone.

ACKNOWLEDGEMENTS

I have been blessed to have mentors, family and friends whose support empowered me to continue on this long journey, and those past years studying at UT Arlington have been the most precious life experience to me. Now at this moment, I am glad that I have reached an important milestone in this amazing journey and I would like to take this opportunity to thank them.

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Won Hwa Kim, whose exceptional advice, guidance and support has been invaluable throughout my doctoral studies. Dr. Kim taught me how to do proper research, writing good papers, making informative slides, basically everything that will be needed and play an important role in my future career. In addition, I would like to sincerely thank my committee members: Dr. Gautam Das, Dr. Dajiang Zhu and Dr. Amal Isaiah whose assistance and support have helped me tremendously. I am grateful for their interest in my research and for taking their valuable time to serve in my committee. It is my great pleasure to hear their suggestions and comments on my research.

I would like to extend my greatest appreciation to the amazing faculty and staff at the Computer Science and Engineering department at the University of Texas at Arlington for their precious support. I am very thankful to receive the prestigious GAANN fellowship in CSE by the U.S. Department of Education, which is principally investigated by Dr. Ishfaq Ahmad. I was fortunate to have received invaluable advise and instructions from Dr. Bahram Khalili, as well as assistance and help from other CSE advising staff memebers, especially Ms. Ginger Dickens, Ms. Sherri Gotcher and Ms. Pamela Mcbride. I would also like to thank my fellow doctoral students and friends I met there for making those years a most

remarkable and memorable experience, and their presence was one of the reasons that my time there was always enjoyable.

Last but not the least, I would like to express my earnest gratitude to my beloved for their unwavering love to support me and my dreams. Their redeeming love has been the main source of encouragement and true inspiration for every step I take during my whole life. Without their patience and support, it would not have been possible to reach this stage in the journey of my life. Their love and support define who I am today. I am extremely grateful, lucky and at the same time humbled to be a part of such a lovely family.

April 2022

ABSTRACT

Machine Learning Methods for Statistical Analysis and Representation Learning on Neuroimaging Data

Fan Yang, Ph.D.

The University of Texas at Arlington, 2022

Supervising Professor: Dr. Won Hwa Kim

With the recent advance and widespread adoption of imaging technological innovations, clinical practitioner and scientists can easily acquire and store a large amount of various neuroimaging modalities, such as Diffusion Tensor Imaging (DTI), Magnetic Resonance Imaging (MRI), resting-state functional MRI (rs-fMRI) and Positron Emission Tomography (PET), etc. These novel imaging data sources cover a rich amount of factors that influence patients' cognitive health, offer an objective view of patients at unprecedented multi-resolution for the understanding of brain structure and function, and have the significant potential to improve healthcare by aiding better decision-making in diagnosing, monitoring and treating diseases. Machine Learning methods have emerged as the state-of-the-art in learning from the large-scale neuroimaging data. While their use for medical applications is interesting and insightful, it is often very challenging in practice. Some of the major challenges we encounter in the adoption of Machine Learning methods for neuroscience tasks are that examining the association between the socioeconomic characteristic and brain clinical measurements is difficult given the subtle variations between groups with different socioeconomic status, that effectively characterizing the early symptoms of the

Alzheimer's disease (AD) is in many cases not possible, that forecasting and capturing the disease-related dynamics of clinical measurements is necessary to better understand the progression of AD, and that modelling the dynamic associations between lengthy sequences of multivariate variables for brain connectivity analysis is computational expensive. To take care of these challenges, we propose multiple novel Machine Learning methods for providing a multi-scale representation of the original measurement to enhance the sensitivity of downstream statistical analysis, for integrating graph structure and diagnostic label information to characterize early symptoms of AD, for incorporating time-dependent label information to better understand the progression of AD, and for efficiently estimate and predict dynamic covariances on large-scale time series data. We demonstrate our developed methods on the challenging real-word data from various clinical studies in the neuroscience domain, including Adolescent Brain Cognitive Development (ABCD) study, Alzheimer's Disease Neuroimaging Initiative (ADNI) study, and Human Connectome Project (HCP). Our contributions advance the state-of-the art in regard to leveraging Machine Learning methods for neuroscience applications and accentuate the foreground in which artificial intelligence on large-scale neuroimaging data can improve healthcare with better decision-makings.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

Introduction

Machine learning methods have emerged as the de-facto state-of-the-art for learning from large-scale neuroimaging data, however, their use on neuroscience applications is customarily challenging in practice. In this thesis, we focus on developing novel machine learning techniques and models for the purpose of handling various real-world neuroimaging modalities, such as Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), resting-state functional MRI (rs-fMRI), and Positron Emission Tomography (PET), etc. Our contributions have advanced the state-of-the art in regard to utilizing machine learning methods for neuroscience applications, which exemplifies the significant potential that artificial intelligence and machine learning on large-scale neuroimaging data can improve healthcare tasks with better decision-makings. In this chapter, we will first discuss the motivations in Section 1.1 which includes what kind of neuroimaging data we are interested in and how those data look like. Sections 1.2 and 1.3 discuss what kind of technical challenges we are facing and how those data inspire our methodologies development. Lastly, we display an outline of the structure of this thesis in Section 1.4.

1.1 Motivations

With the advent of state-of-the-art machine learning methods, it has been proven tremendously successful on a wide range of application problems from various domains, such as recommendation engines in e-commerce, self-driving in computer vision and automatic translation in applied linguistics. Machine learning, the study of inferring and learning patterns from data, represents an automated learning pipeline over massive datasets,

which makes the learning process more efficient and cost-effective. In addition, machine learning methods also benefit from exponentially growing computational resources and data acquisition techniques. Due to the aforementioned reasons, machine learning methods have enjoyed wide popularity in both academic research and industry communities. However, how to effectively extract knowledge from and efficiently exploit neuroimaging data in the neuroscience domain is still under-explored.

There are various forms of neuroimaging modalities used in clinical practice and scientific research which characterize different aspects of the brain, as shown in Figure 1.1. Broadly speaking, magnetic resonance imaging (MRI) is one of the most commonly used imaging techniques which is based on the magnetization properties of atomic nuclei in the brain to provide high-resolution imaging of brain structure and physiology [3]. T1-weighted (T1-w) scans are the most standard of MRI sequences whose scans have two major uses: localization of brain regions and estimation of tissue density [1]. Diffusion MRI (dMRI) measures the strength and direction of water molecules' movement during the scan [4]. When more volumes (e.g., above 30 directions) are acquired to increase directional resolution and allow tensor-based analyses, such sequences will be referred to as diffusion-tensor imaging (DTI) [1]. The aforementioned MRI sequences all focus on capturing and imaging the brain structures. On the contrary, functional MRI (fMRI) images the functions by using a mechanism called the blood-oxygen-level-dependent (BOLD) signal [1, 5]. Moreover, positron emission tomography (PET) scans utilize radioactive tracers that are injected in the body to characterize metabolic changes at the cellular level in the brain tissue.

With this wealth of information, a natural question to ask in this context is whether and how we can potentially leverage these vast amounts of messages from the neuroimaging data to help with disease diagnosis, disease progression prediction and statistical analyses that lead to better decision-making in healthcare. Our primary objective is to advance the state-of-the-art in the use of machine learning methods and neuroimaging datasets for neuroscience

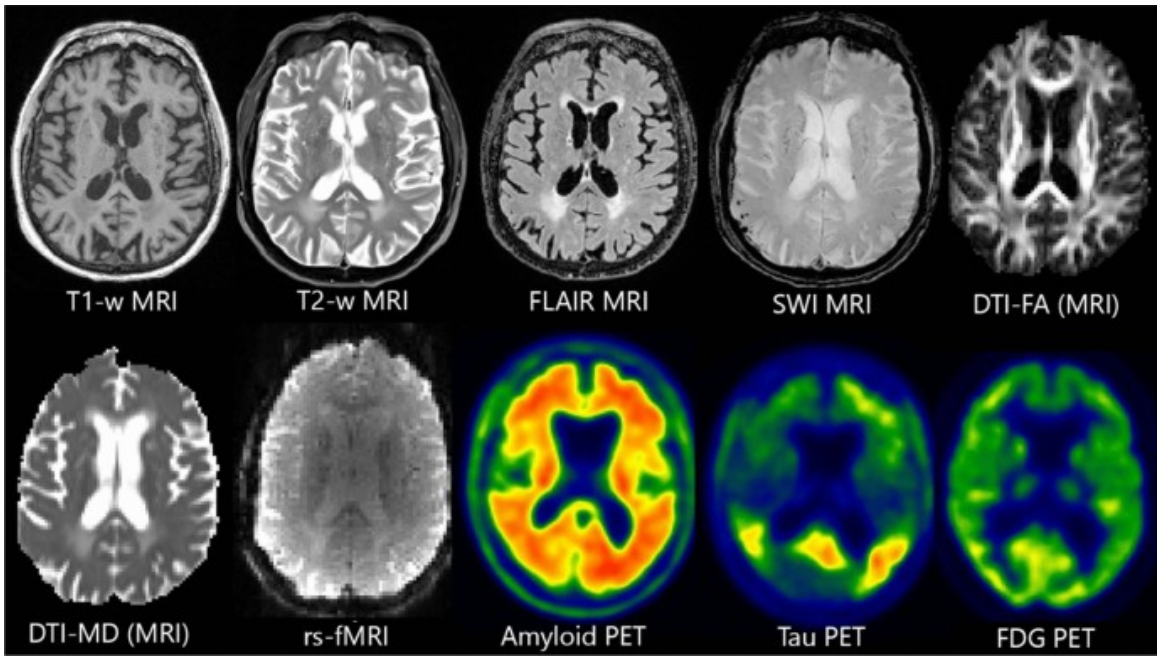


Figure 1.1: An example of various neuroimaging modalities [1].

applications. Specifically, our goals are to make novel methodological contributions that enable the use of machine learning methods on neuroimaging data to extract rich feature representations from raw data to help with the downstream tasks. We additionally aim to evaluate these methodological contributions on real-world datasets from the neuroscience domain.

1.2 Technical Challenges

The focus of this thesis lies primarily in addressing the technical challenges associated with the use of machine learning in the neuroscience domain. Precisely, we aim to develop effective and efficient machine learning approaches to handle neuroimaging data from various imaging modalities for applications in the neuroscience domain. Some of the main challenges we are confronted by in applying machine learning methods for neuroscience problems are described as below.

First of all, we study the problem of multi-resolutional statistical analysis on the Adolescent Brain Cognitive Development (ABCD) study dataset, in which we can perform a better statistical analysis based on different multi-resolutions of feature representations. Specifically, we aim to examine the association between household income (one major socioeconomic characteristic) and fractional anisotropy (one of microstructural characteristics) of cortical regions from the ABCD study. While ABCD offers sufficient data (i.e., a large sample size), it is still difficult to capture the subtle variations in fractional anisotropy (FA) between closely-spaced groups within the spectrum of socioeconomic strata. To detect even the subtle variations in the brain, it is imperative to have a more sensitive method that transforms the data into a new domain where the differences between the groups can be captured better. Inspired from wavelet transform in traditional signal processing, in order for a sensitive method, we aim to develop a novel transform that derives a multi-scale feature from structured data $X \in \mathbb{R}^{N \times P}$ with N samples and P features to improve the downstream analysis. As we know, wavelet transform will transform a given signal $f(x)$ in the Euclidean space to the frequency space and yield a multi-scale representation of the original signal $f(x)$ [6]. The central concept from wavelet transform is that wavelets behave as band-pass filters in the frequency space [7], thus if we can design a filter in the frequency space using a set of orthogonal bases (e.g., Fourier bases), we are able to develop a novel transform that derives multi-scale representation of the original signal even in a complex domain [8, 9]. However, our main barrier is that for the given structured data (i.e., FA measures on regions of interest (ROIs) from N subjects) in the ABCD study, the underlying space of the data remains unknown to us.

Secondly, we examine the problem of disentangled representation learning on the Alzheimer's Disease Neuroimaging Initiative (ADNI) study dataset, in which we can develop a disentangled representation learning framework to extract meaningful dynamic and static features from neuroimaging data. Particularly, we first aspire to be able to effectively

characterize the early symptoms of the Alzheimer's Disease (AD) based on the longitudinal cortical thickness measures from the ADNI study. AD is a progressive neurodegenerative condition which is characterized by the neurodegenerations in the brain [10, 11, 12, 13, 14], hence it is of great importance that the symptoms of the disease can be effectively characterized early. In this regard, we would like to analyze the longitudinal cortical thickness measures over structural brain networks with diagnostic labels of AD, and it is expected that disentangling those longitudinal measures with time-invariant and time-varying components could provide unique insights on the characterization of the early disease symptoms. Specifically, we aim to develop a framework that learns a latent disentangled representation comprising time-varying and time-invariant components of the observations to characterize the early symptoms of the disease. However, although recent works [15, 16] on variational autoencoders [17] are able to learn various representations in an unsupervised manner, they do not introduce supervision at all when dealing with temporal data, nor do they consider the structural brain connectivity from DTI which can provide a prior knowledge on the topology of specific ROIs in the brain. Second, we further seek to capture the disease-related dynamics of β-amyloid and forecast amyloid depositions at future timestamps from the ADNI study, to better early characterize the progression of AD. Unfortunately, due to the fact that only limited timestamps (i.e., less than 3 timestamps on average) are available per subject, it is very difficult to learn the complex dynamics of β-amyloid over brain networks. Moreover, the observations from imaging scans can be complexly affected by many factors both time-varying and time-invariant, such as anatomical structure and disease effect. Hence, we aim to disentangle the observations in the latent space that composed of time-invariant component (e.g., anatomical information) and time-varying component (e.g., dynamics changes), which will help make the model more explainable and easier to control the conditional data generation. However, prior works [18, 19] on disentangled representations mainly focused on video or audio data, the research for representation disen-

tanglement of neuroimaging sequential data has been under-explored, let alone incorporating the time-dependent label information and graph structure and performing the forecasting task as well.

Lastly, we investigate the problem of dynamic covariance modeling on the HCP dataset, in which we can develop a non-parametric statistical model for the dynamic covariance modeling yielding a multi-scale descriptor feature which is significantly informative to clinical behavior scores. Specifically, we aim to study modelling temporal associations between a sequence of multivariate variables in a large-scale functional Magnetic Resonance Imaging (fMRI) dataset from Human Connectome Project (HCP). Typically for brain connectivity analyses in Neuroimaging, for the ease of convenience, in the conventional connectivity constructions, the associations or covariance between a knot of timeseries measurements across different ROIs in the brain are assumed to be static in time [20, 21, 22]. Nevertheless, it has been demonstrated in several studies that brain connectivities change over time whose dynamic variation may be significant [23, 24]. Therefore, to investigate the time-varying coactivation patterns in brain activities [23, 25, 26], modelling such dynamic changes of covariance between ROIs is an essential problem in both Machine Learning and Neuroscience. However, the recent works based on Wishart process (WP) [27, 28] are limited as they often have computational issues due to the computing burden induced from latent Gaussian processes.

1.3 Our Approaches

In this thesis, we address these technical challenges by developing novel methodological approaches leveraging machine learning on neuroimaging data, and we demonstrate the potential of applying these approaches to the challenging problems in the neuroscience domain. This thesis is an interdisciplinary work which includes novel contributions both in

terms of methodological advances in machine learning domain, as well as in the applications of machine learning methods to challenging problems in neuroscience domain. The main contributions of this work are summarized as follows.

First of all, to deal with the problem of designing a framework which yields a multi-scale representation of the original data for improving the statistical sensitivity for downstream analyses, we leverage a precision matrix which is the inverse of covariance matrix. Since it is symmetric and positive definite, it has a set of orthonormal eigenvectors which can be used to develop an orthogonal transform with the filter function defined over the spectrum of eigenvalues. Thus, we can define a multi-scale descriptor based on the developed transform which provides a multi-scale representation of the original measurement to enhance the downstream statistical analysis. We validate our framework on the ABCD dataset to demonstrate the significant performance improvements of our framework over raw measurements in identifying clinically meaningful cortical ROIs which are susceptible to socioeconomic inequality.

Secondly, to cope with the problem of disentangled representation learning for better early characterization of AD, we first propose an innovative Semi-supervised Sequential Graph Autoencoder model, which leverages ideas from the sequential variational autoencoder, graph convolution and semi-supervising frame-work, to learn a latent disentangled representation of the observations that are composed of time-varying and time-invariant components. We not only incorporate the label information as a supervision to balance between extraction of underlying structure and accurate prediction of class labels, but also integrate graph structure to help robustly learn the disentangled latent space. Our proposed method is validated on the longitudinal cortical thickness data from ADNI study to demonstrate its benefits for effective latent representation on longitudinal data for diagnostic label prediction and longitudinal data generation. Second, to further characterize the longitudinal β-amyloid over the structural brain network for better understanding of AD progression, we develop a

framework that learns a latent disentangled representation composed of time-varying and time-invariant latent variables to capture the disease-related dynamics of β-amyloid and as well as forecast the future amyloid depositions using the disentangled representation. We not only incorporate the time-dependent label as a supervision in the model to characterize longitudinal effect with more effective representation learning, but also integrate a brain network to make the framework more robust to the subject-wise heterogeneous dynamics when learning the disentangling latent representation, as it could consider the arbitrary topology of brain networks across the population. Furthermore, we validate this framework to longitudinal β-amyloid data on brain networks with diagnostic labels of AD from ADNI. The experimental results suggests a significant potential that this framework will facilitate clinical research by overcoming the limitation of amyloid data collection and help physicians better understand the role of amyloid in the progression of AD before the disease symptoms manifest.

Lastly, to tackle the problem of modelling dynamic changes of covariance, we develop a Predictive Wishart Process ($\mathcal{PWP}$) which is a novel parsimonious stochastic process providing a collection of positive semi-definite random matrices indexed by input variables. We thoroughly study its stochastic properties and propose a posterior inference associated with our hierarchical models. We also provide a multi-task learning framework using our proposed $\mathcal{PWP}$ to jointly model multiple large-scale signals. The reconstructive performance and predictive performance for dynamic covariances are demonstrated on a large-scale real fMRI dataset from HCP, which empirically prove the efficiency and practicality of our framework.

1.4 Organization

In a nutshell, various novel machine learning methods are proposed throughout this thesis which can facilitate the use of machine learning on neuroimaging data to extract meaningful feature representations to help with the downstream tasks. In the following paragraphs, we briefly provide an overview of this thesis which consists of three parts as described below.

Part I focuses on the work on the multi-resolutional statistical analysis on the ABCD dataset including Chapter 2, in which we present a novel covariance-based wavelet-like transform (COVLET) yielding a multi-scale representation of the original feature measures that increases the performance of downstream analyses [29].

Part II discusses the work on the disentangled representation learning on the ADNI dataset containing Chapter 3 and Chapter 4. In Chapter 3, we propose an innovative and ground-breaking Disentangled Sequential Graph Autoencoder which learns a latent disentangled representation composed of time-variant and time-invariant latent variables to characterize the longitudinal measurements [30]. Chapter 4 presents a supervised sequential graph variational autoencoder to capture the disease-related dynamics and as well as forecast the future measurements using the learned disentangled representation [31].

Part III focuses on the work on the dynamic covariance modeling on the HCP dataset including Chapter 5, in which we develop a novel parsimonious stochastic process named as Predictive Wishart Process ($\mathcal{PWP}$) which provides a collection of positive semi-definite random matrices indexed by input variables and a model based multi-scale descriptor feature.

With the proposal of novel frameworks and extensive experimental results on real-world datasets, we advance and demonstrate the state-of-the-art in the use of machine learning methods on neuroimaging data. The full description on those proposed machine learning approaches are introduced in the subsequent chapters.

Part I

Multi-resolutional Statistical Analysis on ABCD Data

CHAPTER 2

COVLET: Covariance-based Wavelet-like Transform for Statistical Analysis of Brain Characteristics in Children

Adolescence is a period of substantial experience-dependent brain development. A major goal of the Adolescent Brain Cognitive Development (ABCD) study is to understand how brain development is associated with various environmental factors such as socioeconomic characteristics. While ABCD study offers a large sample size, it still requires a sensitive method to detect subtle associations when studying typically developing children. Therefore, we propose a novel transform, i.e. covariance-based multi-scale transform (COVLET), which derives a multi-scale representation from a structured data (i.e., P features from N samples) that increases performance of downstream analyses. The theory driving our work stems from wavelet transform in signal processing and orthonormality of the principal components of a covariance matrix. Given the microstructural properties of brain regions from children enrolled in the ABCD study, we demonstrate a multi-variate statistical group analysis on family income using the multi-scale feature derived from brain structure and validate improvement in the statistical outcomes. Furthermore, our multi-scale descriptor reliably identifies specific regions of the brain that are susceptible to socioeconomic disparity.

2.1 Introduction

Adolescence is a period of rapid brain development shaped by genetic, physiologic and socioeconomic variables [32, 33]. While previous studies have utilized techniques largely focusing on macrostructural properties of the cerebral cortex such as its thickness, surface

area and volume [34], the microstructural characteristics such as diffusion of water along the tracts of the neuronal fibers may provide insights into the functional properties of the brain [35]. These properties, measured with high-resolution Diffusion Tensor Imaging (DTI), have previously been used to study the association of socioeconomic disadvantage with brain structure and function in children. Fractional anisotropy (FA) from DTI, representing the diffusion of water perpendicular to the orientation of the neuronal fibers, has significant potential to identify association of the brain characteristics with neurobehavioral outcomes such as cognitive development [36, 37, 38].

The Adolescent Brain and Cognitive Development (ABCD) study [39], a longitudinal assessment of nearly 12,000 children commencing at the age of 9-11 years through adulthood, provides an unprecedented opportunity to explore development of the brain in adolescence with novel statistical tools. Therefore, we used the baseline dataset from the ABCD study (version 2.0.1) to examine the association between major socioeconomic characteristics, household income and microstructural characteristics (i.e., FA) of cortical regions. While ABCD provides a large sample size, the subtle variations in FA between closely-spaced groups within the spectrum of socioeconomic strata are still difficult to capture, requiring a more sensitive method that transforms the data into a new domain where the differences between the groups can be ascertained better.

Here, we have developed a novel transform that derives a multi-scale feature from structured data $X \in \mathbb{R}^{N \times P}$ with N samples and P features, which can improve its down-stream analyses. The technical core of our method is inspired from wavelet transform in traditional signal processing. Wavelet transform transforms a signal $f(x)$ in x (in the Euclidean space) to the frequency space and its wavelet representation yields "multi-scale" representation of the original signal $f(x)$ [6]. Such a multi-scale representation of signalshas provided successful results in Computer Vision for providing efficient features for robust comparisons of images [40, 41, 42], and shown benefits for statistical group analysis [43, 44].

Wavelets behave as band-pass filters in the frequency space where the scales are defined by the bandwidth covered by the filters [7]. Therefore, if such a filter can be designed in a dual space (e.g., frequency space) defined by orthogonal bases (e.g., Fourier bases), we can design a novel transform that derives multi-scale representation of signals even in a complex domain [8, 9].

The main barrier in our setting is that we are given with a structured data (e.g., FA of regions of interest (ROIs) from N subjects) instead of images, where the underlying space of the data is unknown. In this scenario, we utilize a precision matrix, i.e., inverse covariance matrix, which is symmetric and positive definite and thus has a set of "orthonormal" eigenvectors. The orthonormality lets us define an orthogonal transform, and together with filter functions on the spectrum of eigenvalues, we can design a novel multi-scale representation of the original measurement X. With a premise that multi-scale comparison of data can enhance downstream inference [45], we define a multi-scale descriptor based on the developed transform to increase sensitivity of a statistical analysis.

