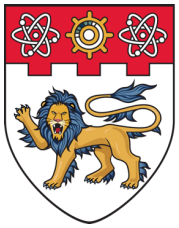
Statistical Modeling and Inference in Generalized Graph Signal Processing



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Jian Xingchao

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

10 Aug 2024
.....
Date


.....
Jian Xingchao

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

10 Aug 2024

Date

Prof. Tay Wee Peng

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed journals and 1 paper submitted as preprint in which I am listed as an author.

Chapter 3 is published as X. Jian and W. P. Tay, "Wide-sense stationarity in generalized graph signal processing," IEEE Trans. Signal Process., vol. 70, pp. 3414–3428, 2022.

The contributions of the co-authors are as follows:

- I proposed the problem formulation and solution, derived the theoretical results, and prepared the manuscript drafts.
- Prof. Tay Wee Peng proposed the initial problem, discussed the theoretical results with me, and revised the manuscript.

Chapter 4 is published as X. Jian, W. P. Tay, and Y. C. Eldar, "Kernel based reconstruction for generalized graph signal processing," IEEE Trans. Signal Process., vol. 72, pp. 2308–2322, Apr. 2024.

The contributions of the co-authors are as follows:

- I proposed the problem formulation and solution, derived the theoretical results, and prepared the manuscript drafts.
- Prof. Tay Wee Peng discussed the theoretical results with me, and revised the manuscript.
- Prof. Yonina C. Eldar reviewed the manuscript.

Chapter 5 is submitted as X. Jian, M. Gölz, F. Ji, W. P. Tay, and A. M. Zoubir, "A Graph Signal Processing Perspective of Network Multiple Hypothesis Testing with False Discovery Rate Control," arXiv preprint arXiv:2408.03142, 2024.

The contributions of the co-authors are as follows:

- I proposed the problem formulation and solution, derived the theoretical results, and prepared the manuscript drafts.
- Prof. Tay Wee Peng and Dr. Ji Feng discussed the theoretical results with me and revised the manuscript.
- This work is inspired by Martin Gölz and Prof. Abdelhak M. Zoubir's work. Martin Gölz discussed the experimental setting with me.

10 Aug 2024

Date

Jian Xingchao

Acknowledgements

This thesis was completed with the help of many people to whom I owe my deepest gratitude. First and foremost, I would like to express my greatest gratitude to my supervisor, Prof. Tay Wee Peng. He recognized my potential for research even though I had not conducted systematic academic research before joining NTU. During my PhD period, he patiently taught me how to conduct solid, useful and professional research, and guided me step by step in experiments and paper writing. I have also learned from him how to collaborate with other scholars.

I am grateful to the Ministry of Education (MOE), which provided the project funding and scholarship for my research. The research experience in NTU strengthened my belief that there is nothing mysterious about the world of technology, and I am capable of making my own novel contributions to existing work.

I have received a lot of help from other researchers, especially our group members. This thesis would not have been finished without their kind assistance. The foundation of this thesis is largely based on Dr. Ji Feng's work in 2019. I really appreciate his pioneer work and deep insights. I also benefit a lot from the discussions with Dr. Wang Chongxiao, who often shared interesting and novel ideas in machine learning with me. Prof. Liu Rui from Nankai University also provided invaluable help by explaining key concepts and problems I met in functional analysis. Besides theoretical guidance, I am thankful to Mr. Zhao Kai, who very patiently taught me how to use PyCharm and work remotely on a workstation. Furthermore, I would like to extend my thanks to Dr. Song Yang, Dr. Zhang Wei, Dr. Wang Chengcheng, Dr. Geng Tianyu, Dr. Xu Yi, Dr. She Rui, Dr. Kang Qiyu, Mr. Wang Sijie, Dr. Luo Wenhao, Dr. Li Xiang, Dr. Xie Yihang and Ms. Zhao Yanan, among many others. Life during these years would have been boring and lonely without their companionship and friendship.

I owe a profound debt of gratitude to my family, who have made immense efforts to raise and educate me. During the pandemic years, I could not return home due

to the uncertain situations and my limited vacation time. It was a great regret that for about three years I could not see my family in person. My grandparents passed away during that time, and I learned of their passing only over the phone. Their kindness and bravery have continually encouraged me to face challenges, undertake duties, and help people around me.

Finally, I would like to express my deep and sincere gratitude to Ms. Chen Qi. She has always encouraged me to confront the ideas, thoughts, desires, and inner demons within myself. Time with her is a brand new page filled with flowers, sunshine and music.

"Algebra is like sheet of music. The important thing isn't can you read music, it's can you hear it. Can you hear the music, Robert?"

—Niels Bohr to J. Robert Oppenheimer, Oppenheimer (2023)

To my dear family

Abstract

In modern signal processing, one of the greatest challenges is the more and more complex domain structure of the signal. One typical structure is a graph, or network, where vertices represent entities and edges denote the relationships between them. For signals supported on the graph structures, the filtering, sampling and reconstruction techniques require proper use of the graph. Graph signal processing (GSP) has emerged as a powerful tool for addressing these problems by leveraging the domain knowledge of graphs. It extends the traditional signal processing domain from discrete-time stamps to vertices in a network. GSP has found wide applications in various domains, including point cloud denoising, image processing, and brain network analysis, etc.

Traditional GSP regards the signal on each vertex as a scalar. In this thesis, apart from the graph structure, we consider an additional domain which is a measure space. This space is the domain of the signal observation on each vertex of the graph. For example, it can be a subset of a Euclidean space, or the index set of a feature vector. In the former case, when the subset is an interval, it represents the scenario where each vertex observes a signal on this interval. By leveraging the product structure of the graph and measure space, signal processing techniques are developed for more general signals over the graph. This class of signal is called generalized graph signal, and we refer to this framework as generalized graph signal processing (GGSP). We refer to the product of the graph and the measure space as joint domain. In this thesis, we develop systematic statistical signal processing techniques for generalized graph signal in three parts: modeling, estimation and hypothesis testing.

Firstly, in order to analyze random generalized graph signal, we model it as random element in a Hilbert space of functions, called graph random process (GRP). We provide conditions for a stochastic process on a graph to be a GRP. We introduce the notion of joint wide-sense stationarity (JWSS) for GRP, which allows us to characterize a stochastic process on graph as a combination of uncorrelated oscillation modes across the joint domain. We elucidate the relationship between the

notions of wide-sense stationarity in different domains, and derive the Wiener filters for denoising and signal completion under this framework. Provided a training set, the power spectral density can be estimated, hence the Wiener filter can be applied to the test dataset.

Secondly, we consider the estimation (reconstruction) of generalized graph signal by means of the reproducing kernel approach. We formulate GGSP signal reconstruction as a kernel ridge regression (KRR) problem. By devising an appropriate kernel, we show that this problem has a solution that can be evaluated in a distributed way. We interpret the problem and solution using both deterministic and Bayesian perspectives and link them to existing graph signal processing and GGSP frameworks. From the Bayesian perspective, our approach can be seen as the maximum a posteriori estimator with a JWSS Gaussian prior. We then provide an online implementation via random Fourier features. Under the Bayesian framework, we investigate the statistical performance under the asymptotic sampling scheme. Compared to the aforementioned Wiener filter, this approach does not require a noiseless and incomplete training set for estimation of power spectral density.

Finally, we consider a multiple hypothesis testing problem in a network over the joint domain. We assume a hypothesis test and an associated p-value for every sample point in the joint domain. Our goal is to determine which points have true alternative hypotheses. By parameterizing the unknown alternative distribution of p-values and the prior probabilities of null hypotheses with a bandlimited generalized graph signal, we obtain consistent estimates for them. Consequently, we also obtain an estimate of the local false discovery rates (lfdr). We prove that by using a step-up procedure on the estimated lfdr, we can achieve asymptotic false discovery rate control at a pre-determined level.

Contents

List of Figures

Symbols and Acronyms

Symbols

In this thesis, plain lower cases like a denote scalars and scalar-valued functions. Bold lower cases like $\mathbf{a}$ denote vectors and vector-valued functions. Bold upper cases like $\mathbf{A}$ denote matrices and operators.

G	graph structure
$\mathcal{V}$	vertex set of the graph
$\mathcal{E}$	edge set of the graph
$\mathbf{L}_G$	Laplacian matrix of graph G
$\mathbf{W}_G$	adjacency matrix of graph G
$\mathbf{A}_G$	graph shift operator of graph G
$\mathbf{A}_\mathcal{H}$	shift operator for Hilbert space $\mathcal{H}$
$\mathbf{A}_\mathcal{J}$	shift operator in joint domain
$\mathfrak{F}$	joint Fourier transform
λ_i	the i-th eigenvalue of the graph shift operator $\mathbf{A}_G$
$\boldsymbol{\phi}_i$	the i-th graph Fourier basis
ψ_i	the i-th basis of space $L^2(\mathcal{T})$
$\mathcal{N}_d(v)$	the d-hop neighborhood of the vertex v
$\odot$	the Hadamard (component-wise) product
$\otimes$	the tensor product
$\langle \cdot, \cdot \rangle$	inner product
$\circ$	the composition of functions
∇	the gradient operator
$\nabla \cdot$	the divergence operator
$\mathbb{I}_\mathcal{X}$	the indicator function of the set $\mathcal{X}$
$\lvert\mathcal{X}\rvert$	the cardinality of the set $\mathcal{X}$

$\mathbf{\Pi}_{\mathcal{X}}$	the projection operator on the space $\mathcal{X}$
$\mathbf{\Pi}_{\mathbf{x}}$	the projection operator on the one-dimensional space span$\{\mathbf{x}\}$
$\mathbf{1}$	all-ones column vector with proper dimension
$\mathfrak{i}$	the imaginary unit
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbb{P}$	probability measure
$\mathbb{E}$	expectation
cov	covariance
$\mathbf{e}_i$	the i-th standard 0-1 basis vector, all entries but the i-th are zero
$\mathbf{I}$	the identity operator

Acronyms

GSO	Graph Shift Operator
GSP	Graph Signal Processing
GGSP	Generalized Graph Signal Processing
GFT	Graph Fourier Transform
JFT	Joint Fourier Transform
WSS	Wide-sense Stationary
JWSS	Joint Wide-Sense Stationary
VWSS	Vertex Wide-Sense Stationary
HWSS	Hilbert Space Wide-Sense Stationary
PSD	Power Spectral Density
JPSD	Joint Power Spectral Density
MSE	Mean-Square Error
DFT	Discrete Time Fourier Transform
BLUE	Best Linear Unbiased Estimator
AWGN	Additive White Gaussian Noise
LCE	Linear Conditional Expectation
ALCC	Average Linear Conditional Operator
GRP	Graph Random Process
MAP	Maximum A Posteriori
GAE	Graph Auto-Encoder
GCN	Graph Convolutional Neural network

GNN	Graph Neural Network
RKHS	Reproducing Kernel Hilbert Space
MKL	Multi-Kernel Learning
KRR	Kernel Ridge Regression
GP	Gaussian Process
GPG	Gaussian Process over a graph
RBF	Radial Basis Function
SGD	Stochastic Gradient Descent
RFF	Random Fourier Feature
GTRSS	Graph Signal Reconstruction via Sobolev Smoothness
GRIN	Graph Recurrent Imputation Network
MHT	Multiple Hypothesis Testing
MLE	Maximum Likelihood Estimator
lfdr	local false discovery rate
FDR	False Discovery Rate
BH	Benjamini-Hochberg
pdf	probability density function
cdf	cumulative distribution function
a.s.	almost surely
r.c.d.	regular conditional distribution
PCA	Principal Component Analysis
DCT	Dominated Convergence Theorem
BIC	Bayesian Information Criterion
GLM	Generalized Linear Model
KKT	Karush–Kuhn–Tucker
WLLN	Weak Law of Large Numbers

Chapter 1

Introduction

1.1 Background

In real-world signal processing, the signal of interest is usually associated with a network. One of the most prominent examples is the data records from sensor networks, which are usually high-dimensional and have relationship with the network structure. Graph signal processing (GSP) techniques have been proposed to analyze this class of signals by accommodating the network structure [1–4]. By extending the core concepts like shift operator, frequency and Fourier transform from discrete time domain to graph, a new set of signal processing techniques such as filtering [5–7], sampling [8, 9], representation [10], and reconstruction [11, 12] have been developed. Besides its theoretical success, GSP has been widely applied in brain signal analysis [13–15], image or point cloud processing [16, 17], and recommendation systems [18].

Traditional GSP makes use of information in the graph or vertex domain rather than the time domain. To leverage information in both the graph and discrete-time domains, a time-vertex GSP theory [19, 20] was introduced. It includes the concepts of joint shift operator, joint Fourier transform (JFT), joint wide-sense stationary (JWSS), joint power spectral density (JPSD), and methods like sampling strategies [21]. The time-vertex approach is more flexible and powerful than traditional GSP since it exploits information from both the vertex and time domains.

In [22], the time-vertex framework is further generalized to include vertex signals from a possibly infinite dimensional, separable Hilbert space, which further developed

the concepts of JFT, filtering and sampling over the joint signal space constructed by the graph and Hilbert space. This framework encompasses a broad range of vertex signals including multichannel signals and continuous-time signals. In the sequel, we refer to this as the GGSP framework. One of the most important features of GGSP is that, since this framework enables analyzing general functional signals on every vertex, it allows for asynchronous sampling on every vertex, and also enables convergence analysis when the number of samples tends to infinity. In contrast, the time-vertex framework assumes the time stamps on every vertex are the same and fixed.

In this thesis, we develop a statistical model and inference methods for GGSP. We derive the conditions under which stochastic processes on graphs can be modeled as random elements in Hilbert spaces, and we define their moments. We call these random elements as graph random process (GRP). We focus on a class of GRPs whose second moment has an exchangeable relationship with the joint shift operator, and we define them as JWSS. We define their energy on each frequency component as JPSD. These concepts form the basic statistical model for our analysis of stochastic generalized graph signals. We explore the inference problem from two aspects: estimation and testing. For estimation, we study the signal reconstruction problem with reproducing kernel. From a deterministic view, it can be seen as a function reconstruction approach. From the statistical view, our approach can be seen as a Bayesian estimation of generalized graph signals with a JWSS prior distribution. For testing, we study the MHT problem in the joint vertex-measure space domain. MHT means that we need to make decisions on multiple hypotheses simultaneously. Using generalized graph signal as a parameter of p-value distribution, we develop a method with asymptotic FDR control.

1.2 Motivation

In this section, we highlight the motivations of statistical model and methods for GGSP in this thesis.

1.2.1 Joint Wide-sense Stationarity Model

The GGSP framework by [22] proposed to reconstruct the generalized graph signal via the assumption that the original signal is bandlimited [22, Section VI.A]. In practice, to determine the bandwidth, we usually need a training set to estimate it. However, given a training set, we can utilize much more information beyond just the bandwidth. Recall that in discrete-time signal processing, we can estimate the power spectral density (PSD) of a wide-sense stationary signal. The PSD provides prior information of Fourier coefficients, allowing for more accurate estimation of the target signal. In the oracle situation where the ground truth of PSD is known, applying Wiener filter yields the best linear estimator with the smallest mean-square error (MSE).

This theoretical line has been studied in GSP literature [23–25] and time-vertex GSP [26, 27], but not in GGSP. Since the time-vertex GSP can be regarded as a product graph, the notions of stationarity and power spectral density can be simply understood as the corresponding notions on the product graph. However, due to the infinite-dimensional nature of general Hilbert spaces, the definitions of wide-sense stationarity and power spectral density are more technical and intrinsically different from the aforementioned literature. Besides, from an application perspective, the time-vertex JWSS only considers discrete-time signals on each vertex, whereas the GGSP allows for more general signals, such as multichannel signals, on each vertex. Therefore, considering JWSS in GGSP can lead to more flexible and universal approaches for statistical GSP.

In light of this motivation, we fill this research gap by introducing statistical elements to the GGSP framework. We first introduce the notion of GRP, which is the generalized concept of random vector in Euclidean space to infinite-dimensional space. Under mild conditions, stochastic generalized graph signals are GRP. Second, we define JWSS as a special class of GRP whose second moment commutes with the joint shift operator. Based on these concepts, we extend the definition of JPSD and Wiener filters from traditional GSP to JWSS signals. Using this generalized Wiener filter, we aim to reconstruct not only discrete-time signals over graphs but also multidimensional feature vectors over graphs.

1.2.2 Kernel Based Reconstruction

As mentioned in Section 1.2.1, the Wiener filter for GGSP [28] relies on knowledge of the signal's JPSD, which can be derived from a noiseless training set without missing values. However, this may not be practical. Besides, although the Wiener filter can theoretically reconstruct unbandlimited signals, in practice, we are limited to using a bandlimited approximation. Thus, we consider the more challenging case where the training set is small, noisy and incomplete, and the target signal is not bandlimited, hence the method in [28] cannot be applied. To address this more challenging problem, recall that in GGSP, the signal on each vertex is a scalar-valued function[1]. Also recall that, for a general function reconstruction problem, by utilizing an appropriate kernel, we are able to reconstruct the signal with good fidelity as long as the target signal is in the corresponding reproducing kernel Hilbert space (RKHS), which can be infinite-dimensional.

In view of this motivation, we propose a kernel based method to reconstruct functions supported on the joint domain which is a product space of the vertex set and a measure space. The core of this method is to design a proper kernel. We propose to design this kernel as a product kernel of a graph-based kernel and a kernel in the measure space. Compared to the method in [22], our proposed method is able to utilize infinitely many features. This makes it more flexible and has better signal representation capability.

To illustrate our motivation, consider the Intel lab temperature dataset[2] which consists of temperature records from 54 sensors in a lab, collected between February and April 2004. The ground truth records and incomplete noisy observations on two connected sensors labeled as vertex 1 and 2 are shown in Fig. 1.1. Our goal is to reconstruct the signal at vertex 1. In the time interval $[0, 40000]$, there is a lack of observations on vertex 1. As shown in Fig. 1.1, the isolated kernel ridge regression (KRR) method fails to reconstruct this part. On the other hand, our proposed approach, referred to as KRR-GGSP, utilizes the graph structure to incorporate the observations from a vertex's neighbor to improve reconstruction. This example motivates the need for a new KRR framework under GGSP. Unlike the methods based on wide-sense stationary (WSS) or JWSS assumptions [27, 29], KRR-GGSP

[1] In this context, for simplicity we only consider real-valued functions on vertices.
[2] `http://db.csail.mit.edu/labdata/labdata.html`

does not require knowledge of the PSD of the signal, which can be hard to estimate when there are only noisy and incomplete samples in the training set.

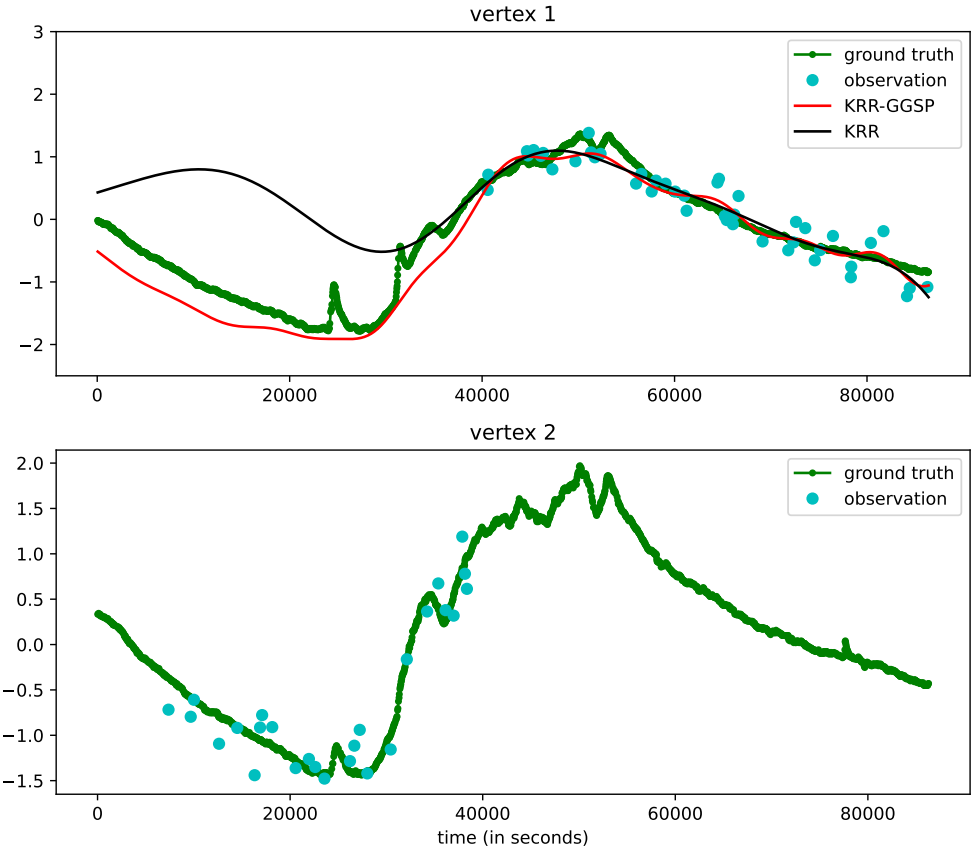


FIGURE 1.1: The upper and lower plots represent observations and ground truths from two connected vertices 1 and 2, respectively. Green curves are ground truth signals, and the cyan dots represent the observations on each vertex. Note that reconstruction based on the single vertex 1's observations using KRR is much worse compared to the proposed KRR-GGSP approach.

1.2.3 Multiple Hypothesis Testing in the Joint Domain

The MHT problem has been considered for identifying interested regions on a graph. Given a hypothesis testing problem and the corresponding p-value on each vertex, the target is to identify those vertices on which the alternative hypotheses should be correct. The papers [30, 31] proposed methods assuming that the detection results should inhibit a grouped structure over the graph. In other words, we expect to see similar decision results on adjacent vertices. In these papers, this grouped structure is imposed by adding sparsity terms to the optimization of the

log-likelihood functions. In [32], the authors proposed an approach for matching the performance of the global Benjamini-Hochberg (BH) method in a distributed way in a network. In the paper [33], the authors considered the MHT problem in spatial domain, which assumes that the p-values can be divided into different sets, and those in the same set have the same distribution.

The MHT problem is usually interpreted in a Bayesian framework such that the ground truth of the hypotheses are randomly generated from a Bernoulli distribution. If the alternative hypothesis is true, the p-value is assumed to follow a certain distribution that is more likely to observe small p-values. We notice that, the existing methods [30–33] assume these distributions to be constant, piecewise constant, or with finite choices. We claim that in practice this may not be the case, and parameterize them using an underlying generalized graph signal, resulting in a finer model. Besides, among these methods, only [31, 32] provides conditions for FDR control. We claim that these conditions can be too restrictive in practice. Finally, to our knowledge, there are no studies regarding MHT on the joint domain of a graph and a measure space.

In view of these research gaps, we model the underlying generalized graph signal as a bandlimited signal defined by a finite number of parameters. This strategy allows us to use the maximum likelihood estimator (MLE), which is known to be consistent. We prove the asymptotic control of the FDR by applying the local false discovery rate (lfdr) based approach with MLE. Our method is more suitable for real-world problems, and not restricted by uncontrollable parameters, making it an improvement over existing approaches.

1.3 Related Work

1.3.1 Algebraic and Statistical Graph Signal Processing Schemes

Central to the theory of algebraic signal processing is the concept of a shift operator and its resulting Fourier transform. The works [1, 34, 35] have extended these algebraic signal processing concepts from the discrete-time domain to the graph (or vertex) domain, yielding the new concepts called graph shift operator (GSO)

and graph Fourier transform (GFT). GFT is the "frequency" representation of a graph signal in a basis induced by the GSO (typically, we assume a complete orthonormal eigenbasis consisting of eigenvectors of the GSO). A GSO and its corresponding GFT are graph-based counterparts of the translation operator and the discrete Fourier transform for a discrete time series. There are multiple choices of GSOs and GFTs, which admit different properties and interpretations. For example, when the graph Laplacian is chosen as the GSO, its eigenvectors represent signals of certain frequencies, i.e., each eigenvector is a representer of what a graph signal with a specific degree of "smoothness" along its edges looks like, with lower frequencies corresponding to smoother signals. The related concepts of algebraic signal processing then follow. For example, a graph signal is *bandlimited* if its GFT support is within a proper subset of frequencies.

Building on these basic concepts, the papers [23–25] studied statistical signal processing on graphs. The authors extended the concept of a WSS signal from discrete-time domain to the vertex domain. To be specific, the set of WSS graph signals is a family of signals with statistically uncorrelated spectral modes. The PSD of a graph signal characterizes its energy distribution among different frequency components, and is also a crucial ingredient for constructing Wiener filters.

Traditional GSP deals with information in the graph or vertex domain rather than the time domain. In practice, the signal on every vertex can be discrete-time signals. To exploit information in both the graph and discrete-time domains, the papers [19, 26, 27] developed the time-vertex GSP theory. The idea was to construct the Cartesian product graph of the original graph and a cyclic graph, the latter of which represents the discrete-time domain. Through this construction, the discrete-time signals on graphs are modeled as a graph signal on the product graph. Hence the concepts and techniques in traditional GSP such as GFT, WSS and PSD can be extended as JFT, JWSS and JPSD. The Wiener filter of the time-vertex framework is also defined as the graph Wiener filter on the product graph.

To form the discrete time series on each vertex from a general signal like a continuous-time signal, the time-vertex framework implicitly assumes the same sampling rate on every vertex and requires a perfectly synchronous sampling scheme on the graph. To overcome these weaknesses, [22, 36] introduced a GGSP framework, which assumes that the vertex signal space is a possibly infinite dimensional Hilbert space. The GGSP framework generalizes GSP to a much wider range of scenarios

and applications, including the analysis of vertex signals in a continuous domain, which allows asynchronous sampling on different vertices. Besides, by considering multidimensional feature spaces, this framework incorporates more complex signals than discrete-time signals on each vertex.

1.3.2 Reconstruction Methods for Signal over Graphs

When the PSD (or JPSD) is known, the works [23–27] developed statistical optimal filters (i.e., Wiener filters) for graph signals or time-vertex signals. PSD represents the energy of random signal on different frequency components. Therefore, the estimation of PSD requires a noiseless and complete training set of signals. In practice, this may not be available. In this case, the following methods can be used to reconstruct signal over graphs.

Regarding the signal reconstruction problem, the most common assumption is that the target signal is bandlimited. The works [37, 38] proposed perfect recovery condition to reconstruct a bandlimited graph signal. The work [22] also proposed to recover generalized graph signal in the GGSP framework under the bandlimited assumption, and also developed a sampling theory for bandlimited signal in GGSP. Besides these approaches, the work [11] proposed a method to estimate graph signals using a kernel based approach without the bandlimited assumption. By designing a kernel on the vertex set, this method reconstructs the graph signal in a finite-dimensional RKHS.

In the time-vertex signal literature, the work [39] proposed to reconstruct signals by penalizing the first-order difference of discrete-time signal on every vertex and the total variation of graph signal on every instance. The work [40] generalized this method by replacing the graph total variation with a more general Sobolev smoothness.

1.3.3 Multiple Hypothesis Testing on Graph

In statistics, the problem of determining which hypotheses to reject among a large set of hypotheses based on their associated p-values is known as the MHT problem [41]. The seminal work [42] proposed an adaptive thresholding method, also known

as the BH method, which controls the FDR to ensure that the identified set of alternative hypotheses does not contain too many false positives. This method was further improved in [43] by estimating the proportion of null hypotheses. In recent years, there have been more MHT works that incorporate structural information. For example, [31] proposed a weighted BH method that assigns weights to p-values based on prior knowledge of structural information, such as low variation over a graph. The resulting FDR is shown to be related to the Rademacher complexity of the feasible set of weights.

The paper [30] proposed a method for solving the MHT problem over a graph. This method leverages the graph information by encoding the prior probabilities of being null as a function of a low variation signal on the graph. The paper [44] developed a method to iteratively threshold, mask, and reveal the p-values to control the FDR at a predetermined level. This approach allows for the incorporation of prior knowledge during the thresholding step. Furthermore, the paper proved that the optimal rejection strategy, in terms of both marginal power and marginal FDR, is to set a threshold on the lfdr values. Another work, [45], demonstrated that this strategy can asymptotically control the FDR at a pre-determined level under certain consistency conditions. The authors also proposed an EM algorithm for estimating the unknown probability density functions (pdfs) of p-values and the prior probabilities of being null. The work [33] proposed a method of estimating a Beta mixture density, and applied it to the detection problem on spatial signals. The work [32] proposed distributed methods for approximating the BH method over graphs. This method incorporates the case where the prior probabilities of being null vary over the graph.

1.4 Thesis Contribution and Outline

The main contributions of this thesis are summarized as follows:

In Chapter 3, we establish a probabilistic model for the GGSP framework by introducing the concept of a graph random process. We define the notion of joint WSS with respect to (w.r.t.) the GGSP shift operator. This framework includes the notion of graph or vertex WSS and joint WSS under the traditional GSP and

time-vertex frameworks, respectively, as special cases. We show that joint WSS implies WSS in both the vertex and Hilbert space domains.

We derive analytical forms for the Wiener filters for denoising and signal completion, which are the best linear unbiased estimator (BLUE) for signal recovery under the joint WSS assumption. We also derive the theoretical MSE of the Wiener filters as a criterion for sample set selection.