In summary, we propose a framework which utilizes a precision matrix to provide a multi-scale descriptor on the original features. Our main contributions are: 1) We develop a novel **cov**ariance-based wave**let**-like transform (**COVLET**) which delivers a multi-scale representation of the original feature measures; 2) We conduct extensive experiments on ABCD dataset, which demonstrates significant performance improvements over raw measurements; 3) We identify clinically meaningful cortical ROIs, which are susceptible to socioeconomic inequality.

2.2 Continuous Wavelets Transform in the $L^2(\mathbb{R})$ Space

Conventional wavelets transform is well understood in the $L^2(\mathbb{R})$ space and is fundamental to our proposed framework. To make this paper self-contained, we provide a brief review of wavelets transform in this section.

Wavelets transform transforms a signal $f(x)$ to the frequency space by decomposing the signal $f(x)$ as a linear combination of oscillating basis functions and their coefficients [6]. Although similar to Fourier transform, however, wavelet transform make use of a localized basis function, i.e., providing a compact finite support that is centered at a specific position. This contrasts wavelet transform from Fourier transform which uses $sin()$ as a basis with infinite duration.

Wavelet transforms require a mother wavelet $\psi_{s,a}$ as the basis. The scale parameter s and translation parameter a control the dilation and location of the mother wavelet respectively. A set of mother wavelets is formalized as

$$\psi_{s,a} = \frac{1}{s}\psi(\frac{t-a}{s}). \tag{2.1}$$

The forward wavelet transform of a signal $f(x)$ is defined as an inner product of these wavelets with the signal $f(x)$, which yields wavelet coefficients $W_f(s,a)$ as

$$W_f(s,a) = \langle f, \psi_{s,a} \rangle = \frac{1}{s}\int_{-\infty}^{\infty} f(x)\psi^*(\frac{x-a}{s})dx, \tag{2.2}$$

where ψ^* is complex conjugate of ψ. Moreover, defining the scaling in the Fourier domain let us further express the wavelet coefficients as

$$W_f(s,a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega a}\widehat{\psi^*}(s\omega)\hat{f}(\omega)d\omega, \tag{2.3}$$

where $\hat{f}(\omega)$ denotes the Fourier representation of the $f(x)$ in the frequency space ω [7]. Briefly, (2.3) suggests that filtering $\hat{f}$ at multiple scales at s with the mother wavelet $\hat{\psi}$ offers a multi-scale view of the original signal $f(x)$.

2.3 Multi-scale Analysis via Covariance: COVLET

Let $X \in \mathbb{R}^{P\times N}$ be a standardized (zero-mean) feature matrix with N samples, each of which has P features. Computing a covariance matrix from X yields $\Sigma_{P\times P} = \frac{1}{N}XX^T$, and a precision matrix is defined as the inverse convariance matrix, $\Omega = \Sigma^{-1}$. In the multivariate normal distribution setting, the precision matrix reveals conditional independence relations across different variables, i.e., pair-wise features, as a graphical model [46]. Specifically, $\Omega_{ij} = 0$ implies that features x_i and x_j are conditionally independent given other features $\{x_k\}_{k\neq i,j}$.

The precision matrix $\Omega_{P\times P}$ is symmetric and positive definite (p.d.), and thus has a set of positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_P$ with corresponding orthonormal eigenvectors $\nu_1, \nu_2, \ldots, \nu_P$ as bases. Then we define a Hilbert space H on $\mathbb{R}^P$ with inner product such that $\langle f, h\rangle = \sum_{i=1}^{P} f_i h_i \in \mathbb{R}$ for any $f, h \in H$.

With the background above, we can define an orthogonal transform for any signal/measurement $f \in H$, where the transformed signal $\hat{f}$ is given by

$$\hat{f}(\ell) = \langle \nu_\ell, f\rangle = \sum_{p=1}^{P} \nu_\ell(p) f(p), \tag{2.4}$$

and its inverse transform expresses the original f as an expansion using the $\hat{f}$ as

$$f(p) \;\; = \;\; \sum_{\ell=1}^{P} \hat{f}(\ell)\nu_\ell(p). \tag{2.5}$$

Due to the orthonormality of the bases ν_ℓ, a Parseval relation exists such that $\langle f, h\rangle = \langle \hat{f}, \hat{h}\rangle$. Based on the precision matrix Ω, we define a linear bounded operator $T_g^s \in B(H, H)$ at scale s such that

$$T_g^s \nu_\ell = g(s\lambda_\ell)\nu_\ell, \tag{2.6}$$

for any eigenvector ν_ℓ, where g is a bounded operator from $\mathbb{R}$ to $\mathbb{R}$ (i.e., a kernel function as a band-pass filter). We term it as Covariance-wavelet (Covlet) operator. Based on the definition

of the Covlet on eigenvectors, it can be naturally extended on the whole eigenspace. Since eigenvector bases are complete on $\mathbb{R}^P$, T_g^s is well defined on $\mathbb{R}^P$.

Lemma 2.3.0.1. *T_g^s is a self-adjoint operator, i.e., $\langle T_g^s f, h\rangle = \langle f, T_g^s h\rangle$.*

The self-adjoint property from Lemma A.1.0.1, whose proof is given in supplement, together with (2.4) consequently implies that

$$\widehat{T_g^s f}(\ell) \quad = \quad \langle \nu_\ell, T_g^s f\rangle = \langle T_g^s \nu_\ell, f\rangle = g(s\lambda_\ell)\hat{f}(\ell), \tag{2.7}$$

which means that the operator is equivalent to applying a filter function $g(\cdot)$ on top of coefficients $\hat{f}$. According to equation (2.7) and the orthonormal property, applying the inverse transform in (2.5) then shows

$$T_g^s f(p) = \sum_{\ell=1}^{P} g(s\lambda_\ell)\hat{f}(\ell)\nu_\ell(p), \tag{2.8}$$

where the operator T_g^s is applied on the p-th feature. This operation in (2.8) is defined as the Covlet transform of an original signal $f(p)$ (i.e. feature) as

$$C_f(s,p) = \langle T_g^s \delta_p, f\rangle, \tag{2.9}$$

which yields Covlet coefficients $C_f(s,p)$ where δ_p denotes a Dirac delta function at p. As claimed by Lemma A.1.0.1, the self-adjoint property implies that

$$C_f(s,p) \quad = \quad \langle \delta_p, T_g^s f\rangle = T_g^s f(p). \tag{2.10}$$

We observe a close analogy between our Covlet operator and the conventional wavelets operator as they both define the mapping through bases. However, as indicated by equations (2.3) and (2.8), they utilize different sets of bases according to eigenvectors of the precision matrix and Fourier bases, respectively. Furthermore, such a transform delivers a multi-scale view of signals defined on each features by repeating this procedure for multiple scales. Therefore, we define the Covlet Multi-scale Descriptor (CMD) as a set of Covlet coefficients on each feature p for each scale in $\mathcal{S} = \{s_1, s_2, \cdots, s_{|S|}\}$,

$$\text{CMD}_f(p) = \{C_f(s,p) | s \in \mathcal{S}\}. \tag{2.11}$$

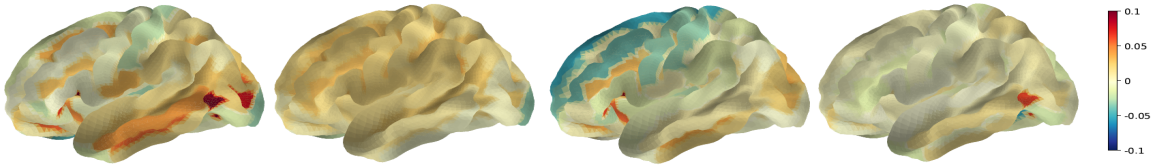


Figure 2.1: Multi-scale FA. First: FA measures, Second: CMD at scale $s_1 = 2.11e-05$, Third: CMD at scale $s_2 = 4.27e-06$, Fourth: CMD at scale $s_3 = 8.67e-07$.

This CMD is a multi-scale feature that is derived from the original univariate measurement/signal by performing multi-scale filtering in a dual space spanned by the eigenvectors ν, i.e., PCA. An example of (standardized) CMD from FA measure is shown in Fig. 2.1. It captures local context along the geometry of the manifold where the data X are defined. When the geometry is given as a graph, the Ω in our framework can be replaced by graph Laplacian and will be formulated as Spectral Graph Wavelet Transform (SGWT) in [7].

2.4 Identifying Changes in Microstructure of Neuron Tracts with Family Income

2.4.1 Experimental Design

Dataset. The ABCD study is the largest long-term study on brain development [39] and child health in the U.S. supported by the National Institutes of Health (NIH). The dataset included 11,873 children enrolled by October 2018, which is also pre-packaged and publicly available (version 2.0.1) on the National Institute of Mental Health Data Archive (NDA) under the data use agreement.

Table 2.1: Demographics of the ABCD study.

Demographics	BP	NP	H	M	L
# of Subjects	954	8883	4208	2794	2835
Gender (M/F)	488/466	4610/4273	2000/2208	1345/1449	1394/1441
Age (mean,std)	118.5 ± 7.3	119.1 ± 7.5	119.4 ± 7.5	118.8 ± 7.5	118.7 ± 7.5

BP: below-poverty, NP: non-poverty, H: high, M: middle, L: low; Age is measured in months.

For our experiments, children were grouped based on household income level. In the first analysis, they were separated into two groups by the poverty criteria from U.S. Census Bureau; the threshold in the U.S. for a single parent family was $\$16,910$ [47]. We defined a below-poverty (BP) group with the subjects with the family income level below level 4 in the dataset (i.e., $< \$16,000$), and a non-poverty (NP) group with the remaining subjects. For the second analysis, subjects were divided into three groups based on the following household income bracket [48]: we regarded household income below $\$50,000$ as Low, between $\$50,000$ and $\$100,000$ as Middle, and above $\$100,000$ as High income groups. $9,837$ children were included following exclusion of missing data. The demographic characteristics of the children are presented in Table 4.1.

For brain structural characteristics, we specifically analyzed the mean FA at each ROI from the diffusion tensor imaging (DTI), which represents a fundamental microstructural property of the brain. FA values were obtained from 148 distinct regions of interest (ROIs) based on the Destrieux atlas [49].

Group Comparison / Parameters. We performed group analyses to demonstrate that CMD enhances downstream statistical analysis and identify income-related ROIs. For the baseline, we used a general linear model (GLM) on the original univariate FA measure to correct for covariates (i.e., age, biological sex at birth, and scanner serial number), and obtained p-values at each ROI. We then applied a multivariate general linear model (MGLM) on CMD, which is a multi-variate feature derived using the Covlet, and resultant p-values were adjusted for the covariates. For both analyses, multiple comparisons were corrected with Bonferroni correction at $\alpha = 0.01$, and the final p-values and ROIs that met the defined threshold were compared.

For the kernel function $g(\cdot)$ for CMD, we used a spline function defined in [7]. We used total of 4 scales for the BP vs. NP analysis and 5 scales for Low/Middle/High income comparisons. The scales were defined in the spectrum of the precision matrix, i.e., $[0, \lambda_P]$.

2.4.2 Group Analysis Results

Table 2.2 summaries the number of significant ROIs whose p-values met the multiple comparisons correction. A larger number of significant ROIs were obtained utilizing CMD compared to the results obtained utilizing the raw FA values, also demonstrating an improvement in statistical sensitivity in every category.

Below-poverty vs. Non-poverty. Comparing the BP and NP groups, we detected only 6 ROIs that met the Bonferroni correction at $\alpha = 0.01$ using the raw FA values. However, using CMD, we identified 22 different ROIs that met the same Bonferroni correction with improved p-values. Interestingly, the 6 ROIs that were discovered by the baseline were subsumed by the ROIs found with CMD. The list of surviving ROIs from BP vs. NP analysis and corresponding p-values are given in Table 2.3, and these ROIs and their p-values are further demonstrated in Fig. 2.2 on a cortical surface of the brain. Looking at the top 11 ROIs in the left column of Table 2.3, we observed that many top ROIs with the lowest p-values are located within the frontal region in the brain, and detailed interpretation of this observation will be given in section 2.4.4.

High vs. Middle vs. Low income groups. While the group differences between High and Low income groups were discernible (Table 2.2), the number of significant ROIs for the comparisons between High and Middle, and Middle and Low groups were small. However, using CMD, we found two ROIs for High vs. Middle group analysis, and 7 ROIs from the Middle vs. Low group comparisons. In Table 2.3, we present the list of significant ROIs when comparing Middle income versus Low income groups. Again, we identified

Table 2.2: Number of ROIs showing variation based on family income level

Feature	BP vs NP	H vs L	H vs M	M vs L
Original FA	6	7	0	1
CMD (COVLET)	22	48	2	7

BP: below-poverty, NP: non-poverty, H: high, M: middle, L: low

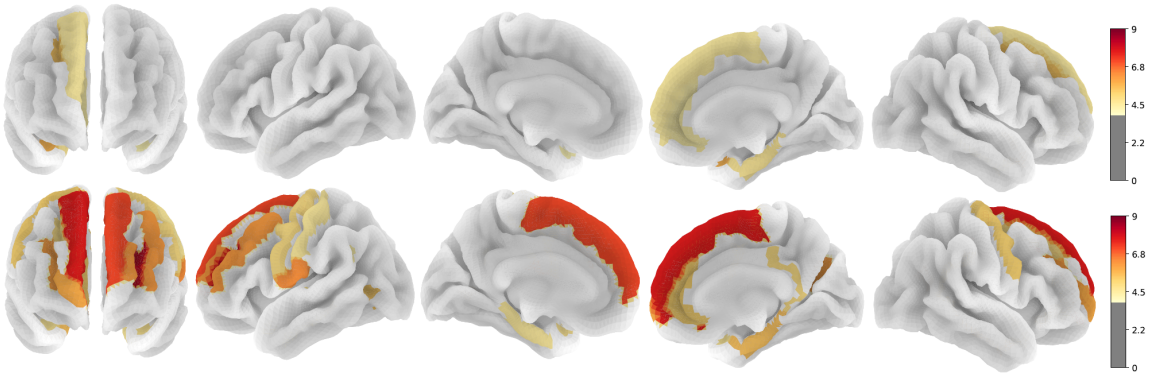


Figure 2.2: Group analysis results from Below-poverty vs. Non-poverty family income groups. p-value maps in $-\log_{10}$ scale are shown on a brain surface. Top: using original FA values, Bottom: using CMD. Notice much stronger signal in the bottom.

only two regions using the raw FA values, but seven ROIs using CMD with subsequent improvement in statistical outcomes. From these comparisons, we concluded that CMD enables the underlying signal to become more detectable, even with subtle differences.

2.4.3 Family Income Group Classification

We further performed classification of family income groups with 10-fold cross-validation using Elastic Net [50]. The purpose was to see if CMD from FA improves prediction performance over the raw FA measures, especially when it has shown improved statistical outcomes in section 2.4.2. Due to class imbalance, we used NearMiss under-sampling [51]. Classification performances were evaluated by accuracy, precision and recall metrics which are summarized in Table 2.4. Using CMD, the accuracy improves by 8% and precision gets increased by 6% in binary case, and similarly for 3-class case, both accuracy and precision improved by 6%. These results show that CMD improves the prediction ability over the raw measurements, controlling for Type-1 error with increased precision.

2.4.4 Neuroscientific Interpretation

The use of a covariance-based wavelet-like (COVLET) transform facilitated more sensitive inference on household income compared to raw FA alone. The finding that majority of the brain ROIs identified in the current study as being within or close to the frontal lobe of the brain is consistent with previous literature that examined the association between socioeconomic characteristics and brain structure. In the largest previous study that included

Table 2.3: Identified ROIs and p-values from BP vs. NP (Left) and Mid vs. Low income (Right) analyses.

(a) BP vs. NP

Idx	ROI	p-value	Idx	ROI	p-value
1	s.front.middle.lh	6.9e-09	12	g.precentral.rh	5.7e-06
2	g.front.sup.rh	2.1e-08	13	g.precentral.lh	1.1e-05
3	g.front.sup.lh	7.1e-08	14	g.cing.post.dorsal.rh	1.1e-05
4	g.and.s.subcentral.lh	4.6e-07	15	g.cing.post.ventral.rh	1.2e-05
5	g.front.middle.lh	6.7e-07	16	s.front.sup.rh	1.4e-05
6	s.front.middle.rh	7.6e-07	17	s.temp.transverse.lh	1.7e-05
7	s.parieto.occipital.rh	7.7e-07	18	g.temp.sup.plan.polar.lh	1.9e-05
8	g.and.s.transv.frontopol.rh	1.4e-06	19	g.temp.sup.plan.polar.rh	2.2e-05
9	g.oc.temp.med.parahip.rh	3.2e-06	20	g.postcentral.lh	2.3e-05
10	s.occipital.ant.lh	4.2e-06	21	g.oc.temp.med.parahip.lh	2.3e-05
11	g.and.s.cingul.ant.rh	5.6e-06	22	s.interm.prim.jensen.lh	4.2e-05

(b) Mid vs. Low income

Idx	ROI	p-value
1	g.front.sup.lh	6.0e-07
2	s.front.sup.rh	6.6e-07
3	s.occipital.ant.lh	1.0e-06
4	s.front.middle.rh	2.0e-06
5	g.front.sup.rh	1.3e-05
6	s.parieto.occipital.rh	4.2e-05
7	s.interm.prim.jensen.lh	5.5e-05

Table 2.4: Classification performance measurements.

	2-Class			3-Class		
Measures	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Original FA	0.77	0.76	0.89	0.41	0.40	0.42
CMD (COVLET)	0.85	0.82	0.91	0.47	0.46	0.46

*Note that macro average precision and recall are reported in 3-class case.

1,100 children, household income accounted for significant variation in regions within the bilateral frontal, temporal and parietal lobes [52]. In smaller studies that assessed the relationship between white matter integrity and socioeconomic characteristics, individuals earning higher incomes demonstrated higher FA as well as cognitive ability [53].

Despite the dominance of ROIs within the frontal lobe as identified in the current study, previous studies do not appear to consistently demarcate ROIs in a standardized fashion, in part due to the varied impact of developmental processes on cortical columns, synaptic formation and pruning [54]. Indeed, the frontal lobe is central to executive functioning–a domain that spans cognition and behavior–and therefore remains vulnerable to stressors during childhood development [55, 56]. Previous studies have generally focused on macrostructural characteristics of the brain such as cortical thickness, surface area and volume, which could be susceptible to overlapping developmental influences. Importantly, the current study provides the first ever large scale evidence that the associations between brain and socioeconomic status extend to their microstructural properties.

2.5 Conclusion

To increase sensitivity in statistical inference/prediction methods for structured data, we proposed a novel transform that utilizes its covariance structure, i.e., Covlet. The Covlet captures local context information along the geometry of precision matrix and provide a multi-scale feature, which lets us compare samples with different labels more robustly. We

performed statistical analysis and classification of children based on family income using large-scale ABCD dataset and demonstrated quantitative improvements in their outcomes. As qualitative results, we identified several ROIs whose microstructure is susceptible to socioeconomic inequality, which were not identifiable with conventional approaches.

Part II

Disentangled Representation Learning on ADNI Data

CHAPTER 3

Representation Learning I: Disentangled Sequential Graph Autoencoder for Preclinical Alzheimer's Disease Characterizations from ADNI study

Given a population longitudinal neuroimaging measurements defined on a brain network, exploiting temporal dependencies within the sequence of data and corresponding latent variables defined on the graph (i.e., network encoding relationships between regions of interest (ROI)) can highly benefit characterizing the brain. Here, it is important to distinguish time-variant (e.g., longitudinal measures) and time-invariant (e.g., gender) components to analyze them individually. For this, we propose an innovative and ground-breaking Disentangled Sequential Graph Autoencoder which leverages the Sequential Variational Autoencoder (SVAE), graph convolution and semi-supervising framework together to learn a latent space composed of time-variant and time-invariant latent variables to characterize disentangled representation of the measurements over the entire ROIs. Incorporating target information in the decoder with a supervised loss let us achieve more effective representation learning towards improved classification. We validate our proposed method on the longitudinal cortical thickness data from Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Our method outperforms baselines with traditional techniques demonstrating benefits for effective longitudinal data representation for predicting labels and longitudinal data generation.

3.1 Introduction

Representation learning is at the core of Image Analysis. Lots of recent attentions are at a disentangled representation of data, as the individual disentangled representations are

highly sensitive to a specific factor whereas indifferent to others [57, 58, 59, 16, 60]. A typical disentangling method would find a low-dimensional latent space for high-dimensional data whose individual latent dimensions correspond to independent disentangling factors. For longitudinal data, one can expect to decompose the longitudinal data into time-invariant factors and time-variant factors by obtaining the "disentangled" representation as longitudinal observations are affected by both time-variant and static variables [18, 61, 62]. In the context of neuroimaging studies, the disentangled representation would be able to separate time-independent concepts (e.g. anatomical information) from dynamical information (e.g. modality information) [63], which may offer effective ways of compression, conditional data generation, classification and others.

Recent advances in variational autoencoders (VAE) [17] have made it possible to learn various representations in an unsupervised manner for neuroimaging analysis [15, 16]. Moreover, various vibrant of autoencoders are also proposed to model temporal data; for example, [18] introduced the factorised hierarchical variational auto-encoder (FHVAE) for unsupervised learning of disentangled representation of time series. Sequential variational autoencoder was proposed in [61] benefiting from the usage of the hierarchical prior. It disentangles latent factors by factorizing them into time-invariant and time-dependent parts and applies an LSTM sequential prior to keep a sequential consistency for sequence generation. [62] modeled the time-varying variables via LSTM in both encoder and decoder for dynamic consistency.

There are two major issues with current approaches. First, while these methods can deal with temporal nature of the data, they do not necessarily introduce supervision at all. Moreover, from a neuroscience perspective, the domain knowledge tells us that the regions of interest (ROIs) in the brain are highly associated to each other both functionally and structurally [64, 65, 66, 67]. This association provides a prior knowledge on connection between the ROIs as a graph; for example, structural brain connectivity from tractography

on Diffusion Tensor Imaging (DTI) provides a path for anisotropic variation and diffusion of structural changes in the brain such as atrophy of cortical thickness. Most of the existing methods do not consider this arbitrary topology of variables, if there is any, into account, which can provide significant benefit for downstream tasks. To summarize, learning with (either full or partial) supervision on longitudinal neuroimaging measurements on a brain network is still **under-explored**.

Given longitudinal observations (e.g., cortical thickness) on specific ROIs in the brain and a structural brain network characterized by bundles of neuron fiber tracts, our aim is to develop a framework to learn a latent disentangled representation of the observations that are composed of time-variant and time-invariant latent variables. For this, we propose an innovative Semi-supervised Sequential Graph Autoencoder model which leverages ideas from the sequential variational autoencoder (SVAE), graph convolution and semi-supervising framework. The core idea is to incorporate target information as a supervision in the decoder with a supervised loss, which let us achieve more effective representation for downstream tasks by balancing extraction of underlying structure as well as accurately predicting class labels.

Our proposed framework learns a latent disentangled representation composed of time-variant and time-invariant latent variables to characterize the longitudinal measurements over the entire structural brain network that consists of ROIs. Our **contributions** are as summarized follows: our model can 1) learn an ideal disentangled representation which separates time-independent content or anatomical information from dynamical or modality information and conditionally generate synthetic sequential data; 2) perform semi-supervised tasks which can jointly incorporate supervised and unsupervised data for classification tasks; 3) leverage graph structure to robustly learn the disentangling latent structure. Using our framework, we analyzed longitudinal cortical thickness measures on brain networks with diagnostic labels of Alzheimer's Disease (AD) from Alzheimer's Disease Neuroimaging

Initiative (ADNI) study. As AD is a progressive neurodegenerative condition characterized by neurodegeneration in the brain caused by synthetic factors [10, 11, 12, 13, 14], it is important to effectively characterize early symptoms of the disease. We expect that disentangling ROI measures with time-variant and static components can provide unique insights.

3.2 Background

Our proposed framework involves two important concepts: 1) graph convolutions and 2) SVAE. Hence, we begin with brief reviews of their basics.

3.2.1 Graph Convolutions

Let $G = \{\mathbb{V}, \mathbb{E}, A\}$ be an undirected graph, where $\mathbb{V}$ is a set of nodes with $|\mathbb{V}| = n$, $\mathbb{E}$ is a set of edges and A is an adjacent matrix that specify connections between the nodes. Graph Fourier analysis relies on the spectral decomposition of graph Laplacian defined as $\mathcal{L} = D - A$, where D is a diagonal degree matrix with $D_{i,i} = \sum_j A_{i,j}$. The normalized Laplacian is defined as $L = I_n - D^{-1/2}AD^{-1/2}$, where I_n is the identity matrix. Since L is real and positive semi-definite, it has a complete set of orthonormal eigenvectors $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ with corresponding non-negative real eigenvalues $\{\lambda_l\}_{l=1}^n$. Eigenvectors associated with smaller eigenvalues carry slow varying signals, indicating that connected nodes share similar values. In contrast, eigenvectors associated with larger values carry faster varying signals across the connected nodes. We are interested in the smallest eigenvalues due to the negation used to compute the Laplacian matrix in terms of the Euclidean Commute Time Distance [68]. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a signal defined on the vertices of the graph. The graph

Fourier transform of $\boldsymbol{x}$ is defined as $\hat{\boldsymbol{x}} = U^T\boldsymbol{x}$, with inverse operation given by $\boldsymbol{x} = U\hat{\boldsymbol{x}}$. The graphical convolution operation between signal $\boldsymbol{x}$ and filter $\boldsymbol{g}$ is

$$\begin{aligned} \boldsymbol{g} * \boldsymbol{x} &= U((U^T\boldsymbol{g}) \odot (U^T\boldsymbol{x})) \\ &= U\hat{G}U^T\boldsymbol{x}. \end{aligned} \tag{3.1}$$

Here, $U^T\boldsymbol{g}$ is replaced by a filter $\hat{G} = \mathrm{diag}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta} \in \theta^n$ in Fourier domain. Unfortunately, eigendecomposition of $\boldsymbol{L}$ and matrix multiplication with U are expensive. Motivated by the Chebyshev polynomials approximation in [7], [64] introduced a Chebyshev polynomial paramterization for ChebyNet that offers fast localized spectral filtering. Later, [66] provided a simplified version of ChebyNet by considering a second order approximation such that $\boldsymbol{g} * \boldsymbol{x} \approx \theta(I_n + D^{-1/2}AD^{-1/2})\boldsymbol{x}$ and illustrate promising model performance in graph-based semi-supervising learning tasks, and GCN is deeply studied in [67]. Then, FastGCN was proposed in [69] which approximates the original convolution layer by Monte Carlo sampling, and recently, [70] leveraged graph wavelet transform to address the shortcomings of spectral graph convolutional neural networks.

3.2.2 Sequential Variational Autoencoder

Variational autoencoder (VAE), initially introduced in [17] as a class of deep generative mode, employs a reparameterized gradient estimator for a evidence lower bound (ELBO) while applying amortized variational inference to an autoencoder. It simultaneously trains both a probabilistic encoder and decoder for elements of a data set $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$ with latent variable $\boldsymbol{z}$. Sequential variational autoencoders (SVAEs) extend VAE to sequential data $\mathcal{D}$, where each data are $\boldsymbol{x}_{1:T} = (x_1, \ldots, x_T)$ [61, 62]. SVAEs factorize latent variables into two disentangled variables: the time-invariant variable $\boldsymbol{f}$ and time-varying variable $\boldsymbol{z}_{1:T} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$. Accordingly, decoder is casted as a conditional probabilistic density $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{f}, \boldsymbol{z}_{1:T})$ and encoder is used to approximate the posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{x})$ as

$q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{x})$ that is referred to as an "inference network" or a "recognition network". $\boldsymbol{\theta}$ refer to the model parameters of generator and $\boldsymbol{\phi}$ refer to the model parameters of encoder. SVAEs are trained to maximize the following ELBO:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}) = \mathbb{E}_{\hat{p}(\boldsymbol{x}_{1:T})}\Big[\mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}, \boldsymbol{f}|\boldsymbol{x}_{1:T})} \ln p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{f}, \boldsymbol{z}_{1:T}) - \mathrm{KL}(q_\phi(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T}), p_\theta(\boldsymbol{f}, \boldsymbol{z}_{1:T}))\Big], \tag{3.2}$$

where $\hat{p}(\boldsymbol{x}_{1:T})$ is the empirical distribution with respect to the data set $\mathcal{D}$, $q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{x}_{1:T})$ is the variational posterior, $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{1:T}|\boldsymbol{f}, \boldsymbol{z}_{1:T})$ is the conditional likelihood and $p_{\boldsymbol{\theta}}(\boldsymbol{f}, \boldsymbol{z}_{1:T})$ is prior over the latent variables.

3.3 Proposed Model

Let us first formalize the problem setting. Consider a dataset consists of shared graph G, and M unsupervised data points $\mathcal{D} = \{\boldsymbol{X}_i\}_{i=1}^{M}$ and M^{sup} supervised data points $\mathcal{D}^{sup} = \{\boldsymbol{X}_i, y_i\}_{i=1}^{M^{sup}}$ as pairs. $\boldsymbol{X}_i = (X_{i,1}, \dots X_{i,T_i})$ refer to the i-th sequential observations on N nodes of a graph G with C input channels, i.e., $X_{i,t} \in \mathbb{R}^{N \times C}$, and y_i is the corresponding class label such as diagnostic labels.

We propose a semi-supervised sequential variational autoencoder model, and for convenience we omit the index i whenever it is clear that we are referring to terms associated with a single data point and treat individual data as $(\boldsymbol{X}, y)$.

3.3.1 Objective Function

Typical semi-supervised learning pipelines for deep generative models, e.g., [71, 72], define an objective function for optimization as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}, \mathcal{D}^{sup}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}) + \tau \mathcal{L}^{sup}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}^{sup}). \tag{3.3}$$

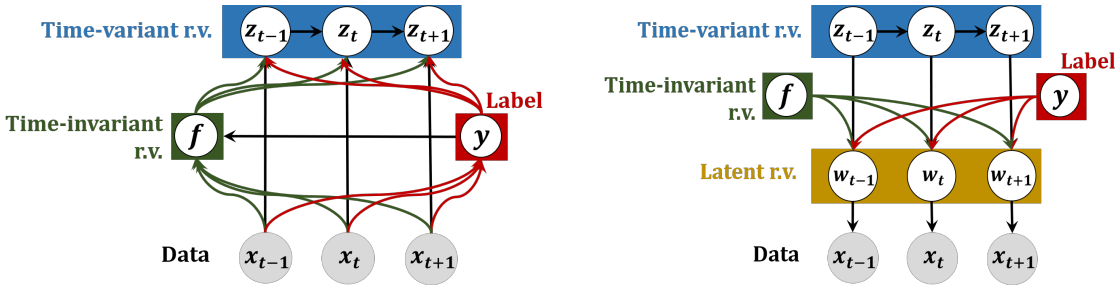


Figure 3.1: A graphical model visualisation of the encoder (left) and decoder (right). In the encoder, label y is inferred by data $\boldsymbol{x}$ and time-invariant r.v. $\boldsymbol{f}$ are inferred by label y and data $\boldsymbol{x}$, and time-varying r.v. $\boldsymbol{z}$ are sequentially inferred by label y, time-invariant r.v. $\boldsymbol{f}$ and data $\boldsymbol{x}$. In the decoder, data are sequentially generated from time-invariant random variable (r.v.) $\boldsymbol{f}$, time-varying r.v. $\boldsymbol{z}$ and label y via latent r.v. $\boldsymbol{w}$.

Similarly, our approach jointly models unsupervised and supervised collections of terms over $\mathcal{D}$ and $\mathcal{D}^{sup}$. The formulation in Eq. 3.3 introduces a constant τ to control the relative strength of the supervised term. As the unsupervised term in Eq. 3.3 is exactly same as that of Eq. 3.2, we focus on the supervised term $\mathcal{L}^{sup}$ in Eq. 3.3 expanded below. Incorporating a weighted component as in [71],

$$\begin{aligned} &\mathcal{L}^{sup}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}^{sup}) \\ &= \mathbb{E}_{\hat{p}(\boldsymbol{X},y)}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{f},\boldsymbol{z}|\boldsymbol{X},y)}\left[\ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{X}, y, \boldsymbol{f}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}|\boldsymbol{X}, y)}\right] + \alpha \ln q_{\boldsymbol{\phi}}(y|\boldsymbol{X})\right] \end{aligned} \tag{3.4}$$

where α balances the classification performance and reconstruction performance. Discussions on generative and inference model will continue in the later sections.

3.3.2 Generative Model

This section discusses modeling conditional probabilistic density $p_{\boldsymbol{\theta}}(\boldsymbol{X}|\boldsymbol{f}, \boldsymbol{z}, y)$ with its corresponding prior. We incorporate the topology information of the graph G into the generative process using a graph convolution. Specifically, we assume that sequences $\boldsymbol{X}$ are generated from P-dimensional latent vectors $\boldsymbol{W} = (W_1, \ldots, W_T)$ and $W_t \in \mathbb{R}^{N \times P}$ via

$$X_t = \hat{A} W_t \Theta, \tag{3.5}$$

where $\hat{A} = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$, $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. A is the adjacent matrix for the graph G and Θ is the trainable weight matrix. Then we assume the latent variables $\boldsymbol{W}$ are generated from two disentangled variables: the time-invariant (or static) variable $\boldsymbol{f}$ and the time-varying (or dynamic) variables $\boldsymbol{z}$, as well as the label y, as shown in Figure 3.1. A joint for the generative model is given as

$$\begin{aligned} & p_{\boldsymbol{\theta}}(\boldsymbol{X}, y, \boldsymbol{z}, \boldsymbol{f}) \\ & = p_{\boldsymbol{\theta}}(\boldsymbol{f}) p_{\boldsymbol{\theta}}(y) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t | \boldsymbol{z}_{<t}) p_{\boldsymbol{\theta}}(X_t | y, \boldsymbol{f}, \boldsymbol{z}_t). \end{aligned} \tag{3.6}$$

The prior of $\boldsymbol{f}$ is defined as a Gaussian distribution: $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, I)$. Time-varying latent variables $\boldsymbol{z}_{1:T}$ follow a sequential prior $\boldsymbol{z}_t | \boldsymbol{z}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t^2))$, where $[\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t]$ are estimated by a recurrent network, such as LSTM [73] or GRU [74], in which the hidden states are updated temporally. The generating distribution of W_t is conditional on y, $\boldsymbol{f}$ and $\boldsymbol{z}_t$: $\mathrm{vec}(W_t) | y, \boldsymbol{f}, \boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_{w,t}, \mathrm{diag}(\boldsymbol{\sigma}_{w,t}^2))$, where $[\boldsymbol{\mu}_{w,t}, \boldsymbol{\sigma}_{w,t}] = \psi^{Decoder}(y, \boldsymbol{f}, \boldsymbol{z}_t)$. This decoder $\psi^{Decoder}$ can be any flexible neural network such as multilayer perceptron (MLP). The $\boldsymbol{f}$ will be capable of modelling global aspects of the whole sequences which are time-invariant, while $\boldsymbol{z}_{1:T}$ will model time-varying features. As mentioned in [61], to separate the static and dynamic information, smaller dimension of $\boldsymbol{z}_t$ is preferred. In the context of ADNI study, $\boldsymbol{z}_t$ would encode how ROIs at timestamp t is morphed into those at timestamp $t + 1$. In the context of generative model, we employ LSTM as the prior for $\boldsymbol{z}$ and use MLP for the conditional probabilistic density, and we set the dimension $P = 1$.

3.3.3 Inference Model

The developed SVAE within our framework proposes a recognition model $q_{\boldsymbol{\phi}}(y, \boldsymbol{f}, \boldsymbol{z}|\boldsymbol{X}) = q_{\boldsymbol{\phi}}(y|\boldsymbol{X})q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}|y, \boldsymbol{X})$ to approximate the posterior $p_{\boldsymbol{\theta}}(y, \boldsymbol{f}, \boldsymbol{z}|\boldsymbol{X})$. The recognition model is formulated as

$$\begin{aligned} y &\sim \mathrm{Cat}(\mathrm{Softmax}(\boldsymbol{p}_y))\,, \\ \boldsymbol{f} &\sim \mathcal{N}(\boldsymbol{\mu}_f, \mathrm{diag}(\boldsymbol{\sigma}_f^2))\,, \\ \boldsymbol{z}_t &\sim \mathcal{N}(\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t^2))\,, \end{aligned} \tag{3.7}$$

where $\boldsymbol{p}_y = \psi_y^{Encoder}(X_{1:T})$, $[\boldsymbol{\mu}_f, \boldsymbol{\sigma}_f] = \psi_f^{Encoder}(y, X_{1:T})$ and $[\boldsymbol{u}_t, 2\log\boldsymbol{\sigma}_t] = \psi_R^{Encoder}(y, X_{\leq t})$. It implies that the label y and the time-invariant variable $\boldsymbol{f}$ are conditional on the whole sequence via $\psi_y^{Encoder}$ and $\psi_f^{Encoder}$, while the time-dependent variable $\boldsymbol{z}_t$ is inferred by the sequence before time t, $X_{\leq t}$. The inference model is visualized in Figure 3.1 and is factorized as

$$\begin{aligned} &q_{\boldsymbol{\phi}}(y, \boldsymbol{z}_{1:T}, \boldsymbol{f}|X_{1:T}) \\ &= q_{\boldsymbol{\phi}}(y|X_{1:T})q_{\boldsymbol{\phi}}(\boldsymbol{f}|y, X_{1:T}) \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|\boldsymbol{f}, y, X_{\leq t})\,. \end{aligned} \tag{3.8}$$

In the context of our inference model, we employ three independent LSTMs for three conditional probabilistic densities of y, $\boldsymbol{f}$ and $\boldsymbol{z}$.

3.4 Experimental Results

We conducted experiments on structural brain connectivity from DTI in ADNI. DTI images were processed by tractography, which extracted neuron fiber tracts and longitudinal cortical thickness measures registered at Destrieux atlas [49] with 148 ROIs. The dataset had five labels; we merged control (CN), Significant Memory Concern (SMC) and Early Mild

Cognitive Impairment (EMCI) groups as Pre-clinical AD group, and Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD) as Prodromal AD group to ensure sufficient sample size. The dataset included N=140 subjects with the Pre-AD group (93 subjects/330 records) and the Pro-AD group (47 subjects/170 records). The mean (std) of ages and sex ratio (Male:Famale) in Pre-AD group and Pro-AD group are 74.02(6.72)/ (185:145) and 74.87(6.92)/ (95:75), respectively. An overall graph was obtained by taking the average of the adjacency matrices. Experiments for disentangle representation and quantitative analysis were performed given below.

3.4.1 Disentangled Representation

In this experiment, we randomly took 100 subjects' records for training, 20 subjects' records for validation and the other 20 subjects' records for testing. We set the dimension size of $\boldsymbol{f}$ as 8 and the dimension size of $\boldsymbol{z}$ as 32. We also set the size of hidden states in LSTMs as 32.

We randomly selected two subjects with more than three records (i.e., time-points), where subject 1 belongs to Prodromal AD group and subject 2 belongs to Pre-clinical AD

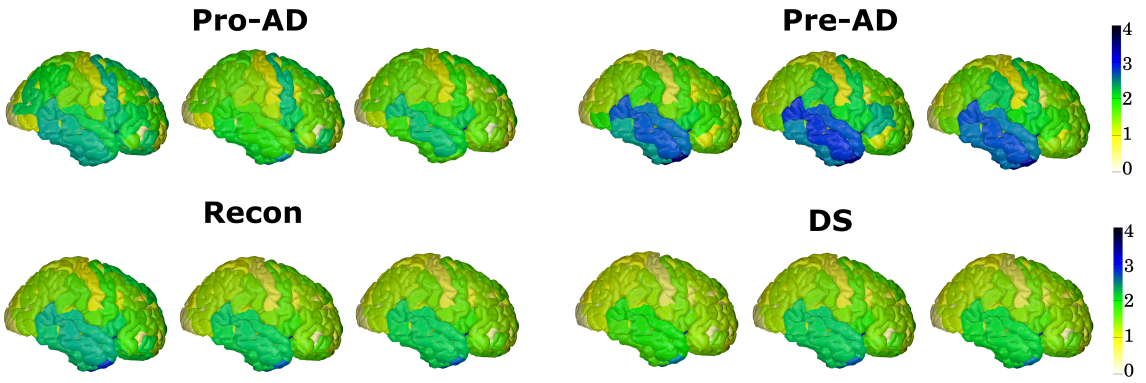


Figure 3.2: Top panel shows the true brain surfaces at timestamp t_0, t_1 and t_2 for subject 1 (Pro-AD) and subject 2 (Pre-AD), respectively. Bottom panel shows the reconstructed brain surfaces for subject 1 (Recon) and subject 1's brain surfaces through the dynamic swapping (DS). Drawings generated using BrainPainter [2].

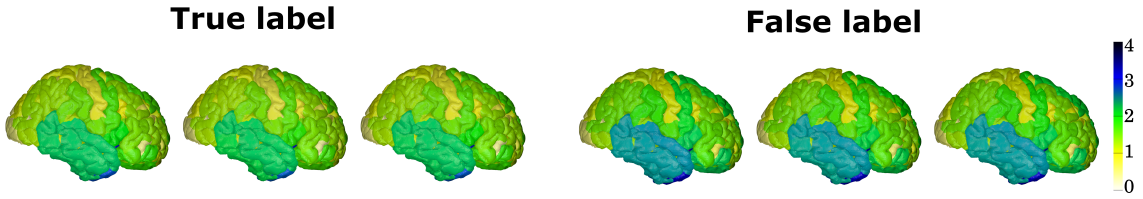


Figure 3.3: Label swapping task. Left panel shows generated brain surfaces for subject 1 (Pro-AD) based on the true label at timestamp t_0, t_1 and t_2, respectively. Right panel shows generated brain surfaces for the same subject 1 but based on the false label.

group. Suppose that the two subjects' sequential records are given for anatomical information and modality information denoted by R_1 and R_2. Our method performs the reconstruction task and the dynamic swapping task in which the record generation is based on the true y, $\boldsymbol{f}$ from R_1 and $\boldsymbol{z}$ from R_2 as in Figure 4.2. It shows that the reconstruction captures both anatomical information and modality information, and figures generated from the dynamic swapping task illustrate that time-varying latent variables $\boldsymbol{z}$ succeed to learning the modality information.

In Figure 3.3, we show results from the label swapping task on subject 1, where we generate cortical thickness based on the $\boldsymbol{f}$ from R_1, $\boldsymbol{z}$ from R_1 and true/false labels y. Comparing the generated measures of subject 1 with the true measures in Figure 4.2, we found that generated measures based on the true label are more similar to the true measures and that based on the false label has totally different patterns but similar to the true measures of subject 2 in Figure 4.2. It suggests that the decoder in our model correctly learns the label.

To understand the disentangled representation on the time invariant latent variable $\boldsymbol{f}$, we carry out latent traversals in $\boldsymbol{f}$ as in [75]. Specifically, we first computed the average Kullback–Leibler divergence for $\boldsymbol{f}$ with its prior. Then we selected the two dimensions in $\mathbf{f}$ with the largest two values (the 1st and 3rd elements), which refer to the two most informative dimensions and then traverse a single latent dimension on 10 equally spaced grids on $[-3, 3]$. For better visualization, we chose the first image as baseline and subtracted

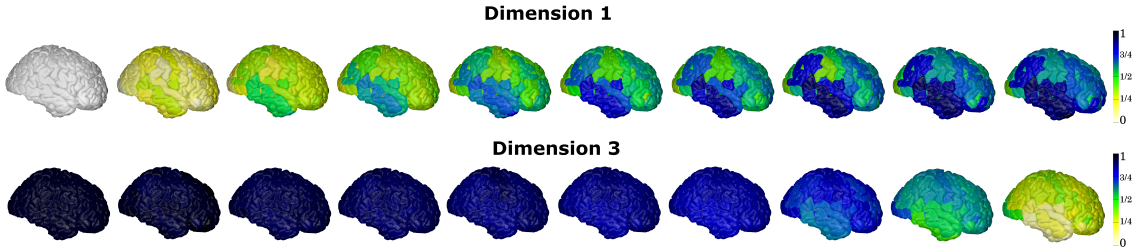


Figure 3.4: Latent traversals task. Top: the latent brain surfaces for dim-1 on subject 1 (pro-AD). Bottom: the latent brain surfaces for dim-3 on subject 1 .

the baseline of image from all generated images. Then we normalized those images in a unit region $[0, 1]$ shown in Figure 3.4.

3.4.2 Quantitative Analysis

We carry out 7-fold cross validation (CV) in which we take six folds for training (one fold for validation from the training set) and one fold for testing. We set the dimension size of $\boldsymbol{f}$ as 8 and the dimension size of $\boldsymbol{z}$ as 4. We also set the size of hidden states in LSTMs as 8. We compared our model with S3VAE model [62], which has a generator as in Figure 3.1 but without a probabilistic model on label y. As S3VAE is unsupervised, we cannot directly compare our model with it. Instead, we tackle the classification task via a two-stage approach. Specifically, we train S3VAE to obtain latent $\boldsymbol{f}$ and train a naive neural network for the label classification. As for testing, we first get $\boldsymbol{f}$ from trained S3VAE

Table 3.1: Reconstruction and classification performance with 7-fold cross validation.