We verify our proposed framework on real and synthetic datasets and demonstrate the utility of working with more general assumptions under our framework. In particular, the Hilbert space shift operator under our proposed framework can be learned from the data, instead of having to be fixed in advance.

In Chapter 4, we construct an appropriate kernel and formulate the signal reconstruction in GGSP as a KRR problem. We interpret it as an extension of existing kernel-based frameworks. This approach does not require estimation of JPSD, since the parameters can be tuned on noisy and incomplete data.

We present an online and distributed approach for generalized graph signal reconstruction. When using radial basis function (RBF) to construct the kernel, the KRR problem can be approximated by a linear regression problem by utilizing random Fourier feature (RFF).

We compute the limit and asymptotic upper bound for conditional MSE of reconstruction under the Bayesian framework. These theoretical results reveal that the posterior estimation MSE will converge to the oracle MSE.

We present numerical case studies to illustrate the utility of our proposed method in several applications. By extending the existing approaches in graph signal reconstruction, our approach is more flexible, and can adapt to more general sampling situations.

In Chapter 5, we propose a method for estimation of p-values' pdfs which are homogeneous over the joint domain. By utilizing the domain knowledge, this method yields consistent estimation.

We provide an optimal MHT strategy with simultaneously varying null probability and alternative distribution under the GGSP framework. We derive the conditions

for asymptotic FDR control and derive the asymptotic power of the proposed method.

We validate the utility of the proposed strategy in practical sensing applications. The numerical results demonstrate that by leveraging the information in the joint domain, our approach achieves the best detection power.

The thesis is organized as follows.

In Chapter 2 we review the framework of GGSP and functional analysis contents like Hilbert space and Bochner integral. We introduce the statistical concepts related to Hilbert space, such as random element and linear conditional expectation. Besides, we also introduce the KRR approach and its interpretation. These are the basic theoretical tools we will use in this thesis. In Chapter 3, we model a random graph signal as a random element in a specific Hilbert space under the GGSP framework. We extend the JWSS and JPSD concepts from the time-vertex framework to the GGSP framework. We provide Wiener filter design for denoising and signal completion. In Chapter 4, we formulate the signal reconstruction problem in GGSP. We derive the solution, discuss its interpretation and compare it with the existing methods. We also provide an online version of the reconstruction problem. We analyze the statistical performance of our reconstruction approach under the asymptotic case. In Chapter 5, we propose an approach for simultaneous estimation of the prior probability of hypotheses being null and the distribution of p-values under alternative hypotheses in a sensor network over the joint spatial-time domain. We prove the asymptotic control of the FDR by applying the lfdr-based approach using this estimation.

Chapter 2

Preliminaries

2.1 Generalized Graph Signal Processing

In this section, we introduce the GGSP framework, which is the basic signal model we will discuss in this thesis. We first introduce the basic concepts in GSP, and then explain their extended versions to GGSP.

2.1.1 Graph Signal Processing

The goal of GSP is to analyze signals residing on the vertices of a graph. A graph G is defined by a tuple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the vertex set and edge set, respectively. Suppose we label the vertices as $\mathcal{V} = \{1, \ldots, N\}$, and $\mathcal{E} \subset \{(i,j) : 1 \leq i, j \leq N\}$. We write $i \sim j$ if $(i,j) \in \mathcal{E}$. The set of neighbors of vertex i is given by $\mathcal{N}_i = \{j : i \sim j\}$. The adjacency matrix $\mathbf{W}_G$ of G is defined as

$$\mathbf{W}_G(i,j) = \begin{cases} w_{ij} & \text{if } i \sim j, \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where w_{ij} denotes the weight of the edge (i,j). Every edge weight $w_{ij} = 1$ for unweighted graphs G. The degree matrix is the diagonal matrix $\mathbf{D}_G = \text{diag}(d_1 \ldots, d_N)$, where $d_i = |\mathcal{N}_i|$ is the degree of the i-th vertex. The graph Laplacian $\mathbf{L}_G$ is defined as $\mathbf{L}_G = \mathbf{D}_G - \mathbf{W}_G$. In this thesis, we mainly consider undirected graphs without self-loops, i.e., $\mathbf{W}_G$ is symmetric with zero diagonal elements.

A graph signal x is defined as a function on the vertex set:

$$\begin{aligned} x : \mathcal{V} &\to \mathcal{M} \\ v &\mapsto x(v), \end{aligned} \tag{2.2}$$

where $\mathcal{M}$ denotes the space of vertex signals, and $x(v)$ denotes the signal associated with the vertex v. For ease of notation, we alternatively write graph signal as a vector $\mathbf{x} = (x(1), \cdots, x(N))^\intercal$. In this subsection, we assume that $\mathcal{M} = \mathbb{R}$.

GSP starts with the definition of GSO and GFT. These definitions lay the foundation for basic GSP theory and inspire further generalizations of GSP. From the analytical and algebraic perspectives of signal processing, there are two major choices for the GSO.

The analytical perspective comes from signal analysis on $\mathbb{R}^n$ [1]. The Laplacian on $L^2(\mathbb{R}^n)$ is defined as

$$\mathbf{\Delta}(f) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial t_i^2},$$

where $\mathbf{t} = (t_1, \cdots, t_n)^\intercal \in \mathbb{R}^n$. It can also be written as the divergence $\nabla\cdot$ of the gradient ∇, i.e., $\mathbf{\Delta}(f) = \nabla \cdot \nabla(f)$. Note that the Fourier sinusoids $\{e^{\mathrm{i}\boldsymbol{\omega}^\intercal \mathbf{t}} : \boldsymbol{\omega} \in \mathbb{R}^n\}$ are the eigenfunctions of the Laplacian operator:

$$\mathbf{\Delta}(e^{\mathrm{i}\boldsymbol{\omega}^\intercal \mathbf{t}}) = \sum_{i=1}^{n} \frac{\partial^2 e^{\mathrm{i}\boldsymbol{\omega}^\intercal \mathbf{t}}}{\partial t_i^2} = -\left\|\boldsymbol{\omega}\right\|^2 e^{\mathrm{i}\boldsymbol{\omega}^\intercal \mathbf{t}}. \tag{2.3}$$

For a graph signal $\mathbf{x}$, the gradient can be defined as

$$\nabla(\mathbf{x}) = \mathbf{B}_G \mathbf{x} = (x(i) - x(j))_{(i,j)\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|},$$

where $\mathbf{B}_G \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{V}|}$ is the incidence matrix. It has elements given by

$$\mathbf{B}_G((i,j), v) = \begin{cases} -1 & v = i, \\ 1 & v = j, \\ 0 & \text{otherwise}, \end{cases}$$

for each $(i, j) \in \mathcal{E}$ and $v \in \mathcal{V}$. It is known that the negative divergence is the adjoint operator of the gradient, thus

$$\nabla \cdot = -\nabla^{\intercal} = -\mathbf{B}_G^{\intercal}. \tag{2.4}$$

The Laplacian on graph G is then defined as $\mathbf{L}_G = -\nabla \cdot \nabla = \mathbf{B}_G^{\intercal}\mathbf{B}_G$ (the negative sign is added so that $\mathbf{L}_G$ is positive semidefinite), which coincides with the previous definition $\mathbf{L}_G = \mathbf{D}_G - \mathbf{W}_G$. Since the graph Laplacian $\mathbf{L}_G$ is the discrete analog of $\boldsymbol{\Delta}$, it is natural to define $\mathbf{L}_G$ as the GSO with the resulting orthonormal eigenvectors as the Fourier basis. To be specific, note that the following quadratic form represents the total variation (i.e., smoothness) of a signal $\mathbf{x}$ on graph G:

$$\mathrm{TV}_{\mathbf{L}_G}(\mathbf{x}) = \mathbf{x}^{\intercal}\mathbf{L}_G\mathbf{x} = \sum_{(i,j)\in\mathcal{E}} w_{ij}(x(i) - x(j))^2. \tag{2.5}$$

Therefore, suppose λ_k is the k-th smallest eigenvalue of $\mathbf{L}_G$, and $\boldsymbol{\phi}_k$ is the corresponding eigenvector. Then we have $\boldsymbol{\phi}_k^{\intercal}\mathbf{L}_G\boldsymbol{\phi}_k = \lambda_k$. This indicates that an eigenvalue with a larger magnitude implies a higher frequency, which is similar to (2.3), where larger components of $\boldsymbol{\omega}$ are associated with higher frequencies in that dimension.

The algebraic perspective comes from signal analysis on discrete-time signals [34, 35]. Recall that the discrete Fourier basis consists of the eigenvectors of the translation operator:

$$\begin{aligned} \mathbf{W}_T : \mathbb{R}^T &\to \mathbb{R}^T \\ (x(1), x(2), \ldots, x(T)) &\mapsto (x(T), x(1), \ldots, x(T-1)), \end{aligned} \tag{2.6}$$

where $\mathbf{W}_T$ is the time shift operator on a discrete-time signal. Note that $\mathbf{W}_T$, represented as an $T \times T$ matrix, coincides with the cyclic graph's adjacency matrix with T vertices, denoted as C_T (see Fig. 2.1). Therefore, the discrete-time signal can be regarded as a graph signal on a cyclic graph with GSO $\mathbf{W}_T$. In the same vein, for a general graph G, we can choose its adjacency matrix $\mathbf{W}_G$ as the GSO. This can be understood via another definition of total variation: the shifted result of a low-frequency signal should be close to the signal itself. In other words, the quantity

$$\mathrm{TV}_{\mathbf{W}_G}(\mathbf{x}) = \|\mathbf{x} - \mathbf{W}_G\mathbf{x}\|_1 = \sum_{i=1}^{N} \left| x(i) - \sum_{j\in\mathcal{N}_i} x(j) \right| \tag{2.7}$$

should be small. To construct an energy-preserving shift, we normalize $\mathbf{W}_G$ by its operator norm, so that its eigenvalues $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_N|$ satisfy $|\lambda_N| = 1$. For each eigenvalue λ_k, let its corresponding eigenvector be $\boldsymbol{\phi}_k$, normalized such that $\|\boldsymbol{\phi}_k\|_1 = 1$. We observe that

$$\mathrm{TV}_{\mathbf{W}}(\boldsymbol{\phi}_k) = |1 - \lambda_k| \,. \tag{2.8}$$

Therefore, an eigenvalue of $\mathbf{W}_G$ close to 1 (after normalization) on the real line (or complex plane) is associated with a smooth eigenvector on the graph.

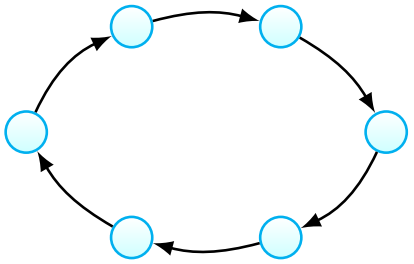


FIGURE 2.1: Cyclic graph with 6 vertices.

In the rest of this thesis, we assume a GSO has been fixed and we denote it as $\mathbf{A}_G$ when it is not important to specify a concrete choice. The eigendecomposition of $\mathbf{A}_G$ is $\mathbf{A}_G = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^{\mathsf{T}}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues and $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \ \ldots \ \boldsymbol{\phi}_N]$ is the matrix whose columns are the corresponding eigenvectors. We assume that the set $\{(\lambda_k, \boldsymbol{\phi}_k) \,:\, k = 1, \ldots, N\}$ is listed in ascending order of frequency. From the above discussion, we see that the eigenvectors of GSO represent graph signals with different total variations or frequencies. This observation leads to the natural definition of GFT as the inner product with these eigenvectors, which are also called the *graph Fourier basis*:

$$\begin{aligned} \mathfrak{F}_k : \mathbb{R}^N &\to \mathbb{R}^N \\ \mathbf{x} &\mapsto \langle \boldsymbol{\phi}_k, \mathbf{x} \rangle. \end{aligned}$$

The frequency domain, containing the frequencies $\{\lambda_k\}$, is discrete. Note that the matrix $\boldsymbol{\Phi}^{\mathsf{T}}$ maps $\mathbf{x}$ to its Fourier coefficients, so we also call it GFT. The inverse GFT is then defined as the matrix $\boldsymbol{\Phi}$. We refer to the eigenvalues $\{\lambda_k \,:\, k = 1, \ldots, N\}$ as graph frequencies. A graph signal $\mathbf{x}$ is said to be *K-bandlimited* if $\mathfrak{F}_k(\mathbf{x}) = 0$ for all $k > K$.

Graph filters are linear operators on the space of graph signals. In GSP, since the signal space is $\mathbb{R}^N$, the filter space is $\mathbb{R}^{N\times N}$. A subspace of filters is identified as convolution filters. Recall that in traditional signal processing, the convolution filter is defined as the pointwise multiplication with a certain transfer function on the frequency domain. In GSP, a *convolution filter* is defined in the same way:

$$\mathbf{H} = \mathbf{\Phi} h(\mathbf{\Lambda}) \mathbf{\Phi}^{\intercal}, \tag{2.9}$$

where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ contains the graph frequencies in the diagonal entries and $h(\mathbf{\Lambda}) = \mathrm{diag}(h(\lambda_1), \ldots, h(\lambda_N))$ is the *frequency response* of $\mathbf{H}$. Note that (2.9) consists of a GFT, a pointwise multiplication, and an inverse GFT.

Another subspace of filters is the space of polynomial filters. Let $h(s) = \sum_{i=0}^{L-1} c_i s^i$ be a polynomial of degree $L-1$. A *polynomial filter* has the form

$$\mathbf{H} = h(\mathbf{A}_G) = \sum_{i=0}^{L-1} c_i \mathbf{A}_G^i. \tag{2.10}$$

From the practical aspect, the implementation of $\mathbf{A}_G$ is distributed because all the aforementioned GSO choices encode the graph topology in their entries, i.e., $\mathbf{A}_G(i,j) = 0$ whenever $(i,j) \notin E$. To be specific, computing $\mathbf{y} = \mathbf{A}_G\mathbf{x}$ amounts to computing the sum

$$y(i) = \sum_{j\sim i} \mathbf{A}_G(i,j) x(j),$$

which only requires information exchange between neighboring vertices. If a graph filter $\mathbf{H}$ commutes with $\mathbf{A}_G$, we call it a *shift-invariant filter*. We see that (2.9) and (2.10) are shift-invariant filters. Besides, it can be shown by [46, 5F, Section 5.2] that if all $\{\lambda_k\}$ are distinct and have multiplicity one, then all shift-invariant filters are polynomial filters.

2.1.2 Generalized Graph Signal Processing

In traditional GSP, a graph signal assigns a number to each vertex in $\mathcal{V}$, i.e., $\mathcal{M} = \mathbb{R}$ or $\mathbb{C}$ in (2.2). By a generalized graph signal, a function is assigned to each vertex. The appropriate mathematical language is the theory of Hilbert spaces. Suppose

$\mathcal{M} = \mathcal{H}$ is a *Hilbert space* (i.e., a complete inner product space). We further assume that $\mathcal{H}$ is separable. A generalized graph signal is a function $x : \mathcal{V} \to \mathcal{H}$. For example, if $\mathcal{H} = \mathbb{R}$ or $\mathbb{C}$, we are in the realm of traditional GSP theory. However, it can be fruitful to consider $\mathcal{H} = L^2([a, b])$, the space of square-integrable functions on an interval $[a, b]$. We refer the reader to Section 2.2 for an introduction of Hilbert spaces.

To make the discussion here self-contained, we briefly review some concepts from the theory of Hilbert spaces. A separable Hilbert space $\mathcal{H}$ is equipped with an inner product and has a countable basis, denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Without loss of generality, we can assume $\mathcal{H} = L^2(\mathcal{T})$ for some measurable space $\mathcal{T}$ with a measure τ [47]. Then, the space of generalized graph signals with vertex signals from $\mathcal{H}$ can be denoted by $L^2(\mathcal{V} \times \mathcal{T})$. For each $x \in L^2(\mathcal{V} \times \mathcal{T})$ and $u \in \mathcal{V}$, we have $x(u) \in L^2(\mathcal{T})$ and $x(u)(\mathbf{t}) \in \mathbb{C}$ for each $\mathbf{t} \in \mathcal{T}$. A useful point of view is that we can treat x as a function of two variables and write $x(u, \mathbf{t})$ for $x(u)(\mathbf{t})$.

For a finite dimensional Euclidean vector space $\mathbb{C}^N$, we may form the *tensor product* [48] $\mathbb{C}^N \otimes \mathcal{H}$ as the set of finite linear sums $\sum_{i=1}^{N} \mathbf{b}_i \otimes h_i$, with $\mathbf{b}_i \in \mathbb{C}^N$ and $h_i \in \mathcal{H}$ for all $i = 1, \ldots, N$, such that the following holds for any $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b} \in \mathbb{C}^N$ and any $h_1, h_2, h \in \mathcal{H}$:

- $\mathbf{b}_1 \otimes h + \mathbf{b}_2 \otimes h = (\mathbf{b}_1 + \mathbf{b}_2) \otimes h$;
- $\mathbf{b} \otimes h_1 + \mathbf{b} \otimes h_2 = \mathbf{b} \otimes (h_1 + h_2)$;
- $r\mathbf{b} \otimes h = \mathbf{b} \otimes rh$ for $r \in \mathbb{C}$.

The tensor product $\mathbb{C}^N \otimes \mathcal{H}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathbb{C}^N \otimes \mathcal{H}}$ induced (linearly) by:

$$\langle \mathbf{b}_1 \otimes h_1, \mathbf{b}_2 \otimes h_2 \rangle_{\mathbb{C}^N \otimes \mathcal{H}} = \langle \mathbf{b}_1, \mathbf{b}_2 \rangle_{\mathbb{C}^N} \langle h_1, h_2 \rangle_{\mathcal{H}}, \tag{2.11}$$

which defines a metric on $\mathbb{C}^N \otimes \mathcal{H}$. As $\mathbb{C}^N$ is finite dimensional, $\mathbb{C}^N \otimes \mathcal{H}$ is complete and hence a Hilbert space. The signal space $L^2(\mathcal{V} \times \mathcal{T})$ can be identified with $\mathbb{C}^N \otimes \mathcal{H}$ by the following map:

$$x \mapsto \sum_{u \in \mathcal{V}} \mathbf{e}_u \otimes x(u) \tag{2.12}$$

for each $x \in L^2(\mathcal{V} \times \mathcal{T})$, where $u \mapsto \mathbf{e}_u$ identifies $\mathcal{V}$ with the standard basis of $\mathbb{C}^N$ [22]. We shall use this identification extensively in the sequel.

Suppose $\mathbf{\Phi}$ is the graph Fourier basis and $\Psi = \{\psi_l : l \geq 1\}$ is an orthonormal basis of $\mathcal{H}$. Then $\mathbf{\Phi} \otimes \Psi = \{\boldsymbol{\phi}_k \otimes \psi_l : 1 \leq k \leq N, l \geq 1\}$ forms an orthonormal basis of $\mathbb{C}^N \otimes \mathcal{H}$. For each $x \in L^2(\mathcal{V} \times \mathcal{T})$, $\boldsymbol{\phi}_k \in \mathbf{\Phi}$ and $\psi_l \in \Psi$, the JFT is defined as:

$$\mathfrak{F}_{k,l}(x) = \langle x, \boldsymbol{\phi}_k \otimes \psi_l \rangle \triangleq \langle \sum_{u \in \mathcal{V}} \mathbf{e}_u \otimes x(u), \boldsymbol{\phi}_k \otimes \psi_l \rangle_{\mathbb{C}^N \otimes \mathcal{H}}. \tag{2.13}$$

Since $\mathbf{\Phi} \otimes \Psi$ is an orthonormal basis, given a sequence of numbers (g_{kl}) such that $\sum_{k,l} |g_{kl}|^2 < \infty$, the *inverse* JFT is given by

$$\mathfrak{F}^{-1}((g_{kl})) = \sum_{k,l} g_{kl} \cdot \boldsymbol{\phi}_k \otimes \psi_l. \tag{2.14}$$

In the above definition of JFT, the graph information is encoded in the orthonormal basis $\mathbf{\Phi}$. Recall that $\mathbf{\Phi}$ is an orthonormal eigenbasis of $\mathbf{A}_G$. On the other hand, there can be bounded linear operators, such as integral operators, on $\mathcal{H} = L^2(\mathcal{T})$. Let $\mathbf{A}_{\mathcal{H}}$ be such an operator and we assume that it is self-adjoint and compact. The basis Ψ is usually chosen as an eigenbasis of such an operator $\mathbf{A}_{\mathcal{H}}$. The tensor product $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}} : \mathbf{b} \otimes h \mapsto \mathbf{A}_G(\mathbf{b}) \otimes \mathbf{A}_{\mathcal{H}}(h) \in \mathbb{C}^N \otimes \mathcal{H} \cong L^2(\mathcal{V} \times \mathcal{T})$ induces a bounded linear map on $L^2(\mathcal{V} \times \mathcal{T})$. We call this operator $\mathbf{A}_{\mathcal{J}} := \mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ the joint shift operator in joint domain.

Analogous to Fourier theory, the domain of JFT is related to the notion of "frequency". For $\boldsymbol{\phi}_k \in \mathbf{\Phi}$ and $\psi_l \in \Psi$, we use λ_k and ν_l to denote their corresponding eigenvalues. The set $\{(\lambda_k, \nu_l^{-1}) \in \mathbb{R} \times \mathbb{R} : \psi_l \in \Psi\}$ is defined as the frequency domain. For $x \in L^2(\mathcal{V} \times \mathcal{T})$, the frequency range of x is defined to be $\{(\lambda_k, \nu_l^{-1}) \in \mathbb{R} \times \mathbb{R} : \mathfrak{F}_{k,l}(x) \neq 0, \boldsymbol{\phi}_k \in \mathbf{\Phi}, \ \psi_l \in \Psi\}$.

In traditional GSP, a filter is a linear transformation, which is purely an algebraic concept. However, as in functional analysis, additional analytic conditions should be taken into consideration when we discuss filters in the GGSP framework. A *filter* is a bounded (or equivalently continuous) linear operator $\mathbf{H} : L^2(\mathcal{V} \times \mathcal{T}) \to L^2(\mathcal{V} \times \mathcal{T})$. A few families of filters are introduced in [22], including weakly shift invariant and shift invariant filters, the limit of finite rank filters, compact filters, convolution

filters, and bandpass filters. Among these, the two largest families are the weakly shift invariant filters and the limit of finite rank filters.

A filter $\mathbf{H}$ is *weakly shift invariant* if it commutes with $\mathbf{A}_{\mathcal{J}}$, i.e., $\mathbf{H} \circ \mathbf{A}_{\mathcal{J}} = \mathbf{A}_{\mathcal{J}} \circ \mathbf{H}$. It is *shift invariant* if $\mathbf{H}$ commutes with both operators $\mathbf{A}_G \otimes \mathbf{I}$ and $\mathbf{I} \otimes \mathbf{A}_{\mathcal{H}}$. As $\mathbf{A}_G$ and $\mathbf{A}_{\mathcal{H}}$ are interpreted as shifts in their respective domains, the defining algebraic properties explains the term "shift invariant". As the names suggest, a shift invariant filter is weakly shift invariant.

To give some examples, let $P(x) = a_0 + a_1 x + \ldots + a_p x^p$ be a polynomial of degree $p < \infty$. Then $P(\mathbf{A}_{\mathcal{J}})$ is a shift invariant filter, analogous to traditional GSP. Such a filter can be used to approximate other filters due to the Stone-Weierstrass theorem. However, unlike traditional GSP, not every shift invariant filter is in the polynomial form as above. Let $a \in \mathbb{R}$ be a positive real number larger than the spectral radius of $\mathbf{A}_{\mathcal{H}}$. Then $(\mathbf{I} - a^{-1}\mathbf{A}_{\mathcal{H}})^{-1}$ is a bounded linear transformation as it has a convergent power series expansion in $\mathbf{A}_{\mathcal{H}}$. The filter $\mathbf{H} = \mathbf{A}_G \otimes (\mathbf{I} - a^{-1}\mathbf{A}_{\mathcal{H}})^{-1}$ is shift invariant but in general not in the above polynomial form. This difference with traditional GSP is essentially due to infinite dimensionality of $\mathcal{H}$.

We now turn to the other large filter family. A filter $\mathbf{H}$ is of finite rank if its range is finite-dimensional, and $\mathbf{H}$ is a "limit of finite rank filters" if $\mathbf{H} = \lim_{i \to \infty} \mathbf{H}_i$ (converge in operator norm) with each $\mathbf{H}_i$ of finite rank. Such a filter is more manageable as it can be approximated by essentially finite dimensional objects.

As the intersection of the above two filter families, we have the space of convolution filters that plays an important role, analogous to its traditional counterpart. Specifically, a convolution $\mathbf{H} = g*$ is defined by pointwise multiplication in the frequency domain with the JFT of a predefined signal $g \in L^2(\mathcal{V} \times \mathcal{T})$. This means that for any signal $x \in L^2(\mathcal{V} \times \mathcal{T})$, the signal $g * f$ is determined uniquely by the identity $\mathcal{F}_{g*f} = \mathcal{F}_g \mathcal{F}_f$ on JFT.

2.2 Hilbert Spaces and the Bochner Integral

In this section, we provide a brief overview of some concepts related to Hilbert spaces and the Bochner integral. Readers are referred to [47, 49–51] for further details.

A Hilbert space $\mathcal{H}$ is a complete normed space (i.e., a Banach space), whose norm $\|\cdot\|$ is induced by an inner product $\langle\cdot,\cdot\rangle$ so that $\|\mathbf{u}\| = \sqrt{\langle\mathbf{u},\mathbf{u}\rangle}$ for $\mathbf{u} \in \mathcal{H}$. Examples of Hilbert spaces include $L^2(\mathcal{T})$ and Euclidean spaces. Here $\mathcal{T}$ can be any measure space.

The norm $\|\cdot\|$ naturally induces a topology on $\mathcal{H}$ whose topological basis consists of the open balls centered at each $\mathbf{y} \in \mathcal{H}$ with radius $\delta > 0$, denoted as

$$B(\mathbf{y}, \delta) = \{\mathbf{u} \in \mathcal{H} \,:\, \|\mathbf{u} - \mathbf{y}\| < \delta\}.$$

This topology defines the Borel σ-algebra $\mathcal{B}$ of $\mathcal{H}$ as the smallest σ-algebra containing all the open subsets of $\mathcal{H}$, so that $(\mathcal{H}, \mathcal{B})$ becomes a measurable space.

Two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are said to be isomorphic ($\mathcal{H}_1 \cong \mathcal{H}_2$) if there exists a bijective map $\psi : \mathcal{H}_1 \mapsto \mathcal{H}_2$ such that:

$$\begin{aligned}\psi(a\mathbf{u}_1 + \mathbf{u}_2) &= a\psi(\mathbf{u}_1) + b\psi(\mathbf{u}_2),\\ \langle\psi(\mathbf{u}_1), \psi(\mathbf{u}_2)\rangle &= \langle\mathbf{u}_1, \mathbf{u}_2\rangle,\end{aligned}$$

for arbitrary $a, b \in \mathbb{C}$, $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}_1$.

An operator $\mathbf{C}$ on $\mathcal{H}$ is a linear map from $\mathcal{H}$ to itself. In this thesis we only consider bounded operators, i.e.,

$$\sup_{\|\mathbf{u}\|=1} \|\mathbf{C}(\mathbf{u})\| < \infty.$$

An operator $\mathbf{C}$ is bounded if and only if it is continuous. When $\dim \mathcal{H} = d < \infty$, an operator can be represented as a $d \times d$ matrix $\mathbf{C}$. Symmetric matrices form a notable class of matrices in linear algebra, since its eigenvectors form an orthonormal basis for the whole space. The counterparts of these in Hilbert spaces are compact self-adjoint operators, defined as follows.

Definition 2.1. An operator $\mathbf{C}$ is *compact* if for any bounded sequence $(\mathbf{x}_k)_{k\geq 1} \subset \mathcal{H}$, there exists a subsequence $(\mathbf{x}_{k_i})_{i\geq 1}$ such that $(\mathbf{C}(\mathbf{x}_{k_i}))_{i\geq 1}$ converges.

Definition 2.2. The operator $\mathbf{C}^*$ defined by

$$\langle\mathbf{u}_1, \mathbf{C}^*(\mathbf{u}_2)\rangle = \langle\mathbf{C}(\mathbf{u}_1), \mathbf{u}_2\rangle$$

for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}$ is called the *adjoint operator* of $\mathbf{C}$. An operator $\mathbf{C}$ is called *self-adjoint* if $\mathbf{C} = \mathbf{C}^*$.

If an operator $\mathbf{C}$ satisfies both Definition 2.1 and Definition 2.2, then it admits an eigendecomposition.