	RMSE	Accuracy	Precision	Recall
Our model ($\alpha = 1$)	0.257(0.041)	0.657(0.168)	0.416(0.367)	0.446(0.349)
Our model ($\alpha = 10$)	0.258(0.046)	0.736(0.151)	0.541(0.346)	0.492(0.337)
S3VAE (Supervised)	0.263(0.042)	0.664(0.164)	0.000(0.000)	0.000(0.000)
S3VAE (Two stages)	0.254(0.043)	0.664(0.164)	0.000(0.000)	0.000(0.000)

and classify $\boldsymbol{f}$. Also, we propose a supervised loss based on the latent time-invariant f for S3VAE as one competitor. The generative model is modified as

$$p_{\boldsymbol{\theta}}(\boldsymbol{X}, y, \boldsymbol{z}, \boldsymbol{f}) \tag{3.9}$$
$$= p_{\boldsymbol{\theta}}(\boldsymbol{f}) p_{\boldsymbol{\theta}}(y|\boldsymbol{f}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{<t}) p_{\boldsymbol{\theta}}(X_t|\boldsymbol{f}, \boldsymbol{z}_t),$$

where we employ a fully connected network following a softmax activation function for $p_{\boldsymbol{\theta}}(y|\boldsymbol{f})$. We treat the pro-AD as positive result and then report three classification measures, accuracy, precision and recall. We also report root mean square error (RMSE) as a reconstruction measure for testing data in Table 3.1. As for our proposed model, we consider the regularization weights $\alpha = 1$ and $\alpha = 10$. We find that our model has a better reconstruction performance in comparison to the supervised S3VAE model and performs similarly to the two-stage S3VAE. As for classification, our model with $\alpha = 10$ outperforms other models. We note that S3VAE based methods always categorize patients into pre-AD group, suffering from the imbalance classification issue. Our model resolves this issue and obtains a significantly better classification result according to both higher precision and recall scores. Finally, we note that to get better reconstruction or prediction results, properly tuning the hyperparameter α is important.

3.5 Conclusion

In summary, we propose a novel Sequential Autoencoder model. It incorporates the graph information via graph convolution operation, and it jointly models supervised and unsupervised data. Our model is flexible for data generation and it can conditionally generate sequential data based on label, disentangled time-invariant and time-varying latent variables. Quantitatively, we show that this model has competitive classification and reconstruction performance compared with two modified state-of-the-art S3VAE models.

CHAPTER 4

Representation Learning II: Disentangled Representation of Longitudinal β-Amyloid for AD via Sequential Graph Variational Autoencoder with Supervision

The emergence of Positron Emission Tomography (PET) imaging allows us to quantify the burden of amyloid plaques *in-vivo*, which is one of the hallmarks of Alzheimer's disease (AD). However, the invasive exposure to radiation and high imaging cost significantly restrict the application of PET imaging in characterizing the evolution of pathology burden which often requires longitudinal PET image sequences. In this regard, we propose a proof-of-concept solution to generate the complete trajectory of pathological events throughout the brain based on very limited number of PET scans. We present a novel variational autoencoder model to learn a latent population-level representation of neurodegeneration process based on the longitudinal β-amyloid measurements at each brain region and longitudinal diagnostic stages. As the propagation of pathological burdens follow the topology of brain connectome, we further cast our neural network into a supervised sequential graph VAE, where we use the brain network to guide the representation learning. Experiments show that the disentangled representation can capture the disease-related dynamics of amyloid and forecast the level of amyloid depositions at future time points.

4.1 Introduction

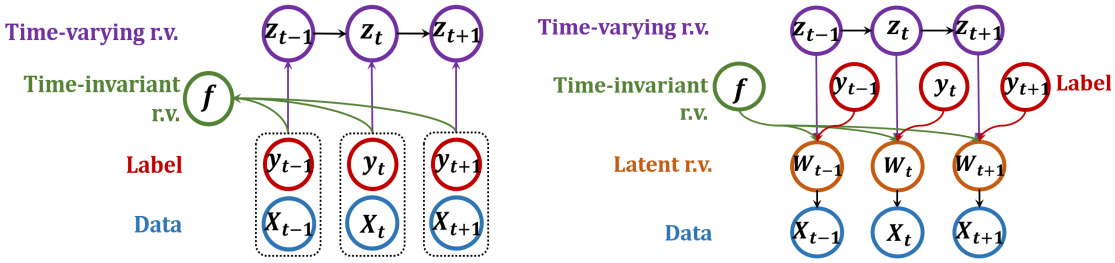
Alzheimer's disease (AD) is a progressive neurodegenerative disorder resulting in memory loss and cognitive impairments that interfere with functions for daily life [76]. In the current pathophysiologic understanding of AD, β-amyloid deposition occurs as the first pathological event in AD, followed by tau fibrillary tangles as the downstream effect of

amyloid, and chronologically lead to the neurodegeneration as the final signal of a series of cognitive disorders [77]. Therefore, accumulation of β-amyloid neuritic plaques is one of the earliest risk factor for potential development of AD [78], and is one of the key proteins to assess in understanding AD [77, 79].

Positron Emission Tomography (PET) is a non-invasive imaging providing a direct measure of in vivo β-amyloid status to better characterize early AD. Despite the improved diagnostic ability from amyloid PET, the cost for PET scans are very expensive ($\sim\$5,000$) which prevents its use for widespread clinical adoption. Under limited amyloid PET data at present, to facilitate clinical research on β-amyloid, a framework that can learn the dynamics of amyloid and forecast future measures based on limited past timestamps may be of great interest, which will benefit understanding of how amyloid functions and its critical roles in neurodegeneration.

Many studies have shown that structural brain connectivity from Diffusion Tensor Imaging (DTI) is highly associated with AD progression [80, 81], whose topology between regions of interest (ROIs) behaves as a path for the amyloid deposition [82]. However, it is very challenging to learn the complex dynamics of β-amyloid over brain networks through noisy data from limited timestamps per subject (e.g., less than 3 timestamps on average). In this work, given the limited and noisy longitudinal β-amyloid PET measures over a structural brain network, our aim is to develop a framework to uncover the disease dynamics or progression pattern of β-amyloid and forecast amyloid depositions at future timestamps to better understand and characterize the progression of AD. Unfortunately, this is not easy as the observations from imaging scans are complexly affected by several variables such as age, gender, anatomical structure, disease effect, etc.

Notice that these variables can be separated as time-varying and time-invariant components, where the progression of AD is a major factor for the time-varying one. Therefore, separating them in a latent space is critical; disentangled representation learning would be



(a) Encoder (b) Decoder

Figure 4.1: A graphical visualisation of the encoder and decoder. (a) Encoder: time-invariant random variable (r.v.) $\boldsymbol{f}$ are inferred by time-dependent labels $\boldsymbol{y}$ and data $\boldsymbol{X}$, and time-varying r.v. $\boldsymbol{z}$ are sequentially inferred by labels $\boldsymbol{y}$ and data $\boldsymbol{X}$; (b) Decoder: data are sequentially generated from time-dependent labels $\boldsymbol{y}$, time-invariant r.v. $\boldsymbol{f}$ and time-varying r.v. $\boldsymbol{z}$ via latent r.v. $\boldsymbol{W}$.

able to disentangle time-invariant contents (e.g., anatomical information) from time-varying contents (e.g., morphological changes, dynamics of amyloid). Such disentangled representations not only help the model become more explainable, but also can benefit conditional data generation for downstream tasks [62, 83], e.g., cross-modality registration and segmentation [84, 85].

To this end, motivated by the schematics of Sequential Graph Variational Autoencoder [18, 19], we develop a framework that learns a latent disentangled representation composed of time-varying and time-invariant latent variables to characterize the longitudinal β-amyloid over the structural brain network. The core idea is to capture the disease-related dynamics of β-amyloid and as well as forecast the future amyloid depositions using the disentangled representation. The major contributions of this work are: 1) We incorporated "time-dependent" label as a supervision in the model to characterize longitudinal effect, 2) We integrated a brain network to make the framework more robust to the subject-wise heterogeneous dynamics when learning the disentangling latent representation, 3) We validated this framework to longitudinal β-amyloid data on brain networks with diagnostic labels of AD from Alzheimer's Disease Neuroimaging Initiative (ADNI). The experimental

results suggests a significant potential that this framework will facilitate clinical research by enriching amyloid data collection and help better understand the role of amyloid in the progression of AD before the disease symptoms manifest.

4.2 Methods

4.2.1 Supervised Sequential Graph VAE Model

Supervised Sequential Graph Variational Autoencoder (SSG-VAE) is designed for learning a disentangled representation composed of time-variant and time-invariant latent variables to capture the dynamics of longitudinal measures and forecast the measure at the future timestamp. We observed sequential data that appear as M^{sup} pairs of supervised data points $\mathcal{D}^{sup} = \{\boldsymbol{X}_i, \boldsymbol{y}_i\}_{i=1}^{M^{sup}}$ over a shared graph G with N nodes, where $\boldsymbol{X}_i = (X_{i,1}, \dots X_{i,T_i})$ refer to the i-th sequential observations, i.e., $X_{i,t} \in \mathbb{R}^N$, and $\boldsymbol{y}_i = (y_{i,1}, \dots y_{i,T_i})$ denote the corresponding time-dependent diagnostic labels. We will leave out the index i wherever it is clear that the terms we are referring to are associated with a single data point. SSG-VAE simultaneously trains a probabilistic encoder and decoder, and factorizes latent variables into two disentangled variables: time-invariant variable $\boldsymbol{f}$ and time-varying variable $\boldsymbol{z}_{1:T} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_T)$. We expect that latent variable $\boldsymbol{f}$ can encode the time-invariant global aspects of the data, while latent variable $\boldsymbol{z}$ will encode how the time-varying information at timestamp t is morphed into that of timestamp $t + 1$. The architectures of decoder and encoder are visualised in Figure 4.1.

4.2.1.1 Objective Function

We design a supervised variational autoencoder framework with an objective function [71, 72] defined over $\mathcal{D}^{sup}$ as

$$\begin{aligned}&\mathcal{L}^{sup}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathcal{D}^{sup}) \\ &= \mathbb{E}_{\hat{p}(\boldsymbol{X},\boldsymbol{y})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{f},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{y})}\left[\ln \frac{p_{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{y},\boldsymbol{f},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{f},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{y})}\right]\right],\end{aligned} \tag{4.1}$$

where the model parameters of the decoder is denoted as $\boldsymbol{\theta}$ and the model parameters of the encoder is denoted as $\boldsymbol{\phi}$. Here, $\hat{p}(\boldsymbol{.})$ denotes the empirical distribution, $p_{\boldsymbol{\theta}}(\boldsymbol{.})$ refers to the decoder distribution and $q_{\boldsymbol{\phi}}(\boldsymbol{.})$ is the variational posterior.

4.2.1.2 Prior

The prior of time-invariant latent variable $\boldsymbol{f}$ is defined as a standard Gaussian distribution, i.e., $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, I)$. We assume that time-varying latent variables $\boldsymbol{z}_{1:T}$ follow a sequential prior $\boldsymbol{z}_t|\boldsymbol{z}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathrm{diag}(\boldsymbol{\sigma}_t^2))$, where $[\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t]$ are the parameters of the prior distribution and is parameterized as a recurrent network LSTM [73], in which the hidden states are updated temporally. Moreover, prior distributions of time-dependent labels $\boldsymbol{y}$ follow the multinomial distribution, i.e., $p_\theta(y_t) = \mathrm{Cat}(y_t|\boldsymbol{\pi})$. Assuming labels $\boldsymbol{y}$, time-invariant $\boldsymbol{f}$ and time-varying $\boldsymbol{z}_{1:T}$ are mutually independent, the joint prior $p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{z}_{1:T})$ can be factorized as

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(y_t) p_{\boldsymbol{\theta}}(\boldsymbol{f}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{<t}). \tag{4.2}$$

We use independent priors to regularize latent variables to be as independent as possible.

4.2.1.3 Generative Model: Decoder

The generative model is formalized by the factorization as

$$
\begin{aligned}
&p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{f}, \boldsymbol{z}_{1:T}) \\
&= \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(y_t) p_{\boldsymbol{\theta}}(\boldsymbol{f}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{z}_t|\boldsymbol{z}_{<t}) p_{\boldsymbol{\theta}}(W_t|y_t, \boldsymbol{f}, \boldsymbol{z}_t) p_{\boldsymbol{\theta}}(X_t|W_t),
\end{aligned} \tag{4.3}
$$

where $\boldsymbol{W} = (W_1, \ldots, W_T)$ denotes the P-dimensional latent vectors and $W_t \in \mathbb{R}^{N \times P}$. Latent vectors $\boldsymbol{W}$ are generated from the time-dependent $\boldsymbol{y}$ and the two disentangled variables, i.e., the time-invariant $\boldsymbol{f}$ and the time-varying $\boldsymbol{z}$. We assume that data sequences $\boldsymbol{X}$ are generated from latent vectors $\boldsymbol{W}$ via the graph convolution as

$$
X_t = \hat{A} W_t \Theta, \tag{4.4}
$$

where Θ is the trainable weight matrix, A is the adjacent matrix of the graph G, $\hat{A} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$, $\tilde{A} = A + I$, and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. That is, we incorporate the topology of the graph G into the generative process using graph convolution.

4.2.1.4 Inference Model: Encoder

We use variational model $q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{X}, \boldsymbol{y})$ to approximate the true posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{X}, \boldsymbol{y})$ over latent variables given data [86, 17]. Our inference model is factorized as

$$
q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{X}, \boldsymbol{y}) = q_{\boldsymbol{\phi}}(\boldsymbol{f}|y_{1:T}, X_{1:T}) \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\boldsymbol{z}_t|y_{\leq t}, X_{\leq t}) \,. \tag{4.5}
$$

We can see that the time-invariant variable $\boldsymbol{f}$ are conditional on the entire time sequences of data and time-dependent labels $\boldsymbol{y}$, while the time-dependent variable $\boldsymbol{z}_t$ is inferred from the sequences before timestamp t, i.e., $X_{\leq t}$ and $y_{\leq t}$. We model both $\boldsymbol{f}$ and $\boldsymbol{z}$ via a recurrent network LSTM.

4.2.2 Predictive Supervised Sequential Graph VAE Model

Although SSG-VAE model in Section 4.2.1 can successfully learn a disentangled representation which decomposes the static, time-varying and label information, it cannot be utilized for data forecasting at future timestamps, due to the lack of the forecasting layer in the model. To conquer this challenge, we extend our model SSG-VAE to the Predictive Supervised Sequential Graph VAE (PSSG-VAE) model, which allows us to forecast the latent variables and output outcomes in the future time stamps.

Assuming the complete data pairs $\boldsymbol{X} = (X_1, \ldots, X_T)$ and $\boldsymbol{y} = (y_1, \ldots, y_T)$, we denote the observed data pairs as $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_{T-1})$ and $\tilde{\boldsymbol{y}} = (y_1, \ldots, y_{T-1})$, with X_T, y_T denotes the forecast data pair. Then we can reformulate the inference model in Section 4.2.1.4 as

$$q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T}|\boldsymbol{X}, \boldsymbol{y}) = q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T-1}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_T|\boldsymbol{z}_{T-1}), \tag{4.6}$$

where $q_\phi(z_T|z_{T-1})$ can be any parametric function and we use a naive linear transition model for the remainder of this work, and $q_{\boldsymbol{\phi}}(\boldsymbol{f}, \boldsymbol{z}_{1:T-1}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$ can be factorized similarly using Eq. 4.5. With the aforementioned reformulation of the inference model, our extended model PSSG-VAE will be capable of data forecasting at future timestamps.

4.3 Experimental Results

4.3.1 ADNI Dataset

Total of N=720 subjects were taken from the ADNI study that contained both amyloid PET and DTI images. Longitudinal β-amyloid data were processed from amyloid PET scans, and structural connectivity matrices (i.e., number of fiber tracts connecting different ROIs) were derived from DTI registered at Destrieux atlas in FreeSurfer [49] using a in-house tractography pipeline. Specifically, standardized uptake value ratio (SUVR) was computed for β-amyloid at each brain region, and an overall graph was obtained by taking the average

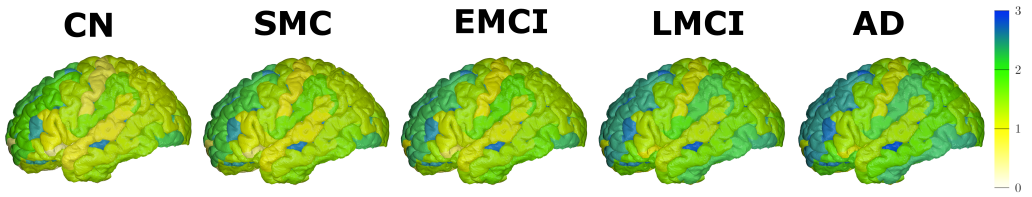


Figure 4.2: Latent Traversals over Labels. From left to right: generated β-amyloid on brain surfaces with latent variables fixed and diagnostic labels varying from CN to AD, respectively. Label-related patterns match with existing knowledge from the neuroscience domain.

of connectivity matrices from healthy subjects. Diagnostic labels of each scan categorize subjects' dementia stage as one of Cognitive Normal (CN), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD). Demographics of the subjects are presented in Table 4.1.

Table 4.1: Demographics of the ADNI dataset.

Demographics	CN	SMC	EMCI	LMCI	AD
# of Subjects	204	80	240	107	89
Gender (M/F)	106:98	27:53	138:102	61:46	44:45
Age (mean,std)	73.33(6.06)	71.06(5.01)	70.91(7.16)	72.04(7.92)	73.22(7.48)

CN: Cognitive Normal; SMC: Significant Memory Concern; EMCI: Early Mild Cognitive Impairment; LMCI: Late Mild Cognitive Impairment; AD: Alzheimer's Disease.

4.3.2 Experimental Setup

We applied the framework to the dataset from Alzheimer's Disease Neuroimaging Initiative(ADNI) including longitudinal β-amyloid data on brain networks with diagnostic stage labels of AD. We conducted three experiments to validate the performance of our framework as described below. The experiments were performed with 3-fold cross validation. Section 4.3.3 displays the task of latent traversals over labels, where we explored how the

patterns of generated amyloid will change corresponding to the variations of diagnostic labels. Section 4.3.4 shows the reconstruction performance on the dynamics of β-amyloid, as compared with the ground truth and visualized on the brain surfaces. Section 4.3.5 demonstrates the forecasting performances at the future timestamp. Here, since each subject has different number of visits, we propose two baseline approaches for the comparison with our approach. One is the averaging where the amyloid measure at the last timestamp is estimated as the average of all historical measures at previous timestamps (assuming no time-varying effect), the other is the linear regression by leveraging the average from all past timestamps as the predictors. Root mean square error (RMSE) and Mean Absolute Error (MAE) between the ground truth and the predicted are used as the metrics for the evaluation of forecasting performance.

4.3.3 Latent Traversals over Labels

To explore the relation between labels and generated amyloid, we conducted the latent traversals task over labels of diagnostic stages using our proposed SSG-VAE model in Section 4.2.1. Specifically, we fixed a time-invariant variable $\boldsymbol{f}$ and a time-varying variable $\boldsymbol{z}$, and we varied the corresponding label from 0 to 4, which represents the label from CN to AD respectively. We reconstructed the β-amyloid with those different labels but the same other latent variables shown in Figure 4.2. It is clear that as the status of disease stage becomes worse, the values of amyloid measurement become larger, matching the existing knowledge from the neuroscience domain.

4.3.4 Reconstruction on the Dynamics of β-Amyloid

Here we illustrate that our SSG-VAE model can learn the complex dynamics of β-amyloid by showing the reconstruction results on the testing data. We show the true amyloid and reconstructed amyloid on brain surfaces in Figure 4.3. It visually demonstrates that the

reconstruction not only captures the anatomical information but also successfully learns the true dynamics from the limited and noisy longitudinal data, as it resembles the patterns of amyloid on brain surfaces across timestamps (see the color changes in Figure 4.3).

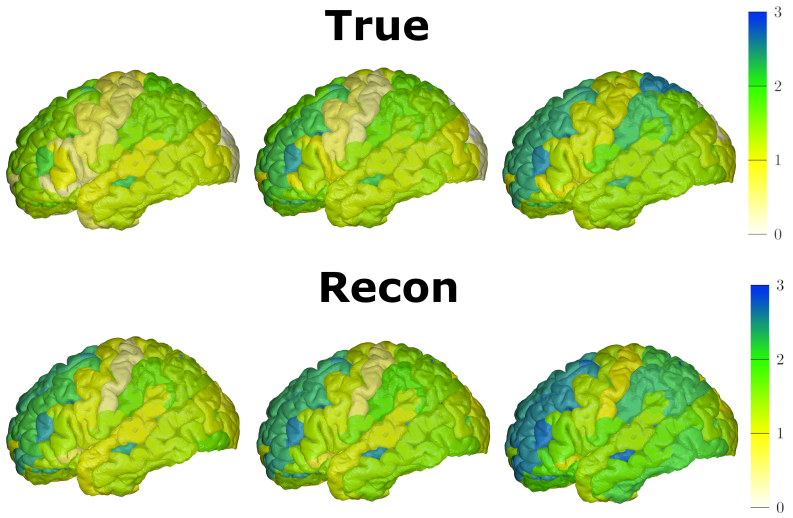


Figure 4.3: Top: true brain surfaces for a randomly selected subject at timestamp t_0, t_1 and t_2 (True). Bottom: reconstructed brain surfaces for the same subject at timestamp t_0, t_1 and t_2 (Recon). Drawings generated using BrainPainter [2].

4.3.5 Forecasting β-Amyloid at the Future Timestamp

We also applied PSSG-VAE model in Section 4.2.2 on the same dataset. The overall RMSEs on all testing data for average approach, regression approach and our approach are 0.38, 0.22 and 0.19, respectively. Our approach attains the lowest overall RMSE on the forecasting. We also summarised the RMSEs and MAEs across all the diagnostic labels in Table 4.2. It can be seen that our approach performs significantly better than the regression approach on the earliest CN stage, which shows the advantage of our approach that it can capture the earliest sign of cognitive decline at the preclinical stage. Moreover, we visualized RMSEs in the boxplot for average approach, regression approach and our proposed approach in Figure 4.4. It shows that the forecasts from our approach is more robust and has smaller

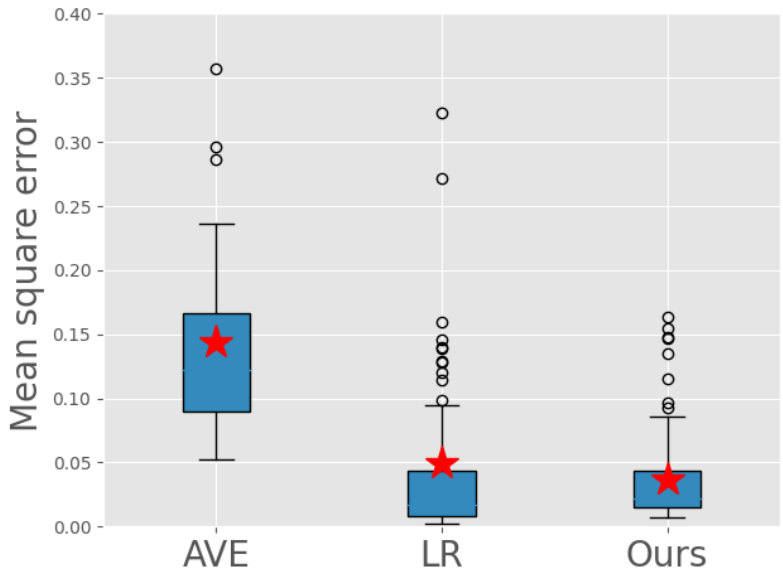


Figure 4.4: Boxplot of forecasting performance at the future timestamp visualizing RMSEs for average approach, regression approach and our approach. Our approach yields the lowest overall RMSE and has smaller variation when compared to other approaches.

variation, as demonstrated by the smaller interquartile range and fewer extreme outliers in the boxplot.

Table 4.2: Forecasting Performance across diagnostic stages.

		CN	SMC	EMCI	LMCI	AD
Average	RMSE	0.42	0.39	0.35	0.37	0.36
	MAE	0.32	0.31	0.28	0.30	0.28
Regression	RMSE	0.30	0.20	0.17	0.17	0.20
	MAE	0.21	0.16	0.13	0.13	0.15
Ours	RMSE	0.20	0.20	0.17	0.15	0.19
	MAE	0.15	0.15	0.13	0.11	0.14

CN: Cognitive Normal; SMC: Significant Memory Concern; EMCI: Early Mild Cognitive Impairment; LMCI: Late Mild Cognitive Impairment; AD: Alzheimer's Disease.

4.4 Conclusion

Understanding the dynamics of β-amyloid and forecasting amyloid depositions at future timestamps will help facilitate the clinical research in the preclinical stages of the disease. In this paper, we developed a novel Supervised Sequential Graph Autoencoder model to learn a latent disentangled representation comprising time-varying and time-invariant information to characterize the longitudinal β-amyloid data over the structural brain network. With the learned disentangled representation of ROI specific measures, our framework can capture the robust dynamics of amyloid and forecast future amyloid depositions from a few past time points.

Part III

Dynamic Covariance Modeling on HCP Data

CHAPTER 5

Dynamic Covariance Estimation via Predictive Wishart Process with an Application on Brain Connectivity Estimation

Modelling the complex dependence in multivariate time series data is a fundamental problem in statistics and machine learning. Traditionally, the task has been approached with methods such as multivariate autoregressive models and multivariate generalized autoregressive conditional heteroskedasticity models, and Gaussian process based methods are recently becoming popular by leveraging the flexibility of non-parametric learning. However, few methods exist that directly model the dynamics of the covariance matrices except generalized Wishart process ($\mathcal{GWP}$), and even the generalized Wishart process is limited with applications on small data due to the extremely high computational capacity induced by multiple Gaussian processes. In this regards, we propose a novel stochastic process named as Predictive Wishart Process ($\mathcal{PWP}$), which provides a collection of positive semi-definite random matrices indexed by input variables. $\mathcal{PWP}$ projects process realizations of $\mathcal{GWP}$ to a lower dimensional subspace to efficiently estimate every $\mathcal{GWP}$. We discuss its theoretical properties and design Bayesian and variational inferences for efficiency. The $\mathcal{PWP}$ is empirically tested on synthetically generated time-series data to validate competitive reconstructive performance and efficient predictive performance, and applied on a large-scale real functional magnetic resonance imaging (fMRI) dataset from Human Connectome Project (HCP) to demonstrate its practicality via a multi-task learning framework.

5.1 Introduction

Accurate estimation of associations over a set of variables is a fundamental problem in statistics (and machine learning) with significant interest from diverse domains. Typically, the associations (e.g., covariance) are assumed to be static, and they are often estimated using structural equation models or graphical models [87, 88, 89]. However, when the given data are time-dependent, they often exhibit heteroscedasticity, i.e., the variances and correlations of variables of interest are time-varying [90, 91, 92, **?**, **?**]. Therefore, accounting for both temporal and spatial dependency in the covariance is critical in various motivating applications, e.g., capturing the time-varying volatility of a collection of risky assets in econometrics [93, 27], and modelling spatial variations in correlations for customarily recorded multivariate measurements at a large collection of locations for geoscience [94, 95].

Such a problem routinely arise in brain connectivity analyses in Neuroimaging, which often requires estimating covariance from a knot of measurements (e.g., timeseries) across spatially parcellated Regions of Interest (ROIs) in the brain. Here, the covariance quantifies the level of associations between different ROIs as a functional connectivity [20]. Conventional connectivity constructions assume that the functional associations are static in time over the entire scan period [21, 22]. Nevertheless, several studies demonstrate that the functional connectivities *change over time* whose temporal variation may be significant [23, 24]. Therefore, deriving dynamic associations between ROIs is an important problem for both statistics and neuroscience, which investigates the time-varying co-activation patterns in the brain activities [23, 25, 26].

Unfortunately, modelling such dynamic changes of covariance is quite challenging, because the given data are often in a large scale in length and typically only a single observation is recorded at each time stamp. In the statistical literature, modeling the dynamics of covariance has been tackled with Multivariate Generalized Autoregressive Conditional Heteroskedasticity (MGARCH) models [96], and alternative approaches were proposed

such as Bayesian nonparametric models based on Wishart process (WP) [27, 28]. However, recent works including Generalized Wishart Process ($\mathcal{GWP}$) on Bayesian inference for WP are limited as they often require extremely high computational capacity due to the burden introduced from latent Gaussian processes, and hence makes it difficult to scale down for practical model inference.

To tackle the problem above, we develop Predictive Wishart Process ($\mathcal{PWP}$), which is a novel *parsimonious stochastic process* which approximates the traditional $\mathcal{GWP}$. We thoroughly study the stochastic properties of the $\mathcal{PWP}$ and provide full Bayesian posterior inference, which has been dismissed in previous literature. This framework is *scalable* to generate time-varying covariance $\Sigma(x)$ for a given index x from large-scale data (see Figure 5.1) under rigorous mathematical properties. The complexity of generating time-varying covariance matrices is *linear* with respect to the number of covariance matrices (N) as opposed to $\mathcal{GWP}$ whose complexity of generating latent variables in each GP is *cubic* in N. Due to the parsimony of the predictive process, both Bayesian and variational inferences of the dynamic covariance structure with $\mathcal{PWP}$ become efficient.

The main **contributions** of our work are summarized as:

- **(i)** We introduce a novel matrix variate stochastic process and theoretically demonstrate its desirable properties,
- **(ii)** We propose Markov chain Monte Carlo (MCMC) and variational expectation maximization inference associated with a hierarchical Gaussian model and illustrate both computational benefits and comparable predictive performance of $\mathcal{PWP}$;
- **(iii)** We provide a multi-task learning framework using $\mathcal{PWP}$ to jointly model multiple large-scale signals, and empirically prove the efficiency and practicality of $\mathcal{PWP}$ by tackling a real large-scale problem where conventional methods fails.

Extensive experiments are carried on synthetic experiments (with ground truth) as well as on a large-scale real Neuroimaging study (i.e., Human Connectome Project (HCP))

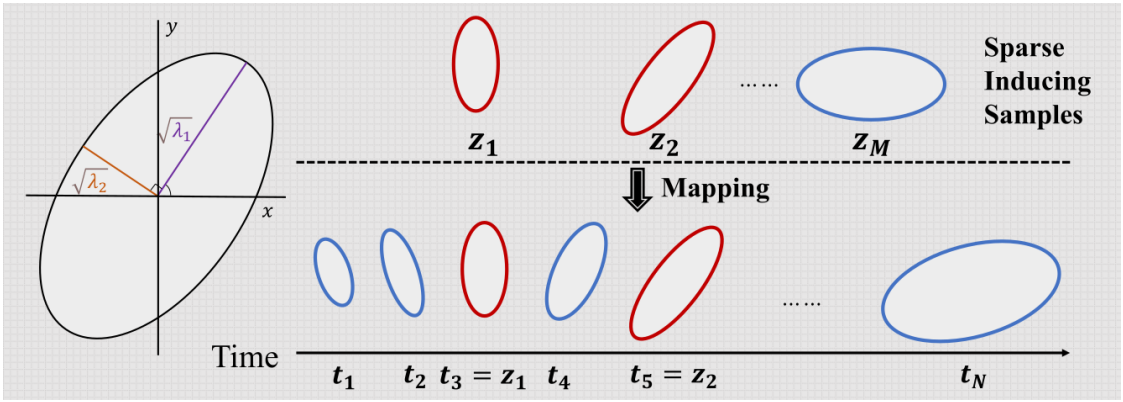


Figure 5.1: A draw from a Predictive Wishart Process ($\mathcal{PWP}$). Each ellipse is a 2×2 covariance matrix index by observed time $\{t_i\}_{t=1}^{N}$ or inducing time $\{z_j\}_{j=1}^{M}$. The rotation indicates the correlation between the two variables, and the major and minor axes scale with the eigenvalues (i.e., λ_1, λ_2) of the matrix. A draw from a PWP consists of two steps: (i), we draw a collection of matrices indexed by inducing time; (ii), we map the collection of matrices to another collection of matrices indexed by observed time.

with resting-state functional MRI (fMRI) [97] for reconstruction and prediction of dynamic covariances. Utilizing $\mathcal{PWP}$ leads to improvement in characterizing behavioral scores with dynamic covariance; our pioneering exploration on modelling dynamic connectivity should be worth pursuing further.

5.2 Related Works

There exists a large body of literature on modeling time-varying covariance matrix, and classical strategies for estimating the covariance rely on standard regression methods with the Cholesky decomposition of the covariance or precision matrices [98, 99]. Alternatively, nonparametric approaches have been proposed in [100, 95].

For modelling multivariate time series, heteroscedastic modeling has a long history, where the main approaches are Multivariate Generalized Autoregressive Conditional Heteroskedasticity (MGARCH) models [96, 101, 102], multivariate stochastic volatility models [103, 104] and Wishart process [105, 28].

In particular, there exist two Wishart process based methods: Wishart autoregressive processes [105] that construct positive definite volatility matrices with latent autogressive (AR) models, and generalised Wishart process ($\mathcal{GWP}$) [28] that utilize Gaussian process to model latent process instead of AR models. Due to the limited expressiveness of AR models, Wishart autoregressive process cannot handle the long temporal dependence. On the other hand, $\mathcal{GWP}$ led to a diverse class of covariance dynamics, but it is not scalable to large datasets due to the expensive computation induced from corresponding latent Gaussian processes. Our approach attains the best of both worlds by utilizing a predictive process to model the dependence within those latent functions.

5.3 Preliminary

In this section, we briefly review a predictive process ($\mathcal{PP}$) [106, 107], as it sets the foundation of our proposed $\mathcal{PWP}$ construction. We begin with distributions over functions $u(x)$ using Gaussian process ($\mathcal{GP}$) as

$$u(x) \sim \mathcal{GP}(m(x), C(x, x')), \tag{5.1}$$

with a mean function $m(x)$ and a covariance function $C(x, x')$ of choice specified with hyper-parameters τ. And we name it parent process.

In the remainder of this paper, we consider a zero-mean Gaussian process, i.e., $m(x) \equiv 0$. Given a collection of inducing inputs $\boldsymbol{z} = (z_1, \ldots, z_M)$, the collection of function values $\boldsymbol{u}$ has a joint Gaussian distribution as

$$\boldsymbol{u} = (u(z_1), \ldots, u(z_M))^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}^*), \tag{5.2}$$

where $\boldsymbol{C}^*$ is the covariance matrix introduced by the covariance function $C(x, x')$ on inducing points $\boldsymbol{z}$.

A predictive process, i.e. $\mathcal{PP}$, is derived from its parent process (5.1) on a completely specified lower dimensional subspace. Specifically, given (5.1), the predictive process is

defined as $\tilde{u}(x) \sim \mathcal{PP}(0, C(x, x')) = \mathcal{GP}(0, \tilde{C}(x, x'))$ that is equivalent to a new specified Gaussian process defined by the covariance function

$$\tilde{C}(x, x') = \boldsymbol{c}^T(x)\boldsymbol{C}^{*-1}\boldsymbol{c}(x')\,, \tag{5.3}$$

where $\boldsymbol{c}(x) = (C(x, z_1), \ldots, C(x, z_M))^T$. Here, two major properties of $\mathcal{PP}$ are given [106]:

$$\tilde{u}(x) = c^T(x)\boldsymbol{C}^{*-1}\boldsymbol{u}\,, \tag{5.4}$$

$$\tilde{C}(x, x) \leq C(x, x)\,. \tag{5.5}$$

Note that (5.4) shows that the predictive process can be treated as a linear projection on the subspace spanned by $\boldsymbol{u}$, and (5.5) reveals that the predictive process will underestimate the variance of its parent process. A modified predictive process proposed in [107] can correct the bias of variances by replacing (5.3) with

$$\tilde{C}(x, x') = \begin{cases} C(x, x') & x = x' \\ \boldsymbol{c}^T(x)C^{*-1}\boldsymbol{c}(x') & x \neq x'. \end{cases} \tag{5.6}$$

In this paper, we construct our $\mathcal{PWP}$ based on the native $\mathcal{PP}$ rather than the modified version to design a concrete predictive process.

5.4 The Predictive Wishart Process

In this section, we first introduce the concept and construction of our proposed $\mathcal{PWP}$, then in the following we discuss its theoretical properties.

5.4.1 Construction of Predictive Wishart Process

Suppose that we have $\mathcal{V} \times \mathcal{D}$ independent predictive process functions with an unit variance in its parent process, i.e. $C(x, x) = 1$ for $x \in \mathcal{X}$, as

$$\tilde{u}_{vd}(x) \overset{ind}{\sim} \mathcal{PP}(0, C_d(x, x')), \tag{5.7}$$

where $v = 1, \ldots, \mathcal{V}$ represents the index of the degrees of freedom $\mathcal{V}$, and $d = 1, \ldots, \mathcal{D}$ is the index of the dimension of the multivariate features. We assume $\mathcal{V} \geq \mathcal{D}$ to ensure our construction is well defined. Here, the objective is to design a collection of positive semi-definite (p.s.d.) random matrices $\Sigma(x)$ (e.g., covariance matrices), indexed by any arbitrary input variable $x \in \mathcal{X}$ (e.g., time). Let $\tilde{\boldsymbol{u}}_v(x) = (\tilde{u}_{v1}(x), \ldots, \tilde{u}_{v\mathcal{D}}(x))^T$, and let $S \in \mathcal{S}^D$ represent a positive definite matrix with its unique lower Cholesky decomposition matrix L such that $LL^T = S$. We denote $\tilde{U}(x) = (\tilde{\boldsymbol{u}}_1(x), \ldots, \tilde{\boldsymbol{u}}_{\mathcal{V}}(x))$.

Predictive Wishart Process ($\mathcal{PWP}$) is defined as a collection of p.s.d. random matrices $\{\Sigma(x)\}$ indexed by $x \in \mathcal{X}$, by modelling the process as

$$\Sigma(x) = L\tilde{U}(x)\tilde{U}(x)^T L^T = \sum_{v=1}^{\mathcal{V}} L\tilde{\boldsymbol{u}}_v(x)\tilde{\boldsymbol{u}}_v^T(x)L^T. \tag{5.8}$$

with all latent processes following independent predictive processes. We denote this process as $\mathcal{PWP}(L, \mathcal{V}, \tau)$ that depends on a lower triangular matrix L and a degree of freedom $\mathcal{V}$. The lower triangular matrix L models the marginal variance-covariance at any fixed timestamp and the degrees of freedom $\mathcal{V}$ describes the flexibility of temporal dependence and the hyper-parameters τ characterize latent processes.

If each predictive process of $\tilde{u}_{vd}(x)$ is replaced by its parent process (5.1), and then this process is formulated as *Generalized Wishart Process* ($\mathcal{GWP}$) [28] which is a generalization of the original Wishart process defined by [108]. The *Predictive Inverse Wishart Process* ($\mathcal{PIWP}$), consequently, can be indirectly defined as $\Omega(x) = \Sigma^{-1}(x)$, given $\Sigma(x) \sim \mathcal{PWP}(L, \mathcal{V}, \tau)$. We note that at any index x, the distribution of $\Omega(x)$ is an inverse Wishart distribution.

5.4.2 Properties of Predictive Wishart Process

We first show that the proposed $\mathcal{PWP}$ at any input x follows a *well-defined* Wishart distribution $\mathcal{W}_{\mathcal{D}}$ in the theorem below.

Theorem 1. *For any input variable x, the distribution of $\Sigma(x) \sim \mathcal{PWP}(L, \mathcal{V}, \tau)$ at x is the Wishart distribution such that $\Sigma(x) \sim \mathcal{W}_{\mathcal{D}}(\mathcal{V}, S^*)$, where $S^* = LBL^T$ and B is the diagonal matrix with elements $b_d = \tilde{C}_d(x, x)$.*

Remarks 1. Theorem 1 shows the marginal distribution of $\mathcal{PWP}$ prior at any input x is a well-defined Wishart distribution, and the distribution of $\Sigma(x)$ in $\mathcal{PWP}$ is different from $\mathcal{GWP}$.

Notice that when the predictive process priors are replaced by modified predictive process priors [106, 107], the distribution of $\Sigma(x)$ at any input variable x is the Wishart distribution such that $\Sigma(x) \sim \mathcal{W}_D(\mathcal{V}, S)$.

For simplicity, in the reminder of paper, we assume that all latent functions $\tilde{u}_{vd}$ share the same covariance function C. We derive expressions for the covariance between elements of $\Sigma(x)$ and $\Sigma(x')$ for any pair of inputs x and x' in Theorem 2, assuming L is diagonal and $\{\tilde{u}_{vd}\}$ have an identical predictive process prior. Proofs of Theorem 1 and 2 will be given in the Appendix.

Theorem 2. *Assume that L is a diagonal matrix and $\{\tilde{u}_{vd}\}$ have an independent identical predictive process priors. For any pair of inputs variables x and x', the covariance between $\Sigma_{ij}(x)$ and $\Sigma_{kl}(x')$ is given as*

$$\begin{aligned} &\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\ &= \begin{cases} 2\mathcal{V} l_i^4 \tilde{C}^2(x, x') & i = j = k = l \\ \mathcal{V} l_i^2 l_j^2 \tilde{C}^2(x, x') & i = k \neq j = l \\ 0 & otherwise. \end{cases} \end{aligned} \tag{5.9}$$

Remarks 2. Theorem 2 discusses the temporal cross-relation of dynamic covariance matrices. The covariance turns out to be proportional to the $\tilde{C}^2(x, x')$ and hence shows that the selection of C undoubtedly plays an important role of controlling the autocorrelations. The covariance relation can be generalized to any lower triangular L.

Remarks 3. Although the priors of $\Sigma(x)$ from $\mathcal{PWP}$ and $\mathcal{GWP}$ both belong to Wishart distribution, they have different scale matrices, $S = LL^T$ for $\mathcal{GWP}$ and $S^* = LBL^T$ for $\mathcal{PWP}$. Because $b_d(x) = \tilde{C}(x,x)$ and $C(x,x) = I$, this difference depends on how well $\tilde{C}$ approximates C. Notice that $\tilde{C}$ is the Nyström approximation of C in (5.3) [109], and the error $\|\tilde{C} - C\|_F$ under the Frobenious norm has an upper bound which is a polynomial function of the square root of the quantization error $\sum_{n=1}^{N} \|x_i - z_{c(i)}\|$ with c coding each input x_i with the closest inducing input z_j. Therefore, the difference of prior of $\Sigma(x)$ from $\mathcal{PWP}$ and $\mathcal{GWP}$ is determined on the displacement of inducing inputs and quantitatively influenced by the quantization error. We suggest the K-mean sampling method for the displacement of inducing inputs, and the sampling approach is used to minimize the quantization error.

5.5 Hierarchical Gaussian Model with $\mathcal{PWP}$

Given a $\mathcal{D}\times N$ dataset $\boldsymbol{Y} = (\boldsymbol{y}(x_1), \cdots, \boldsymbol{y}(x_N))$ with $\mathcal{D}$-dimensional multivariate features indexed by the input variables $x_1, \cdots, x_N$. We consider a conditional Gaussian model with time-varying covariance modeled by $\mathcal{PWP}$ as

$$\begin{aligned} \boldsymbol{y}_i|\Sigma_i &\sim \mathcal{N}(\mathbf{0}, \Sigma_i), \\ \Sigma(x) &\sim \mathcal{PWP}(L, \mathcal{V}, \tau), \end{aligned} \tag{5.10}$$

where $\boldsymbol{y}_i = \boldsymbol{y}(x_i)$ and $\Sigma_i = \Sigma(x_i)$. We propose two inference approaches: 1) Bayesian and 2) Variational inferences. Specifically, Bayesian inference is a Markov Chain Monte Carlo method (MCMC), which accurately provides the samples of posterior distributions. As MCMC can be computationally expensive because it would take long time to converge, we also propose a variational inference which is well suited for large datasets. Moreover, in practice, learning the uncertainty of model parameters L and τ is not of interest and thus we treat them as hyper-parameters to relieve computational burden. Two inference methods are

briefly summarized in Table 5.1 and will be described in details in the following sections respectively.

Table 5.1: A summary of inference approaches for $\mathcal{PWP}$s. Here, $\boldsymbol{w}$, τ, L refer to the inducing variables, input-dependent hyper-parameters and input-independent hyper-parameters, respectively.

Inference [+]	Parameters		
	$\boldsymbol{w}$	τ	L
PWP-MCMC	MCMC	MCMC	MCMC
PWP-VB	VB	(optimized)	(optimized)

[+] PWP-MCMC: Bayesian inference, PWP-VB: Variational inference.

5.5.1 Bayesian Inference Approach

This section discusses a Bayesian inference with $\mathcal{PWP}$. In the context of (5.10), the objective is to infer the posterior $p(\Sigma(x)|\boldsymbol{y})$ using Gibbs sampling [110], which is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations in cycles from the conditional distribution of one parameter with the remaining parameters fixed to their current values.

For the sampling, we rewrite (5.10) as a hierarchical model:

$$\boldsymbol{y}_i|L, \tilde{U}_i \sim \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\tilde{U}_i\tilde{U}_i^T L^T), \tag{5.11}$$

$$\tilde{\boldsymbol{u}}_{vd} = \boldsymbol{c}^T \boldsymbol{C}^{*-1} \boldsymbol{w}_{vd}, \tag{5.12}$$

$$\boldsymbol{w}_{vd}|\tau \sim \mathcal{N}(\boldsymbol{w}_{vd}|\boldsymbol{0}, C^*), \tag{5.13}$$

where $\tilde{U}_i = \tilde{U}(x_i)$, $\tilde{\boldsymbol{u}}_{vd} = (\tilde{u}_{vd}(x_1), \ldots, \tilde{u}_{vd}(x_N))^T$, $\boldsymbol{w}_{vd} = (u_{vd}(z_1), \ldots, u_{vd}(z_M))^T$. Here u_{vd} refers to the function of the parent process with respect to $\tilde{u}_{vd}(x)$. On the other hand, C^* refers to covariance between $\{z_i\}_{i=1}^M$ and $\boldsymbol{c}$ is cross covariance between $\{x_i\}_{i=1}^N$ and $\{z_i\}_{i=1}^M$.

As for the prior specification, we set prior of hyper-parameters of GPs $\tau \sim \pi(\tau)$ and the prior of the lower triangular matrix $L \sim \pi(L)$. The prior of τ is chosen based on the choice of covariance function C. In the experiments, we consider two types of covariance functions, one for periodic covariance function and the other for square exponential function. We put a flat normal distribution as a prior of the log of lengthscale parameters. And for $\pi(L)$, we put independent standard Gaussian priors for the entries on or below the diagonal of L. We then design a Gibbs sampling procedure as

$$p(\boldsymbol{w}|\boldsymbol{Y}, \tau, L) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)p(\boldsymbol{w}|\tau), \tag{5.14}$$

$$p(\tau|\boldsymbol{Y}, \boldsymbol{w}, L) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)p(\boldsymbol{w}|\tau)\pi(\tau), \tag{5.15}$$

$$p(L|\boldsymbol{Y}, \boldsymbol{w}, \tau) \propto p(\boldsymbol{Y}|\boldsymbol{w}, L, \tau)\pi(L), \tag{5.16}$$

where $\boldsymbol{w}$ represent the vector of functions evaluated from the inducing points, τ denote the input-dependent hyper-parameters in $\mathcal{PWP}$ and they are also the hyper-parameters in the covariance function C, and L denote the input-independent hyper-parameters in $\mathcal{PWP}$. Furthermore, we present the details of parameter initialization, posterior sampling and inducing point selection regarding the MCMC implementation for the Bayesian inference approach.