Proposition 2.1. *[47, Corollary 4.10.2, 4.10.3] Let $\mathbf{C}$ be a compact self-adjoint operator on $\mathcal{H}$. Then $\mathcal{H}$ has an orthonormal basis $\{\boldsymbol{\psi}_k\}_{k=1}^{\infty}$ consisting of the eigenvectors of $\mathbf{C}$. Furthermore, we have*

$$\mathbf{C}(\mathbf{u}) = \sum_{k=1}^{\infty} \sigma_k \langle \mathbf{u}, \boldsymbol{\psi}_k \rangle \boldsymbol{\psi}_k.$$

In other words, $\mathbf{C}$ can be decomposed into finite-rank projection operators:

$$\mathbf{C} = \sum_{k}^{\infty} \sigma_k \boldsymbol{\Pi}_{\boldsymbol{\psi}_k}.$$

Definition 2.3. An operator $\mathbf{C}$ on $\mathcal{H}$ is *trace-class* if

$$\operatorname{tr}(\mathbf{C}) = \sum_{k=1}^{\infty} \left\langle (\mathbf{C}^*\mathbf{C})^{1/2} \tilde{\boldsymbol{\psi}}_k, \tilde{\boldsymbol{\psi}}_k \right\rangle$$

converges for some orthonormal basis $\{\tilde{\boldsymbol{\psi}}_k\}_{k=1}^{\infty}$ of $\mathcal{H}$. $\operatorname{tr}(\mathbf{C})$ is known as the *operator trace* of $\mathbf{C}$.

It can be shown that Definition 2.3 is independent of the choice of orthonormal basis $\{\tilde{\boldsymbol{\psi}}_k\}_{k=1}^{\infty}$.

For a function taking values in a Hilbert space, its Bochner integral is defined by the limit of a series of simple functions, which is similar to the definition of the Lebesgue integral. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. In the following, we define the Bochner integral of a measurable function $\mathbf{f} : \Omega \mapsto \mathcal{H}$. We start with the definition of a simple function.

Definition 2.4. A *simple function* $\mathbf{g}$ is a linear combination of characteristic functions:

$$\mathbf{g} = \sum_{k=1}^{n} \mathbf{u}_k \mathbf{1}_{\{A_k\}},$$

wherein $A_k \in \mathcal{F}$, $A_k \cap A_l = \varnothing$ for $k \neq l$, and $\mathbf{u}_k \in \mathcal{H}$, for all $k \geq 1$. The Bochner integral of $\mathbf{g}$ is defined to be

$$\int_\Omega \mathbf{g}\, \mathrm{d}\mu = \sum_{k=1}^{n} \mu(A_k)\mathbf{u}_k \in \mathcal{H}.$$

Definition 2.5. A measurable function $\mathbf{f} : \Omega \mapsto \mathcal{H}$ is Bochner integrable if there exists simple functions $\{\mathbf{f}_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} \int_\Omega \|\mathbf{f} - \mathbf{f}_k\|\, \mathrm{d}\mu = 0.$$

Its integral on any set $E \in \mathcal{F}$ is defined as

$$\int_E \mathbf{f}\, \mathrm{d}\mu := \lim_{k\to\infty} \int_E \mathbf{f}_k\, \mathrm{d}\mu.$$

It can be shown that Definition 2.5 is independent of the sequence of converging simple functions chosen. One may also note that neither Definition 2.4 nor Definition 2.5 directly makes use of the inner product but only the norm. In fact, the Bochner integral is also applicable to functions taking values in Banach spaces. Alternative definitions and properties of Bochner integral can be found in [52–54].

2.3 Random Elements in a Hilbert Space

In this section, we introduce the definition of a random element and its moments. We provide interpretation of a random element's moments via the Bochner integral, and explain why they generalize the mean and covariance in a Euclidean space to a Hilbert space.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ a σ-algebra and $\mathbb{P}$ a probability measure. Let $\mathcal{H}$ be a complex separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Since $\mathcal{H}$ is a Banach space, it is naturally endowed with the norm-induced topology. This topology defines the Borel σ-algebra $\mathcal{B}$ as the smallest σ-algebra containing all the open subsets of $\mathcal{H}$, so that $(\mathcal{H}, \mathcal{B})$ is a measurable space. A random element is defined as a measurable map $\mathbf{x} : \Omega \to \mathcal{H}$, which induces a

probability measure $\mathbb{P}_{\mathbf{x}}$ on $(\mathcal{H}, \mathcal{B})$ given by

$$\mathbb{P}_{\mathbf{x}}(B) = \mathbb{P}\big(\mathbf{x}^{-1}(B)\big), \ \forall\, B \in \mathcal{B}.$$

A sufficient and necessary condition for $\mathbf{x}$ to be measurable (i.e., $\mathbf{x}^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$) is as follows.

Proposition 2.2. *[49, Theorem 7.1.2]* $\mathbf{x}$ *is measurable if and only if* $\langle \mathbf{x}, \mathbf{u} \rangle$ *is measurable for all* $\mathbf{u} \in \mathcal{H}$.

Proposition 2.2 enables us to verify the measurability of $\mathbf{x}$ by investigating that of a family of complex-valued functions. In Section 3.1, we utilize Proposition 2.2 to model continuous-time graph processes as random elements.

In the paper, Assumption 3.1 ensures the existence of a random element's mean and covariance operators:

$$\langle \mathbf{m}_{\mathbf{x}}, \mathbf{u} \rangle = \mathbb{E}[\langle \mathbf{x}, \mathbf{u} \rangle] = \int_{\mathcal{H}} \langle \mathbf{z}, \mathbf{u} \rangle \, \mathrm{d}\mathbb{P}_{\mathbf{x}}(\mathbf{z}), \tag{2.15}$$

$$\begin{aligned} \langle \mathbf{C}_{\mathbf{x}} \mathbf{u}_1, \mathbf{u}_2 \rangle &= \mathbb{E}\Big[\overline{\langle \mathbf{x} - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_1 \rangle} \langle \mathbf{x} - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_2 \rangle\Big] \\ &= \int_{\mathcal{H}} \overline{\langle \mathbf{z} - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_1 \rangle} \langle \mathbf{z} - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_2 \rangle \, \mathrm{d}\mathbb{P}_{\mathbf{x}}(\mathbf{z}). \end{aligned} \tag{2.16}$$

By the Riesz representation theorem [47, Theorem 3.7.7], $\mathbf{m}_{\mathbf{x}}$ and $\mathbf{C}_{\mathbf{x}}(\mathbf{u})$ exist uniquely and are hence well-defined. Suppose two random elements $\mathbf{x} : \Omega \to \mathcal{H}_1$ and $\mathbf{y} : \Omega \to \mathcal{H}_2$ induce a joint probability measure $\mathbb{P}_{\mathbf{xy}}$ on $\mathcal{H}_1 \times \mathcal{H}_2$, and $\mathbb{E}\,\|(\mathbf{x}, \mathbf{y})\|^2 < \infty$, their cross-covariance operator is given by

$$\begin{aligned} \langle \mathbf{C}_{\mathbf{xy}} \mathbf{u}_1, \mathbf{u}_2 \rangle &= \mathbb{E}\Big[\overline{\langle \mathbf{y} - \mathbf{m}_{\mathbf{y}}, \mathbf{u}_1 \rangle} \langle \mathbf{x} - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_2 \rangle\Big] \\ &= \int_{\mathcal{H}_1 \times \mathcal{H}_2} \overline{\langle \mathbf{z}_1 - \mathbf{m}_{\mathbf{y}}, \mathbf{u}_1 \rangle} \langle \mathbf{z}_2 - \mathbf{m}_{\mathbf{x}}, \mathbf{u}_2 \rangle \, \mathrm{d}\mathbb{P}_{\mathbf{xy}}(\mathbf{z}_2, \mathbf{z}_1). \end{aligned} \tag{2.17}$$

We say that $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated if $\mathbf{C}_{\mathbf{xy}} = \mathbf{0}$.

Let $\mathbf{u}_1 \circledast \mathbf{u}_2$ denote the operator $\mathbf{A}$ that maps $\mathbf{u}$ to $\mathbf{A}\mathbf{u} = \langle \mathbf{u}, \mathbf{u}_2 \rangle \mathbf{u}_1$. This is the Kronecker product matrix $\mathbf{A} = \mathbf{u}_1 \mathbf{u}_2^*$ of $\mathbf{u}_1$ and $\mathbf{u}_2$ when both are finite dimensional column vectors. An alternative and more direct way to define the mean element

and covariance operator is via the Bochner integral (cf. Definition 2.5 and [49]):

$$\mathbf{m}_{\mathbf{x}} = \int_{\Omega} \mathbf{x} \, \mathrm{d}\mathbb{P}, \tag{2.18}$$

$$\mathbf{C}_{\mathbf{x}} = \int_{\Omega} (\mathbf{x} - \mathbf{m}_{\mathbf{x}}) \circledast (\mathbf{x} - \mathbf{m}_{\mathbf{x}}) \, \mathrm{d}\mathbb{P}. \tag{2.19}$$

We can interpret $\mathbf{m}_{\mathbf{x}}$ as the (generalized) mean of the random element or vector $\mathbf{x}$. For $\mathbf{C}_{\mathbf{x}}$, we note that $\mathbb{E}[(\mathbf{x} - \mathbf{m}_{\mathbf{x}}) \circledast (\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{u})] = \mathbb{E}[\langle \mathbf{u}, \mathbf{x} - \mathbf{m}_{\mathbf{x}} \rangle (\mathbf{x} - \mathbf{m}_{\mathbf{x}})]$, which is nothing but the application of the matrix $\mathbb{E}[(\mathbf{x} - \mathbf{m}_{\mathbf{x}})(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^*]$ to $\mathbf{u}$ when $\mathbf{x}$ is a finite dimensional random vector in a Euclidean space. It can be shown that covariance operators are bounded, trace-class and self-adjoint (cf. Definitions 2.2 and 2.3).

The cross-covariance operator can also be defined via Bochner integral. If $\mathbb{E} \left\| (\mathbf{x}, \mathbf{y}) \right\|^2 < \infty$, the cross-covariance operator $\mathbf{C}_{\mathbf{xy}} : \mathcal{H}_2 \mapsto \mathcal{H}_1$ is defined as

$$\mathbf{C}_{\mathbf{xy}} = \int_{\Omega} (\mathbf{x} - \mathbf{m}_{\mathbf{x}}) \circledast (\mathbf{y} - \mathbf{m}_{\mathbf{y}}) \, \mathrm{d}\mathbb{P}. \tag{2.20}$$

Similarly, it can be verified that this definition also degenerates to the standard definitions of mean and covariance of random vectors in finite dimensional Hilbert spaces. By Assumption 3.1, the covariance and cross-covariance operators are well-defined as Bochner integrals.

Due to the fact that $\mathbf{C}_{\mathbf{x}}$ is self-adjoint and trace-class, its trace can be computed given an arbitrary orthonormal basis $\{\tilde{\boldsymbol{\psi}}_k\}_{k=1}^{\infty}$of $\mathcal{H}$:

$$\mathrm{tr}(\mathbf{C}_{\mathbf{x}}) = \sum_{k=1}^{\infty} \langle \mathbf{C}_{\mathbf{x}}(\tilde{\boldsymbol{\psi}}_k), \tilde{\boldsymbol{\psi}}_k \rangle. \tag{2.21}$$

The trace of the covariance matrix of a random vector in a finite dimensional space equals to the expectation of its squared norm. Analogously, this fact holds true for a random element, as shown below. Proposition 2.3 is needed in the MSE analysis in the main paper.

Proposition 2.3. *For a random element* $\mathbf{x}$ *with* $\mathbf{m}_{\mathbf{x}} = 0$, $\mathbb{E} \left\| \mathbf{x} \right\|^2 = \mathrm{tr}(\mathbf{C}_{\mathbf{x}})$.

Proof. Suppose that $\{\tilde{\boldsymbol{\psi}}_k\}_{k=1}^{\infty}$ is an orthonormal basis of $\mathcal{H}$. Then,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{x}\|^2 &= \int_{\mathcal{H}} \sum_{k=1}^{\infty} |\langle \mathbf{u}, \tilde{\boldsymbol{\psi}}_k \rangle|^2 \,\mathrm{d}\mathbb{P}_{\mathbf{x}}(\mathbf{u}) \\ &= \sum_{k=1}^{\infty} \int_{\mathcal{H}} |\langle \mathbf{u}, \tilde{\boldsymbol{\psi}}_k \rangle|^2 \,\mathrm{d}\mathbb{P}_{\mathbf{x}}(\mathbf{u}) \\ &= \sum_{k=1}^{\infty} \langle \mathbf{C}_{\mathbf{x}} \tilde{\boldsymbol{\psi}}_k, \tilde{\boldsymbol{\psi}}_k \rangle \\ &= \mathrm{tr}(\mathbf{C}_{\mathbf{x}}),\end{aligned}$$

where the second inequality follows from the monotone convergence theorem, and the third equality from (2.16). □

For a random vector $\mathbf{x}$ in a finite dimensional space, one can obtain principal axes by eigendecomposition of its covariance matrix $\mathbf{C}_{\mathbf{x}}$ to perform principal component analysis (PCA). By projecting $\mathbf{x}$ onto different principal axes, it is decomposed into uncorrelated components. For a random element $\mathbf{x}$ we have the same result as follows.

Proposition 2.4. *[49, Theorem 7.2.6, Theorem 7.2.7]* $\mathbf{C}_{\mathbf{x}}$ *admits the eigendecomposition*

$$\mathbf{C}_{\mathbf{x}} = \sum_{i=1}^{\infty} \sigma_i \mathbf{h}_i \circledast \mathbf{h}_i,$$

where $\{\sigma_i\}_{i=1}^{\infty}$ *are eigenvalues of* $\mathbf{C}_{\mathbf{x}}$ *and* $\{\mathbf{h}_i\}_{i=1}^{\infty}$ *the corresponding orthonormal eigenvectors. The eigenvectors* $\{\mathbf{h}_i\}_{i=1}^{\infty}$ *form an orthonormal basis for the closure of the image of* $\mathbf{C}_{\mathbf{x}}$, $\overline{\mathrm{im}\,\mathbf{C}_{\mathbf{x}}}$. *The random element* $\mathbf{X} \in \overline{\mathrm{im}\,\mathbf{C}_{\mathbf{x}}}$ *almost surely, and can be written as*

$$\mathbf{x} = \sum_{i=1}^{\infty} \langle \mathbf{x}, \mathbf{h}_i \rangle \mathbf{h}_i,$$

where $\{\langle \mathbf{x}, \mathbf{h}_i \rangle\}_{i=1}^{\infty}$ *are uncorrelated random variables with zero means and variances* σ_i.

The covariance operator behaves similarly as the covariance matrix under linear transformation. We list a few of their properties here.

Proposition 2.5. *Suppose* $\mathbf{x} : \Omega \to \mathcal{H}_1$ *and* $\mathbf{y} : \Omega \to \mathcal{H}_2$ *have zero means, and* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *are bounded linear operators on* $\mathcal{H}_1$ *and* $\mathcal{H}_2$*, respectively.*

(a) Let $\mathbf{z}_1 = \mathbf{T}_1\mathbf{x}$*. Then* $\mathbf{C}_{\mathbf{z}_1} = \mathbf{T}_1\mathbf{C}_\mathbf{x}\mathbf{T}_1^*$.

(b) Let $\mathbf{z}_2 = \mathbf{T}_2\mathbf{y}$*. Then* $\mathbf{C}_{\mathbf{z}_1\mathbf{z}_2} = \mathbf{T}_1\mathbf{C}_{\mathbf{xy}}\mathbf{T}_2^*$.

(c) If $\mathbf{x}$ *and* $\mathbf{y}$ *are independent, then* $\mathbf{C}_{\mathbf{x}+\mathbf{y}} = \mathbf{C}_\mathbf{x} + \mathbf{C}_\mathbf{y}, \mathbf{C}_{\mathbf{xy}} = \mathbf{0}$.

Proof. From the definition of a covariance operator, for any $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{H}$ we have

$$\begin{aligned}
\langle \mathbf{C}_{\mathbf{z}_1}\mathbf{u}_1, \mathbf{u}_2\rangle &= \mathbb{E}\left[\overline{\langle \mathbf{z}_1, \mathbf{u}_1\rangle}\langle \mathbf{z}_1, \mathbf{u}_2\rangle\right] \\
&= \mathbb{E}\left[\overline{\langle \mathbf{T}_1\mathbf{x}, \mathbf{u}_1\rangle}\langle \mathbf{T}_1\mathbf{x}, \mathbf{u}_2\rangle\right] \\
&= \mathbb{E}\left[\overline{\langle \mathbf{x}, \mathbf{T}_1^*\mathbf{u}_1\rangle}\langle \mathbf{x}, \mathbf{T}_1^*\mathbf{u}_2\rangle\right] \\
&= \langle \mathbf{C}_\mathbf{x}\mathbf{T}_1^*\mathbf{u}_1, \mathbf{T}_1^*\mathbf{u}_2\rangle \\
&= \langle \mathbf{T}_1\mathbf{C}_\mathbf{x}\mathbf{T}_1^*\mathbf{u}_1, \mathbf{u}_2\rangle,
\end{aligned}$$

yielding the result of claim (a). The proof of claim (b) is similar and omitted here.

For claim (c), due to the independence of $\mathbf{x}$ and $\mathbf{y}$, we have

$$\begin{aligned}
\langle \mathbf{C}_{\mathbf{xy}}\mathbf{u}_1, \mathbf{u}_2\rangle &= \mathbb{E}\left[\overline{\langle \mathbf{y}, \mathbf{u}_1\rangle}\langle \mathbf{x}, \mathbf{u}_2\rangle\right] \\
&= \overline{\mathbb{E}[\langle \mathbf{y}, \mathbf{u}_1\rangle]}\mathbb{E}[\langle \mathbf{x}, \mathbf{u}_2\rangle] \\
&= 0.
\end{aligned}$$

This implies that $\mathbf{C}_{\mathbf{xy}} = \mathbf{0}$. By plugging this result into the definition of $\mathbf{C}_{\mathbf{x}+\mathbf{y}}$ to eliminate the cross terms, we obtain $\mathbf{C}_{\mathbf{x}+\mathbf{y}} = \mathbf{C}_\mathbf{x} + \mathbf{C}_\mathbf{y}$. □

The conditional expectation and covariance of a random element are defined as follows [51, Section II.4.1], [55]:

Definition 2.6. Suppose the random element $\mathbf{x}$ takes values in a separable Hilbert space $\mathcal{H}$, $\mathbb{E}[\|\mathbf{x}\|] < \infty$, and $\mathcal{F}'$ is a sub σ-algebra of $\mathcal{F}$. The conditional expectation of $\mathbf{x}$ w.r.t. $\mathcal{F}'$ is the random element $\mathbf{x}_{\text{cond}} \in \mathcal{F}'$ such that $\mathbb{E}[\|\mathbf{x}_{\text{cond}}\|] < \infty$ and

$$\mathbb{E}[\mathbf{x}_{\text{cond}}\mathbb{I}_A] = \mathbb{E}[\mathbf{x}\mathbb{I}_A], \ \forall\, A \in \mathcal{F}', \tag{2.22}$$

where $\mathbb{I}_A$ is the indicator function on the set A. We denote $\mathbf{x}_{\text{cond}}$ by $\mathbb{E}[\mathbf{x} \,|\, \mathcal{F}']$. According to [51, Proposition 4.1], $\mathbb{E}[\mathbf{x} \,|\, \mathcal{F}']$ always exists.

The conditional covariance is defined as

$$\operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2 \,|\, \mathcal{F}') = \mathbb{E}[(\mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1 \,|\, \mathcal{F}']) \circledast (\mathbf{x}_2 - \mathbb{E}[\mathbf{x}_2 \,|\, \mathcal{F}']) \,|\, \mathcal{F}']$$

We write $\operatorname{cov}(\mathbf{x}, \mathbf{x} \,|\, \mathcal{F}')$ as $\operatorname{cov}(\mathbf{x} \,|\, \mathcal{F}')$ for simplicity.

By the defining property (2.22) of conditional expectation it can be shown that

$$\begin{aligned} \langle \mathbb{E}[\mathbf{x} \,|\, \mathcal{F}'], \mathbf{h} \rangle &= \mathbb{E}[\langle \mathbf{x}, \mathbf{h} \rangle \,|\, \mathcal{F}'], \\ \langle \mathbb{E}[\mathbf{x}_1 \circledast \mathbf{x}_2 \,|\, \mathcal{F}'](\mathbf{h}_2), \mathbf{h}_1 \rangle &= \mathbb{E}[\langle \mathbf{x}_1, \mathbf{h}_1 \rangle \langle \mathbf{x}_2, \mathbf{h}_2 \rangle \,|\, \mathcal{F}'], \end{aligned} \tag{2.23}$$

for all $\mathbf{h} \in \mathcal{H}$, $\mathbf{h}_1 \in \mathcal{H}_1$, $\mathbf{h}_2 \in \mathcal{H}_2$. From (2.23) we know that $\mathbb{E}[\mathbf{x} \,|\, \mathcal{F}']$ is uniquely defined. Let $\mathcal{F}''$ be a sub σ-algebra of $\mathcal{F}'$. Like random variables, the random elements also satisfy the property [51, Section II.4.1]:

$$\mathbb{E}[\mathbb{E}[\mathbf{x} \,|\, \mathcal{F}'] \,|\, \mathcal{F}''] = \mathbb{E}[\mathbf{x} \,|\, \mathcal{F}''].$$

Let $(\mathcal{I}, \mathcal{F}_{\mathcal{I}}, \mu_{\mathcal{I}})$ be a σ-finite measure space. The stochastic process $\{f(\omega, \boldsymbol{\xi}) : \omega \in \Omega, \boldsymbol{\xi} \in \mathcal{I}\}$ can be modeled as a random element if it satisfies regularity conditions:

Theorem 2.1. *[56, Theorem 2] Suppose*

1. *f is a $\mu \times \mu_{\mathcal{I}}$-measurable stochastic process.*
2. *the paths of f are in $L^2(\mathcal{I})$.*

Then the map

$$\begin{aligned} \Omega &\to L^2(\mathcal{I}) \\ \omega &\mapsto f(\omega, \cdot) \end{aligned} \tag{2.24}$$

is a random element with mean element $\mathbb{E}[f(\boldsymbol{\xi})] \in L^2(\mathcal{I})$. Its covariance operator $\mathbf{C}_f$ is the integral operator with kernel $\operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2))$. Specifically, if f is Gaussian process (GP), then (2.24) is a Gaussian random element, i.e., composing any linear functional with it will yield a Gaussian random variable.

If we further assume that $\mathcal{I}$ is a compact metric space and $\mu_{\mathcal{I}}$ is a strictly positive Borel measure, and the function $\operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2))$ is continuous on $\mathcal{I} \times \mathcal{I}$, then it can be shown by Mercer's theorem [57] that

$$\operatorname{tr}(\mathbf{C}_f) = \int_{\mathcal{I}} \operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2)) \, \mathrm{d}\mu_{\mathcal{I}}. \tag{2.25}$$

In this thesis we will also make use of the following theorem which is more general than Theorem 2.1. The proof of it is included in Section 2.6 for completeness.

Theorem 2.2. *Suppose a stochastic process f satisfies condition 1 and condition 2 in Theorem 2.1. $\mathcal{F}'$ is a sub σ-algebra of the underlying probability space. Suppose f and $\mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'] \in L^2(\Omega \times \mathcal{I})$. Then*

$$\mathbb{E}[f \,|\, \mathcal{F}'] = \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], \tag{2.26}$$

$$\operatorname{cov}(f \,|\, \mathcal{F}') : L^2(\mathcal{I}) \to L^2(\mathcal{I}),$$

$$g(\cdot) \mapsto \int_{\mathcal{I}} \operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}') g(\boldsymbol{\xi}_2) \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_2). \tag{2.27}$$

Let $\operatorname{var}(f(\boldsymbol{\xi}) \,|\, \mathcal{F}')$ be the conditional variance of the random variable $f(\boldsymbol{\xi})$. If we further assume that $\mathcal{I}$ is a compact metric space, and $\operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}')$ is continuous w.r.t. $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, then we have

$$\begin{aligned} \operatorname{tr}(\operatorname{cov}(f \,|\, \mathcal{F}')) &= \mathbb{E}\left[\| f - \mathbb{E}[f \,|\, \mathcal{F}'] \|^2 \,\middle|\, \mathcal{F}' \right] \\ &= \int_{\mathcal{I}} \operatorname{var}(f(\boldsymbol{\xi}) \,|\, \mathcal{F}') \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}). \end{aligned} \tag{2.28}$$

In the above formulas, the left-hand side (L.H.S.) are defined by moments of f as a random element. The moments in right-hand side (R.H.S.) are defined pointwise, as functions on $\mathcal{I}$ or $\mathcal{I} \times \mathcal{I}$.

In this thesis, the index set $\mathcal{I}$ can be $\mathcal{V} \times \mathcal{T}$ or a subset of $\mathcal{V} \times \mathcal{T}$. We always assume that the conditions in Theorem 2.1 are met for the stochastic processes in concern.

2.4 Linear Conditional Expectation in Hilbert space

In this section, we present the concept of a linear conditional expectation (LCE) in Hilbert spaces, and theorems from [55] that characterize it. The contents are simplified to fit this thesis. Readers are referred to [55] for details.

Let $L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{H})$ be the space of random elements taking value in $\mathcal{H}$. Consider two random elements $\mathbf{x}$ and $\mathbf{y}$ belonging to $L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{H})$. Let $\mathsf{L}(\mathcal{H}; \mathcal{H})$ be the set of all bounded linear operators on $\mathcal{H}$. We first define several specific operator spaces:

$$\begin{aligned}
\mathsf{A}(\mathcal{H}; \mathcal{H}) &:= \{\mathbf{L} : \mathcal{H} \mapsto \mathcal{H} \,|\, \mathbf{L}(\mathbf{h}) = \mathbf{b} + \mathbf{A}(\mathbf{h}) \text{ for some } \mathbf{b} \in \mathcal{H}, \mathbf{A} \in \mathsf{L}(\mathcal{H}; \mathcal{H})\}, \\
\mathsf{L}_{\mathbf{y}}(\mathcal{H}, \mathcal{H}) &:= \left\{\mathbf{L} : \mathcal{H} \mapsto \mathcal{H} \,\middle|\, \mathbf{L} \text{ is linear and } \mathbf{L}(\mathbf{y}) \in L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{H})\right\}, \\
\mathsf{A}_{\mathbf{y}}(\mathcal{H}; \mathcal{H}) &:= \{\mathbf{L} : \mathcal{H} \mapsto \mathcal{H} \,|\, \mathbf{L}(\mathbf{h}) = \mathbf{b} + \mathbf{A}(\mathbf{h}) \text{ for some } \mathbf{b} \in \mathcal{H}, \mathbf{A} \in \mathsf{L}_{\mathbf{y}}(\mathcal{H}; \mathcal{H})\}.
\end{aligned}$$

Definition 2.7. The *LCE* $\mathbb{E}^{\mathsf{A}}[\mathbf{x} \,|\, \mathbf{y}]$ is defined to be

$$\mathbb{E}^{\mathsf{A}}[\mathbf{x} \,|\, \mathbf{y}] = \mathbf{\Pi}_{\overline{\mathsf{A}_{\mathcal{H}}(\mathbf{y})}}(\mathbf{x}),$$

wherein $\overline{\mathsf{A}_{\mathcal{H}}(\mathbf{y})}$ is the closure of the space $\{\mathbf{z} \,|\, \mathbf{z} = \mathbf{L}(\mathbf{y}) \text{ for some } \mathbf{L} \in \mathsf{A}(\mathcal{H}; \mathcal{H})\}$ in $L^2(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{H})$. Recall that $\mathbf{\Pi}_{\overline{\mathsf{A}_{\mathcal{H}}(\mathbf{y})}}$ denotes the projection operator onto $\overline{\mathsf{A}_{\mathcal{H}}(\mathbf{y})}$.

Proposition 2.6. $\mathbb{E}^{\mathsf{A}}[\mathbf{x} \,|\, \mathbf{y}]$ *is of the form* $\mathbf{G}(\mathbf{y})$*, where* $\mathbf{G} \in \mathsf{A}_{\mathbf{y}}(\mathcal{H}; \mathcal{H})$*.*

If the range inclusion $\mathrm{im}(\mathbf{C}_{\mathbf{yx}}) \subset \mathrm{im}(\mathbf{C}_{\mathbf{y}})$ holds, we call it the compatible case.

Proposition 2.7 (Formula for the LCE: compatible case)**.** *Under the compatible case, the LCE has the explicit formula*

$$\mathbf{G}(\mathbf{u}) = \mathbf{m}_{\mathbf{x}} + (\mathbf{C}_{\mathbf{y}}^{\dagger}\mathbf{C}_{\mathbf{yx}})^{*}(\mathbf{u} - \mathbf{m}_{\mathbf{y}}). \tag{2.29}$$

Proposition 2.8. *The condition* $\mathrm{im}(\mathbf{C}_{\mathbf{yx}}) \subset \mathrm{im}(\mathbf{C}_{\mathbf{y}})$ *holds if* $\mathrm{im}(\mathbf{C}_{\mathbf{y}})$ *is closed. This condition is trivially met when* $\dim \mathcal{H} < \infty$*.*

In the non-compatible case, $\mathbb{E}^{\mathsf{A}}[\mathbf{x} \,|\, \mathbf{y}]$ can be approximated by a sequence of finite-rank operators composed with $\mathbf{y}$. According to Proposition 2.4, suppose $\mathbf{C}_{\mathbf{y}}$ admits

the eigen-decomposition

$$\mathbf{C}_{\mathbf{y}} = \sum_{i=1}^{\infty} \sigma_i \mathbf{h}_i \circledast \mathbf{h}_i.$$

For every $m \in \mathbb{N}$, let $\mathcal{H}^{(m)} := \text{span}\{\mathbf{h}_1, \cdots, \mathbf{h}_m\}$, and $\mathbf{y}^{(m)} := \mathbf{\Pi}_{\mathcal{H}^{(m)}}\mathbf{y}$. We define a sequence of operators $\{\mathbf{G}^{(m)}\}_{m\geq 1}$ as

$$\mathbf{G}^{(m)}(\mathbf{u}) := \mathbf{m}_{\mathbf{x}} + (\mathbf{C}^{\dagger}_{\mathbf{y}^{(m)}}\mathbf{C}_{\mathbf{y}^{(m)}\mathbf{x}})^*(\mathbf{u} - \mathbf{m}_{\mathbf{y}}).$$

It can be shown that the sequence $\{\mathbf{G}^{(m)}(\mathbf{y})\}_{m\geq 1}$ is a good approximation of $\mathbb{E}^{\mathsf{A}}[\mathbf{x}\,|\,\mathbf{y}]$, as captured in the following result.

Proposition 2.9 (Formula for the LCE: incompatible case)**.** *$\{\mathbf{G}^{(m)}(\mathbf{y})\}_{m\geq 1}$ converges to $\mathbb{E}^{\mathsf{A}}[\mathbf{x}\,|\,\mathbf{y}]$ in L^2 norm, i.e.,*

$$\mathbb{E}\left\|\mathbb{E}^{\mathsf{A}}[\mathbf{x}\,|\,\mathbf{y}] - \mathbf{G}^{(m)}(\mathbf{y})\right\|^2 \to 0,$$

as $m \to \infty$. Let $\mathbb{P}_{\mathbf{y}}$ denote the probability measure induced by $\mathbf{y}$ on $\mathcal{H}$. Then,

$$\left\|\mathbf{G}^{(m)}(\mathbf{u}) - \mathbb{E}^{\mathsf{A}}[\mathbf{x}\,|\,\mathbf{y} = \mathbf{u}]\right\| \xrightarrow{\text{a.s.}} 0$$

$\mathbb{P}_{\mathbf{y}}$-almost surely.

We can define and measure the estimation error of the LCE as the average linear conditional operator (ALCC).

Definition 2.8. Let $R^{\mathsf{A}}[\mathbf{x} \mid \mathbf{y}] := \mathbf{x} - \mathbb{E}^{\mathsf{A}}[\mathbf{x}\,|\,\mathbf{y}]$. The ALCC of $\mathbf{x}$ given $\mathbf{y}$ is defined as

$$\text{cov}^{\mathsf{A}}_{\mathbf{y}}[\mathbf{x}] := \mathbb{E}\big[R^{\mathsf{A}}[\mathbf{x} \mid \mathbf{y}] \circledast R^{\mathsf{A}}[\mathbf{x} \mid \mathbf{y}]\big].$$

In the compatible case, $\text{cov}^{\mathsf{A}}_{\mathbf{y}}[\mathbf{x}]$ can be computed as

$$\text{cov}^{\mathsf{A}}_{\mathbf{y}}[\mathbf{x}] = \mathbf{C}_{\mathbf{x}} - \mathbf{C}_{\mathbf{xy}}\mathbf{C}^{\dagger}_{\mathbf{y}}\mathbf{C}_{\mathbf{yx}}.$$

2.5 KRR Reconstruction and Interpretation

KRR is a supervised learning approach that aims to learn a map from $\mathcal{X}$ to $\mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$. Given a set of training inputs and outputs, it searches for the best fitting function in a RKHS. Given a symmetric positive semi-definite kernel

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\to \mathcal{Y} \\ (\mathbf{x}, \mathbf{x}') &\mapsto k(\mathbf{x}, \mathbf{x}'), \end{aligned}$$

the associated RKHS $\mathcal{H}_k$ is defined as the Hilbert space satisfying [58, Definition 1]:

1. $k(\cdot, \mathbf{x}) \in \mathcal{H}_k$ for all $\mathbf{x} \in \mathcal{X}$.
2. $\langle g, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k} = g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and $g \in \mathcal{H}_k$.

According to the Moore-Aronszajn theorem [58, Theorem 3], there exists a unique Hilbert space $\mathcal{H}_k$ satisfying these conditions. When $\mathcal{X}$ is a subset of Euclidean space, typical choices for k include the polynomial kernel ($k(\mathbf{x}, \mathbf{x}') = (a\mathbf{x}^\intercal \mathbf{x}' + 1)^b$ with parameters $a \in \mathbb{R}$, $b \in \mathbb{N}$), linear kernel (polynomial kernel with $a = 1, b = 1$), and RBF kernel ($k(\mathbf{x}, \mathbf{x}')$ is a function of $\|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}}$).

Given a training set $\{(\mathbf{x}_m, y_m) : \mathbf{x}_m \in \mathcal{X}, y_m \in \mathcal{Y}, m = 1, \ldots, M\}$, KRR searches for an optimal function in $\mathcal{H}_k$ to fit the data by solving for

$$\hat{f} = \arg\min_{\tilde{f} \in \mathcal{H}_k} \sum_{m=1}^{M} \left| \tilde{f}(\mathbf{x}_m) - y_m \right|^2 + \mu J(\left\| \tilde{f} \right\|_{\mathcal{H}_k}), \tag{2.30}$$

where $J(\cdot)$ is an increasing function, and μ is a penalty weight. The representer theorem [59, Theorem 4.2] states that the optimal solution to (2.30) takes the form

$$\hat{f} = \sum_{m=1}^{M} c_m k(\cdot, \mathbf{x}_m), \tag{2.31}$$

where c_m, $m = 1, \ldots, M$, are coefficients to be determined. By substituting (2.31) into (2.30), the problem (2.30) becomes an optimization over $\{c_m\}_{m=1}^M$. Specifically, when $J(\cdot) = (\cdot)^2$, problem (2.30) is quadratic and its solution is given by

$$(c_1, \ldots, c_M)^\intercal = (\mathbf{K} + \mu \mathbf{I}_M)^{-1} \mathbf{y}, \tag{2.32}$$

where $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^M \in \mathbb{R}^{M\times M}$ and $\mathbf{y} = (y_1, \ldots, y_M)^\intercal$. In the sequel, we assume $J(\cdot) = (\cdot)^2$ unless otherwise stated. When k is chosen as the linear kernel, (2.30) is equivalent to learning a linear function from $\mathcal{X}$ to $\mathcal{Y}$, i.e., linear regression.

It is natural to consider whether we can recover any continuous function pointwise to within arbitrary fidelity with a sufficiently large number of samples by KRR. This is achievable by employing a *universal* kernel k [60]. Let $\mathcal{X}$ be a Hausdorff topological space (e.g., $\mathbb{R}$) and $\mathcal{Z} \subset \mathcal{X}$ be a compact subset (e.g., $[a, b]$). Let $\mathcal{C}(\mathcal{Z})$ be the space of continuous functions on $\mathcal{Z}$ with the supremum norm. Define $\mathcal{K}(\mathcal{Z}) := \overline{\text{span}}\{k(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}\}$, where the closure is taken w.r.t. the norm in $\mathcal{C}(\mathcal{Z})$. The kernel k is said to be universal if $\mathcal{K}(\mathcal{Z}) = \mathcal{C}(\mathcal{Z})$ for any compact $\mathcal{Z} \subset \mathcal{X}$. In other words, $\text{span}\{k(\cdot, \mathbf{x}) : \mathbf{x} \in \mathcal{Z}\}$ is dense in $\mathcal{C}(\mathcal{Z})$.

Problem (2.30) has a Bayesian interpretation. Consider a GP w with mean function zero and covariance function $k(\mathbf{x}, \mathbf{x}')$, denoted as $w \sim \mathcal{GP}(0, k)$. Given the noisy observations $y_m = w(\mathbf{x}_m) + \epsilon_m$, $\epsilon_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mu)$, the maximum a posteriori (MAP) estimator of $w(\mathbf{x})$ is $\hat{f}(\mathbf{x})$ as defined in (2.31) and (2.32) for any $\mathbf{x} \in \mathcal{X}$.

The readers are referred to [58, 61] for more detailed discussions on RKHS and KRR.

2.6 Appendix: Proof of Theorem 2.2

Proof. We first prove (2.26) and (2.27). According to (2.23), it suffices to prove that for any $A \in \mathcal{F}'$,

$$\mathbb{E}[\langle \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h(\boldsymbol{\xi})\rangle 1_A] = \mathbb{E}[\langle f, h\rangle 1_A], \tag{2.33}$$

$$\begin{aligned}&\mathbb{E}[\langle f(\boldsymbol{\xi}) - \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h_1(\boldsymbol{\xi})\rangle \langle f(\boldsymbol{\xi}) - \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h_2(\boldsymbol{\xi})\rangle 1_A] \\ &= \mathbb{E}[\langle f - \mathbb{E}[f \,|\, \mathcal{F}'], h_1\rangle \langle f - \mathbb{E}[f \,|\, \mathcal{F}'], h_2\rangle 1_A],\end{aligned} \tag{2.34}$$

where 1_A is the indicator function of the set A. We first prove (2.33) as follows:

$$\begin{aligned}\mathbb{E}[\langle \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h(\boldsymbol{\xi})\rangle 1_A] &= \int_A \int_{\mathcal{I}} \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'] h(\boldsymbol{\xi}) \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}) \, \mathrm{d}\mathbb{P} \\ &= \int_{\mathcal{I}} \int_A \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'] h(\boldsymbol{\xi}) \, \mathrm{d}\mathbb{P} \, \mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi})\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{I}} \mathbb{E}[\mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}']1_A] h(\boldsymbol{\xi}) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}) \\
&= \int_{\mathcal{I}} \int_A f(\boldsymbol{\xi}) h(\boldsymbol{\xi}) \,\mathrm{d}\mathbb{P} \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}) \\
&= \int_A \int_{\mathcal{I}} f(\boldsymbol{\xi}) h(\boldsymbol{\xi}) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}) \,\mathrm{d}\mathbb{P} \\
&= \mathbb{E}[\langle f, h \rangle 1_A].
\end{aligned}
$$

The integrals are exchangeable by Fubini's theorem. Similarly we prove (2.34) as follows:

$$
\begin{aligned}
&\mathbb{E}[\langle f(\boldsymbol{\xi}) - \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h_1(\boldsymbol{\xi}) \rangle \langle f(\boldsymbol{\xi}) - \mathbb{E}[f(\boldsymbol{\xi}) \,|\, \mathcal{F}'], h_2(\boldsymbol{\xi}) \rangle 1_A] \\
&= \int_A \int_{\mathcal{I}} \int_{\mathcal{I}} (f(\boldsymbol{\xi}_1) - \mathbb{E}[f(\boldsymbol{\xi}_1) \,|\, \mathcal{F}']) h_1(\boldsymbol{\xi}_1) \\
&(f(\boldsymbol{\xi}_2) - \mathbb{E}[f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}']) h_2(\boldsymbol{\xi}_2) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_1) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_2) \,\mathrm{d}\mathbb{P} \\
&= \int_{\mathcal{I}} \int_{\mathcal{I}} \mathbb{E}[(f(\boldsymbol{\xi}_1) - \mathbb{E}[f(\boldsymbol{\xi}_1) \,|\, \mathcal{F}'])(f(\boldsymbol{\xi}_2) - \mathbb{E}[f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}'])1_A] \\
&h_1(\boldsymbol{\xi}_1) h_2(\boldsymbol{\xi}_2) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_1) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_2) \\
&= \int_A \int_{\mathcal{I}} \int_{\mathcal{I}} (f(\boldsymbol{\xi}_1) - \mathbb{E}[f(\boldsymbol{\xi}_1) \,|\, \mathcal{F}'])(f(\boldsymbol{\xi}_2) - \mathbb{E}[f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}']) \\
&h_1(\boldsymbol{\xi}_1) h_2(\boldsymbol{\xi}_2) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_1) \,\mathrm{d}\mu_{\mathcal{I}}(\boldsymbol{\xi}_2) \,\mathrm{d}\mathbb{P} \\
&= \mathbb{E}[\langle f - \mathbb{E}[f \,|\, \mathcal{F}'], h_1 \rangle \langle f - \mathbb{E}[f \,|\, \mathcal{F}'], h_2 \rangle 1_A].
\end{aligned}
$$

To prove (2.28), we first note that

$$
\begin{aligned}
\mathbb{E}\Big[\|f - \mathbb{E}[f \,|\, \mathcal{F}']\|^2 \, 1_A\Big] &= \int_A \sum_{i=1}^{\infty} \langle f - \mathbb{E}[f \,|\, \mathcal{F}'], e_i \rangle^2 \,\mathrm{d}\mathbb{P} \\
&= \sum_{i=1}^{\infty} \mathbb{E}\big[\langle f - \mathbb{E}[f \,|\, \mathcal{F}'], e_i \rangle^2 1_A\big] \\
&= \sum_{i=1}^{\infty} \mathbb{E}[\langle \mathrm{cov}(f \,|\, \mathcal{F}') e_i, e_i \rangle 1_A] \\
&= \mathbb{E}[\mathrm{tr}(\mathrm{cov}(f \,|\, \mathcal{F}'))1_A].
\end{aligned}
$$

Thus, $\mathbb{E}\big[\|f - \mathbb{E}[f \,|\, \mathcal{F}']\|^2 \,\big|\, \mathcal{F}'\big] = \mathbb{E}[\mathrm{tr}(\mathrm{cov}(f \,|\, \mathcal{F}')) \,|\, \mathcal{F}'] = \mathrm{tr}(\mathrm{cov}(f \,|\, \mathcal{F}'))$.

On the other hand, by Mercer's theorem [57], $\operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}')$ and $\operatorname{cov}(f \,|\, \mathcal{F}')$ can be decomposed as

$$\begin{aligned}\operatorname{cov}(f(\boldsymbol{\xi}_1), f(\boldsymbol{\xi}_2) \,|\, \mathcal{F}') &= \sum_{i=1}^{\infty} \varrho_i \psi_i(\boldsymbol{\xi}_1)\psi_i(\boldsymbol{\xi}_2), \\ \operatorname{cov}(f \,|\, \mathcal{F}')(h) &= \sum_{i=1}^{\infty} \varrho_i \langle \psi_i, h\rangle \psi_i, \forall\, h \in L^2(\mathcal{I}),\end{aligned}$$

where the convergence is uniform and absolute. $\varrho_i \geq 0$. Besides, $\{\psi_i\}$ forms an orthonormal system in $L^2(\mathcal{I})$. Therefore,

$$\operatorname{tr}(\operatorname{cov}(f \,|\, \mathcal{F}')) = \sum_{i=1}^{\infty} \langle \operatorname{cov}(f \,|\, \mathcal{F}')\psi_i, \psi_i\rangle = \sum_{i=1}^{\infty} \varrho_i,$$

and

$$\begin{aligned}\int_{\mathcal{I}} \operatorname{var}(f(\boldsymbol{\xi}) \,|\, \mathcal{F}') &= \int_{\mathcal{I}} \sum_{i=1}^{\infty} \varrho_i \psi_i(\boldsymbol{\xi}_1)\psi_i(\boldsymbol{\xi}_2)\mu_{\mathcal{I}}(\boldsymbol{\xi}) \\ &= \sum_{i=1}^{\infty} \varrho_i \int_{\mathcal{I}} \psi_i(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi})\mu_{\mathcal{I}}(\boldsymbol{\xi}) \\ &= \sum_{i=1}^{\infty} \varrho_i,\end{aligned}$$

which concludes the proof of (2.28). □

Chapter 3

Joint Wide-Sense Stationarity and Wiener Filters

3.1 GGSP Statistical Model

In this section, we present our model and assumptions. We define a GRP in the GGSP framework as a random element and provide conditions under which it exists. Under the same conditions, the covariance operator of the GRP is an integral operator. As a matrix kernel, its closed form enables us to interpret its trace via the generalized kernel method [62].

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected (weighted) graph, where $\mathcal{V} = \{1, \dots, N\}$ is the vertex set, and $\mathcal{E} \subset \{(i, j) \in \mathcal{V} \times \mathcal{V} : i < j\}$ denotes the edge set. Suppose that each vertex of G is associated with an element from a separable Hilbert space $\mathcal{H}$. A generalized graph signal [22] has the form $\mathbf{x} = (x_1, x_2, \dots, x_N)$ where for each $v \in \mathcal{V}$, we have $x_v \in \mathcal{H}$. For example, if $\mathcal{H} = L^2[a, b]$ the space of square integrable functions on a bounded interval $[a, b]$ (with the Borel σ-algebra and a given measure), then at each vertex v of the graph G, we have associated with it a function $x_v = x_v(\cdot) \in L^2[a, b]$.

It is shown in [22] that the space of all generalized graph signals can be identified with the Hilbert space $\mathbb{C}^N \otimes \mathcal{H}$ via the isomorphism

$$\mathbf{x} \cong \sum_{v=1}^{n} \mathbf{e}_v \otimes x_v,$$

where $\{\mathbf{e}_v : v = 1, \dots, N\}$ is the standard basis in $\mathbb{C}^N$. Let $|||$ denote the norm of $\mathbb{C}^N \otimes \mathcal{H}$.

Suppose $\mathcal{H} = L^2(\mathcal{T})$ for a measure space $\mathcal{T}$. For $\mathbf{x} \in \mathbb{C}^N \otimes \mathcal{H}$ and each $\mathbf{t} \in \mathcal{T}$, we write

$$\mathbf{x}(\mathbf{t}) = (x_1(\mathbf{t}), x_2(\mathbf{t}), \dots, x_N(\mathbf{t}))^\intercal, \tag{3.1}$$

a vector-valued function.

We next utilize the notion of a random element on a Hilbert space (cf . Section 2.3, [49–51]) to define a GRP.

Definition 3.1. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space, $\mathcal{F}$ a σ-algebra and $\mathbb{P}$ a probability measure. A *GRP* $\mathbf{x}$ is a measurable map (i.e., a random element) from $(\Omega, \mathcal{F})$ to $(\mathbb{C}^N \otimes \mathcal{H}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra induced by the norm of $\mathbb{C}^N \otimes \mathcal{H}$.

Definition 3.1 formally introduces the concept of random elements on the generalized graph signal space $\mathbb{C}^N \otimes \mathcal{H}$, which was not studied in [22].

To ensure the existence of a GRP $\mathbf{x}$'s first and second moments, we only consider those $\mathbf{x}$ satisfying the following assumption. $\mathbf{x}$ induces a probability measure $\mathbb{P}_\mathbf{x}$ on $(\mathbb{C}^N \otimes \mathcal{H}, \mathcal{B})$ given by

$$\mathbb{P}_\mathbf{x}(B) = \mathbb{P}\big(\mathbf{x}^{-1}(B)\big), \ \forall\, B \in \mathcal{B}.$$

Assumption 3.1.

$$\mathbb{E}\,\|\mathbf{x}\|^2 = \int_{\mathbb{C}^N \otimes \mathcal{H}} \|\mathbf{z}\|^2 \, \mathrm{d}\mathbb{P}_\mathbf{x}(\mathbf{z}) < \infty.$$

Under Assumption 3.1, $\mathbf{x}$'s mean element $\mathbf{m}_\mathbf{x}$ and covariance operator $\mathbf{C}_\mathbf{x}$ are defined as in (2.15) and (2.16). To preclude the pathological cases (e.g., an uncorrelated complex random process' real and imaginary parts can be correlated), in this chapter we only consider those random elements that are proper, i.e., satisfy the equivalent conditions in [63, Theorem 1]. Without loss of generality, we assume that any GRP $\mathbf{x}$ of interest has mean element $\mathbf{m}_\mathbf{x} = 0$ unless otherwise specified. Filters on $\mathbb{C}^N \otimes \mathcal{H}$, as in Section 2.1.2, are defined as bounded linear operators on $\mathbb{C}^N \otimes \mathcal{H}$.

As an example, in Definition 3.1, $\mathbf{x}$ can represent a multichannel signal. For instance, each vertex in a graph may correspond to a sensor station that records PM2.5, temperature and humidity levels, so that the recorded signal at each vertex is in $\mathcal{H} = \mathbb{R}^3$. If we take $\mathbf{x}$ to be the random vector observations at all vertices, it satisfies Definition 3.1 and is a GRP.

From the perspective of time-vertex analysis, it is also natural to suppose that each vertex signal can be a discrete- or continuous-time signal. We next investigate conditions under which the following stochastic process on G can be modeled as a GRP as in Definition 3.1: let $\mathcal{H} = L^2(\mathcal{T})$, where $\mathcal{T}$ is a set like $[a, b]$, and

$$\begin{aligned} \mathbf{x} : \Omega \times \mathcal{T} &\mapsto \mathbb{C}^N, \\ (\omega, \mathbf{t}) &\mapsto \mathbf{x}(\omega, \mathbf{t}). \end{aligned} \tag{3.2}$$

If $\mathbf{t}$ is treated as an index so that $\mathbf{x}(\cdot, \mathbf{t})$ is a function of ω, and $\mathbf{x}(\cdot, \mathbf{t})$ is measurable (w.r.t. ω) for every $\mathbf{t}$, then $\{\mathbf{x}(\cdot, \mathbf{t}) : \mathbf{t} \in [a, b]\}$ is a family of N-dimensional random vectors. We abbreviate $\mathbf{x}(\cdot, \mathbf{t})$ as $\mathbf{x}(\mathbf{t})$.

On the other hand, if ω is treated as an index so that $\mathbf{x}(\omega, \cdot)$ is a function of $\mathbf{t}$, $\mathbf{x}(\omega, \cdot)$ can be interpreted as an n-dimensional trajectory under the outcome ω. An immediate problem arises when no restrictions are imposed on $\mathbf{x}$. In this case, the trajectory $\tilde{\mathbf{x}}(\omega) = \mathbf{x}(\omega, \cdot)$ can be arbitrarily irregular, and is not guaranteed to belong to $\mathbb{C}^N \otimes L^2(\mathcal{T})$. Besides, even if the trajectories are restricted to $\mathbb{C}^N \otimes L^2(\mathcal{T})$, without further constraints, $\tilde{\mathbf{x}}(\omega)$ is not necessarily a measurable map from Ω to $\mathbb{C}^N \otimes L^2(\mathcal{T})$, and hence may not fit the definition of a GRP in Definition 3.1. To overcome these problems, we provide a sufficient condition in Theorem 3.1 to model $\tilde{\mathbf{x}}$ as a random element in $\mathbb{C}^N \otimes L^2(\mathcal{T})$. In Theorem 3.2, we derive the mean element and covariance operator. The proofs of Theorem 3.1 and Theorem 3.2 are in Appendix 3.A.

Theorem 3.1. *Suppose $\mathcal{H} = L^2(\mathcal{T}, \mathcal{A}, \tau)$, where $\mathcal{A}$ is a σ-algebra and τ a measure. The map $\tilde{\mathbf{x}}(\omega) = \mathbf{x}(\omega, \cdot)$ for $\omega \in \Omega$, where $\mathbf{x}(\cdot, \cdot)$ is a map of the form (3.2), is a GRP for $\mathbb{C}^N \otimes \mathcal{H}$ if the following conditions hold:*

(a) $\mathbf{x}(\omega, \mathbf{t})$ is jointly measurable on $\Omega \times \mathcal{T}$ w.r.t. the product measure $\mathbb{P} \times \tau$.

(b) $\mathbf{x}(\omega, \cdot) \in \mathbb{C}^N \otimes \mathcal{H}$ for every $\omega \in \Omega$.

(c) *For every* $\mathbf{s}, \mathbf{t} \in \mathcal{T}$*, the pointwise mean* $\mathbf{m}_{\mathbf{x}}(\mathbf{t}) := \mathbb{E}[\mathbf{x}(\mathbf{t})] \in \mathbb{C}^N$ *and cross-covariance* $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) := \operatorname{cov}(\mathbf{x}(\mathbf{s}), \mathbf{x}(\mathbf{t})) \in \mathbb{C}^{N \times N}$ *are well-defined. Every entry of* $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})$ *as a function of* $(\mathbf{s}, \mathbf{t}) \in \mathcal{T} \times \mathcal{T}$ *belongs to* $L^1(\mathcal{T} \times \mathcal{T})$.

Theorem 3.1 encompasses a broad range of signals that can be modeled as GRPs on $\mathbb{C}^N \otimes \mathcal{H}$. For instance, let $\mathcal{T} = \mathbb{R}_+$, $\mathcal{A}$ the Borel σ-algebra and τ the Lebesgue measure. Let $\{\mathbf{x}(\omega, t) : t \geq 0\}$ be a family of random variables indexed by t. If $\mathbf{x}(\omega, \cdot)$ is continuous and condition (c) in Theorem 3.1 is met, then all conditions in Theorem 3.1 are satisfied (see [64, Proposition 1.13]). In another example, if $\mathcal{T} = \{1, 2, \ldots, d\}$ (i.e., $\mathcal{H} = \mathbb{R}^d$), the conditions in Theorem 3.1 are met as long as $\mathbf{x}(\mathbf{t})$ is a random vector with finite second moments for each $t = 1, \ldots, d$. Hereafter where there is no confusion, we will not distinguish the maps $\tilde{\mathbf{x}}$ and $\mathbf{x}$, and simply write $\mathbf{x}$ for both of them.

We next show that the mean element $\mathbf{m}_{\mathbf{x}}$ and covariance operator $\mathbf{C}_{\mathbf{x}}$ of $\mathbf{x}$ follow from Theorem 3.1. In particular, $\mathbf{C}_{\mathbf{x}}$ is an integral operator with kernel $\mathbf{K}_{\mathbf{x}}$.

Theorem 3.2. *Suppose the conditions in Theorem 3.1 hold. Then, the mean element of* $\mathbf{x}$ *as defined in* (2.15) *coincides with* $\mathbf{m}_{\mathbf{x}}(\mathbf{t})$ *in Theorem 3.1, and the covariance operator of* $\mathbf{x}$ *as defined in* (2.16) *is given by*

$$(\mathbf{C}_{\mathbf{x}}\mathbf{f})(\mathbf{s}) = \int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})\mathbf{f}(\mathbf{t})\, \mathrm{d}\tau(\mathbf{t})$$

for all $\mathbf{f} \in \mathbb{C}^N \otimes \mathcal{H}$*, where* $\mathbf{f}(\mathbf{t})$ *in the integral is in the form* (3.1)*. Furthermore, suppose* $\mathcal{T}$ *is a separable metric space and* τ *is a finite measure on the Borel* σ*-algebra* $\mathcal{B}(\mathcal{T})$ *of* $\mathcal{T}$*. If* $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})$ *is continuous in* $(\mathbf{s}, \mathbf{t})$*, the operator trace of* $\mathbf{C}_{\mathbf{x}}$ *agrees with the integral of its trace, i.e.,*

$$\operatorname{tr}(\mathbf{C}_{\mathbf{x}}) = \int_{\mathcal{T}} \operatorname{tr}(\mathbf{K}_{\mathbf{x}}(\mathbf{t}, \mathbf{t}))\, \mathrm{d}\tau(\mathbf{t}).$$

From Proposition 2.3, the deviation of a random element from its mean can be measured by $\operatorname{tr}(\mathbf{C}_{\mathbf{x}})$. To interpret Theorem 3.2, suppose $\mathcal{T}$ represents time domain. Then we have shown that this deviation amounts to performing integration of the pointwise variance on the time domain first, and then summing them up over all vertices, which also naturally measures the overall deviation.

3.2 Generalized Joint Wide-sense Stationarity

In this section, we develop the notion of WSS w.r.t. a shift operator for a GRP. We define WSS in different domains and study their relationships.

3.2.1 Joint WSS

Before we define the concept of joint WSS for a GRP, we briefly review the analogous concept of WSS for a time domain scalar-valued stochastic process $\mathbf{x} = (x_1, \ldots, x_m)$. This stochastic process $\mathbf{x}$ is said to be WSS if $\mathrm{cov}(x_i, x_j)$ only depends on $j - i$. Let $\mathbf{A}_T$ be the one-step shift right operator, with wrapping to the front. The eigendecomposition of $\mathbf{A}_T$ yields the discrete time Fourier transform (DFT) matrix with columns being its eigenvectors. Then WSS can be equivalently defined as x's covariance matrix being diagonalizable by the DFT matrix.

In the same spirit as the scalar case and noting that $\mathbf{A}_T$ is the adjacency matrix of a directed cyclic graph, the analogous concept of graph wide-sense stationarity (GWSS)[23–25] based on the GSO [34, 65] have been proposed. For GWSS, [24] provided equivalent definitions analogous to the aforementioned statements for WSS of time domain scalar-valued signals. In [25], GWSS is defined by localization of a graph kernel.

In this thesis, to motivate a reasonable shift operator on $\mathbb{C}^N \otimes \mathcal{H}$, we first investigate the product graph model. This model assumes the signal on each vertex to be a graph signal, hence is a special case where $\mathcal{H}$ is a space of graph signals. This model allows for parallelized and vectorized implementations, and reduces the computational complexity for filters [66], and is thus an important model in practice. For a (weighted) graph G, we let $\mathbf{D}_G$ be its (diagonal) degree matrix.