5.5.1.1 Parameter Initialization

According to Theorem 1 that $\Sigma(x) \sim \mathcal{W}(\mathcal{V}, S^*)$, the prior expectation of covariance matrix $\Sigma(x)$ equals $\mathcal{V}S^*$. In the initialization step, we assume that $\Sigma(x_1), \ldots, \Sigma(x_N)$ are independent, then the covariance matrix $\Sigma(x)$ has an unbiased estimate $\hat{\Sigma}(x) = \frac{1}{N-1}\sum_{i=1}^{N} \boldsymbol{y}_i\boldsymbol{y}_i^T$. Consequently, $\hat{L}$ can be estimated by the Cholesky decomposition of $\frac{\hat{\Sigma}}{\mathcal{V}}$ by assuming that S^* is close to S. Then following (5.10), we estimate the $\boldsymbol{w}$ and τ by maximizing the log likelihood of $\boldsymbol{Y}$ given $\hat{L}$.

5.5.1.2 Details on Posterior Sampling

We first sample the $\tilde{\boldsymbol{u}}_{vd}$ from its posterior distribution via indirect sampling of $\boldsymbol{w}_{vd}$ using (5.14). Given the property of predictive process (5.4) and $\boldsymbol{w}_{vd}$, $\tilde{\boldsymbol{u}}_{vd}$ is generated via $\tilde{\boldsymbol{u}}_{vd} = \boldsymbol{c}C^{*-1}\boldsymbol{w}_{vd}$. As for the sampling of $\boldsymbol{w}$, we employ the Elliptical Slice sampling and this sampling procedure requires to computing the posterior of $\boldsymbol{w}$, taking $\mathcal{O}(M^2N)$ time complexity where N is the number of observations and M is the number of inducing points. Therefore the sampling complexity is linear to the number of observations N. In contrast to $\mathcal{GWP}$ in which sampling the latent function from the posterior would take $\mathcal{O}(N^3)$ time complexity, $\mathcal{PWP}$ is much more efficient, especially when the number of inducing points M is significantly smaller N, i.e., $M \ll N$. Then, we sample τ using (5.15) and sample L using (5.16). Since the posterior of τ and L do not have a closed-form expression, we leverage Metropolis Hastings for sampling.

5.5.1.3 Inducing Points Selection

For selecting the inducing points, we take equal-spaced points $\{z_i\}_{i=1}^{M}$ over the whole input space $\mathcal{X}$ to ensure the better prediction performance over the whole input space. These z are fed in (5.2) that leads to the definition the $\mathcal{PWP}$. While Bayesian inference yields the true posterior for better estimation of covariance, it is often intractable due to slow convergence with exhaustive sampling. We therefore propose an efficient variational inference in the following.

5.5.2 Variational Expectation Maximization

Variational inference provides an alternative efficient inference approach at the price of precision of the posterior approximation. It is a Bayesian technique of approximating the posterior which has emerged as an important tool [86, 111]. We consider the same

hierarchical model from (5.11, 5.12, 5.13), and L and τ are treated as hyper-parameters as opposed to Bayesian inference. This is because learning the posterior distribution of those hyper-parameters is not of interest in practice, and it would save computation in training.

Given above specifications, the evidence lower bound (ELBO), a lower bound of the log marginal likelihood is derived with Shannon entropy H as

$$\log p(\boldsymbol{Y}) \geq \mathrm{E}_{q(\boldsymbol{w})}[\log p(\boldsymbol{Y}, \boldsymbol{w})] + H(q(\boldsymbol{w})) = \mathrm{ELBO}, \tag{5.17}$$

where $q(w)$ is a variational distribution of w.

We assume $q(\boldsymbol{w})$ belongs to normal distribution. Instead of directly maximizing the ELBO (5.17) with respect to $q(\boldsymbol{w})$ and (L, τ) via stochastic gradient descend, we iteratively and conditionally update $q(\boldsymbol{w})$ and (L, τ) until they converge. It is called variational expectation maximization (VEM) inference [**?**]. Specifically, given (L, τ), maximizing the ELBO (5.17) is equivalent to minimizing the Kullback-Leibler divergence between the variational distribution $q(\boldsymbol{w})$ and the posterior distribution $p(\boldsymbol{w}|\boldsymbol{y})$. Due to the Gaussian assumption in $q(\boldsymbol{w})$, we approximately update $q(\boldsymbol{w})$ via the Laplace approximation [112] $q^*(\boldsymbol{w})$. On the other hand, given a $q(\boldsymbol{w})$, (L, τ) are updated by

$$\begin{aligned} L^*, \tau^* &= \arg\max_{L,\tau} \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{w})}[\log \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\tilde{U}_i\tilde{U}_i^T L^T)] + R \\ &= \arg\max_{L,\tau} \sum_{i=1}^{N} [\log \mathcal{N}(\boldsymbol{y}_i|\boldsymbol{0}, L\langle\tilde{U}_i\rangle\langle\tilde{U}_i\rangle^T L^T)] + R \\ &= \arg\max_{L,\tau} \sum_{i=1}^{N} \mathcal{L}_i + R, \end{aligned} \tag{5.18}$$

where both regularization term $R = \mathrm{KL}(q(\boldsymbol{w})\|p(\boldsymbol{w}))$ and latent variables $\tilde{U}_i$ depend on τ, and $\langle\cdot\rangle = \mathrm{E}_{q(\boldsymbol{w})}[\cdot]$. We iteratively update $q(w)$ and (L, τ) until they converge.

5.5.3 Prediction of Covariance at New Timestamp

For both Bayesian and variational EM inferences, given a new time stamp x^*, we extract posterior samples $\{\boldsymbol{w}^{(s)}, \tau^{(s)}, L^{(s)}\}_{s=1}^{S}$ from MCMC or variational distributions, then we sample the corresponding $\tilde{u}_{vd}^* = \tilde{u}_{vd}(x^*)$ using

$$\tilde{u}_{vd}^{*(s)} = \boldsymbol{c}^{*T} C^{*-1} \boldsymbol{w}_{vd}^{(s)}, \tag{5.19}$$

where $\boldsymbol{c}^*$ denotes the vector of covariance functions evaluated between the new time stamp x^* and inducing inputs $\{z_i\}_{i=1}^{M}$, i.e. $\boldsymbol{c}(x^*)$, and $\boldsymbol{w}_{vd}^{(s)}$ represents the s^{th} posterior sample. Consequently, according to the construction (5.8), we obtain the posterior predictive samples of $\Sigma^* = \Sigma(x^*)$ by

$$\Sigma^{*(s)} = \sum_{v=1}^{V} L^{(s)} \tilde{u}_{vi}^{*(s)} \tilde{u}_{vj}^{*(s)} L^{(s)T}. \tag{5.20}$$

At last, we estimate Σ^* using the posterior predictive mean of the samples $\{\Sigma^{*(s)}\}_{s=1}^{S}$.

5.6 Multi-task Learning with $\mathcal{PWP}$

In this section, we consider a scenario of feature selection for multiple tasks, where each task is assigned with unique features. Assume that we have N tasks in which the i^{th} task consists of a multivariate time series with length N_i, i.e. $\boldsymbol{Y}_i = \{\boldsymbol{y}_{i,j}\}_{j=1}^{N_i}$. The corresponding time stamps are denoted as $\boldsymbol{x}_i = \{x_{i,j}\}_{j=1}^{N_i}$ and each observation $\boldsymbol{y}_{i,j} \in \mathbb{R}^M$ is assigned to the time stamp $x_{i,j}$. A hierarchical model is formulated as

$$\begin{aligned} \boldsymbol{y}_{i,j} | \Sigma_{i,j} &\sim \mathcal{N}(0, \Sigma_{i,j}), \\ \Sigma_i(x) &\sim \mathcal{PWP}(L_i, \mathcal{V}, \tau), \end{aligned} \tag{5.21}$$

where $\Sigma_{i,j} = \Sigma_i(x_{i,j})$. We assume that the model of $\Sigma_i(\cdot)$ shares the same degreee of freedom $\mathcal{V}$ and the same hyper-parameters in GPs τ, but has individual effect modeled by the task-specified lower triangular matrix L_i for the ith task. This specification suggests that covariances across tasks share the same latent temporal process prior, and covariances

within each task share a task-specified correlation structure modeled by the lower triangular matrix L_i. Thus, we take the L_i as a feature for task i which directly refers to task-specific feature.

To find out task-specific features, we estimate $L_i(\tau)$ for each task i under different settings of τ where τ can be treated as different scale and $L_i(\tau)$ is the feature at the scale τ. Because in the multi-task learning context, extracting task-specified feature is of interest and thus we treat $L_i(\tau)$ as model parameters. Specifically, we consider a square exponential covariance function in $\mathcal{PWP}$ where τ is the length scale parameter, and we define a $\mathcal{PWP}$ Multi-scale Descriptor ($\mathcal{PWPMD}$) as

$$\mathcal{PWPMD}_\tau(i) = \{L_i^*; L_i^*, q^*(\boldsymbol{w}) = \arg\max_{L_i, q(\boldsymbol{w})} (\text{ELBO}|\tau)\}. \tag{5.22}$$

Here, under each setting of τ, L_i^* becomes a feature for the i^{th} task. It has the same size of the feature of each task regardless of the number of observations N_i, and can be used for downstream prediction tasks. To infer the multi-scale descriptor, we propose a variational EM algorithm and describe it in Algorithm 1.

Algorithm 1: Variational Expectation Maximization Algorithm for Multitask Learning

Input : Observations $\boldsymbol{Y}$; Hyper-parameters of covariance functions $\boldsymbol{\tau}$

Output : Variational distribution $q(\boldsymbol{w})$; Task-specified features $\{L_i\}_{i=1}^N$

do

  Fix all task-specified features $\{L_i\}_{i=1}^N$ and update the variational distribution $q(\boldsymbol{w})$ by the Laplace approximation on $p(\boldsymbol{w}|\boldsymbol{Y}, \{L_i\}_{i=1}^N)$;

  for $i \leftarrow 1$ **to** N **do**

    Fix the variational distribution $q(\boldsymbol{w})$ and update the L_i by maximizing the term in ELBO that is only related to L_i:

$$L_i^* = \arg\max_{L_i} \sum_{j=1}^{N_i} [\log \mathcal{N}(\boldsymbol{y}_{i,j}|\mathbf{0}, L_i \langle \tilde{U}_{ij} \rangle \langle \tilde{U}_{ij} \rangle^T L_i^T)] \tag{5.23}$$

    where $\langle \cdot \rangle = E_{q(\boldsymbol{w})}[\cdot]$.

  end

while *Both* $q(\boldsymbol{w})$ *and* $\{L_i\}_{i=1}^N$ *converge*;

5.7 Simulation Study

In this section, we performed an experiment on the synthetic multivariate time-series data which were generated based on ground truth covariance matrices Σs to validate both covariance reconstruction and predictive performance of $\mathcal{PWP}$.

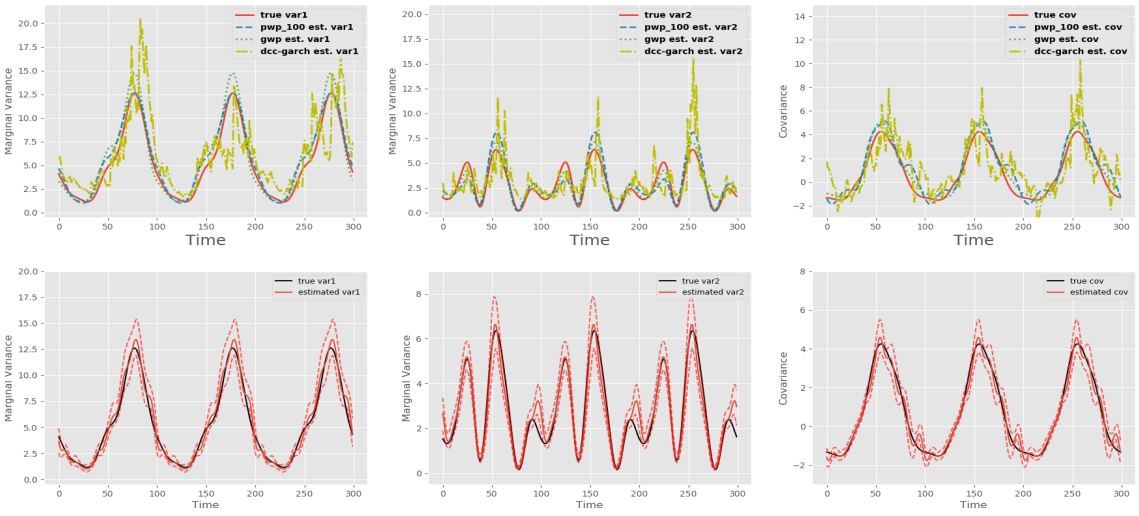
5.7.1 Experimental Setup

Synthetic Data Generation. We generated multivariate time series data using the $\mathcal{GWP}$ model with a periodic covariance function for all $\{u_{vd}(x)\}$ such that $k(x, x') = \sigma^2 e^{-2\sin(\pi * (x-x')/p)^2}$, with a scale parameter σ and a period parameter p. Specifically, $N = 350$, $\mathcal{D} = 2$ and $\mathcal{V} = 3$, L was chosen as an identity matrix and hyper-parameters were set as $\sigma = 1, p = 100$,

Table 5.2: Parameter posterior credible intervals 50 (2.5, 97.5), RMSE of the reconstruction for Σs, NLML with mean (standard deviation) and corresponding average inference time for 100 iterations.

	True	$\mathcal{GWP}$	$\mathcal{PWP}(\mathcal{B})_{20}$	$\mathcal{PWP}(\mathcal{B})_{50}$	$\mathcal{PWP}(\mathcal{B})_{100}$	$\mathcal{PWP}(\mathcal{VI})_{20}$	$\mathcal{PWP}(\mathcal{VI})_{50}$	$\mathcal{PWP}(\mathcal{VI})_{100}$	$\mathcal{DCC}$
L_{00}	1	1.12(0.98,1.28)	1.36(0.97.1.68)	1.17(1.06,1.43)	**0.99**(0.84, 1.08)	1.63	1.55	1.57	-
L_{01}	0	**-0.02**(-0.04,0.07)	-0.06(-0.19, -0.01)	0.04(-0.03, 0.15)	**0.02**(-0.02,0.08)	0.33	0.31	0.31	-
L_{11}	1	1.04(0.92,1.11)	1.05(0.87, 1.21)	1.12(1.02,1.26)	**1.02**(0.78,1.46)	1.16	1.10	1.10	-
RMSE (Σ_{00})	-	1.15	1.23	1.10	**0.55**	0.84	0.81	0.95	2.71
RMSE (Σ_{01})	-	**0.48**	0.95	0.74	0.90	0.67	0.65	0.70	1.67
RMSE (Σ_{11})	-	0.53	**0.46**	0.61	0.49	0.95	0.85	0.88	1.50
NLML	-	1098.94(2.88)	1105.88(6.69)	1096.82(3.37)	1105.27(4.19)	1082.05	1083.90	**1081.53**	-
Time (sec)	-	50.24	**25.39**	35.97	43.81	-	-	-	-

Subscript indicates the number of inducing points used in each model. (B) refers to Bayesian inference and (VI) refers to Variation inference. For all Bayesian inference, we have informative initialization on all latent variables based on the true values. We also provide the ground true parameters L.



(a) Marginal Var. of y_1 (b) Marginal Var. of y_2 (c) Covariance

Figure 5.2: Top: Reconstruction of Σs; Bottom: 95% confident intervals (shown in red dashed lines) in the reconstruction with $\mathcal{PWP}_{100}$, (a) the marginal variances at the first dimension (1st diagonal element of Σs), (b) the marginal variances at the second dimension (2nd diagonal element of Σs), (c) the covariances (symmetric off-diagonal element of Σ)s. Our proposed $\mathcal{PWP}$ delivers smoother estimations compared with $\mathcal{DCC}$ and also provides a comparable fitting performance compared to $\mathcal{GWP}$.

assuming that the period of the time series is 100. The first 300 data points were used for training and the following 50 samples were used for testing.

Baselines. Most recent methods such as $\mathcal{GWP}$ and zero-mean multivariate GARCH models, i.e., Dynamic Conditional Correlation ($\mathcal{DCC}$) [113], were chosen as the baseline methods.

Setup. For $\mathcal{PWP}$, different number of inducing points (i.e., $M = 20, 50,$ and 100) with the same type of periodic covariance function were investigated. We implemented both Bayesian inference and variational EM inference for $\mathcal{PWP}$. We fixed the hyper-parameter $p = 100$ since that is difficult to learn.

For Bayesian inference, we initialized L at the values near the true values in $\mathcal{GWP}$, latent variables $\boldsymbol{w}$ at the estimates via the inverse of (5.12) with the true $\tilde{U}$. This yields informative initialization to identify the property of the global optima in $\mathcal{GWP}$ and $\mathcal{PWP}$ for inferences. During the Bayesian inference of $\mathcal{PWP}$, we used 5000 samples whose first 2500 samples were burned-in. For variational EM inference, L and $\boldsymbol{w}$ were randomly initialized.

Evaluation Metric. In Table 5.2, we displayed the root mean square error (RMSE) of parameters for L as the evaluation of inference. We displayed the RMSE between true variance-covariance matrices and corresponding reconstruction as the evaluation of covariance reconstruction. Moreover we also provided the negative log marginal likelihood (NLML) to evaluate the model fitting.

In Table 5.3, We showed the predictive performance of $\mathcal{PWP}$ with i-step ahead forecast, where observations until the last timestamp x in training data are considered to predict $\Sigma(x + i)$ and $i = 1, \ldots, 50$.

5.7.2 Results and Discussions

Parameter estimation and model fitting results in Table 5.2 illustrate that $\mathcal{PWP}$ has a significantly better covariance matrix estimation performance than the $\mathcal{DCC}$ model due to the notably smaller RMSE. Comparing with the $\mathcal{GWP}$, with a suitable number of inducing points, $\mathcal{PWP}$ has a competitive result for both parameter estimation and covariance matrix estimation. As for the computational benefits, the computation time of $\mathcal{PWP}$ is significantly lowered compared with $\mathcal{GWP}$ in the same Bayesian setting.

As for the predictive performance, we conducted Bayesian inference for $\mathcal{PWP}$ as a fair comparison with the Bayesian inference in $\mathcal{GWP}$. We reported the RMSEs of predicted covariance matrices and true covariance matrices for $\mathcal{GWP}$, $\mathcal{PWP}$ with 20, 50 and 100 inducing points and $\mathcal{DCC}$ models in Table 5.3. The averaged RMSEs over all entries for the five models are 0.53, 0.52, 0.70, 0.70 and 2.53. It shows that $\mathcal{PWP}$ has a comparable performance compared with $\mathcal{GWP}$ and significantly outperforms the $\mathcal{DCC}$. Moreover, we visualized the ground truth for Σs and the reconstruction of Σs in $\mathcal{PWP}_{100}$ in Figure 5.2 and showed the uncertainty quantification of covariance matrices in $\mathcal{PWP}$, illustrating that $\mathcal{PWP}$ achieves a great uncertainty quantification in the sense that the confident intervals cover almost the true values with a narrow band-width.

With respect to the computational benefits, we find that as the number of inducing points decrease the computation time would be significantly shorter than that from $\mathcal{GWP}$. It matches the theoretical analysis of the computational complexity which is linear to the number of observations N in contrast to the $\mathcal{O}(N^3)$ in $\mathcal{GWP}$.

Table 5.3: RMSE between predicted $\hat{\Sigma}^*$ and true Σ^* element-wisely for the next 50 timestamps. $\mathcal{PWP}$ has a comparable performance with $\mathcal{GWP}$ even with much less inducing points.

Model[+]	Variance 1	Variance2	Covariance
$\mathcal{GWP}$	0.72	**0.49**	0.45
$\mathcal{PWP}_{20}$	**0.50**	0.84	**0.36**
$\mathcal{PWP}_{50}$	0.93	0.70	0.58
$\mathcal{PWP}_{100}$	0.75	0.95	0.55
$\mathcal{DCC}$	5.10	2.19	1.41

[+] Subscript indicates the number of inducing points used for $\mathcal{PWP}$. For $\mathcal{GWP}$ and $\mathcal{PWP}$, Bayesian inference and informative initialization on all latent variables based on the true values were used.

The results in Table 5.2 show that: For Bayesian inference, as the number of inducing inputs (M) increases, the parameter estimates of L become closer to the true values. However, the performance of covariance reconstruction and data fitting does not always improve as M increases in our setting. This may be caused by the efficiency of sampling the inducing variables $\boldsymbol{w}$. Even with an efficient elliptical slice sampling, as the size of $\boldsymbol{w}$ increases, the sampling step suffers from the slow mixing of sampling and cause undesirable fitting performance. It demonstrates that $\mathcal{PWP}$ becomes more expressive with more inducing points but fitting becomes more difficult, which emphasizes that the importance of the selection of inducing points.

On the other hand, $\mathcal{PWP}$ has a comparable prediction performance with $\mathcal{GWP}$ even with less inducing points. This may be because the learning of Gaussian processes in $\mathcal{GWP}$ is affected by over-fitting, while the learning of predictive processes in $\mathcal{PWP}$ resists this issue.

As for the variational EM inference of $\mathcal{PWP}$, it would provide biased estimates on L but we find that those estimates are consistently robust under different settings of the inducing points. Beside that, the variational EM inference provides comparable performance on both covariance reconstruction and model fitting.

5.8 Analysis of Dynamic Brain Connectivity

We performed two experiments on dynamic functional brain connectivity using real brain imaging data to confirm the practicality of $\mathcal{PWP}$. As $\mathcal{GWP}$ was not scalable for the real data, we compared $\mathcal{PWP}$ with $\mathcal{DCC}$-GARCH models for the individual analysis of dynamic functional connectivity. Then, we performed a multi-task learning task on multiple rs-fMRI timeseries via variational EM algorithm to identify associations between functional connectivity and behavioral scores.

5.8.1 Experimental Setup

Human Connectome Data. The pre-processed resting-state functional MRI (rs-fMRI) data used in this experiment were obtained from the Human Connectome Project (HCP) S1200 data release [114] for 812 subjects whose fMRI data were complete and reconstructed using the improved $r227$ recon algorithm. Timeseries data were generated through the HCP preprocessing pipeline [97] which yielded one representative timeseries across 4800 timestamps per independent component analysis (ICA) component for each subject at several different dimensionalities. Specifically, we used the rs-fMRI timeseries from 15 ICA components with a length of 4800.