Example 3.1 (Product graph)**.** Suppose each vertex of the graph $G = (\mathcal{V}, \mathcal{E})$ observes a graph signal, which is defined on yet another graph $G' = (\mathcal{V}', \mathcal{E}')$. This can be interpreted in the traditional GSP framework as signals on the vertices of a product graph $G_{\mathcal{J}}$. There are multiple ways to construct the product graph $G_{\mathcal{J}}$, including the tensor product graph $G \otimes G'$ and the Cartesian product graph $G \times G'$. In the case where $G_{\mathcal{J}} = G \otimes G'$, the adjacency matrix and graph Laplacian of $G_{\mathcal{J}}$

can be written respectively as [67, 68]:

$$\mathbf{W}_{G\otimes G'} = \mathbf{W}_G \otimes \mathbf{W}_{G'}, \tag{3.3}$$

$$\mathbf{L}_{G\otimes G'} = \mathbf{D}_G \otimes \mathbf{L}_{G'} + \mathbf{L}_G \otimes \mathbf{D}_{G'} - \mathbf{L}_G \otimes \mathbf{L}_{G'}. \tag{3.4}$$

In the case where $G_{\mathcal{J}} = G \times G'$, they become

$$\mathbf{W}_{G\times G'} = \mathbf{W}_G \otimes \mathbf{I}_{|\mathcal{V}'|} + \mathbf{I}_{|\mathcal{V}|} \otimes \mathbf{W}_{G'}, \tag{3.5}$$

$$\mathbf{L}_{G\times G'} = \mathbf{L}_G \otimes \mathbf{I}_{|\mathcal{V}'|} + \mathbf{I}_{|\mathcal{V}|} \otimes \mathbf{L}_{G'}. \tag{3.6}$$

Suppose we are given self-adjoint and compact operators $\mathbf{A}_G$ on G and $\mathbf{A}_{\mathcal{H}}$ on $\mathcal{H}$. (From the Hilbert-Schmidt theorem on spectral decomposition, the choice of a self-adjoint and compact operator leads to an orthonormal eigenbasis consisting of eigenvectors of the chosen operator, which then allows us to define the Fourier transform. For finite dimensional spaces, compactness trivially holds.) Taking motivations from Example 3.1, a shift operator on $\mathbb{C}^N \otimes \mathcal{H}$ can be chosen as either $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ or $\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}} + \mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}$. The operator $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ is the preferred over $\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}} + \mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}$ because the former is compact and self-adjoint, while the latter is not for infinite dimensional $\mathcal{H}$. However, the eigenvectors of $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ are also the eigenvectors of $\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}} + \mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}$, hence the Fourier transform induced by both operators are the same. In the rest of this thesis, we adopt $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ as the shift operator for $\mathbb{C}^N \otimes \mathcal{H}$.

Suppose $\mathbf{A}_G$ has eigenvalues $\lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$ with corresponding eigenvectors $\{\boldsymbol{\phi}_k\}_{k=1}^N$, and $\mathbf{A}_{\mathcal{H}}$ has eigenvalues $\{\nu_l\}_{l=1}^\infty$ with corresponding eigenvectors $\{\psi_l\}_{l=1}^\infty$. Then, $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ has eigenvalues $\{\lambda_k \nu_l \,:\, k = 1, \ldots, N, \, l \geq 1\}$ and eigenvectors $\{\boldsymbol{\phi}_k \otimes \psi_l \,:\, k = 1, \ldots, N, \, l \geq 1\}$.

To simplify the exposition and to obtain a unique orthonormal eigenbasis (up to multiples of ± 1), similar to most of the GSP literature [24, 34], we make the following assumption throughout this chapter.

Assumption 3.2. The geometric multiplicity of each eigenvalue of $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ is one.

The convolution filter with coefficients $\{g_{k,l} : k = 1, \ldots, N, l \geq 1\}$ is defined as the pointwise multiplication operator in the frequency domain:

$$\mathbf{G}(\mathbf{x}) = \sum_{k,l} g_{k,l} \mathfrak{F}_{k,l}(\mathbf{x}) \boldsymbol{\phi}_k \otimes \psi_l = \sum_{k,l} g_{k,l} \boldsymbol{\Pi}_{\boldsymbol{\phi}_k \otimes \psi_l}(\mathbf{x}), \tag{3.7}$$

where we recall that $\boldsymbol{\Pi}_{\mathbf{w}}$ is the projection operator onto the subspace spanned by $\mathbf{w}$.

Definition 3.2. A GRP $\mathbf{x}$ is JWSS if

$$\mathbf{C}_{\mathbf{x}} \circ (\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}) = (\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}) \circ \mathbf{C}_{\mathbf{x}}. \tag{3.8}$$

In other words, from Assumption 3.2 and [69, Chapter 4, Exercise 35 (a)], $\mathbf{C}_{\mathbf{x}}$ and the shift operator $\mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ have the same complete orthonormal set of eigenvectors with

$$\mathbf{C}_{\mathbf{x}} = \sum_{k,l} p_{\mathbf{x}}(k,l) \boldsymbol{\Pi}_{\boldsymbol{\phi}_k \otimes \psi_l}, \tag{3.9}$$

where $\{p_{\mathbf{x}}(k,l) : k = 1, \ldots, N, l \geq 1\}$ are the eigenvalues of $\mathbf{C}_{\mathbf{x}}$, also known as the *JPSD* of $\mathbf{x}$. If $\{(k,l) : p_{\mathbf{x}}(k,l) > 0\}$ is a finite set, $\mathbf{x}$ is said to be *bandlimited*.

In the following, we give two examples to illustrate Definition 3.2.

Example 3.2 (Time-vertex model)**.** Suppose each vertex of a graph G observes a time series with T discrete time steps. Then, the signal can be represented as a graph signal whose underlying graph $G_{\mathcal{J}}$ is the Cartesian product graph of G and a directed cyclic graph G' with T vertices [19]. Under this framework, we choose as the GSO the Laplacian matrix $\mathbf{L}_{\mathcal{J}}$ of $G_{\mathcal{J}}$. In [27], a time-vertex WSS signal is defined as WSS on the graph $G_{\mathcal{J}}$ w.r.t. $\mathbf{L}_{\mathcal{J}}$.

From (3.6), the product graph Laplacian is $\mathbf{L}_{\mathcal{J}} = \mathbf{L}_G \otimes \mathbf{I}_T + \mathbf{I}_N \otimes \mathbf{L}_{G'}$, which is different from the shift operator used in (3.8). Furthermore, $\mathbf{L}_{G'}$ is not a self-adjoint operator on $\mathbb{C}^T$. To be consistent with Definition 3.2, let $\widetilde{G'}$ denote the undirected cyclic graph with T vertices. Now we take $\mathbf{L}_G \otimes \mathbf{L}_{\widetilde{G'}}$ as the shift operator. We first observe that $\mathbf{L}_{\widetilde{G'}}$ is self-adjoint and has the DFT matrix columns as eigenvectors. Therefore, the Fourier transform induced by $\mathbf{L}_G \otimes \mathbf{L}_{\widetilde{G'}}$ is the same as that induced by $\mathbf{L}_{\mathcal{J}}$. Thus, a time-vertex WSS signal as defined in [27] fits Definition 3.2.

The time-vertex model is a special case in which the signal on each vertex belongs to a Euclidean space. In the next example, we will see that in a more general setting, the shift operator in the joint domain can also be defined in a meaningful way.

Example 3.3 (Euclidean-vertex model)**.** Suppose each vertex v observes a d-dimensional random vector $\mathbf{x}_v$ having identical distribution over vertices $v \in \mathcal{V}$. The signal can be written as a matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{C}^{d \times N}$, where the v-th column is the observation by vertex v while the i-th row represents the i-th measurement or feature among all vertices. We choose the graph shift operator $\mathbf{A}_G = \mathbf{L}_G$, and $\mathbf{A}_{\mathcal{H}}$ as the covariance matrix $\mathbf{C}_{\mathcal{H}} = \mathbb{E}[\mathbf{x}_v \mathbf{x}_v^*]$ of $\mathbf{x}_v$. In the context of Definition 3.2, we let the shift operator be $\mathbf{L}_G \otimes \mathbf{C}_{\mathcal{H}}$.

In this case, we note that the shift operator $\mathbf{L}_G \otimes \mathbf{C}_{\mathcal{H}}$ has a physical meaning. Suppose $\mathbf{L}_G = \mathbf{\Phi} \mathbf{\Lambda}_G \mathbf{\Phi}^\intercal$ and $\mathbf{C}_{\mathcal{H}} = \mathbf{\Psi} \mathbf{\Lambda}_{\mathcal{H}} \mathbf{\Psi}^*$ are eigendecompositions. The joint Fourier transform on the signal $\mathbf{X}$ can be written as

$$\mathfrak{F}(\mathbf{X}) = \mathbf{\Psi}^* \mathbf{X} \mathbf{\Phi}, \tag{3.10}$$

which in vectorized form $\operatorname{vec}(\mathfrak{F}(\mathbf{X})) = \mathbf{\Phi}^\intercal \otimes \mathbf{\Psi}^* \operatorname{vec}(\mathbf{X})$ corresponds to (2.13). The transform (3.10) can be interpreted as a two-step Fourier transform: first we multiply $\mathbf{\Psi}^*$ on $\mathbf{X}$ on the left-hand side, so that each vertex signal $\mathbf{x}_v$ is transformed into its principal component scores. Next, we multiply $\mathbf{\Phi}$ on $\mathbf{\Psi}^* \mathbf{X}$ on the right-hand side to transform each row into the graph frequency domain.

If $\mathbf{X}$ is JWSS w.r.t. $\mathbf{L}_G \otimes \mathbf{C}_{\mathcal{H}}$, the covariance $\mathbf{C}_{\mathbf{X}} = \mathbb{E}[\operatorname{vec}(\mathbf{X}) \operatorname{vec}(\mathbf{X})^*]$ of $\mathbf{X}$ have eigenvectors given by the eigenbasis $\mathbf{\Phi} \otimes \mathbf{\Psi}$. From Proposition 2.4, the entries of $\mathfrak{F}(\mathbf{X})$ are uncorrelated random variables. This implies that $\mathbf{X}$ can be decomposed into different oscillation modes of its vertex signals' principal component scores along the edges of G. In practice, $\mathbf{C}_{\mathcal{H}}$ can be estimated by the sample covariance of the vertex signal observations. Therefore, the shift operator and Fourier transform are deduced from both the graph topology and training data. Compared to the traditional GSP and time-vertex framework that construct the principal axes from the domain structures, this framework is closer to a data-driven approach.

Although we have assumed that every vertex observes a d-dimensional random vector, our model is not equivalent to the product graph model, because $\mathbf{L}_G \otimes \mathbf{C}_{\mathcal{H}}$ does not correspond to any kind of graph product in Example 3.1. However, one may adopt the shift operator $\mathbf{W}_G \otimes \mathbf{C}_{\mathcal{H}}$ to obtain a tensor product graph model.

In this case, $\mathbf{C}_{\mathcal{H}}$ is treated as the adjacency matrix of a weighted graph with d vertices. But in this case, this model's physical meaning is unclear compared to the construction in this example.

Finally, we note that this model can be generalized such that the vertex signals $\mathbf{x}_v$, $v \in \mathcal{V}$, are not identically distributed. In fact, we only need to assume that every $\mathbf{x}_v$ has the same set of principal axes, and regard $\mathbf{L}_G \otimes \mathbf{C}_{x_1}$ as the shift operator. It can be shown that this yields the same Fourier transform and physical meaning as discussed above.

An example with infinite dimensional $\mathcal{H} = L^2[-\pi, \pi]$ is presented in Section 3.2.2. In the numerical experiments in Section 3.4, we employ the models in Example 3.2 and Example 3.3 to compare the effectiveness of their corresponding Wiener filters, which are discussed in Section 3.3.

3.2.2 WSS in different domains

In the time-vertex framework of [26, 27], stationarity in the time domain (TWSS), vertex domain and joint domain are related in the sense that JWSS implies stationarity in both the time and vertex domains. In this subsection, the corresponding concepts and relations are generalized to the GGSP framework.

Definition 3.3. Given a GRP $\mathbf{x}$ satisfying the conditions in Theorem 3.1, we say that $\mathbf{x}$ is vertex wide-sense stationary (VWSS) if

$$\mathbf{K}_{\mathbf{x}}(\mathbf{t}, \mathbf{t})\mathbf{A}_G = \mathbf{A}_G\mathbf{K}_{\mathbf{x}}(\mathbf{t}, \mathbf{t})$$

τ-almost everywhere (a.e.).

Note that Definition 3.3 implicitly requires $\mathcal{H} = L^2(\mathcal{T})$ (an assumption in Theorem 3.1) so that $\mathbf{K}_{\mathbf{x}}(\mathbf{t}, \mathbf{t})$ can be defined. This does not result in loss of generality since we work only with separable Hilbert spaces. Definition 3.3 requires each measurement $\mathbf{x}(\mathbf{t})$ at each "time" $\mathbf{t} \in \mathcal{T}$ to be WSS as a vector-valued graph signal. For example, if $\mathcal{T}$ represents the time domain, this definition implies that the graph signal observed at every time instance is VWSS as a random vector.

Definition 3.4. A GRP $\mathbf{x}$ is said to be Hilbert space wide-sense stationary (HWSS) if

$$\boldsymbol{\delta}_m^{\mathcal{H}}\mathbf{C}_{\mathbf{x}}\boldsymbol{\delta}_m^{\mathcal{H}} \circ (\mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}) = (\mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}) \circ \boldsymbol{\delta}_m^{\mathcal{H}}\mathbf{C}_{\mathbf{x}}\boldsymbol{\delta}_m^{\mathcal{H}} \tag{3.11}$$

for all $m = 1, \ldots, N$. Here, $\boldsymbol{\delta}_m^{\mathcal{H}} := \mathrm{diag}(\mathbf{e}_m) \otimes \mathbf{I}_{\mathcal{H}}$ where $\mathbf{e}_m$ is the m-th standard basis vector in $\mathbb{C}^N$ consisting of all zeros except a one at the m-th entry.

In Definition 3.4, the operator $\boldsymbol{\delta}_m^{\mathcal{H}}$ keeps the vertex signal at vertex m unchanged while nullifying the other vertex signals. Suppose $\mathbf{x}$ satisfies all conditions in Theorem 3.1. For $\mathbf{y}(\cdot) = (y_1(\cdot), \ldots, y_N(\cdot))^\intercal \in \mathbb{C}^N \otimes L^2(\mathcal{T})$, the left-hand side of (3.11) can be computed step by step as follows:

$$\begin{aligned} \mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}(\mathbf{y})(\mathbf{t}) &= (\mathbf{A}_{\mathcal{H}}(y_1)(\mathbf{t}), \ldots, \mathbf{A}_{\mathcal{H}}(y_N)(\mathbf{t}))^\intercal, \\ \boldsymbol{\delta}_m^{\mathcal{H}}\mathbf{C}_{\mathbf{x}}\boldsymbol{\delta}_m^{\mathcal{H}} \circ (\mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}(\mathbf{y}))(\mathbf{s}) &= \int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,m}\mathbf{A}_{\mathcal{H}}(y_m)(\mathbf{t})\, \mathrm{d}\tau(\mathbf{t}) \cdot \mathbf{e}_m, \end{aligned} \tag{3.12}$$

where $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,n}$ is the (m, n)-th entry of $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \in \mathbb{C}^{N \times N}$. Similarly, the right-hand side of (3.11) can be derived as

$$(\mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}) \circ \boldsymbol{\delta}_m^{\mathcal{H}}\mathbf{C}_{\mathbf{x}}\boldsymbol{\delta}_m^{\mathcal{H}}(\mathbf{y})(\mathbf{s}) = \mathbf{A}_{\mathcal{H}}\left(\int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\cdot, \mathbf{t})_{m,m} y_m(\mathbf{t})\, \mathrm{d}\tau(\mathbf{t})\right)(\mathbf{s}) \cdot \mathbf{e}_m. \tag{3.13}$$

Let $\mathbf{K}_m$ be the integral operator with kernel $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,m}$ in the right-hand side of (3.13). Comparing (3.12) and (3.13), for $\mathbf{x}$ to be HWSS, the operator $\mathbf{A}_{\mathcal{H}}$ commutes with the integral operator $\mathbf{K}_m$.

Consider the example where $\mathcal{T} = [-\pi, \pi]$ with $\mathbf{A}_{\mathcal{H}}$ being a convolution operator,

$$\mathbf{A}_{\mathcal{H}}(y) = \sum_{l=-\infty}^{\infty} \nu_{|l|}\boldsymbol{\Pi}_{\psi_l},$$

where $\{\nu_l\}_{l=0}^{\infty}$ is a positive sequence converging to 0, and $\{\psi_l : l \in \mathbb{Z}\}$ denotes the standard Fourier basis $\{e^{\mathrm{i}lt}/\sqrt{2\pi} : l \in \mathbb{Z}\}$, where $\mathrm{i} = \sqrt{-1}$. Suppose $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})$ is continuous on $\mathcal{T} \times \mathcal{T}$. In this case, both $\mathbf{A}_{\mathcal{H}}$ and $\mathbf{K}_m$ are compact and self-adjoint operators on $L^2(\mathcal{T})$. If $\mathbf{A}_{\mathcal{H}}$ commutes with $\mathbf{K}_m$, then since the standard Fourier

basis $\{\psi_l\}_{k=0}^{\infty}$ are eigenvectors of $\mathbf{A}_{\mathcal{H}}$, they are also the eigenvectors of $\mathbf{K}_m$. Since

$$\sum_{i,j}^{n} c_i \bar{c}_j \mathbf{K}_{\mathbf{x}}(s_i, s_j)_{m,m} = \operatorname{var}\left(\sum_{i}^{n} c_i x_m(s_i)\right) \geq 0$$

for arbitrary $\{c_i\}_{i=1}^{n} \subset \mathbb{C}$ and $\{s_i\}_{i=1}^{n} \subset \mathcal{T}$, $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,m}$ is a positive definite symmetric kernel on $\mathcal{T}$. Utilizing Mercer's theorem we obtain that

$$\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,m} = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \sigma_l e^{\mathrm{i}ls} e^{-\mathrm{i}lt} = \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \sigma_l e^{\mathrm{i}l(s-t)},$$

where $\{\sigma_l \ : \ l \in \mathbb{Z}\}$ are $\mathbf{K}_m$'s eigenvalues associated with $\{\psi_l \ : \ l \in \mathbb{Z}\}$. This infinite sum uniformly converges on $\mathcal{T} \times \mathcal{T}$, indicating that $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})_{m,m}$ is actually a univariate function of $s-t$ for each m. This result agrees with the classical definition of stationarity in the time domain.

Definitions 3.3 and 3.4 can be regarded as traditional WSS definitions embedded in the GGSP framework. In the following theorem we show that JWSS implies WSS in both the vertex and Hilbert space domains.

Theorem 3.3. *A JWSS GRP* $\mathbf{x}$ *that satisfies the conditions in Theorem 3.1 is both VWSS and HWSS.*

Proof. We first show that $\mathbf{x}$ is VWSS. Since every $\mathbf{\Pi}_{\phi_k \otimes \psi_l}$ commutes with $\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}}$, from the fact that $\mathbf{x}$ is JWSS, (cf. Definition 3.2), we obtain that $\mathbf{C}_{\mathbf{x}}$ commutes with $\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}}$. To be specific, for any $\mathbf{y} \in \mathbb{C}^N \otimes L^2(\mathcal{T})$,

$$\begin{aligned}
\mathbf{C}_{\mathbf{x}} \circ (\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}})(\mathbf{y})(\mathbf{s}) &= \int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \mathbf{A}_G \mathbf{y}(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}), \\
(\mathbf{A}_G \otimes \mathbf{I}_{\mathcal{H}}) \circ \mathbf{C}_{\mathbf{x}}(\mathbf{y})(\mathbf{s}) &= \mathbf{A}_G \int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \mathbf{y}(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}),
\end{aligned}$$

hence

$$\int_{\mathcal{T}} (\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \mathbf{A}_G - \mathbf{A}_G \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})) \mathbf{y}(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}) = \mathbf{0}.$$

Therefore, $\mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \mathbf{A}_G = \mathbf{A}_G \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t})$ τ-a.e., indicating that $\mathbf{x}$ is VWSS.

We can show that $\mathbf{x}$ is HWSS by a similar approach. It suffices to notice from (3.9) that

$$\boldsymbol{\delta}_m^{\mathcal{H}} \mathbf{C}_{\mathbf{x}} \boldsymbol{\delta}_m^{\mathcal{H}} = \sum_{k,l} p_{\mathbf{x}}(k,l) \cdot \boldsymbol{\delta}_m^{\mathcal{H}} \boldsymbol{\Pi}_{\boldsymbol{\phi}_k \otimes \boldsymbol{\psi}_l} \boldsymbol{\delta}_m^{\mathcal{H}},$$

and each $\boldsymbol{\delta}_m^{\mathcal{H}} \boldsymbol{\Pi}_{\boldsymbol{\phi}_k \otimes \boldsymbol{\psi}_l} \boldsymbol{\delta}_m^{\mathcal{H}}$ commutes with $\mathbf{I}_N \otimes \mathbf{A}_{\mathcal{H}}$. □

From the proof of Theorem 3.3, we note that JWSS is strictly stronger than VWSS, since we actually show that JWSS implies $\mathbf{K}_{\mathbf{x}}(\mathbf{s},\mathbf{t})\mathbf{A}_G = \mathbf{A}_G\mathbf{K}_{\mathbf{x}}(\mathbf{s},\mathbf{t})$, which is a stronger condition than the VWSS condition of $\mathbf{K}_{\mathbf{x}}(\mathbf{t},\mathbf{t})\mathbf{A}_G = \mathbf{A}_G\mathbf{K}_{\mathbf{x}}(\mathbf{t},\mathbf{t})$. In fact, JWSS is strictly stronger than both VWSS and HWSS. This can be seen from [27], which introduces MTWSS (multivariate time WSS) and MVWSS (multivariate vertex WSS) so that JWSS is equivalent to simultaneously satisfying these conditions. These two concepts require not only the covariance but also the cross-covariance in their respective domains to admit certain forms of eigendecomposition. Our definitions of VWSS and HWSS extend the concept of VWSS and TWSS defined in [26], which only require the covariance matrices to satisfy the conditions in Definitions 3.3 and 3.4.

3.3 Wiener Filters

In this section, we investigate the denoising and recovery problems in the GGSP framework. In traditional GSP, these problems are formulated as regularized regression problems, whose regularization terms depend on the PSD values of the signal and noise [25]. This optimization framework is also adopted under the time-vertex framework [26]. The Wiener filters of these frameworks are the BLUE in their respective frameworks. In the GGSP framework, the observed signal on each vertex may come from an infinite-dimensional Hilbert space. In the sequel, we see that the Wiener filter takes the same form as the aforementioned formulations.

3.3.1 Wiener filter for denoising

Consider the model

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\epsilon} \tag{3.14}$$

where $\mathbf{x}$ and $\boldsymbol{\epsilon}$ denote the signal and noise, respectively. Suppose $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are independent JWSS GRPs. From Proposition 2.5, $\mathbf{y}$ is then a JWSS GRP with its JPSD given by $\{p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l) \,:\, k = 1,\dots,N,\, l \geq 1\}$. We further assume that $\mathbf{m}_{\mathbf{x}} = \mathbf{m}_{\boldsymbol{\epsilon}} = \mathbf{0}$ for simplicity.

Deriving the Wiener filter, in this case, amounts to deriving the BLUE for $\mathbf{x}$ given $\mathbf{y}$. In Hilbert space, this corresponds to LCE (see Section 2.4 or [55]). These results enable us to derive an explicit formula for the Wiener filter.

Theorem 3.4. *The Wiener filter* $\mathbf{G}$ *corresponding to the model* (3.14) *is a convolution filter of the form* (3.7) *with coefficients*

$$g_{k,l} = \frac{p_{\mathbf{x}}(k,l)}{p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l)}. \tag{3.15}$$

Proof. From Proposition 2.9, $\mathbf{G}$ can be asymptotically approximated by a sequence of bounded finite-rank operators $\{\mathbf{G}^{(m)}\}_{m\geq 1}$ as follows: Let $\mathcal{X}^{(m)} = \operatorname{span}\{\boldsymbol{\phi}_k \otimes \psi_l \,:\, k = 1,\dots,N,\, l = 1,\dots,m\}$ and $\mathbf{y}^{(m)} = \mathbf{\Pi}_{\mathcal{X}^{(m)}}\mathbf{y}$, then for any $\mathbf{z} \in \mathbb{C}^N \otimes \mathcal{H}$,

$$\mathbf{G}^{(m)}(\mathbf{z}) = (\mathbf{C}^{\dagger}_{\mathbf{y}^{(m)}}\mathbf{C}_{\mathbf{y}^{(m)}\mathbf{x}})^{*}\mathbf{z}, \tag{3.16}$$

where † denotes the Moore-Penrose pseudoinverse. Since $\mathbf{y}$ is JWSS, $\mathbf{C}_{\mathbf{y}}$ has the eigendecomposition

$$\mathbf{C}_{\mathbf{y}} = \sum_{l=1}^{\infty}\sum_{k=1}^{N}(p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l))\mathbf{\Pi}_{\boldsymbol{\phi}_k\otimes\psi_l},$$

Using Proposition 2.5, the terms on the right-hand side of (3.16) can be written as

$$\mathbf{C}^{\dagger}_{\mathbf{y}^{(m)}} = \sum_{l=1}^{m}\sum_{k=1}^{N}(p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l))^{-1}\mathbf{\Pi}_{\boldsymbol{\phi}_k\otimes\psi_l},$$

$$\mathbf{C}_{\mathbf{y}^{(m)}\mathbf{x}} = \mathbf{\Pi}_{\mathcal{X}^{(m)}}\mathbf{C}_{\mathbf{y}\mathbf{x}} = \mathbf{\Pi}_{\mathcal{X}^{(m)}}\mathbf{C}_{\mathbf{x}} = \sum_{l=1}^{m}\sum_{k=1}^{N}p_{\mathbf{x}}(k,l)\mathbf{\Pi}_{\boldsymbol{\phi}_k\otimes\psi_l}.$$

By substituting these into (3.16), we obtain

$$\mathbf{G}^{(m)}(\mathbf{z}) = \sum_{l=1}^{m}\sum_{k=1}^{N} \frac{p_{\mathbf{x}}(k,l)}{p_{\mathbf{x}}(k,l)+p_{\boldsymbol{\epsilon}}(k,l)}\mathbf{\Pi}_{\boldsymbol{\phi}_k\otimes\boldsymbol{\psi}_l}(\mathbf{z}).$$

From Proposition 2.9, $\left\|\mathbf{G}^{(m)}(\mathbf{z}) - \mathbf{G}(\mathbf{z})\right\| \to 0$ as $m \to \infty$ for $\mathbb{P}_{\mathbf{y}}$-almost surely all $\mathbf{z} \in \mathbb{C}^N \otimes \mathcal{H}$. Therefore, the Wiener filter $\mathbf{G}$ can be chosen to be

$$\mathbf{G}(\mathbf{z}) = \lim_{m\to\infty} \mathbf{G}^{(m)}(\mathbf{z}) = \sum_{l=1}^{\infty}\sum_{k=1}^{N} \frac{p_{\mathbf{x}}(k,l)}{p_{\mathbf{x}}(k,l)+p_{\boldsymbol{\epsilon}}(k,l)}\mathbf{\Pi}_{\boldsymbol{\phi}_k\otimes\boldsymbol{\psi}_l}(\mathbf{z}), \tag{3.17}$$

yielding the result in (3.15).

Before concluding the proof, we remark that (3.17) may not converge with respect to the operator norm induced topology. However, this infinite sum is still well-defined in terms of the strong operator topology, i.e., $\|\mathbf{G}^{(m)}(\mathbf{z}) - \mathbf{G}(\mathbf{z})\| \to 0$ for all $\mathbf{z} \in \mathcal{H}$. Finally, it is straightforward to see that $\mathbf{G}$ is a bounded linear operator with $\|\mathbf{G}\| \leq 1$. □

3.3.2 Wiener filter for signal completion

We now consider the case where only signals from a subspace $\mathcal{S} \subset \mathbb{C}^N \otimes \mathcal{H}$ are observable, i.e.,

$$\mathbf{y} = \mathbf{\Pi}_{\mathcal{S}}(\mathbf{x} + \boldsymbol{\epsilon}), \tag{3.18}$$

where $\mathbf{\Pi}_{\mathcal{S}}$ denotes the projection operator onto $\mathcal{S}$.

In this case it may not be possible to give an explicit formula for the Wiener filter. However, if we assume that the signal is bandlimited (cf. Definition 3.2), the noise can be assumed to be bandlimited as well without loss of generality. This is because we can apply the finite-rank projection operator on the observed signal, projecting it to the subspace where the original signal lies in. By doing this, the frequencies that do not involve $\mathbf{x}$ are discarded. Under this assumption, there is an explicit characterization of the Wiener filter as follows.