Setup. We took the whole 4800 observations to estimate covariance matrices and computed the log likelihood at each timestamp. For $\mathcal{PWP}$, we selected 50 inducing points uniformly located in the whole time interval. Squared exponential covariance function was employed here to model the dynamics of covariance matrix of HCP data. We considered a weakly informative prior on the length scale parameter $\log \tau \sim \mathcal{N}(0, 10^2)$ and a data-driven prior on L, $L_{ij} \sim \mathcal{N}(0, 20^2)$ for $i \geq j$. On our server machine with 128G RAM (which is not small), $\mathcal{GWP}$ model failed to run on the HCP dataset due to its lack in scalability. Therefore, we compared our results with four parametric $\mathcal{DCC}$-GARCH models. Three of them employ a autoregression-moving-average model with order (1,1) for the mean but leverage different types of noise following multivariate Normal ($\mathcal{MVN}$), multivariate Student-t ($\mathcal{MVT}$) and multivariate Laplace distributions ($\mathcal{MVL}$). The last $\mathcal{DCC}$-GARCH model sets zero mean and has noise following multivariate Normal distributions ($\mathcal{MVN}0$).

Since as the Markov chain Monte Carlo yields less biased result than variational EM algorithm shown in Table 5.3, we conducted the Markov chain Monte Carlo inference and estimated model parameters using the maximum a posteriori. Given those estimates, we reconstructed covariance matrices on the observed timestamps.

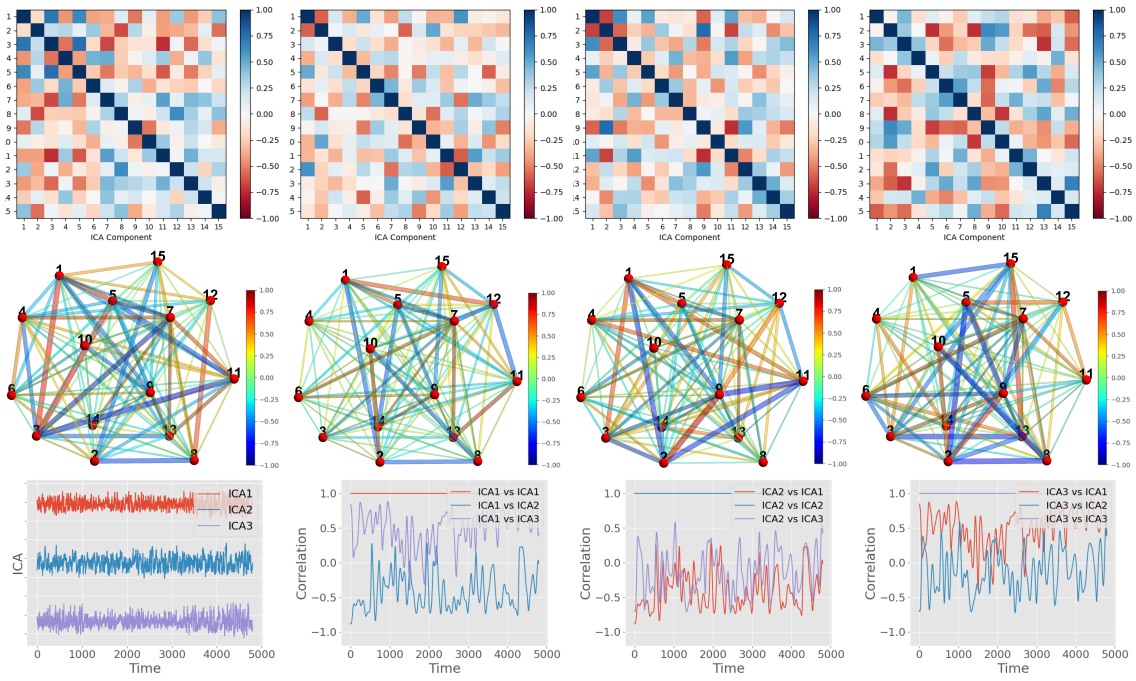


Figure 5.3: Dynamic correlations between ICA components (i.e., dynamic functional connectivity) and corresponding network representations derived from the estimations of $\Sigma(x)$ at $x = 1001, 2001, 3001, 4800$ with HCP timeseries data. Top row: connectivity matrices; Middle row: corresponding network representations (thicker edge represents larger absolute edge values and the colormap renders the value of the edge from low to high); Bottom row: three true ICA components and corresponding inferred dynamic correlation processes.

5.8.2 Individual Functional Connectivity Construction

We randomly selected one participant (ID: 990366) for the demonstration of individual dynamic functional connectivity derivation. The log-likelihood of observation (i.e., ICA) at each timestamp was computed and plotted as a boxplot for all observations in Figure 5.4. We also plotted the same boxplots of log-likelihoods estimated from $\mathcal{DCC}$ models. $\mathcal{PWP}$ and $\mathcal{DCCMVN}0$ assume zero mean, which makes them comparable. The figure shows that $\mathcal{PWP}$ performs relatively worse than $\mathcal{DCC}$ models in terms of the mean of log-likelihood, but it provides more stable results than $\mathcal{DCC}$ models in the sense of less extreme outliers and lower variance.

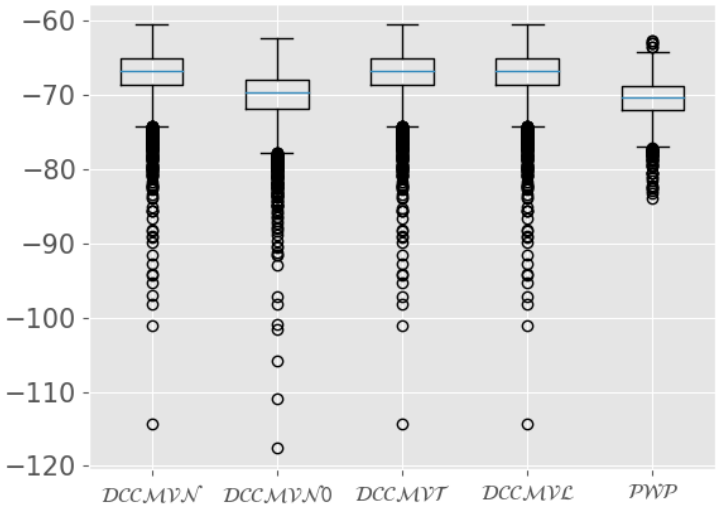


Figure 5.4: Boxplots of log-Likelihood w.r.t. the whole 4800 timestamps (i.e., time) with the reconstructed covariance matrix. $\mathcal{PWP}$ shows better stability than $\mathcal{DCC}$s with less extreme outliers and lower variance.

In Figure 5.3, we presented dynamic correlation matrices and the structural networks derived from the estimated $\Sigma(x)$ at timestamp $x = 1001, 2001, 3001, 4800$ to show the changes of their functional brain connectivity across time. This result proves the hypothesis in [23] that the structure of covariance along time in functional connectivity may be significant, and shows a significant potential that our $\mathcal{PWP}$ is a very powerful tool to visualize the estimate of covariance in time-varying data. Moreover, to directly illustrate the temporal relation, we provided the plot of three ICA components as well as their corresponding inferred correlation processes in Figure 5.3. It illustrates that the correlations between ICA components are not random and they have certain patterns.

5.8.3 Multi-task Learning on HCP Data

In order to show the applicability of our dynamic brain connectivity features, we compared the fitting performances of $\mathcal{PWPMD}$ against baseline features.

Here we considered a three-level multi-scale descriptor from (5.22) where the length scale parameter τ in the squared exponential covariance function is set to $500, 2000$ and 5000. We used the matrix from Cholesky decomposition of the sample covariance matrix as the baseline features for each subject as conducted in [115, 89, 116]. Then we conducted the linear regression between the features extracted from the rs-fMRI timeseries and exogenous variables.

We considered five behavioral scores available in the HCP dataset as exogenous variables: MMSE, PSQI, PainIntens (raw), PainInterf T-score and Mars Log Scores. Specifically, Mini Mental Status Exam (MMSE) [117, 118] is a broad measure of cognitive status, Pittsburgh Sleep Questionnaire (PSQI) [119] is a measure of sleep quality, Pain Intensity Raw Score (PainIntens) [120] consists of a single item measuring immediate (i.e., acute) pain in adults, Pain Interference T-score (PainInterf) [120] measures the degree to which pain interferes with other activities in life in adults, and Mars Contrast Sensitivity Test (Mars) [121, 122, 123] is a brief and reliable measure that assesses color contrast sensitivity.

The resulting R^2 scores from linear model fitting are reported in Table 5.4. It is apparent that the $\mathcal{PWPMD}$ achieves the best fitting performance across all five HCP behavioral measures. Notably, the $\mathcal{PWPMD}$ exhibits better performance by 39% when compared

Table 5.4: R^2 scores of linear model fitting with different features for different exogenous variables.

Feature	Linear Regression				
	MMSE	PSQI	PainIntens	PainInterf	Mars
Baseline features	0.21	0.19	0.16	0.18	0.19
$\mathcal{PWPMD}$	**0.48**	**0.50**	**0.45**	**0.48**	**0.58**

MMSE: Mini Mental Status Exam; PSQI: Pittsburgh Sleep Questionnaire; PainIntens: Pain Intensity Raw Score; PainInterf: Pain Interference T-score; Mars: Mars Contrast Sensitivity Test.

with the baseline feature on behavioral measurement Mars Log Score, and also outperforms the baseline by 27%, 31%, 29%, 30% on MMSE score, PSQI score, PainIntens raw score, PainInterf T-score, respectively. Our experiments illustrate that our proposed dynamic brain connectivity features $\mathcal{PWPMD}$ significantly improve the regression performance as compared with the baseline features. The promising results from these experiments on HCP dataset implicate a great potential for our $\mathcal{PWP}$ for multi-task learning in real-world clinical applications.

5.9 Conclusion

There is a significant interest in modeling time-varying changes of relationships between different variables in both theoretical and application-wise perspectives. As previous stochastic approaches heavily suffer from computational burden, we introduced a novel stochastic process, i.e., $\mathcal{PWP}$, which can model dynamic covariance matrices accurately and efficiently. Not only we provide theoretical guarantee that it is a well defined process, but also illustrate that it is easy to be incorporated into different models such as hierarchical Gaussian model and multi-task model. Moreover, we empirically evaluate our ideas and its usefulness with two independent sets of experiments. Especially for the real experiment on HCP data, features derived from dynamic functional connectivity can be useful for multi-task learning over traditional approaches extracting features from covariance matrices. We believe there is a significant potential that $\mathcal{PWP}$ can further utilized in various areas where time-varying associations between variables need to effectively characterized.

CHAPTER 6

Conclusion

In summary, the main goal of this thesis was to advance the state-of-the-art, develop and prototype machine learning models for neuroimaging applications. Building on the machine learning frameworks, we developed several approaches in this thesis that tackled the fundamental challenges which routinely arose in neuroimaging applications. We evaluated our proposed approaches using several typical yet challenging real-world neuroimaging data from the neuroscience domain, including Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), resting-state functional MRI (rs-fMRI), and Positron Emission Tomography (PET). The experimental results demonstrated that our proposed methods advanced the the state-of-the-art in machine learning for neuroimaging applications, which can provide effective and efficient solutions for problems on real-world neuroimaging data.

In the following paragraphs, we briefly summarize the key contributions of each individual method presented in this thesis:

Multi-resolutional Statistical Analysis on ABCD Data. In Chapter 2, we proposed a novel transform that utilizes the precision matrix for structured data to increase sensitivity in downstream statistical tasks. It can capture the local context information along the geometry of precision matrix and yield a multi-scale feature. We conducted statistical analysis and classification task based on household income using large-scale ABCD dataset and demonstrated significant quantitative improvements. Furthermore, we detected multiple ROIs whose microstructures are susceptible to socioeconomic disparity which were not identifiable with conventional approaches.

Disentangled Representation Learning on ADNI Data. In Chapter 3 and 4, we investigated the problem of disentangled representation learning for neuroimaging applications on ADNI dataset. First, in Chapter 3, we proposed a novel sequential autoencoder model which is flexible for data generation as well as conditionally generate sequential data based on label, disentangled time-varying and time-invariant latent variables. We quantitatively demonstrated that our model has competitive reconstruction and classification performance as compared to two modified versions of unsupervised state-of-the-art S3VAE models. Second, in Chapter 4, we developed a novel supervised sequential graph autoencoder model which learns a latent disentangled representation consisting of time-varying and time-invariant information to early characterize the longitudinal amyloid over the structural brain network. We demonstrated that our model not only can capture the robust dynamics of amyloid but also forecast future amyloid depositions from limited past time points.

Dynamic Covariance Modeling on HCP Data. In Chapter 5, we studied the problem of modeling time-varying changes of relationships between different variables from both theoretical and application-wise perspectives. We introduced a novel stochastic process which can model the dynamic covariance matrices to overcome the limitation of previous stochastic approaches that heavily suffer from the computational burden. We empirically evaluated our model on the real-world HCP dataset and exemplified that features derived from dynamic functional connectivity can be useful for multi-task learning over traditional approaches extracting features from static covariance matrices, which shows a significant potential that our model can be a very powerful tool to estimate the time-varying covariance for real-world clinical applications.

In a nutshell, this thesis advanced the state-of-the-art in the field of leveraging machine learning techniques for neuroimaging applications and highlighted the potential of applying machine learning methods on large-scale real-world neuroimaging data to improve decision-makings in the neuroscience domain. It is expected that the adoption of machine leaning

methods for applications on neuroimaging data may potentially continue to expand in the future, in which machine learning plays a critical role in improving and, ultimately, revolutionizing the decision-makings for the neuroimaging problems from the neuroscience domain.

APPENDIX A

Supplementary Materials for Chapter 2

In this appendix, we present the supplementary materials for Chapter 2.

A.1 Proof of Lemma 1

We provide here the proof of Lemma 1 as follows.

Lemma A.1.0.1. *T_g^s is a self-adjoint operator, i.e., $\langle T_g^s f, h\rangle = \langle f, T_g^s h\rangle$.*

Proof. For any signals $f, h \in H$, because of the uniqueness of basis representation, there exist coefficient $f_1, f_2 \ldots, f_P \in \mathbb{R}$ and $h_1, h_2, \ldots, h_P \in \mathbb{R}$ such that $f = \sum_{\ell=1}^{P} f_\ell \nu_\ell$ and $h = \sum_{\ell=1}^{P} h_\ell \nu_\ell$. Then we have

$$\begin{aligned} \langle T_g^s f, h\rangle &= \langle T_g^s(\sum_{\ell=1}^{P} f_\ell \nu_\ell), \sum_{\ell=1}^{P} h_\ell \nu_\ell\rangle \\ &= \sum_{\ell=1}^{P} g(s\lambda_\ell) f_\ell h_\ell \\ &= \langle \sum_{\ell=1}^{P} f_\ell \nu_\ell, T_g^s(\sum_{\ell=1}^{P} h_\ell \nu_\ell)\rangle \\ &= \langle f, T_g^s h\rangle. \end{aligned} \tag{A.1}$$

We hence complete the proof. □

A.2 Classification Performance

We summarized the values of accuracy, precision and recall across all folds for both 2-calss and 3-class cases in Table A.1. Figure A.1 shows the ROC curves for the binary case on both raw cortical FA and CMD. It can be observed that ROC curve on CMD more tightly approaches the top-left corner. The area under the curve (AUC) goes up by roughly 2% for the ROC curve on CMD as compared with that of the ROC curve for the raw FA measurements.

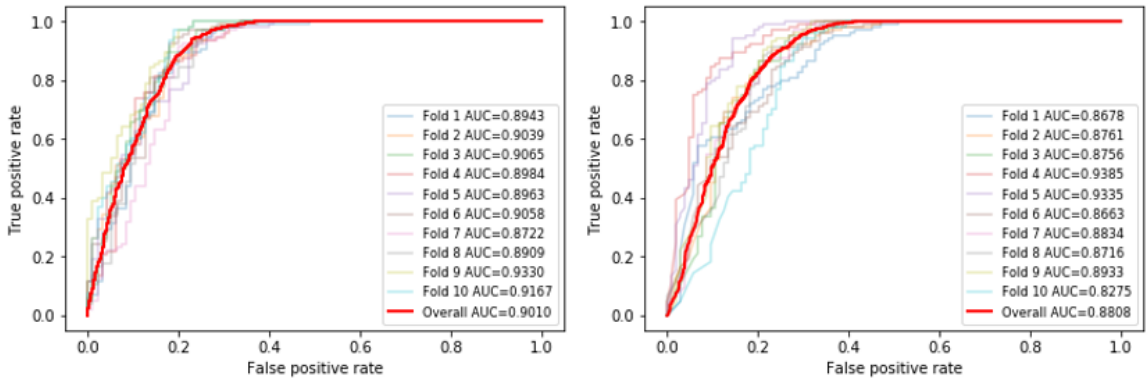


Figure A.1: ROC curves for 2-class case. Left: ROC on CMD, Right: ROC on raw FA measures.

Table A.1: Classification performance measurements across folds.

		2-Class			3-Class		
Measures	Folds	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Original FA	Fold 1	0.72	0.69	0.91	0.41	0.40	0.43
	Fold 2	0.81	0.80	0.91	0.41	0.41	0.43
	Fold 3	0.70	0.68	0.93	0.41	0.40	0.44
	Fold 4	0.81	0.80	0.90	0.39	0.38	0.40
	Fold 5	0.72	0.73	0.87	0.42	0.41	0.42
	Fold 6	0.79	0.78	0.89	0.41	0.41	0.43
	Fold 7	0.76	0.76	0.88	0.40	0.39	0.40
	Fold 8	0.76	0.78	0.89	0.40	0.40	0.43
	Fold 9	0.82	0.82	0.82	0.42	0.41	0.43
	Fold 10	0.79	0.77	0.86	0.41	0.40	0.42
CMD (COVLET)	Fold 1	0.85	0.81	0.91	0.48	0.47	0.47
	Fold 2	0.85	0.82	0.91	0.50	0.50	0.49
	Fold 3	0.86	0.82	0.93	0.48	0.47	0.47
	Fold 4	0.89	0.83	0.97	0.47	0.47	0.46
	Fold 5	0.88	0.86	0.91	0.47	0.47	0.46
	Fold 6	0.84	0.82	0.87	0.49	0.49	0.48
	Fold 7	0.86	0.84	0.88	0.48	0.48	0.47
	Fold 8	0.84	0.81	0.89	0.44	0.44	0.43
	Fold 9	0.85	0.81	0.92	0.46	0.46	0.45
	Fold 10	0.83	0.78	0.92	0.44	0.44	0.43

APPENDIX B

Supplementary Materials for Chapter 5

In this appendix, we present the supplementary materials for Chapter 5.

B.1 Theorem Proving

B.1.1 Proof of Theorem 1

Here we present the proofs of Theorem 1 as below.

Proof. In the construction of $\mathcal{PWP}$, $\{\tilde{u}_{vd}\}$ have independent predictive process priors. Therefore, we have

$$\tilde{\boldsymbol{u}}_v(x) = (\tilde{u}_{v1}(x), \ldots, \tilde{u}_{v\mathcal{D}}(x))^T \sim \mathcal{N}_{\mathcal{D}}(\mathbf{0}, B), \tag{B.1}$$

where B is the diagonal matrix with elements $b_d = \tilde{C}_d(x, x)$. Because of $C_d(x, x) = 1$ and the property (5.5), $b_d \leq 1$ for $d = 1, \ldots, \mathcal{D}$. According to the property of multivariate Gaussian distribution, it immediately follows that

$$L\tilde{\boldsymbol{u}}_v(x) \sim \mathcal{N}_{\mathcal{D}}(\mathbf{0}, S^*), \tag{B.2}$$

where $S^* = LBL^T$. Due to (B.3) and according to the definition of Wishart distribution, we have $\Sigma(x) \sim \mathcal{W}_{\mathcal{D}}(\mathcal{V}, S^*)$. Since $\mathcal{V} \geq \mathcal{D}$ in the construction, this Wishart distribution is well defined. □

B.1.2 Proof of Theorem 2

Here we present the proofs of Theorem 2 as below.

Proof. We denote the diagonal elements of L as $(l_1, \ldots, l_D)$, then according to

$$\begin{aligned} \Sigma(x) &= L\tilde{U}(x)\tilde{U}(x)^T L^T \\ &= \sum_{v=1}^{\mathcal{V}} L\tilde{\boldsymbol{u}}_v(x)\tilde{\boldsymbol{u}}_v^T(x)L^T, \end{aligned} \tag{B.3}$$

the $(i,j)^{\text{th}}$ element of the covariance $\Sigma(x)$ is given as

$$\Sigma_{ij}(x) = \sum_{v=1}^{V} l_i \tilde{u}_{vi} \tilde{u}_{vj} l_j \,. \tag{B.4}$$

According to (5.7), we let $\tilde{u}_{0d} \overset{iid}{\sim} \mathcal{PP}(0, \tilde{C}(x,x'))$, and then we have

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\sum_{v=1}^{\mathcal{V}} l_i l_j l_k l_l \operatorname{cov}(\tilde{u}_{vi}(x)\tilde{u}_{vj}(x), \tilde{u}_{vk}(x')\tilde{u}_{vl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \operatorname{cov}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x), \tilde{u}_{0k}(x')\tilde{u}_{0l}(x')).
\end{aligned} \tag{B.5}$$

Because of the symmetric property of covariance, let $s \neq t$, and we only need to consider three classes summarized as the following three cases:

1. $\operatorname{cov}(\Sigma_{ss}(x), \Sigma_{ss}(x'))$.
2. $\operatorname{cov}(\Sigma_{st}(x), \Sigma_{st}(x'))$ and $\operatorname{cov}(\Sigma_{st}(x), \Sigma_{ts}(x'))$.
3. Otherwise.