Theorem 3.5. *Suppose* $\mathbf{x}$ *and* $\boldsymbol{\epsilon}$ *are bandlimited. The Wiener filter* $\mathbf{G}$ *for signal completion can be written as*

$$\mathbf{G}(\mathbf{z}) = ((\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\boldsymbol{\epsilon}})\mathbf{\Pi}_{\mathcal{S}})^{\dagger}\mathbf{\Pi}_{\mathcal{S}}\mathbf{C}_{\mathbf{x}})^{*}\mathbf{z}, \tag{3.19}$$

for $\mathbf{z} \in \mathbb{C}^N \otimes \mathcal{H}$*. In particular, when* $\mathcal{H}$ *is finite-dimensional,* $\mathbf{x}$ *and* $\boldsymbol{\epsilon}$ *are trivially bandlimited.*

Proof. Since $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are bandlimited, their covariance operators are finite-rank. Notice that $\mathbf{C}_{\mathbf{y}} = \mathbf{\Pi}_{\mathcal{S}}(\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\boldsymbol{\epsilon}})\mathbf{\Pi}_{\mathcal{S}}$, hence $\mathbf{C}_{\mathbf{y}}$'s range is also finite-rank, thus closed. From Proposition 2.8, the formula for the compatible case (cf. Proposition 2.7) can be applied to directly obtain the result in (3.19). □

When $\mathbf{x}$ and $\boldsymbol{\epsilon}$ are not bandlimited, the Wiener filter can be approximated by a sequence of operators as in the proof of Theorem 3.4.

Theorem 3.6. *The Wiener filter* $\mathbf{G}$ *for signal completion can be asymptotically approximated by*

$$\mathbf{G}_m(\mathbf{z}) = ((\mathbf{\Pi}_{\mathcal{X}^{(m)}}\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C}_{\mathbf{x}} + \mathbf{C}_{\boldsymbol{\epsilon}})\mathbf{\Pi}_{\mathcal{S}}\mathbf{\Pi}_{\mathcal{X}^{(m)}})^{\dagger}\mathbf{\Pi}_{\mathcal{X}^{(m)}}\mathbf{\Pi}_{\mathcal{S}}\mathbf{C}_{\mathbf{x}})^{*}\mathbf{z},$$

where $\mathcal{X}^{(m)} = \operatorname{span}\{\boldsymbol{\phi}_k \otimes \psi_l \,:\, k = 1, \ldots, N,\, l = 1, \ldots, m\}$*, and* $\|\mathbf{G}_m(\mathbf{z}) - \mathbf{G}(\mathbf{z})\| \to 0$ *for almost surely all* $\mathbf{z} \in \mathbb{C}^N \otimes \mathcal{H}$*.*

Proof. This theorem is a direct result of Proposition 2.9. □

In the specific case that $\mathcal{S}$ is the subspace of signals that annihilate on a set of vertices (i.e., only signals on a subset of vertices $\mathcal{U}$ is observable), the analytical form of the signal completion error can be computed. To obtain the MSE of $\mathbf{G}$, we define the following matrices $\mathbf{\Lambda}_l$ and $\mathbf{\Gamma}_l$ whose (i,j)-th elements are respectively:

$$(\mathbf{\Lambda}_l)_{i,j} = \begin{cases} p_{\mathbf{x}}(i,l) + p_{\boldsymbol{\epsilon}}(i,l), & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases}$$

$$(\mathbf{\Gamma}_l)_{i,j} = \begin{cases} p_{\mathbf{x}}(i,l)^2, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Let $\mathbf{\Phi}_{\mathcal{U}}$ be the submatrix of $\mathbf{\Phi}$ that contains the rows with index set $\mathcal{U}$.

Theorem 3.7. *Assume that there exists $l' < \infty$ such that whenever $l < l'$, $p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l) > 0$ for all $k = 1, \dots N$; and whenever $l \geq l'$, $p_{\mathbf{x}}(k,l) + p_{\boldsymbol{\epsilon}}(k,l) = 0$ for all $k = 1, \dots N$. Then the MSE of the Wiener filter $\mathbf{G}$ can be written as*

$$\mathbb{E}\,\|\mathbf{G}(\mathbf{y}) - \mathbf{x}\|^2 = \sum_{l=1}^{l'} \sum_{k=1}^{N} p_{\mathbf{x}}(k,l) - \sum_{l=1}^{l'} \operatorname{tr}((\boldsymbol{\Phi}_{\mathcal{U}} \boldsymbol{\Lambda}_l \boldsymbol{\Phi}_{\mathcal{U}}^{\intercal})^{-1} \boldsymbol{\Phi}_{\mathcal{U}} \boldsymbol{\Gamma}_l \boldsymbol{\Phi}_{\mathcal{U}}^{\intercal}). \tag{3.20}$$

Proof. The basic idea of the proof is to characterize the image and kernel space of $\mathbf{C}_{\mathbf{y}}$ first (i.e. Lemma 3.8), and then compute the MSE via Proposition 2.3. See Appendix 3.B for details. □

Theorem 3.7 indicates a method to measure the quality of a sampling strategy. When the statistical properties of the signal and noise are fixed, one can choose the sampling set $\mathcal{U}$ that admits a large value of $\sum_{l=1}^{l'} \operatorname{tr}((\boldsymbol{\Phi}_{\mathcal{U}} \boldsymbol{\Lambda}_l \boldsymbol{\Phi}_{\mathcal{U}}^{\intercal})^{-1} \boldsymbol{\Phi}_{\mathcal{U}} \boldsymbol{\Gamma}_l \boldsymbol{\Phi}_{\mathcal{U}}^{\intercal})$ to obtain a small MSE.

3.4 Numerical Experiments

In this section, we verify the performance of our proposed GRP framework on four datasets. We compare its performance with the time-vertex and traditional GSP frameworks, and demonstrate that with its more general assumptions, the GRP can fit data better, thus achieving better performance. We also illustrate the optimality of the proposed Wiener filter and the criterion of sample set selection through the experiments.

Each dataset used can be organized into samples $\{\mathbf{X}_i : i = 1, \dots, m_s\}$, where $\mathbf{X}_i \in \mathbb{R}^{N \times d}$, N is the number of vertices in a graph G, and d is the dimension of each data feature vector. We assume that $\mathbf{X}_i$ are the matrix form of independent and identically distributed (i.i.d.) realizations of a JWSS GRP $\mathbf{x}$.

Each of the GSP frameworks has different statistical model assumptions, under which PSD estimation or covariance estimation from a training set of size m_a samples is performed. Throughout, we let the graph shift operator $\mathbf{A}_G$ be the graph Laplacian with eigenbasis $\{\boldsymbol{\phi}_k : k = 1, \dots, N\}$. In the following, we present the concrete implementation of the PSD estimator under each framework when $d < \infty$. The case $d = \infty$ is discussed in Section 3.4.3.

1. GRP. The datasets we test on contain vertex signals from finite-dimensional real spaces, which fit the Euclidean-vertex model in Example 3.3. To estimate the vertex signal covariance $\mathbf{C}_{\mathcal{H}}$, we use the sample covariance matrix $\widehat{\mathbf{C}}_{\mathcal{H}} \in \mathbb{R}^{d\times d}$ of a set of training samples. Let the eigenbasis induced by $\widehat{\mathbf{C}}_{\mathcal{H}}$ be $\widehat{\mathbf{\Psi}} = \{\widehat{\psi}_l \,:\, l = 1, \ldots, d\}$.

   The JPSD values of the GRP $\mathbf{x}$ is estimated by its empirical mean squared Fourier coefficients through

$$\widehat{p}_{\mathbf{x}}(k,l) = \frac{1}{m_a}\sum_{i=1}^{m_a} |\boldsymbol{\phi}_k^{\intercal}\mathbf{X}_i\overline{\widehat{\psi}}_l|^2, \tag{3.21}$$

   for $k = 1, \ldots, N$ and $l = 1, \ldots, d$. The covariance $\mathbf{C}_{\mathbf{x}}$ is estimated by

$$\widehat{\mathbf{C}}_{\mathbf{x}} = \sum_{k,l} \widehat{p}_{\mathbf{x}}(k,l)(\boldsymbol{\phi}_k \otimes \widehat{\psi}_l)(\boldsymbol{\phi}_k \otimes \overline{\widehat{\psi}}_l)^{\intercal}. \tag{3.22}$$

2. Time-vertex (TV) [27]. The JWSS model and sample JPSD estimator proposed in [27] are adopted. Specifically, the JPSD and $\mathbf{C}_{\mathbf{x}}$'s estimators are the same as those for the GRP model, with the exception that the eigenbasis $\widehat{\mathbf{\Psi}}$ is replaced by the column vectors of the DFT matrix rather than learned from the sample set.

3. Traditional GSP. Since the number of features $d > 1$ at each vertex, to adopt the traditional GSP framework that assumes scalar-valued vertex signals, we process the d features separately. To be specific, write each sample $\mathbf{X}_i = (\mathbf{f}_{i,1}, \ldots, \mathbf{f}_{i,d})$, where $\mathbf{f}_{i,j} \in \mathbb{R}^N$, $j = 1, \ldots, d$, is the j-th column and contains the j-th feature of all vertices. We use the periodogram [24] to estimate $p_{\mathbf{x}}(k,l)$, $k = 1, \ldots, N$, $l = 1, \ldots, d$:

$$\widehat{p}_{\mathbf{x}}(k,l) = \frac{1}{m_a}\sum_{i=1}^{m_a} |\boldsymbol{\phi}_k^{\intercal}\mathbf{f}_{i,l}|^2.$$

   The covariance estimator for feature $l = 1, \ldots, d$ is constructed similarly as (3.22):

$$\widehat{\mathbf{C}}_{\mathbf{x},l} = \sum_{k} \widehat{p}_{\mathbf{x}}(k,l)\boldsymbol{\phi}_k\boldsymbol{\phi}_k^{\intercal}.$$

   A separate Wiener filter for each feature is then constructed.

In the experiments, all Wiener filter forms are applied with full bandwidth for fair comparison unless otherwise specified.

3.4.1 Wiener filter for denoising

We investigate the denoising performance of the Wiener filter on an epilepsy dataset [70], which is collected by monitoring a patient's brain signal.[1] The dataset is collected from 76 electrodes, during ictal and pre-ictal periods. In our experiments, due to the assumption of stationarity, we only make use of the pre-ictal data, which contains 8 pre-ictal periods. Each period lasts for 10s, and we partition it into non-overlapping 125ms periods due to the intrinsic long-term instability of a brain signal. Since the sampling rate is 400Hz, this partition means that each signal sample $\mathbf{X}_i \in \mathbb{R}^{76\times 50}$, where $i = 1, \ldots, 640$. We use the samples from the first 4 pre-ictal periods as training and the rest of them for testing.

In order to embed a signal sample in a graph structure, we use a simpler but similar strategy as that in [70] to determine the connections between electrodes. Specifically, 4.5-5.5s of data in the first pre-ictal period is extracted. Then the correlation matrix of this sample set is computed. Assuming that a large absolute value of correlation indicates strong connection, we treat each node pair as connected with an edge if the absolute value of their correlation coefficient is larger than 0.75. Otherwise, they are not connected.

To compare the performance of different denoising strategies, the pre-ictal datasets are divided into training and test sets with the same size. Additive white Gaussian noise (AWGN) with different energies is added to both of these sets to obtain different input SNRs. Here, SNR in dB is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\mathbb{E}\,\|\mathbf{x}\|^2}{\mathbb{E}\,\|\widehat{\mathbf{x}}-\mathbf{x}\|^2},$$

where $\mathbf{x}$ is the original signal. For input SNR, $\widehat{\mathbf{x}}$ is the noisy version of $\mathbf{x}$; for output SNR, $\widehat{\mathbf{x}}$ denotes the estimate of $\mathbf{x}$.

By learning the signal spectrum from the training set, we aim to recover the signal from the noisy test set. The corresponding output SNR is taken as a measurement of

[1] https://math.bu.edu/people/kolaczyk/datasets.html

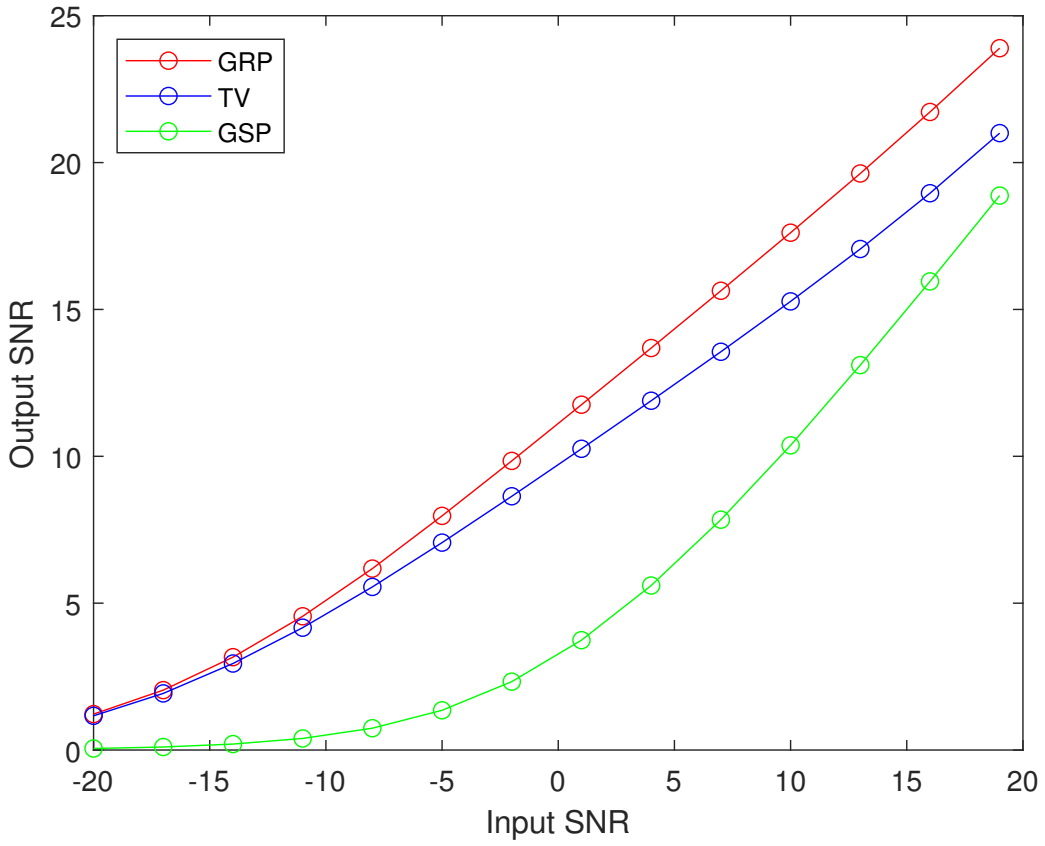


FIGURE 3.1: Denoising performance of Wiener filters under different frameworks. Experiments are repeated 20 times for each input SNR value.

performance. We compare the GRP Wiener filter's performance with the time-vertex joint Wiener filter corresponding to the solution of the optimization framework proposed in [27], which has been shown to outperform both purely time-based and graph-based Wiener filters. In this experiment we also include the traditional GSP Wiener filter as a benchmark method. We have tested the purely time-based Wiener filter on this dataset, but the performance is much worse than the other methods, and is omitted here.

From Fig. 3.1, we observe that the GRP framework produces the highest output SNR compared to the other benchmark methods. This result indicates that the strategy of learning $\widehat{\mathbf{\Psi}}$ from the training set provides a better fit than using the DFT basis.

We want to test the denoising performance of the GRP Wiener filter versus other parameterized filters. A common strategy (see e.g., Example 2 in [1]) in traditional

GSP is to solve the following problem:

$$\widehat{\mathbf{x}}(\mathbf{y}) = \underset{\widetilde{\mathbf{x}}\in\mathbb{R}^n}{\arg\min} \|\widetilde{\mathbf{x}} - \mathbf{y}\|^2 + \rho\widetilde{\mathbf{x}}^\intercal \mathbf{L}_G \widetilde{\mathbf{x}},$$

where ρ controls the smoothness of the recovered signal. On the other hand, the MAP estimator for Gaussian random vectors in $\mathbb{R}^d$ under the model (3.14) is obtained as follows:

$$\widehat{\mathbf{x}}(\mathbf{y}) = \underset{\widetilde{\mathbf{x}}\in\mathbb{R}^d}{\arg\min} \|\widetilde{\mathbf{x}} - \mathbf{y}\|^2 + \sigma^2\widetilde{\mathbf{x}}^\intercal \mathbf{C}_{\mathbf{x}}^\dagger \widetilde{\mathbf{x}},$$

where σ^2 denotes the variance of the noise. Combining the above optimization problems, we obtain the following problem under the GRP model:

$$\begin{aligned}\widehat{\mathbf{x}}(\mathbf{y}) = &\underset{\widetilde{\mathbf{x}}\in\mathbb{R}^{n\times d}}{\arg\min} \|\mathrm{vec}(\widetilde{\mathbf{x}}) - \mathrm{vec}(\mathbf{y})\|^2 \\ &+ \rho\,\mathrm{vec}(\widetilde{\mathbf{x}})^\intercal (\mathbf{L}_G \otimes \mathbf{C}_{\mathcal{H}}^\dagger)\,\mathrm{vec}(\widetilde{\mathbf{x}}),\end{aligned} \tag{3.23}$$

which is equivalent to applying the convolution filter with coefficients

$$\tilde{g}_{k,l} = \frac{\nu_l}{\nu_l + \rho\lambda_k},$$

where $\{\nu_l\}$ and $\{\lambda_k\}$ are the eigenvalues of $\mathbf{C}_{\mathcal{H}}$ and $\mathbf{L}_G$, respectively. We select the optimal parameter ρ with the best performance, and compare it with the Wiener filter $\mathbf{G}$ in Theorem 3.4 under the same setting as Fig. 3.1. From Fig. 3.2, we see that the Wiener filter $\mathbf{G}$ derived in (3.15) outperforms the parameterized filter in (3.23) as it is BLUE.

3.4.2 Wiener filter for signal completion

We next evaluate the Wiener filter for signal completion on the Krakow air quality dataset, which contains air quality data from a sensor network in Krakow, Poland.[2] The network consists of 56 sensors deployed across the city and each taking measurements on an hourly basis throughout the year 2017. We regard each sensor as a vertex in a graph. Each sensor records six measurements, namely the PM1, PM2.5, PM10, temperature, air pressure and humidity values. In this experiment

[2] https://www.kaggle.com/datascienceairly/air-quality-data-from-extensive-network-of-sensors

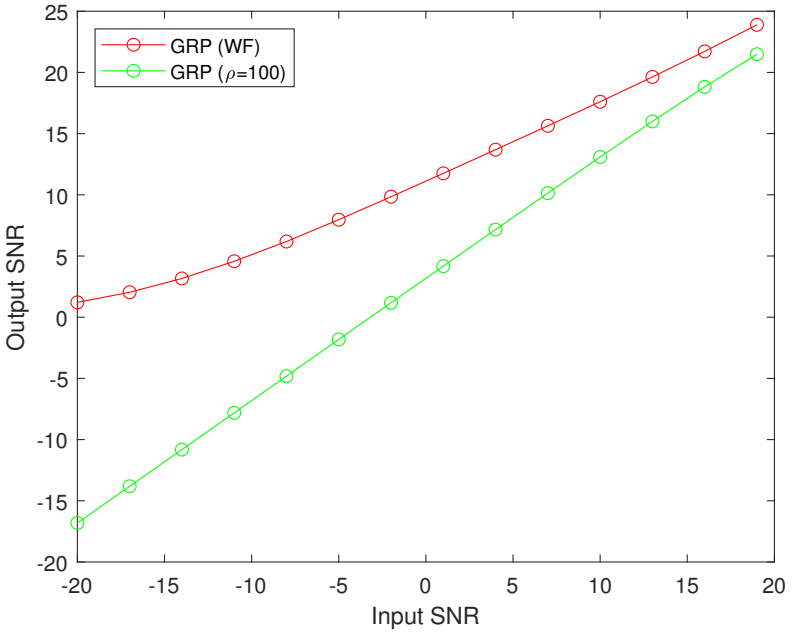


FIGURE 3.2: Denoising performance of different filter forms under GRP. Experiments are repeated 20 times for each input SNR value. "GRP (WF)" denotes the Wiener filter (3.15), and "GRP ($\rho = 100$)" denotes (3.23).

we only consider pollution-related features, i.e., PM1, PM2.5 and PM10. Thus, this dataset can be modeled by the Euclidean-vertex model where $\dim \mathcal{H} = 3$. To keep the statistical properties of the data approximately invariant, we only make use of the data in the winter months December, January, February and March, which contains a total of 114 days of records. Since the original data has missing values, we first omit those sensors with more than 10% missing values so that $n = 30$ sensors are left. Then we fill in the remaining missing values by taking the average of the nearest two days' records.

We embed the sensors in a weighted graph using their geographical coordinates and a K-NN method as in [27]. To be specific, we employ the 5-NN strategy to determine the connectivity between vertices. Next, we accord each edge the weight $\exp(-d(i,j)^2/\sigma^2)$, where $d(i,j)$ is the geographic distance between sensors i and j, and σ^2 the variance of all distances between pairs of sensors.

We randomly choose 57 days' records as the training set. The remaining days' records form the test set. We randomly remove some data, and then try to recover these values. To this end, we use the training set to estimate the JPSD and covariance operators as described at the beginning of Section 3.4, then apply the Wiener filter to recover the missing values in the test set. In this experiment, we take the

normalized error

$$\frac{\mathbb{E}\,\|\widehat{\mathbf{x}}-\mathbf{x}\|}{\mathbb{E}\,\|\mathbf{x}\|}$$

as the performance metric.

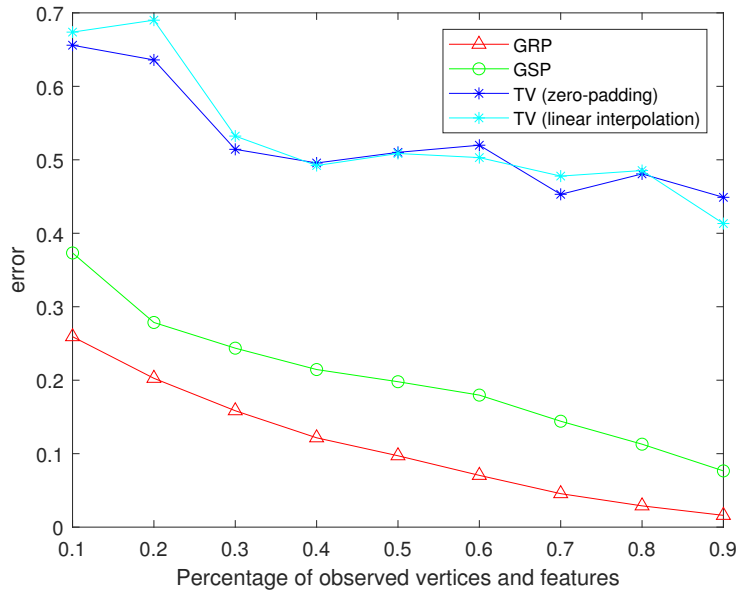
The dataset is indexed by three dimensions: vertex, feature and time in hourly intervals. To illustrate the effectiveness of the JWSS assumption, we test different frameworks under two missing data models:

1. Consecutive missing model. For each day in the dataset, we randomly choose a set of (vertex, feature) indices, and remove their corresponding data during a time period whose length is a geometric random variable. This model reflects the original dataset, which contains missing values over a continuous time period. Since both training and test sets have missing values, we test two methods for the TV framework to recover JPSD: either zero-pad missing values, or perform linear interpolation to recover the training set first and then estimate JPSD. Both TV and GSP process the three features separately.

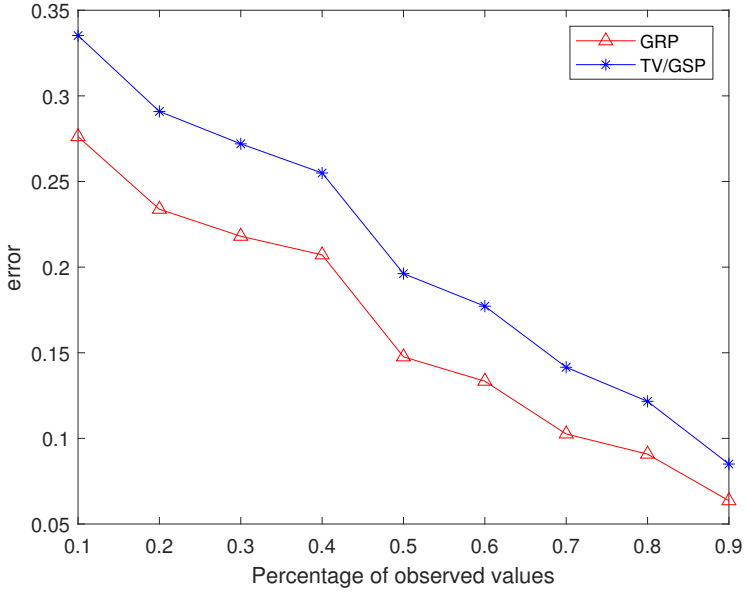
2. Uniform missing model. For each day in the test set, we randomly choose a set of (vertex, feature, time) indices and remove their data. In this model, the training set contains no missing values. Therefore, for GRP, we use the Cartesian product of the spatial graph and time graph (the cyclic graph with 24 vertices) as the underlying graph. On this graph, each vertex contains three features. In this case, applying TV is equivalent to applying GSP on this product graph, in which the three features are processed separately.

From Fig. 3.3a, we observe that the GRP Wiener filter dominates the traditional GSP Wiener filter and TV Wiener filter under the consecutive missing model. This is due to the fact that the GRP incorporates the correlation between features via (3.22) while the other methods regard them as independent. In fact, the features we use (i.e., PM1, PM2.5 and PM10) are strongly correlated as indicated by their correlation coefficient matrix estimated empirically from the whole dataset:

$$\begin{pmatrix} 1.0000 & 0.9954 & 0.9898 \\ 0.9954 & 1.0000 & 0.9955 \\ 0.9898 & 0.9955 & 1.0000 \end{pmatrix}.$$



(A) Consecutive missing model



(B) Uniform missing model

FIGURE 3.3: Recovery performance by Wiener filters under different frameworks. In Fig. 3.3a, the lengths of missing periods are generated by Geo(1/12). Each point is the average of 40 repetitions in Fig. 3.3a, and 50 repetitions in Fig. 3.3b.

Furthermore, the consecutive missing values along the time domain in the training set cause difficulties for TV JPSD estimation. Since the GRP model has the flexibility to ignore the missing period and to use only the complete data to learn JPSD, it is more accurate than TV methods.

From Fig. 3.3b, we observe that the GRP Wiener filter outperforms the TV Wiener filter, which is also due to the incorporation of correlation between features in GRP.

3.4.3 Continuous-time signal recovery

In this subsection, we evaluate the recovery performance on continuous-time graph signals under the GRP framework, and compare it with both TV and the traditional time stationary (TS) framework. The graph is generated by the Erdős-Rényi model with 30 vertices and edge probability 0.5. We enforce the graph to be connected. The generalized graph signal is generated as a randomized linear combination of the tensor products of the graph Fourier basis and sinusoids:

$$\mathbf{x}(t) = \sum_{k,l} d_{kl} \boldsymbol{\phi}_k \otimes \sin(\beta_l t), \ t \in [-\pi, \pi], \tag{3.24}$$

where d_{kl} are independently generated by the Gaussian distribution $\mathcal{N}\left(0, \sigma_{k,l}^2\right)$, and β_l are real-valued constants. We note that $\mathbf{x}(t)$ is not bandlimited as long as there exists a β_l that is not integer. Both training and test sets consist of noisy observations sampled from different vertices at different time instances.

To recover the continuous-time signal, we first apply the variational EM algorithm on the training set to estimate the PSD and noise power. To be specific, no matter what framework we use, the observation model can be written as

$$\mathbf{y} = \mathbf{B}\mathbf{c} + \boldsymbol{\epsilon}, \tag{3.25}$$

where each column of $\mathbf{B}$ contains the values of a basis function at the sample points, and $c = (c_1, c_2, \ldots, c_s)^\intercal$ are the Fourier coefficients that are assumed to be independent and have distribution $\mathcal{N}(0, p_i)$. The observation error vector $\boldsymbol{\epsilon}$ has components independently generated via $\mathcal{N}(0, \sigma^2)$.

For instance, in the GRP framework, the basis $\{\mathbf{f}_{k,l}\}$ is $\{\boldsymbol{\phi}_k \otimes \sin(lt) \ : \ k = 1 \ldots, n, \ l \in \mathbb{N}\} \cup \{\boldsymbol{\phi}_k \otimes \cos(lt) \ : \ k = 1 \ldots, n, \ l \in \mathbb{N}\}$. In practice, we only use a

subset of them so that $1 \leq l \leq m_0$ for some m_0. Suppose the sampled time instances on the ith vertex are $\{t_{i1}, t_{i2}, \ldots, t_{iq}\}$. Then for each basis function $\mathbf{f} \in \{\mathbf{f}_{k,l}\}$, its values at the sample points can be written as a vector

$$\mathbf{b} = (f_1(t_{11}), \ldots, f_1(t_{1q}), \ldots, f_n(t_{n1}), \ldots, f_n(t_{nq}))^\intercal, \tag{3.26}$$

where f_i $(i = 1, \cdots, n)$ denotes the i-th vertex signal of f. (Recall that $\mathbf{f}(t)$ is an N-dimensional vector valued function). By concatenating all b into a matrix we obtain the matrix $\mathbf{B}$.

For the TV framework, we use the basis $\{\boldsymbol{\phi}_k \otimes \psi_l : k = 1, \ldots, N,\ l = 1, \ldots, m_0\}$, where $\{\psi_l\}$ denotes the Fourier basis induced by a cyclic graph. After recovering all values on the time grid, we apply linear interpolation to recover the continuous signal. Under the time stationary assumption, every vertex signal is stationary in the time domain. Each vertex signal is processed separately, with model (3.25) and basis functions $\{\sin(lt) : l = 1, \ldots, m_0\} \cup \{\cos(lt) : l = 0, \ldots, m_0\}$.

Now that $\mathbf{y}$ and $\mathbf{B}$ are known, we apply the EM algorithm to solve for the estimates of $\{p_i\}$ and σ^2. However, since the PSD values $\{p_i\}$ are not assumed to be equal, the posterior $p(c, \sigma^2, p_1, \ldots, p_s \mid y)$ cannot be computed analytically. Therefore, we apply variational EM [71], which admits explicit analytical optimal values at each iterative step, to estimate these hidden values.

On the test set, the continuous-time signals are recovered based on the sample values and the information learned from the training set. Since we have assumed the regression model (3.25) in which $\{c_i\}$ and e are normal random variables, it suffices to use the posterior mean as the estimator of the Fourier coefficients.

In this experiment we consider two sampling schemes:

3.4.3.1 Equally spaced sampling

The samples are collected from a subset of equally spaced points of $[-\pi, \pi]$. To be specific, on each vertex we randomly choose m points from the time grid $\{t_i : t_i = -\pi + 2(i-1)\pi/(2m-1),\ i = 1, \ldots, 2m\}$. The graph signals are generated via (3.24) given $\widetilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \beta_3)$, and $\sigma^2_{k,l} = 10/(kl)$ such that the signals are smooth over the graph.

We investigate the performance of different frameworks under varying grid density (represented by the number of samples m) and noise energy (represented by SNR in dB). For each fixed pair of (m, SNR), we uniformly generate 100 $\widetilde{\boldsymbol{\beta}}$ vectors via Unif $(1, 15)$. Then, for each $\widetilde{\boldsymbol{\beta}}$, we generate a training set and test set, each containing 60 realizations of continuous data from (3.24). To measure the recovery performance, for each $\widetilde{\boldsymbol{\beta}}$, we compute the relative error

$$\frac{\mathbb{E}\,\|\widehat{\mathbf{x}} - \mathbf{x}\|}{\mathbb{E}\,\|\mathbf{x}\|}, \tag{3.27}$$

where the expectation denotes averaging over the test set.

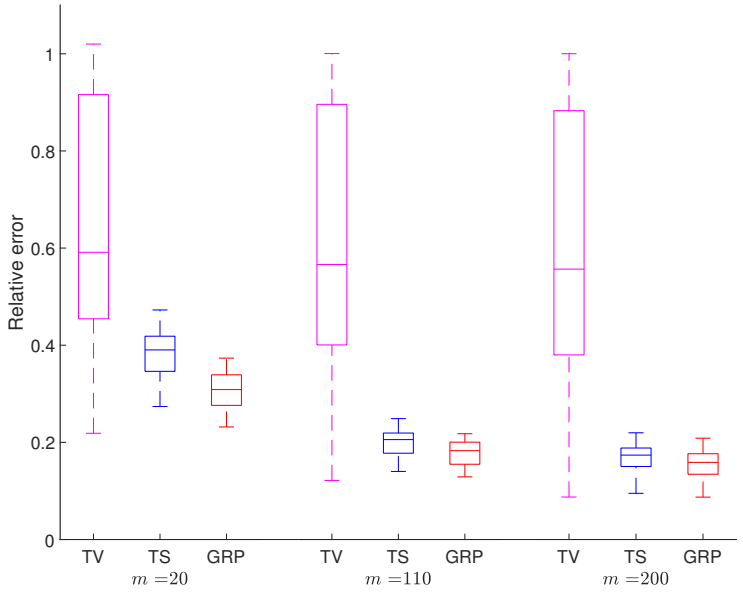
We summarize the results as box plots (Fig. 3.4 and Fig. 3.5), where each box reflects the distribution of recovery error with 100 different $\widetilde{\boldsymbol{\beta}}$. To recover the test signal, we set $m_0 = 20$. From Fig. 3.4, we observe that GRP outperforms both TS and TV. Compared with pure time domain based methods like TS, GRP makes use of the graph structure, which provides additional information. In addition, GRP uses a basis of continuous functions, while the interpolation step in TV fails to recover the high-frequency variations in the signal. Therefore, the performance of TV is largely affected by the choice of $\widetilde{\boldsymbol{\beta}}$.

3.4.3.2 Uniformly distributed sampling

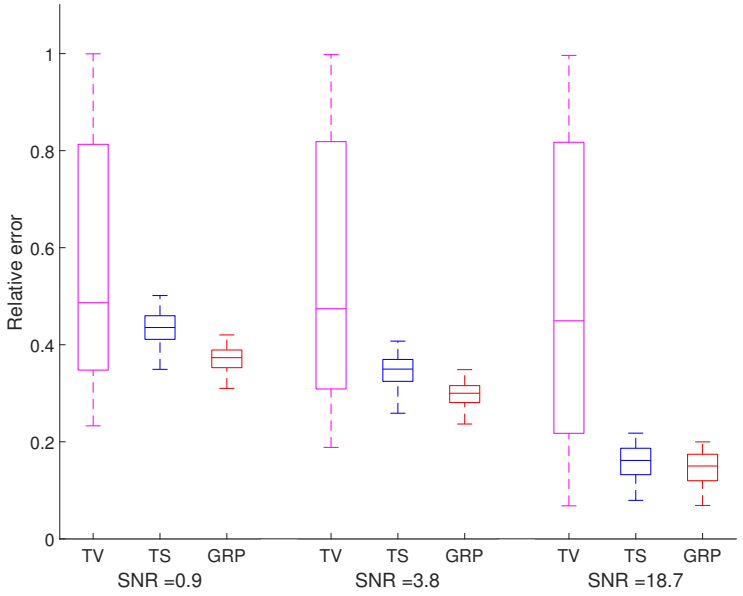
In this setting, the sampled time instances on each vertex are uniformly distributed on $[-\pi, \pi]$. Since the time instances are not placed in a time grid, the TV framework cannot be applied in this setting. The graph signals are generated the same way as Section 3.4.3.1. The performance is measured by (3.27). We investigate the performance of different frameworks under varying number of samples and noise energy. We observe from Fig. 3.5 that GRP outperforms TS.

3.4.4 Comparing sampling strategies

In this subsection, we illustrate the usefulness of Theorem 3.7 in comparing the MSEs of different sampling methods. To this end, we first estimate the JPSD from the training set. On the test set, we randomly generate sampling sets with different sizes. We compute the theoretical MSE values for all the sampling sets via

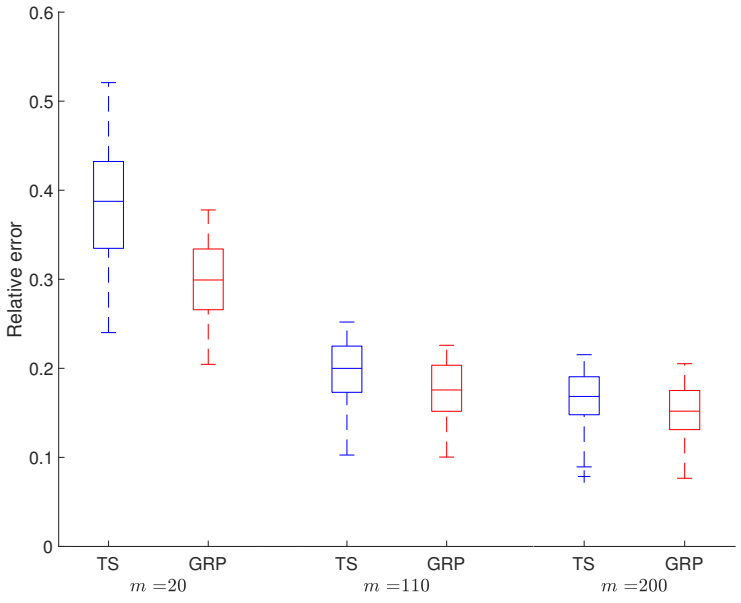


(A) Performance under varying time grid density.

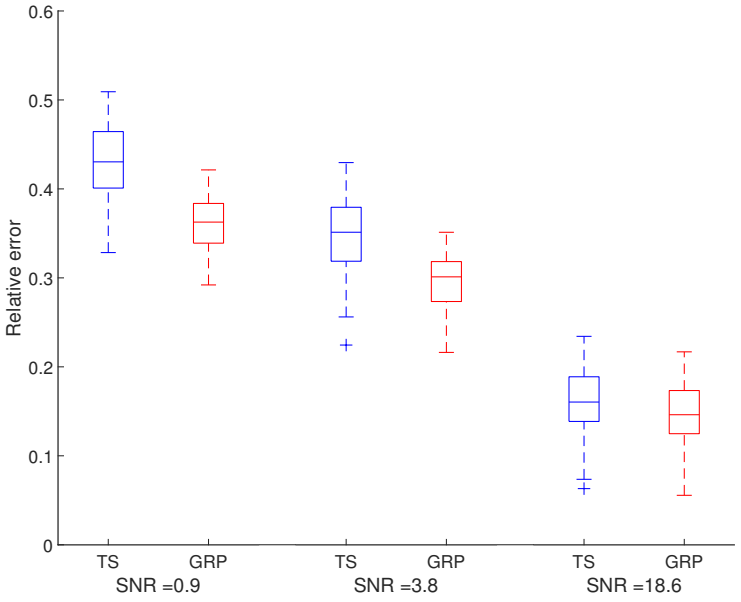


(B) Performance under varying noise energy.

FIGURE 3.4: Boxplots of recovery performance under different frameworks via equally spaced sampling. In Fig. 3.4a, the SNR is 8.6. In Fig. 3.4b, $m = 60$.



(A) Performances under varying number of samples.



(B) Performances under varying noise energy.

FIGURE 3.5: Boxplots of recovery performance under different frameworks via uniformly distributed sampling on $[-\pi, \pi]$. In Fig. 3.5a, the SNR is 8.6. In Fig. 3.5b, the number of samples m on each vertex is 60.

(3.20), and compare them with the empirical MSE. If the empirical MSE is close to the theoretical value (computed based on the estimated JPSD), this experiment demonstrates that it is possible to estimate the quality of a sampling set.

The experiment is done on the Molene weather dataset[3] published by the French national meteorological service. It contains hourly weather records in the region of Brest, France in January 2014. The temperature records we use are measured by 32 stations. We split the data into 31 periods, so that each period contains 24 hourly records. The graph is constructed similarly as in Section 3.4.2: we use 5-NN to connect the vertices according to their geographic distances $d(i,j)$, and assign the edges with weight $\exp(-d(i,j)^2/\sigma^2)$. Here, σ^2 denotes the variance of all distances between pairs of sensors. We randomly sample 20 days' records among 31 days as the training set, and the remaining 11 days' records form the test set. We aim to recover the test set from noisy observations on the sampled vertices. The noise energy is set to be 1/5 of the signal energy. In order to simulate different sampling strategies, we randomly generate the sample sets $\mathcal{U} \subset \mathcal{V}$ with different sizes $(1/5, 2/5, 3/5, 4/5)N$. For each sample size, we generate 20 sample sets. To recover the original signal and compute the theoretical MSE, we employ the Euclidean-vertex model. We observe from Fig. 3.6 that the empirical MSE is aligned with the theoretical MSE, hence (3.20) can be utilized to measure the quality of sampling sets.

3.A Appendix: Proof of Theorem 3.1 and Theorem 3.2

This appendix contains the proofs of Theorems 3.1 and 3.2, which assume $\mathcal{H} = L^2(\mathcal{T})$. The proofs generalize the ones given in [49], which assumes a one-dimensional stochastic process. The condition of continuity of trajectories and compactness of $\mathcal{T}$ are also relaxed here compared to [49].

Proof of Theorem 3.1. From Proposition 2.2 and condition (b), it suffices to show that for any $\mathbf{y} \in \mathbb{C}^N \otimes \mathcal{H}$, the map $\langle \mathbf{x}(\omega, \cdot), \mathbf{y} \rangle : \Omega \mapsto \mathbb{C}$ is measurable. Because $\mathbb{C}^N$ is finite dimensional, it suffices to consider $\mathbf{y}$ such that for each $\mathbf{t} \in \mathcal{T}$, $\mathbf{y}(\mathbf{t}) =$

[3] `https://donneespubliques.meteofrance.fr/donnees_libres/Hackathon/RADOMEH.tar.gz`

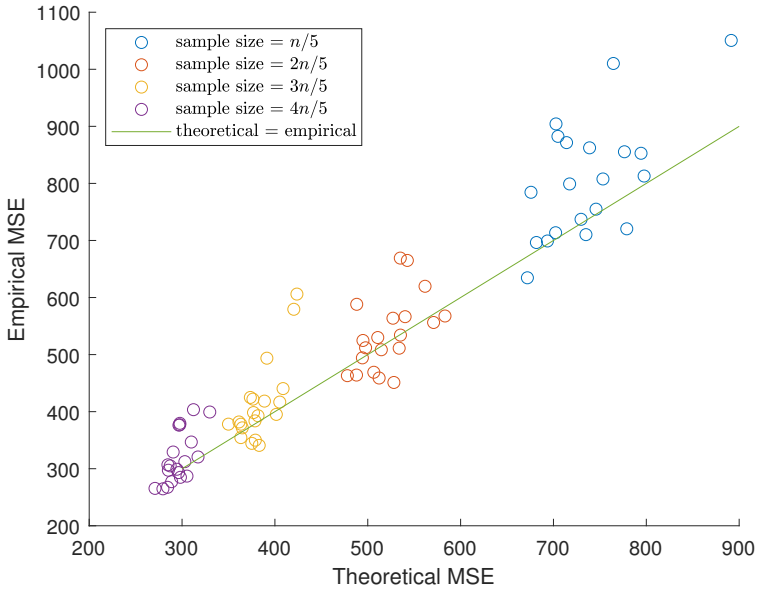


FIGURE 3.6: Theoretical and empirical recovery performances under different sample sets. Each point represents the theoretical and empirical errors of a sample set.

$\mathbf{y}(\cdot, \mathbf{t}) \in \mathbb{C}^N$. From Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\int_{\mathcal{T}} \int_{\Omega} |\mathbf{y}(\mathbf{t})^* \mathbf{x}(\omega, \mathbf{t})| \, \mathrm{d}\mathbb{P}(\omega) \, \mathrm{d}\tau(\mathbf{t}) \\
&\leq \int_{\mathcal{T}} \int_{\Omega} \|\mathbf{x}(\omega, \mathbf{t})\| \, \|\mathbf{y}(\mathbf{t})\| \, \mathrm{d}\mathbb{P}(\omega) \, \mathrm{d}\tau(\mathbf{t}) \\
&= \int_{\mathcal{T}} \mathbb{E}[\|\mathbf{x}(\mathbf{t})\|] \, \|\mathbf{y}(\mathbf{t})\| \, \mathrm{d}\tau(\mathbf{t}) \\
&\leq \left(\int_{\mathcal{T}} \mathbb{E}[\|\mathbf{x}(\mathbf{t})\|]^2 \, \mathrm{d}\tau(\mathbf{t}) \right)^{1/2} \left(\int_{\mathcal{T}} \|\mathbf{y}(\mathbf{t})\|^2 \, \mathrm{d}\tau(\mathbf{t}) \right)^{1/2} \\
&\leq \left(\int_{\mathcal{T}} \mathbb{E}\left[\|\mathbf{x}(\mathbf{t})\|^2\right] \mathrm{d}\tau(\mathbf{t}) \right)^{1/2} \left(\int_{\mathcal{T}} \|\mathbf{y}(\mathbf{t})\|^2 \, \mathrm{d}\tau(\mathbf{t}) \right)^{1/2}. \qquad (3.28)
\end{aligned}$$

From condition (c), we have

$$\int_{\mathcal{T}} \mathbb{E}\left[\|\mathbf{x}(\mathbf{t})\|^2\right] \mathrm{d}\tau(\mathbf{t}) = \int_{\mathcal{T}} \mathrm{tr}(\mathbf{K}_{\mathbf{x}}(\mathbf{t}, \mathbf{t})) \, \mathrm{d}\tau(\mathbf{t}) < \infty.$$

In addition, since $\mathbf{y}(v, \cdot) \in L^2(\mathcal{T})$ for each $v \in \mathcal{V}$, the R.H.S. of (3.28) is finite. By Fubini's theorem, we obtain that the following integral, as a function of ω,

$$\langle \mathbf{x}(\omega, \cdot), \mathbf{y} \rangle = \int_{\mathcal{T}} \mathbf{y}(\mathbf{t})^* \mathbf{x}(\omega, \mathbf{t}) \, \mathrm{d}\tau(\mathbf{t})$$

is measurable (in fact, integrable). This concludes the proof. □

Proof of Theorem 3.2. We first verify that the pointwise mean defined in Theorem 3.1, $\mathbf{m}_{\mathbf{x}}(\cdot) \in \mathbb{C}^N \otimes L^2(\mathcal{T})$ as follows:

$$\int_{\mathcal{T}} \|\mathbf{m}_{\mathbf{x}}(\mathbf{t})\|^2 \, \mathrm{d}\tau(\mathbf{t}) \leq \int_{\mathcal{T}} \mathbb{E}\big[\|\mathbf{x}(\mathbf{t})\|^2\big] \, \mathrm{d}\tau(\mathbf{t}) = \int_{\mathcal{T}} \mathrm{tr}(\mathbf{K}_{\mathbf{x}}(\mathbf{t},\mathbf{t})) \, \mathrm{d}\tau(\mathbf{t}) < \infty.$$

We then have for any $\mathbf{u} \in \mathbb{C}^N \otimes \mathcal{H}$ with $\mathbf{u}(\mathbf{t}) \in \mathbb{C}^N$ written in the form (3.1),

$$\langle \mathbf{m}_{\mathbf{x}}(\cdot), \mathbf{u}(\cdot) \rangle = \int_{\mathcal{T}} \mathbf{u}(\mathbf{t})^* \mathbf{m}_{\mathbf{x}}(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}) = \int_{\mathcal{T}} \int_{\Omega} \mathbf{u}(\mathbf{t})^* \mathbf{x}(\omega, \mathbf{t}) \, \mathrm{d}\mathbb{P}(\omega) \, \mathrm{d}\tau(\mathbf{t}).$$

Using the same argument as in the proof of Theorem 3.1, the integrals can be interchanged, so that

$$\langle \mathbf{m}_{\mathbf{x}}(\cdot), \mathbf{u}(\cdot) \rangle = \int_{\Omega} \int_{\mathcal{T}} \mathbf{u}(\mathbf{t})^* \mathbf{x}(\omega, \mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}) \, \mathrm{d}\mathbb{P}(\omega) = \mathbb{E}[\langle \mathbf{x}, \mathbf{u} \rangle],$$

which is the definition of the mean element of $\mathbf{x}$ in (2.15).

In the rest of the proof, without loss of generality, we assume $\mathbf{m}_{\mathbf{x}} = 0$. By definition (2.16), we have for any $\mathbf{f}, \mathbf{g} \in \mathbb{C}^N \otimes \mathcal{H}$,

$$\langle \mathbf{C}_{\mathbf{x}} \mathbf{f}, \mathbf{g} \rangle = \mathbb{E}\Big[\overline{\langle \mathbf{x}, \mathbf{f} \rangle} \langle \mathbf{x}, \mathbf{g} \rangle\Big] = \mathbb{E}\left[\int_{\mathcal{T} \times \mathcal{T}} \mathbf{f}(\mathbf{t})^{\intercal} \bar{\mathbf{x}}(\cdot, \mathbf{t}) \mathbf{x}(\cdot, \mathbf{s})^{\intercal} \bar{\mathbf{g}}(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{s})\right]. \quad (3.29)$$

In the proof of Theorem 3.1, we have shown that the function $\|\mathbf{x}(\omega, \mathbf{t})\| \, \|\mathbf{y}(\mathbf{t})\|$ is integrable on $\mathcal{T}$. We use this fact with Fubini's theorem to interchange the expectation and integral in (3.29) to obtain

$$\langle \mathbf{C}_{\mathbf{x}} \mathbf{f}, \mathbf{g} \rangle = \int_{\mathcal{T} \times \mathcal{T}} \mathbf{f}(\mathbf{t})^{\intercal} \overline{\mathbf{K}}_{\mathbf{x}}(\mathbf{t}, \mathbf{s}) \bar{\mathbf{g}}(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{s}) = \left\langle \int_{\mathcal{T}} \mathbf{K}_{\mathbf{x}}(\mathbf{s}, \mathbf{t}) \mathbf{f}(\mathbf{t}) \, \mathrm{d}\tau(\mathbf{t}), \mathbf{g}(\mathbf{s}) \right\rangle.$$

Note that the first element in the above inner product is an integral operator on $\mathbf{f}(\mathbf{t})$, therefore $\mathbf{C}_{\mathbf{x}}$ coincides with this operator by definition.

To prove the second part of Theorem 3.2, we employ the generalized Mercer's theorem in terms of a matrix-valued kernel [62]. From [62, Theorem A.1], there exists a sequence $\{\mathbf{f}_i(\mathbf{t}) : i = 1, 2, \ldots\} \subset \mathbb{C}^N \otimes L^2(\mathcal{T})$ such that

1. $\mathbf{C}_{\mathbf{x}} \mathbf{f}_i = \sigma_i \mathbf{f}_i$ with $\sigma_i > 0$;

2. $\{\mathbf{f}_i(\mathbf{t})\}$ forms an orthonormal basis of $\ker(\mathbf{C}_\mathbf{x})^\perp = \overline{\mathrm{im}(\mathbf{C}_\mathbf{x})}$;

3. $\{\mathbf{f}_i(\mathbf{t})\} \subset \mathcal{H}_K$, where $\mathcal{H}_K$ is the reproducing kernel Hilbert space induced by the kernel $\mathbf{K}_\mathbf{x}(\mathbf{s},\mathbf{t})$.

Using [62, Remark 3.3], since $\{\mathbf{f}_i(\mathbf{t})\} \subset \mathcal{H}_K$, $\mathbf{f}_i(\mathbf{t})$ are also continuous w.r.t. the topology induced by $\mathbf{K}_\mathbf{x}(\mathbf{s},\mathbf{t})$. Then according to [62, Theorem 3.4], the kernel function $\mathbf{K}_\mathbf{x}(\mathbf{s},\mathbf{t})$ can be decomposed as follows

$$\mathbf{K}_\mathbf{x}(\mathbf{s},\mathbf{t}) = \sum_{i=1}^{\infty} \sigma_i \bar{\mathbf{f}}_i(\mathbf{s}) \mathbf{f}_i(\mathbf{t})^\intercal,$$

for all $\mathbf{s},\mathbf{t} \in \mathcal{T}$ except on a zero-measure set. Let $\mathbf{f}_i(\mathbf{t}) = (f_i^{(1)}(\mathbf{t}), \ldots, f_i^{(N)}(\mathbf{t}))^\intercal$. Therefore, we can compute the integral of $\mathrm{tr}(\mathbf{K}_\mathbf{x}(\mathbf{t},\mathbf{t}))$ as

$$\begin{aligned}
\int_\mathcal{T} \mathrm{tr}(\mathbf{K}_\mathbf{x}(\mathbf{t},\mathbf{t}))\, \mathrm{d}\tau(\mathbf{t}) &= \int_\mathcal{T} \sum_{l=1}^{N} \sum_{i=1}^{\infty} \sigma_i |f_i^{(l)}(\mathbf{t})|^2 \, \mathrm{d}\tau(\mathbf{t}) \\
&= \int_\mathcal{T} \sum_{i=1}^{\infty} \sum_{l=1}^{N} \sigma_i |f_i^{(l)}(\mathbf{t})|^2 \, \mathrm{d}\tau(\mathbf{t}) \\
&= \sum_{i=1}^{\infty} \int_\mathcal{T} \sigma_i \left\|\mathbf{f}_i(\mathbf{t})\right\|^2 \mathrm{d}\tau(\mathbf{t}) \\
&= \sum_{i=1}^{\infty} \sigma_i \\
&= \mathrm{tr}(\mathbf{C}_\mathbf{x}),
\end{aligned}$$

which concludes the proof. □

3.B Appendix: Proof of Theorem 3.7

In this section, we prove Theorem 3.7. We start off with a lemma. Recall that $\mathbf{y}$ is restricted to a subset of vertices $\mathcal{U}$.

Lemma 3.8. *The null space and image space of* $\mathbf{C}_\mathbf{y}$ *are respectively*

$$\begin{aligned}
\ker \mathbf{C}_\mathbf{y} &= \overline{\mathrm{span}}\{\mathbf{e}_i \otimes \psi_l \,:\, i \in \mathcal{U}^\mathsf{c} \text{ or } l \geq l'\} := V_0, \\
\mathrm{im}\, \mathbf{C}_\mathbf{y} &= \mathrm{span}\{\mathbf{e}_i \otimes \psi_l \,:\, i \in \mathcal{U},\ l < l'\} := V_1.
\end{aligned}$$

Proof. Note that $\mathbf{C}_\mathbf{y} = \mathbf{\Pi}_\mathcal{S}(\mathbf{C}_\mathbf{x} + \mathbf{C}_\epsilon)\mathbf{\Pi}_\mathcal{S}$. First notice that $\mathbf{C}_\mathbf{y}(\mathbf{e}_i \otimes \psi_l) = 0$ when $i \in \mathcal{U}^\mathsf{c}$ or $l \geq l'$. Then the continuity of $\mathbf{C}_\mathbf{y}$ implies that $V_0 \subset \ker \mathbf{C}_\mathbf{y}$. When $i \in \mathcal{U}$ and $l < l'$, we have

$$\begin{aligned}
\mathbf{C}_\mathbf{y}(\mathbf{e}_i \otimes \psi_l) &= \mathbf{\Pi}_\mathcal{S}(\mathbf{C}_\mathbf{x} + \mathbf{C}_\epsilon)(\mathbf{e}_i \otimes \psi_l) \\
&= \mathbf{\Pi}_\mathcal{S}(\sum_{l=1}^{\infty}\sum_{k=1}^{N}(p_\mathbf{x}(k,l) + p_\epsilon(k,l))\mathbf{\Pi}_{\boldsymbol{\phi}_k \otimes \psi_l}(\mathbf{e}_i \otimes \psi_l)) \\
&= \mathbf{\Pi}_\mathcal{S}(\sum_{k=1}^{N}(p_\mathbf{x}(k,l) + p_\epsilon(k,l))\mathbf{\Phi}(i,k)\boldsymbol{\phi}_k \otimes \psi_l) \\
&= \sum_{k=1}^{N}(p_\mathbf{x}(k,l) + p_\epsilon(k,l))\mathbf{\Phi}(i,k)\mathbf{\Pi}_\mathcal{S}(\boldsymbol{\phi}_k \otimes \psi_l) \\
&= \sum_{k=1}^{N}(p_\mathbf{x}(k,l) + p_\epsilon(k,l))\mathbf{\Phi}(i,k)\sum_{j\in\mathcal{U}}\mathbf{\Phi}(j,k)\mathbf{e}_j \otimes \psi_l \\
&= \sum_{j\in\mathcal{U}}(\sum_{k=1}^{N}(p_\mathbf{x}(k,l) + p_\epsilon(k,l))\mathbf{\Phi}(j,k)\mathbf{\Phi}(i,k))\mathbf{e}_j \otimes \psi_l \\
&= \sum_{j\in\mathcal{U}}\mathbf{O}^{(l)}(i,j)\mathbf{e}_j \otimes \psi_l,
\end{aligned}$$

where $\mathbf{O}^{(l)} = \mathbf{\Phi}_\mathcal{U}\mathbf{\Lambda}_l\mathbf{\Phi}_\mathcal{U}^\mathsf{T}$. In order to simplify notations, we let the row and column indices of $\mathbf{O}^{(l)}$ be consistent with $\mathcal{U}$. This result indicates that $\operatorname{im}\mathbf{C}_\mathbf{y} \subset V_1$. We note that when $l < l'$, $\mathbf{O}^{(l)}$ is invertible as a principal submatrix of a positive definite matrix $\mathbf{\Phi}\mathbf{\Lambda}_l\mathbf{\Phi}^\mathsf{T}$. This implies that for $l < l'$, when restricted on

$$\widetilde{V}_l := \operatorname{span}\{\mathbf{e}_i \otimes \psi_l \,:\, i \in \mathcal{U}\}, \tag{3.30}$$

$\mathbf{C}_\mathbf{y}$ can be represented as an invertible matrix $\mathbf{O}^{(l)}$. Therefore, the basis of V_1 is a subset of $\operatorname{im}\mathbf{C}_\mathbf{y}$. Since $\operatorname{im}\mathbf{C}_\mathbf{y}$ is closed, it follows that $V_1 \subset \operatorname{im}\mathbf{C}_\mathbf{y}$, i.e., $\operatorname{im}\mathbf{C}_\mathbf{y} = V_1$. Since $V_0 = V_1^\perp$ and $\ker\mathbf{C}_\mathbf{y} = \operatorname{im}\mathbf{C}_\mathbf{y}^\perp$, we obtain $\ker\mathbf{C}_\mathbf{y} = V_0$. The lemma is now proved. □

We now return to the proof of Theorem 3.7. Define $\widehat{\mathbf{x}} = \mathbf{G}(\mathbf{y})$ as the Wiener filter (i.e., BLUE) and $\mathbf{r} := \widehat{\mathbf{x}} - \mathbf{x}$ as the estimation error. According to Proposition 2.3, $\mathbb{E}\,\|\mathbf{r}\|^2 = \operatorname{tr}(\mathbf{C}_\mathbf{r})$. In the sequel, we are going to compute it as a function of $\mathcal{U}$.