For the first case, without loss of generality, we assume $i = j = k = l$, then we rewrite (B.5) as

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}^2(x)\tilde{u}_{0i}^2(x')) - \mathrm{E}(\tilde{u}_{0i}^2(x))\mathrm{E}(\tilde{u}_{0i}^2(x'))\big) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\tilde{C}(x,x)\tilde{C}(x',x') + 2\tilde{C}^2(x,x') - \tilde{C}(x,x)\tilde{C}(x',x')\big) \\
=&2\mathcal{V} l_i^4 \tilde{C}^2(x,x').
\end{aligned} \tag{B.6}$$

In the second case, without loss of generality, we assume $i = k \neq j = l$, then we rewrite (B.5) as

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0i}(x'))\mathrm{E}(\tilde{u}_{0j}(x)\tilde{u}_{0j}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0i}(x')\tilde{u}_{0j}(x'))\big) \\
=&\mathcal{V} l_i^2 l_j^2 \tilde{C}^2(x, x'). \qquad \text{(B.7)}
\end{aligned}$$

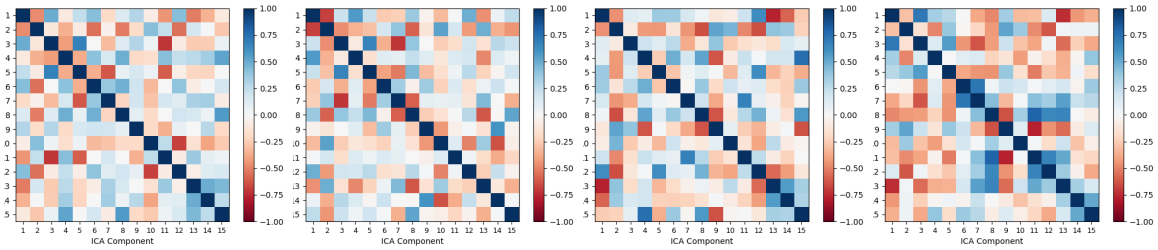
The third case includes two situations: (a) $i \neq j, k, l$, or (b) $i = j \neq k = l$. As for situation (a), (B.5) is rewritten as

$$\begin{aligned}
&\operatorname{cov}(\Sigma_{ij}(x), \Sigma_{kl}(x')) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x)\tilde{u}_{0k}(x')\tilde{u}_{0l}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x)\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0k}(x')\tilde{u}_{0l}(x'))\big) \\
=&\mathcal{V} l_i l_j l_k l_l \big(\mathrm{E}(\tilde{u}_{0i}(x))\mathrm{E}(\tilde{u}_{0j}(x)\tilde{u}_{0k}(x')\tilde{u}_{0l}(x')) \\
&- \mathrm{E}(\tilde{u}_{0i}(x))\mathrm{E}(\tilde{u}_{0j}(x))\mathrm{E}(\tilde{u}_{0k}(x')\tilde{u}_{0l}(x'))\big) = 0. \qquad \text{(B.8)}
\end{aligned}$$

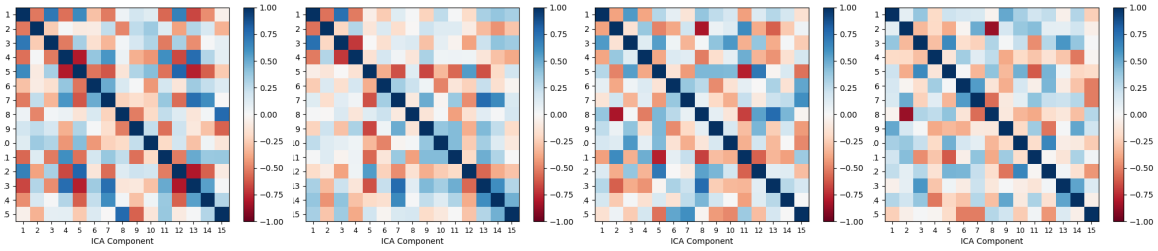
And it is trivial that situation (b) has the same result. □

B.2 Dynamic Correlation Matrices on More Participants

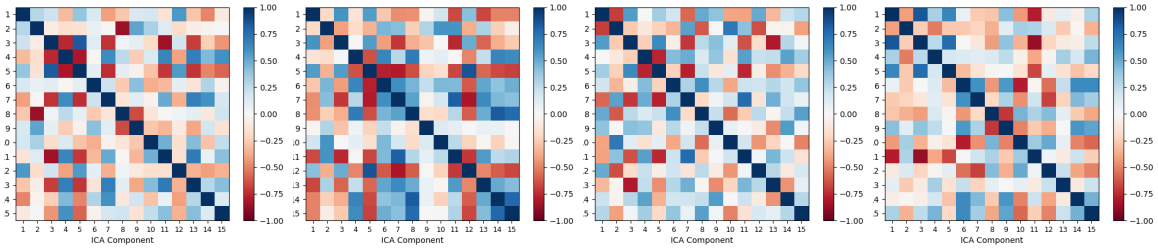
We also display the dynamic correlation matrices derived from the estimated $\Sigma(x)$ at timestamp $x = 1001, 2001, 3001$ and 4800 on more randomly selected participants with IDs 169946, 199958 and 668361 in Figure B.1 part (a), (b) and (c), respectively. These plots show the changes of brain connectivity across time as well and further provide evidences that the structure of covariance/correlation may be significantly time-varying.



(a) Dynamic correlations on participant with ID 169949



(b) Dynamic correlations on participant with ID 199958



(c) Dynamic correlations on participant with ID 668361

Figure B.1: Dynamic correlations (i.e., dynamic functional connectivity between ICA components) derived from the estimations of $\Sigma(x)$ at x =1001, 2001, 3001 and 4800 with HCP timeseries data.

REFERENCES

[1] C. G. Schwarz, "Uses of human mr and pet imaging in research of neurodegenerative brain diseases," *Neurotherapeutics*, vol. 18, no. 2, pp. 661–672, 2021.

[2] R. V. Marinescu, A. Eshaghi, D. C. Alexander, and P. Golland, "Brainpainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*. Springer, 2019, pp. 112–120.

[3] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

[4] D. Le Bihan, E. Breton, D. Lallemand, P. Grenier, and E. Cabanis, "Imagerie de self-diffusion in vivo par résonance magnétique nucléaire," *Innovation et technologie en biologie et médecine*, vol. 7, no. 6, pp. 713–720, 1986.

[5] S. Ogawa, T.-M. Lee, A. S. Nayak, and P. Glynn, "Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields," *Magnetic resonance in medicine*, vol. 14, no. 1, pp. 68–78, 1990.

[6] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *TPAMI*, vol. 11, no. 7, pp. 674–693, 1989.

[7] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Comp. Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[8] ——, "The spectral graph wavelet transform: Fundamental theory and fast computation," in *Vertex-Frequency Analysis of Graph Signals*. Springer, 2019, pp. 141–175.

[9] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Comp. Harmonic Analysis*, vol. 21, no. 1, pp. 53–94, 2006.

[10] P. M. Thompson, K. M. Hayashi, E. R. Sowell, N. Gogtay, J. N. Giedd, J. L. Rapoport, G. I. De Zubicaray, A. L. Janke, S. E. Rose, J. Semple, *et al.*, "Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia," *Neuroimage*, vol. 23, pp. S2–S18, 2004.

[11] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, *et al.*, "The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[12] R. Wolz, V. Julkunen, J. Koikkalainen, E. Niskanen, D. P. Zhang, D. Rueckert, H. Soininen, J. Lötjönen, A. D. N. Initiative, *et al.*, "Multi-method analysis of mri images in early diagnostics of Alzheimer's disease," *PloS one*, vol. 6, no. 10, p. e25446, 2011.

[13] Y. Cho, J.-K. Seong, Y. Jeong, S. Y. Shin, and ADNI, "Individual subject classification for alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data," *Neuroimage*, vol. 59, no. 3, pp. 2217–2230, 2012.

[14] W. H. Kim, A. M. Racine, N. Adluru, *et al.*, "Cerebrospinal fluid biomarkers of neurofibrillary tangles and synaptic dysfunction are associated with longitudinal decline in white matter connectivity: A multi-resolution graph analysis," *NeuroImage: Clinical*, vol. 21, p. 101586, 2019.

[15] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain mr images," in *MICCAI Brainlesion Workshop*. Springer, 2018, pp. 161–169.

[16] Q. Zhao, E. Adeli, N. Honnorat, T. Leng, and K. M. Pohl, "Variational autoencoder for regression: Application to brain aging analysis," in *MICCAI*. Springer, 2019, pp. 823–831.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2014.

[18] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *NeurIPS*, 2017, pp. 1878–1889.

[19] L. Yingzhen and S. Mandt, "Disentangled sequential autoencoder," in *ICML*, 2018, pp. 5670–5679.

[20] S. M. Smith, "The future of fmri connectivity," *Neuroimage*, vol. 62, no. 2, pp. 1257–1266, 2012.

[21] G. Varoquaux, A. Gramfort, J.-b. Poline, and B. Thirion, "Brain covariance selection: better individual functional connectivity models using population prior," in *NeurIPS*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010. [Online]. Available: https://proceedings.neurips.cc/paper/2010/file/db576a7d2453575f29eab4bac787b919-Paper.pdf

[22] B. Chai, D. B. Walther, D. M. Beck, and L. Fei-Fei, "Exploring functional connectivity of the human brain using multivariate information analysis," *NeurIPS*, vol. 22, pp. 270–278, 2009.

[23] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo, *et al.*, "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, 2013.

[24] R. Hindriks, M. H. Adhikari, Y. Murayama, M. Ganzetti, D. Mantini, N. K. Logothetis, and G. Deco, "Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI?" *Neuroimage*, vol. 127, pp. 242–256, 2016.

[25] S. D. Keilholz, "The neural basis of time-varying resting-state functional connectivity," *Brain connectivity*, vol. 4, no. 10, pp. 769–779, 2014.

[26] L. Li, D. Pluta, B. Shahbaba, N. Fortin, H. Ombao, and P. Baldi, "Modeling dynamic functional connectivity with latent factor gaussian processes," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/bf499a12e998d178afd964adf64a60cb-Paper.pdf

[27] E. B. Fox and M. West, "Autoregressive models for variance matrices: Stationary inverse wishart processes," *arXiv preprint arXiv:1107.5239*, 2011.

[28] A. G. Wilson and Z. Ghahramani, "Generalised wishart processes," *arXiv preprint arXiv:1101.0240*, 2010.

[29] F. Yang, A. Isaiah, and W. H. Kim, "Covlet: covariance-based wavelet-like transform for statistical analysis of brain characteristics in children," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 83–93.

[30] F. Yang, R. Meng, H. Cho, *et al.*, "Disentangled sequential graph autoencoder for preclinical Alzheimer's disease characterizations from adni study," in *MICCAI*. Springer, 2021, pp. 362–372.

[31] F. Yang, G. Wu, and W. H. Kim, "Disentangled representation of longitudinal b-amyloid for ad via sequential graph variational autoencoder with supervision," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[32] D. Fuhrmann, L. J. Knoll, and S.-J. Blakemore, "Adolescence as a sensitive period of brain development," *Trends in cognitive sciences*, vol. 19, no. 10, pp. 558–566, 2015.

[33] S.-J. Blakemore, "Imaging brain development: the adolescent brain," *Neuroimage*, vol. 61, no. 2, pp. 397–406, 2012.

[34] T. L. Jernigan, T. T. Brown, D. J. Hagler Jr, *et al.*, "The pediatric imaging, neurocognition, and genetics (ping) data repository," *Neuroimage*, vol. 124, pp. 1149–1154, 2016.

[35] K. Jednoróg, I. Altarelli, K. Monzalvo, J. Fluss, J. Dubois, C. Billard, G. Dehaene-Lambertz, and F. Ramus, "The influence of socioeconomic status on children's brain structure," *PloS one*, vol. 7, no. 8, 2012.

[36] C. Lebel, L. Walker, A. Leemans, *et al.*, "Microstructural maturation of the human brain from childhood to adulthood," *Neuroimage*, vol. 40, no. 3, pp. 1044–1055, 2008.

[37] P. DeRosse, T. Ikuta, K. H. Karlsgodt, *et al.*, "History of childhood maltreatment is associated with reduced fractional anisotropy of the accumbofrontal 'reward' tract in healthy adults," *Brain imaging and behavior*, pp. 1–9, 2020.

[38] C. Lebel and S. Deoni, "The development of brain white matter microstructure," *Neuroimage*, vol. 182, pp. 207–218, 2018.

[39] N. D. Volkow, G. F. Koob, R. T. Croyle, D. W. Bianchi, J. A. Gordon, W. J. Koroshetz, E. J. Pérez-Stable, W. T. Riley, M. H. Bloch, K. Conway, *et al.*, "The conception of the abcd study: From substance use to a broad nih collaboration," *Developmental cognitive neuroscience*, vol. 32, pp. 4–7, 2018.

[40] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, vol. 2. Ieee, 1999, pp. 1150–1157.

[41] Z. Cai, Q. Fan, R. S. Feris, *et al.*, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*. Springer, 2016, pp. 354–370.

[42] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*, 2017, pp. 624–632.

[43] W. H. Kim, D. Pachauri, C. Hatt, *et al.*, "Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination," in *NeurIPS*, 2012, pp. 1241–1249.

[44] W. H. Kim, V. Singh, M. K. Chung, *et al.*, "Multi-resolutional shape features via non-euclidean wavelets: Applications to statistical analysis of cortical thickness," *NeuroImage*, vol. 93, pp. 107–123, 2014.

[45] E. Bullmore, J. Fadili, M. Breakspear, R. Salvador, J. Suckling, and M. Brammer, "Wavelets and statistical analysis of functional magnetic resonance images of the human brain," *Statistical methods in medical research*, vol. 12, no. 5, pp. 375–399, 2003.

[46] W. H. Kim, H. J. Kim, N. Adluru, and V. Singh, "Latent variable graphical model selection using harmonic analysis: applications to the human connectome project (HCP)," in *CVPR*, 2016, pp. 2443–2451.

[47] A. Lee, "Us poverty thresholds and poverty guidelines: What's the difference," *Population Reference Bureau.(2019)*, 2018.

[48] A. T. Marshall, S. Betts, E. C. Kan, *et al.*, "Association of lead-exposure risk and family income with childhood brain outcomes," *Nature Medicine*, vol. 26, no. 1, pp. 91–97, 2020.

[49] C. Destrieux, B. Fischl, A. Dale, *et al.*, "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature," *Neuroimage*, vol. 53, no. 1, pp. 1–15, 2010.

[50] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society*, vol. 67, no. 2, pp. 301–320, 2005.

[51] I. Mani and I. Zhang, "k-NN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, 2003.

[52] K. G. Noble, S. M. Houston, N. H. Brito, *et al.*, "Family income, parental education and brain structure in children and adolescents," *Nature neuroscience*, vol. 18, no. 5, p. 773, 2015.

[53] P. J. Gianaros, A. L. Marsland, L. K. Sheu, K. I. Erickson, and T. D. Verstynen, "Inflammatory pathways link socioeconomic inequalities to white matter architecture," *Cerebral cortex*, vol. 23, no. 9, pp. 2058–2071, 2013.

[54] M. J. Farah, "The neuroscience of socioeconomic status: Correlates, causes, and consequences," *Neuron*, vol. 96, no. 1, pp. 56–71, 2017.

[55] J. M. Fuster, "Frontal lobe and cognitive development," *Journal of neurocytology*, vol. 31, no. 3-5, pp. 373–385, 2002.

[56] E. R. Sowell, D. Delis, J. Stiles, and T. L. Jernigan, "Improved memory functioning and frontal lobe maturation between childhood and adolescence: a structural mri study," *Journal of the International Neuropsychological Society*, vol. 7, no. 3, pp. 312–322, 2001.

[57] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[58] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[59] H. Kim and A. Mnih, "Disentangling by factorising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.

[60] R. Meng and K. Bouchard, "Bayesian inference in high-dimensional time-serieswith the orthogonal stochastic linear mixing model," *arXiv preprint arXiv:2106.13379*, 2021.

[61] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *arXiv preprint arXiv:1803.02991*, 2018.

[62] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3vae: Self-supervised sequential vae for representation disentanglement and data generation," in *CVPR*, 2020, pp. 6538–6547.

[63] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk, "Representation Disentanglement for Multi-modal MR Analysis," *arXiv e-prints*, p. arXiv:2102.11456, Feb. 2021.

[64] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, pp. 3844–3852, 2016.

[65] X. Ma, G. Wu, S. J. Hwang, and W. H. Kim, "Learning multi-resolution graph edge embedding for discovering brain network dysfunction in neurological disorders," in *IPMI*. Springer, 2021, pp. 253–266.

[66] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[67] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *AAAI*, vol. 32, no. 1, 2018.

[68] M. Saerens, F. Fouss, L. Yen, and P. Dupont, "The principal components analysis of a graph, and its relationships to spectral clustering," in *European conference on machine learning*. Springer, 2004, pp. 371–383.

[69] J. Chen, T. Ma, and C. Xiao, "Fastgcn: fast learning with graph convolutional networks via importance sampling," *arXiv preprint arXiv:1801.10247*, 2018.

[70] B. Xu, H. Shen, Q. Cao, Y. Qiu, and X. Cheng, "Graph wavelet neural network," in *International Conference on Learning Representations*, 2019.

[71] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[72] S. N, B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[74] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*. ACL, 2014, pp. 1724–1734.

[75] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in beta-VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[76] M. S. Henry, A. P. Passmore, *et al.*, "The development of effective biomarkers for alzheimer's disease: a review," *International journal of geriatric psychiatry*, vol. 28, no. 4, pp. 331–340, 2013.

[77] R. J. Bateman, C. Xiong, *et al.*, "Clinical and biomarker changes in dominantly inherited alzheimer's disease," *N Engl J Med*, vol. 367, pp. 795–804, 2012.

[78] C. R. Jack Jr, D. A. Bennett, *et al.*, "Nia-aa research framework: toward a biological definition of alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[79] J. Therriault, A. L. Benedet, *et al.*, "Determining amyloid-β positivity using 18f-azd4694 pet imaging," *Journal of Nuclear Medicine*, vol. 62, no. 2, pp. 247–252, 2021.

[80] J. W. Vogel, Y. Iturria-Medina, *et al.*, "Spread of pathological tau proteins through communicating neurons in human alzheimer's disease," *Nature communications*, vol. 11, no. 1, pp. 1–15, 2020.

[81] A. Raj, A. Kuceyeski, and M. Weiner, "A network diffusion model of disease progression in dementia," *Neuron*, vol. 73, no. 6, pp. 1204–1215, 2012.

[82] S. J. Hwang *et al.*, "Associations between positron emission tomography amyloid pathology and diffusion tensor imaging brain connectivity in pre-clinical alzheimer's disease," *Brain connectivity*, vol. 9, no. 2, pp. 162–173, 2019.

[83] J. Ouyang, E. Adeli, *et al.*, "Representation disentanglement for multi-modal mr analysis," *arXiv preprint arXiv:2102.11456*, 2021.

[84] C. Qin, B. Shi, *et al.*, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *IPMI*. Springer, 2019, pp. 249–261.

[85] J. Yang, X. Li, *et al.*, "Cross-modality segmentation by self-supervised semantic alignment in disentangled content space," in *DART and DCL*. Springer, 2020, pp. 52–61.

[86] M. I. Jordan, Z. Ghahramani, *et al.*, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[87] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar MRI," *Magnetic resonance in medicine*, vol. 34, no. 4, pp. 537–541, 1995.

[88] M. Greicius, "Resting-state functional connectivity in neuropsychiatric disorders," *Current opinion in neurology*, vol. 21, no. 4, pp. 424–430, 2008.

[89] B. B. Biswal, "Resting state fMRI: a personal history," *Neuroimage*, vol. 62, no. 2, pp. 938–944, 2012.

[90] M. Dai, Z. Zhang, and A. Srivastava, "Testing stationarity of brain functional connectivity using change-point detection in fMRI data," in *CVPR Workshop*, 2016, pp. 19–27.

[91] C. Seiler and S. Holmes, "Multivariate heteroscedasticity models for functional brain connectivity," *Frontiers in neuroscience*, vol. 11, p. 696, 2017.

[92] Y. Zhu, X. Zhu, M. Kim, D. Kaufer, P. J. Laurienti, and G. Wu, "Characterizing dynamic functional connectivity using data-driven approaches and its application in the diagnosis of Alzheimer's disease," in *Connectomics*. Elsevier, 2019, pp. 181–197.

[93] L. Cappiello, R. F. Engle, and K. Sheppard, "Asymmetric dynamics in the correlations of global equity and bond returns," *Journal of Financial econometrics*, vol. 4, no. 4, pp. 537–572, 2006.

[94] A. E. Gelfand, S. Banerjee, and D. Gamerman, "Spatial process modelling for univariate and multivariate dynamic spatial data," *Environmetrics: The official journal of the International Environmetrics Society*, vol. 16, no. 5, pp. 465–479, 2005.

[95] E. B. Fox and D. B. Dunson, "Bayesian nonparametric covariance regression," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2501–2542, 2015.

[96] R. Engle, "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business & Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002.

[97] H. WU-Minn, "1200 subjects data release reference manual," *URL https://www. humanconnectome. org*, 2017.

[98] M. Pourahmadi, "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation," *Biometrika*, vol. 86, no. 3, pp. 677–690, 1999.

[99] W. Zhang and C. Leng, "A moving average cholesky factor model in covariance modelling for longitudinal data," *Biometrika*, vol. 99, no. 1, pp. 141–150, 2012.

[100] J. Yin, Z. Geng, R. Li, and H. Wang, "Nonparametric covariance model," *Statistica Sinica*, vol. 20, p. 469, 2010.

[101] R. F. Engle and K. F. Kroner, "Multivariate simultaneous generalized arch," *Econometric theory*, vol. 11, no. 1, pp. 122–150, 1995.

[102] R. F. Engle and K. Sheppard, "Theoretical and empirical properties of dynamic conditional correlation multivariate garch," National Bureau of Economic Research, Tech. Rep., 2001.

[103] S. Chib, F. Nardari, and N. Shephard, "Analysis of high dimensional multivariate stochastic volatility models," *Journal of Econometrics*, vol. 134, no. 2, pp. 341–371, 2006.

[104] G. Kastner, S. Frühwirth-Schnatter, and H. F. Lopes, "Efficient bayesian inference for multivariate factor stochastic volatility models," *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 905–917, 2017.

[105] C. Gouriéroux, J. Jasiak, and R. Sufana, "The wishart autoregressive process of multivariate stochastic volatility," *Journal of Econometrics*, vol. 150, no. 2, pp. 167–181, 2009.

[106] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang, "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 4, pp. 825–848, 2008.

[107] A. O. Finley, H. Sang, S. Banerjee, and A. E. Gelfand, "Improving the performance of predictive process modeling for large datasets," *Computational statistics & data analysis*, vol. 53, no. 8, pp. 2873–2884, 2009.

[108] M. F. Bru, "Wishart processes," *Journal of Theoretical Probability*, vol. 4, no. 4, pp. 725–751, 1991.

[109] K. Zhang, I. W. Tsang, and J. T. Kwok, "Improved nyström low-rank approximation and error analysis," in *ICML*, 2008, pp. 1232–1239.

[110] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE TPAMI*, no. 6, pp. 721–741, 1984.

[111] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[112] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[113] E. Orskaug, "Multivariate DCC-GARCH model:-with various error distributions," Master's thesis, Institutt for matematiske fag, 2009.

[114] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, *et al.*, "Resting-state fMRI in the human connectome project," *Neuroimage*, vol. 80, pp. 144–168, 2013.

[115] M. P. Van Den Heuvel and H. E. H. Pol, "Exploring the brain network: a review on resting-state fMRI functional connectivity," *European neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[116] N. Leonardi, J. Richiardi, M. Gschwind, S. Simioni, J.-M. Annoni, M. Schluep, P. Vuilleumier, and D. Van De Ville, "Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest," *NeuroImage*, vol. 83, pp. 937–950, 2013.

[117] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[118] R. M. Crum, J. C. Anthony, S. S. Bassett, and M. F. Folstein, "Population-based norms for the mini-mental state examination by age and educational level," *JAMA*, vol. 269, no. 18, pp. 2386–2391, 1993.

[119] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The pittsburgh sleep quality index: a new instrument for psychiatric practice and research," *Psychiatry research*, vol. 28, no. 2, pp. 193–213, 1989.

[120] R. C. Gershon, M. V. Wagster, H. C. Hendrie, N. A. Fox, K. F. Cook, and C. J. Nowinski, "NIH toolbox for assessment of neurological and behavioral function," *Neurology*, vol. 80, no. 11, pp. S2–S6, 2013.

[121] A. Arditi, "Improving the design of the letter contrast sensitivity test," *Investigative ophthalmology & visual science*, vol. 46, no. 6, pp. 2225–2229, 2005.

[122] B. E. Dougherty, R. E. FLOM, and M. A. Bullimore, "An evaluation of the mars letter contrast sensitivity test," *Optometry and Vision Science*, vol. 82, no. 11, pp. 970–975, 2005.

[123] S. A. Haymes, K. F. Roberts, A. F. Cruess, M. T. Nicolela, R. P. LeBlanc, M. S. Ramsey, B. C. Chauhan, and P. H. Artes, "The letter contrast sensitivity test: clinical evaluation of a new design," *Investigative ophthalmology & visual science*, vol. 47, no. 6, pp. 2739–2745, 2006.