From Definition 2.8, $\mathbf{C_r}$ equals the ALCC $\mathrm{cov}_{\mathbf{y}}^{\mathsf{A}}[\mathbf{x}]$. Therefore, it can be written as

$$\begin{aligned}\mathbf{C_r} &= \mathbf{C_x} - \mathbf{C_{xy}}\mathbf{C_y^\dagger}\mathbf{C_{yx}} \\ &= \mathbf{C_x} - \mathbf{C_x}\mathbf{\Pi}_{\mathcal{S}}(\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x} \\ &:= \mathbf{C_x} - \mathbf{C}_0.\end{aligned}$$

Next, we compute $\mathrm{tr}(\mathbf{C_r})$, the main step of which is to deal with $\mathrm{tr}(\mathbf{C}_0)$. By the operator trace definition in (2.21),

$$\begin{aligned}\mathrm{tr}(\mathbf{C}_0) &= \sum_{l=1}^{\infty}\sum_{k=1}^{N}\langle \mathbf{C}_0(\boldsymbol{\phi}_k\otimes\psi_l), \boldsymbol{\phi}_k\otimes\psi_l\rangle \\ &= \sum_{l=1}^{l'}\sum_{k=1}^{N}\langle (\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l), \mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l)\rangle. \quad (3.31)\end{aligned}$$

To simplify this expression, we compute the elements in the inner products as follows:

$$\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l) = p_{\mathbf{x}}(k,l)\mathbf{\Pi}_{\mathcal{S}}(\boldsymbol{\phi}_k\otimes\psi_l) = p_{\mathbf{x}}(k,l)\sum_{i\in\mathcal{U}}\mathbf{\Phi}(i,k)\mathbf{e}_i\otimes\psi_l,$$

$$(\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l) = p_{\mathbf{x}}(k,l)\sum_{i\in\mathcal{U}}\mathbf{\Phi}(i,k)(\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger(\mathbf{e}_i\otimes\psi_l).$$

Note that $\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}} = \mathbf{C_y}$, hence it is guaranteed to be compact and self-adjoint. To compute $\mathbf{C_y^\dagger}(\mathbf{e}_i\otimes\psi_l)$, we utilize the result from Lemma 3.8, which characterizes the restriction of $\mathbf{C_y}$ on $\widetilde{V}_l$ as an invertible matrix $\mathbf{O}^{(l)}$. Therefore, $\sum_{j\in\mathcal{U}}\mathbf{O}^{(l)-1}(i,j)\mathbf{e}_j\otimes\psi_l$ is the preimage of $\mathbf{e}_i\otimes\psi_l$. Besides, since it also belongs to $(\ker\mathbf{C_y})^\perp$, we have $\mathbf{C_y^\dagger}(\mathbf{e}_i\otimes\psi_l) = \sum_{j\in\mathcal{U}}\mathbf{O}^{(l)-1}(i,j)\mathbf{e}_j\otimes\psi_l$. Hence,

$$\begin{aligned}(\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l) &= p_{\mathbf{x}}(k,l)\sum_{i\in\mathcal{U}}\mathbf{\Phi}(i,k)\mathbf{C_y^\dagger}(\mathbf{e}_i\otimes\psi_l) \\ &= p_{\mathbf{x}}(k,l)\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}\mathbf{\Phi}(i,k)\mathbf{O}^{(l)-1}(i,j)\mathbf{e}_j\otimes\psi_l.\end{aligned}$$

By substituting this result into each term in the double sum of (3.31), we obtain

$$\begin{aligned}&\langle (\mathbf{\Pi}_{\mathcal{S}}(\mathbf{C_x}+\mathbf{C}_\epsilon)\mathbf{\Pi}_{\mathcal{S}})^\dagger\mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l), \mathbf{\Pi}_{\mathcal{S}}\mathbf{C_x}(\boldsymbol{\phi}_k\otimes\psi_l)\rangle \\ &= p_{\mathbf{x}}(k,l)^2\langle\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}\mathbf{\Phi}(i,k)\mathbf{O}^{(l)-1}(i,j)\mathbf{e}_j\otimes\psi_l, \sum_{s\in\mathcal{U}}\mathbf{\Phi}(s,k)\mathbf{e}_s\otimes\psi_l\rangle\end{aligned}$$

$$
\begin{aligned}
&= p_{\mathbf{x}}(k,l)^2 \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{U}} \mathbf{\Phi}(i,k)\mathbf{O}^{(l)-1}(i,j)\mathbf{\Phi}(j,k) \\
&= p_{\mathbf{x}}(k,l)^2 \mathbf{\Phi}_{\mathcal{U}}^{\intercal}(k,:)\mathbf{O}^{(l)-1}\mathbf{\Phi}_{\mathcal{U}}(:,k),
\end{aligned}
$$

where $\mathbf{\Phi}_{\mathcal{U}}^{\intercal}(k,:)$ denotes the k-th row of $\mathbf{\Phi}_{\mathcal{U}}^{\intercal}$.

Then, we have

$$
\begin{aligned}
\operatorname{tr}(\mathbf{C}_0) &= \sum_{l=1}^{l'} \sum_{k=1}^{N} p_{\mathbf{x}}(k,l)^2 \mathbf{\Phi}_{\mathcal{U}}^{\intercal}(k,:)\mathbf{O}^{(l)-1}\mathbf{\Phi}_{\mathcal{U}}(:,k) \\
&= \sum_{l=1}^{l'} \sum_{k=1}^{N} p_{\mathbf{x}}(k,l)^2 \operatorname{tr}(\mathbf{O}^{(l)-1}\mathbf{\Phi}_{\mathcal{U}}(:,k)\mathbf{\Phi}_{\mathcal{U}}^{\intercal}(k,:)) \\
&= \sum_{l=1}^{l'} \operatorname{tr}(\mathbf{O}^{(l)-1} \sum_{k=1}^{N} p_{\mathbf{x}}(k,l)^2 \mathbf{\Phi}_{\mathcal{U}}(:,k)\mathbf{\Phi}_{\mathcal{U}}^{\intercal}(k,:)) \\
&= \sum_{l=1}^{l'} \operatorname{tr}((\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Lambda}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal})^{-1}\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Gamma}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal}).
\end{aligned}
$$

Finally, we obtain the MSE of the Wiener filter $\mathbf{G}$ as

$$
\begin{aligned}
\mathbb{E}\,\|\mathbf{r}\|^2 &= \operatorname{tr}(\mathbf{C}_{\mathbf{x}}) - \sum_{l=1}^{l'} \operatorname{tr}((\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Lambda}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal})^{-1}\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Gamma}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal}) \\
&= \sum_{l=1}^{l'} \sum_{k=1}^{N} p_{\mathbf{x}}(k,l) - \sum_{l=1}^{l'} \operatorname{tr}((\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Lambda}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal})^{-1}\mathbf{\Phi}_{\mathcal{U}}\mathbf{\Gamma}_l\mathbf{\Phi}_{\mathcal{U}}^{\intercal}).
\end{aligned}
$$

The proof is now complete.

Chapter 4

Kernel Based Reconstruction for Generalized Graph Signals

4.1 Problem Formulation

In this section, we formulate the generalized graph signal reconstruction problem.

Consider a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, N\}$ is the vertex set, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the edge set. We use $\mathcal{N}_d(v)$ to denote the d-hop neighborhood of the vertex v and let $\overline{\mathcal{N}}_d(v) = \mathcal{N}_d(v) \cup \{v\}$. We assume that G is a connected undirected graph with no self-loops. One important case in GGSP is where $\mathcal{H}$ is a function space. Specifically, consider the domain of the functions to be a measure space $(\mathcal{T}, \mathcal{A}, \tau)$ and $\mathcal{H} = L^2(\mathcal{T})$ (i.e., the space of square-integrable functions on $\mathcal{T}$). For example, in Intel lab data mentioned in Section 1.2.2, $\mathcal{T} = [0, 86400]$, representing the time duration (in seconds) of one day. Then, a generalized graph signal f can be identified with the map

$$\begin{aligned} f' : \mathcal{V} \times \mathcal{T} &\to \mathbb{R} \\ (v, \mathbf{t}) &\mapsto f(v)(\mathbf{t}). \end{aligned}$$

Thus, the space of generalized graph signals can be also identified with $L^2(\mathcal{V} \times \mathcal{T})$. In this chapter, we will mainly use $L^2(\mathcal{V} \times \mathcal{T})$ to denote the space of generalized graph signals, while references to $\mathbb{R}^N \otimes \mathcal{H}$ are used in explanations and proofs. We

refer to $\mathcal{T}$ colloquially as the *time* domain. However, it is not restricted to subsets of $\mathbb{R}$ and can be a general measure space.

Given noisy observation samples at a subset $\mathcal{S} \subset \mathcal{V} \times \mathcal{T}$ of vertices and time instances, our objective is to recover the generalized graph signal f. To avoid cluttered notations, denote $\mathcal{J} = \mathcal{V} \times \mathcal{T}$. Suppose the sampling set is $\mathcal{S} = \{(v_m, \mathbf{t}_m) : m = 1, \dots, M\} \subset \mathcal{J}$, and the noisy observations are

$$y_m = f(v_m, \mathbf{t}_m) + \epsilon_m, \ m = 1, \dots, M, \tag{4.1}$$

where ϵ_m are i.i.d. zero-mean noise with variance σ^2. In the Bayesian framework, f in (4.1) is further modeled as a Gaussian process. In this case, we will model f as a random element (cf . Section 2.3). The noise terms ϵ_m are assumed to be Gaussian and independent of this process.

The GGSP signal reconstruction problem can be summarized in the following form:

$$\min_{\tilde{f} \in F(\mathcal{J}, \mathbb{R})} \sum_{m=1}^{M} L(\tilde{f}(v_m, \mathbf{t}_m), y_m) + P(\tilde{f}), \tag{4.2}$$

where $F(\mathcal{J}, \mathbb{R})$ is an appropriate space of functions from $\mathcal{J}$ to $\mathbb{R}$, $L(\cdot)$ is a loss function measuring the fitness of $\tilde{f}$ on the observations. Typical choices include the ℓ_1 and ℓ_2 losses. The regularization term $P(\tilde{f})$ imposes a smoothness constraint on $\tilde{f}$ over the vertex and time domains. To design proper $F(\mathcal{J}, \mathbb{R})$ and $P(\tilde{f})$, we employ the KRR technique, which we briefly review in Section 2.5.

The existing time-vertex methods [39, 40] have already addressed the reconstruction problem for time series on graphs. However, these methods are based on the assumption that the signals are evenly sampled with the same sampling rate on all vertices. In contrast, from (4.2), we observe that our formulation does not require synchronous samples from each vertex and applies even in the case where the sampling frequencies differ across vertices, or where the signal is not evenly sampled. In addition, compared to the time-vertex techniques, this formulation is not sensitive to the sampling rate since it makes use of the true time stamps. We refer the reader to the detailed discussion in Section 4.2.2.

4.2 KRR Reconstruction in GGSP

In this section, we derive the KRR reconstruction solution for GGSP. We interpret this method under both deterministic and Bayesian models and connect our technique with existing kernel-based frameworks in GSP and graph signal reconstruction approaches. We also propose an online approach based on RFF that results in a distributed implementation.

To reconstruct a generalized graph signal $f \in L^2(\mathcal{J})$, we use a kernel $k : \mathcal{J} \times \mathcal{J} \to \mathbb{R}$ that is the multiplication of two kernels $k_G : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and $k_{\mathcal{T}} : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$:

$$\begin{aligned} k : \mathcal{J} \times \mathcal{J} &\to \mathbb{R} \\ ((u, \mathbf{s}), (v, \mathbf{t})) &\mapsto k_G(u, v) k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}). \end{aligned} \tag{4.3}$$

The RKHS associated with the kernel (4.3) is $\mathcal{H}_k = \mathcal{H}_{k_G} \otimes \mathcal{H}_{k_{\mathcal{T}}}$ [58, Theorem 13]. In this paper, we construct k_G based on a GSO $\mathbf{A}_G$ of the graph G. In particular, we focus on the case where the matrix $\mathbf{K}_G := (k_G(i, j)) \in \mathbb{R}^{N \times N}$ takes the following form (cf. [11, (14)]):

$$\mathbf{K}_G = \mathbf{\Phi} \operatorname{diag}(r(\lambda_1), \ldots, r(\lambda_N)) \mathbf{\Phi}^{\mathsf{T}}, \tag{4.4}$$

where $\{\lambda_i\}$ are the eigenvalues of the GSO $\mathbf{A}_G$, $r(\cdot)$ is a non-negative function such that $r(\lambda_1) \geq \cdots \geq r(\lambda_N)$,[1] and $\mathbf{\Phi}$ is the matrix formed by the eigenvectors of $\mathbf{A}_G$. When $\mathcal{T}$ is a subset of Euclidean space, we can usually choose $k_{\mathcal{T}}$ as the RBF kernel, e.g., $k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \exp(-\left\|\mathbf{s} - \mathbf{t}\right\|_2^2 / \beta_{\text{scale}})$ (Gaussian kernel) or $k_{\mathcal{T}}(\mathbf{s}, \mathbf{t}) = \exp(-\left\|\mathbf{s} - \mathbf{t}\right\|_1 / \beta_{\text{scale}})$ (Laplacian kernel), where β_{scale} is a tunable parameter.

Following the standard KRR formulation (2.30), we specify the reconstruction problem (4.2) as follows:

$$\hat{f} = \arg\min_{\tilde{f} \in \mathcal{H}_k} \sum_{m=1}^{M} \left| \tilde{f}(v_m, \mathbf{t}_m) - y_m \right|^2 + \mu \left\| \tilde{f} \right\|_{\mathcal{H}_k}^2. \tag{4.5}$$

[1] Recall that $\{\lambda_i\}$ are indexed in increasing order of graph frequencies. Also note that [11, (14)] uses $r^{\dagger}(\mathbf{\Lambda})$ instead of $r(\mathbf{\Lambda})$ in the definition (4.4).

Let $\mathbf{K}(\mathcal{S},\mathcal{S}) = (k((v_m,\mathbf{t}_m),(v_{m'},\mathbf{t}_{m'})))_{m,m'=1}^{M} \in \mathbb{R}^{M\times M}$ and $\mathbf{y}(\mathcal{S}) = (y_1,\dots,y_M)^\intercal$. Using the representer theorem, the optimal solution to (4.5) is

$$\begin{aligned} \hat{f} &= \sum_{m=1}^{M} c_m k(\cdot,(v_m,\mathbf{t}_m)), \\ (c_1,\dots,c_M)^\intercal &= (\mathbf{K}(\mathcal{S},\mathcal{S}) + \mu\mathbf{I}_M)^{-1}\mathbf{y}(\mathcal{S}). \end{aligned} \tag{4.6}$$

Henceforth, we refer to the problem (4.5) and its solution (4.6) as KRR-GGSP. In this chapter, we assume that all eigenvalues of $\mathbf{A}_G$ are distinct. By construction (4.4), $\mathbf{K}_G$ is a polynomial of $\mathbf{A}_G$ for some degree $L < N$, i.e., it suffices to consider $r(\cdot)$ as a polynomial whose degree is smaller than N, thus $k_G(u,v) = 0$ as long as $u \notin \overline{\mathcal{N}}_L(v)$. Therefore, the evaluation of $\hat{f}(v,\mathbf{t})$ only requires information from $\overline{\mathcal{N}}_L(v)$:

$$\begin{aligned} \hat{f}(v,\mathbf{t}) &= \sum_{m=1}^{M} c_m k((v,\mathbf{t}),(v_m,\mathbf{t}_m)) \\ &= \sum_{m=1}^{M} c_m k_G(v,v_m) k_{\mathcal{T}}(\mathbf{t},\mathbf{t}_m) \\ &= \sum_{v_m\in\overline{\mathcal{N}}_L(v)} c_m k_G(v,v_m) k_{\mathcal{T}}(\mathbf{t},\mathbf{t}_m). \end{aligned} \tag{4.7}$$

Note that when $\mathcal{T}$ is a singleton (i.e., the vertex signal space is one-dimensional), the KRR-GGSP framework degenerates to the GSP recovery problem [11]. In addition, when $\mathbf{K}_G = \mathbf{I}_N$, it degenerates to separately solving KRR problems on each vertex using the kernel $k_{\mathcal{T}}$. To see this, suppose on each vertex v we have M_v samples. We relabel $\mathcal{S}$ and $\{y_m\}$ such that $\mathcal{S} = \bigcup_{v\in\mathcal{V}} \{(v,\mathbf{t}_i^{(v)}) : i = 1,\dots,M_v\}$, $\{y_m\} = \bigcup_{v\in\mathcal{V}} \{y_i^{(v)} : i = 1,\dots,M_v\}$. We also relabel the coefficients as $c_i^{(v)}$, so that (4.6) can be rewritten as

$$\hat{f}(u,\mathbf{t}) = \sum_{v=1}^{N} \delta(u,v) \sum_{i=1}^{M_u} c_i^{(v)} k_{\mathcal{T}}(\mathbf{t},\mathbf{t}_i^{(v)})$$

for each $u \in \mathcal{V}$ and $\mathbf{t} \in \mathcal{T}$, where $\delta(u,v) = 1$ when $u = v$ and $\delta(u,v) = 0$ otherwise. Note that $\hat{f}(u,\mathbf{t}) = \sum_{i=1}^{M_u} c_i^{(u)} k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}_i^{(u)})$ and

$$\left\|\hat{f}\right\|_{\mathcal{H}_k}^2 = \sum_{u=1}^{N}\sum_{i,j=1}^{M_u} c_i^{(u)} k_{\mathcal{T}}(\mathbf{t}_i^{(u)}, \mathbf{t}_j^{(u)}) c_j^{(u)} = \sum_{u=1}^{N} \left\|\hat{f}(u,\cdot)\right\|_{\mathcal{H}_{k_{\mathcal{T}}}}^2 .$$

Then problem (4.5) becomes

$$\hat{f} = \arg\min_{\tilde{f}\in\mathcal{H}_k} \sum_{u=1}^{N}\sum_{i=1}^{M_u} \left|\tilde{f}(u,\mathbf{t}_i^{(u)}) - y_i^{(u)}\right|^2 + \mu \sum_{u=1}^{N} \left\|\tilde{f}(u,\cdot)\right\|_{\mathcal{H}_{k_{\mathcal{T}}}}^2 , \tag{4.8}$$

and each $\hat{f}(u,\cdot)$ can be solved separately using the samples on the vertex u.

In the rest of this chapter, we make the following assumption.

Assumption 4.1. For the measure space $(\mathcal{T}, \mathcal{A}, \tau)$, $\mathcal{T}$ is a compact metric space, $\mathcal{A}$ is the Borel σ-algebra, and τ is a strictly positive finite Borel measure. The kernel $k_{\mathcal{T}}$ is a continuous symmetric positive definite kernel and $\mathbf{K}_G$ is a positive definite matrix.

4.2.1 Deterministic Interpretation

In this subsection, we consider the case where f in (4.1) is deterministic. Under Assumption 4.1, by Mercer's theorem [57], there exists an orthonormal sequence $\{\xi_i : i \geq 1\}$ in $L^2(\mathcal{T})$ such that:

$$\int_{\mathcal{T}} k_{\mathcal{T}}(\mathbf{s},\mathbf{t})\xi_i(\mathbf{s})\,\mathrm{d}\tau(\mathbf{s}) = \gamma_i \xi_i(\mathbf{t}),$$
$$\int_{\mathcal{T}} \xi_i(\mathbf{s})\xi_j(\mathbf{s})\,\mathrm{d}\tau(\mathbf{s}) = \delta(i,j),$$
$$k_{\mathcal{T}}(\mathbf{s},\mathbf{t}) = \sum_{i=1}^{\infty} \gamma_i \xi_i(\mathbf{s})\xi_i(\mathbf{t}),$$

where the sum converges absolutely and uniformly on $\mathcal{T}$ and γ_i, $i \geq 1$, are non-negative eigenvalues. Let $\phi_n(u)$ be the (n,u)-th element of $\mathbf{\Phi}$. Since k_G is given by

(4.4), it can be decomposed in the same way:

$$k_G(u,v) = \sum_{n=1}^{N} r(\lambda_n)\phi_n(u)\phi_n(v).$$

By definition of k in (4.3), we then have

$$k((u,\mathbf{s}),(v,\mathbf{t})) = \sum_{n=1}^{N}\sum_{i=1}^{\infty} r(\lambda_n)\gamma_i \cdot \phi_n(u)\xi_i(\mathbf{s}) \cdot \phi_n(v)\xi_i(\mathbf{t}).$$

Note that $\{\phi_n(\cdot)\xi_i(\cdot) : n = 1,\dots,N, i \geq 1\}$ is an orthonormal sequence in $L^2(\mathcal{J})$. Following the same argument as [72], $\mathcal{H}_k$ is a subset of $L^2(\mathcal{J})$ where the functions $\tilde{f}$ satisfy the following condition:

$$\begin{aligned} &\tilde{f}(v,\mathbf{t}) = \sum_{n=1}^{N}\sum_{i=1}^{\infty} c_{n,i} \cdot \phi_n(v)\xi_i(\mathbf{t}) \\ &\text{s. t. } \left\|\tilde{f}\right\|_{\mathcal{H}_k}^2 = \sum_{n=1}^{N}\sum_{i=1}^{\infty} \frac{c_{n,i}^2}{r(\lambda_n)\gamma_i} < \infty. \end{aligned} \tag{4.9}$$

By the definition of JFT (cf. (2.13)), it can be shown that $c_{n,i} = \mathfrak{F}_{n,i}(\tilde{f})$ where $\mathfrak{F}_{n,i}$ represents the (n,i)-th JFT coefficient. Therefore, penalizing on $\left\|\tilde{f}\right\|_{\mathcal{H}_k}$ is the same as penalizing on the energy of $\mathfrak{F}_{n,i}(\tilde{f})$ with weights $\dfrac{1}{r(\lambda_n)\gamma_i}$. Note that $r(\cdot)$ is non-increasing so that the Fourier coefficients associated with larger graph frequencies are more heavily penalized.

It is worth noting that if we construct $k_{\mathcal{T}}$ and k_G such that

$$k_{\mathcal{T}}(\mathbf{s},\mathbf{t}) = \sum_{i=1}^{B'} \gamma_i \xi_i(\mathbf{s})\xi_i(\mathbf{t}) \tag{4.10}$$

for some $B' < \infty$, and $r(\lambda_n) = 0$ for all $n > B''$ in (4.4), then problem (4.5) is equivalent to the bandlimited signal reconstruction in [22, Section VI.A] with an additional ridge penalty. To see this, we first note that $\mathcal{H}_k = \text{span}\{\phi_n(\cdot)\xi_i(\cdot) : n = 1,\dots,B'', i = 1,\dots,B'\}$, i.e., the signal space used for reconstruction is a bandlimited space. We substitute (4.10) into (4.9) to obtain the optimization

problem

$$\hat{f}(v,\mathbf{t}) = \underset{\tilde{f}\in\mathcal{H}_k}{\arg\min} \sum_{m=1}^{M} \left|\tilde{f}(v_m,\mathbf{t}_m) - y_m\right|^2 + \mu \sum_{n=1}^{B''}\sum_{i=1}^{B'} \frac{c_{n,i}^2}{r(\lambda_n)\gamma_i}, \tag{4.11}$$

which coincides with the bandlimited signal reconstruction problem formulated in [22] but with an additional penalty term. This indicates that if $k_{\mathcal{T}}$ is not a combination of finite functions, then $\dim(\mathcal{H}_k) = \infty$. This implies that the algorithm is able to capture more features than that of bandlimited signals. An example is the Gaussian kernel [73, Section 4.3.1].

Finally, we discuss the universality (see Section 2.5 for definition) of the kernel k in the following theorem. The definition of universality requires defining a topology on $\mathcal{J}$. In this paper, we equip $\mathcal{V}$ with the discrete topology and $\mathcal{J} = \mathcal{V} \times \mathcal{T}$ the product topology.

Theorem 4.1. *If $k_{\mathcal{T}}$ is a universal kernel on $\mathcal{T}$, then k is universal on $\mathcal{J}$.*

Proof. Consider an arbitrary compact set $\mathcal{Z}_J \subset \mathcal{V} \times \mathcal{T}$, and define $\mathcal{Z}_v$ such that $\{v\} \times \mathcal{Z}_v = \mathcal{Z}_J \cap (\{v\} \times \mathcal{T})$. By using the finite-cover definition of a compact set, we note that $\mathcal{Z}_v$ is compact in $\mathcal{T}$. Consider an arbitrary $h \in \mathcal{C}(\mathcal{Z}_J)$, where $\mathcal{C}(\mathcal{Z}_J)$ is the space of continuous functions on $\mathcal{Z}_J$ equipped with the supremum norm. Let $h_v := h|_{\{v\}\times\mathcal{Z}_v}$. Due to the universality of $k_{\mathcal{T}}$, for any $\epsilon > 0$, there exists $h'_v \in \operatorname{span}\{k_{\mathcal{T}}(\cdot,\mathbf{t}) : \mathbf{t} \in \mathcal{Z}_v\}$ such that $\|h'_v - h_v(v,\cdot)\|_{\mathcal{C}(\mathcal{Z}_v)} < \epsilon$. Let $h' := \sum_{v=1}^{N} \delta(v,\cdot)h'_v$. Then, we have $\|h' - h\|_{\mathcal{C}(\mathcal{Z}_J)} < \epsilon$. On the other hand, since $\mathbf{K}_G$ is positive definite, $\mathbf{K}_G$ is invertible. Therefore, there exists $\{a_{v,n}\}$ such that $\delta(v,\cdot) = \sum_{n=1}^{N} a_{v,n}k_G(n,\cdot)$, i.e., $\delta(v,\cdot) \in \operatorname{span}\{k_G(n,\cdot) : n = 1,\ldots,N\}$. Let $\mathcal{K}(\mathcal{Z}_J)$ be the closure of $\operatorname{span}\{k(\cdot,(u,\mathbf{s})) : (u,\mathbf{s}) \in \mathcal{V}\times\mathcal{T}\}$ in $\mathcal{C}(\mathcal{Z}_J)$. By combining the above results, we conclude that $h' \in \mathcal{K}(\mathcal{Z}_J)$ and the universality of k follows by $\mathcal{K}(\mathcal{Z}_J) = \mathcal{C}(\mathcal{Z}_J)$. □

The universality discussed in Theorem 4.1 is different from that in [74, Theorem 2], which established universality for the following operator-valued kernels:

$$\begin{aligned}\mathbf{K} : \mathcal{X}\times\mathcal{X} &\to \mathcal{L}(\mathcal{Y})\\ (\mathbf{x}_1,\mathbf{x}_2) &\mapsto k_s(\mathbf{x}_1,\mathbf{x}_2)\mathbf{T}\end{aligned}$$

where $\mathcal{X} \subset L^2(\mathcal{J})$ and $\mathcal{Y} \subset L^2(\mathcal{J})$, $k_s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a real-valued RBF kernel, $\mathcal{L}(\mathcal{Y})$ is the space of linear operators on $\mathcal{Y}$, and $\mathbf{T} \in \mathcal{L}(\mathcal{Y})$.

We note that Theorem 4.1 cannot be derived from [74, Theorem 2] and vice versa. First, the kernel domain in this paper is in $\mathcal{J} \times \mathcal{J}$ instead of $L^2(\mathcal{J}) \times L^2(\mathcal{J})$. For simplicity, consider the case where $\mathcal{T}$ is a singleton, so that $\mathcal{V} \times \mathcal{T}$ can be identified with $\mathcal{V}$. If we use the kernel in [74, Theorem 2] and let $\mathcal{X} = \mathcal{V}$, then it is required that $\mathcal{V}$ is a real (or complex) separable Hilbert space. However, as long as $1 < |\mathcal{V}| < \infty$, this is impossible. Second, the output of the kernel in this paper is in $\mathbb{R}$ instead of an operator space, hence none of these two formulations encompasses the other.

4.2.2 Bayesian Interpretation

We now turn to the Bayesian interpretation where $f \sim \mathcal{GP}(0, k)$ and $\epsilon_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ in (4.1). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space, where $\mathcal{F}$ stands for the σ-algebra of the space. We regard $\mathcal{J} = \mathcal{V} \times \mathcal{T}$ as a measure space whose measure is the product measure of counting measure on $\mathcal{V}$ and the measure τ on $\mathcal{T}$. We denote this product measure as ζ. To be specific, f is a stochastic process $\{f((v, \mathbf{t}), \omega) : (v, \mathbf{t}) \in \mathcal{J}, \omega \in \Omega\}$. We make the following assumptions:

Assumption 4.2.

i) $f((v, \mathbf{t}), \omega)$ is jointly measurable w.r.t. the product measure $\zeta \times \mathbb{P}$.

ii) $f(\cdot, \omega) \in L^2(\mathcal{J})$ for all $\omega \in \Omega$.

Under Assumption 4.2, f is a Gaussian random element (cf. Theorem 2.1). Henceforth, we abbreviate $f((v, \mathbf{t}), \omega)$ as $f(v, \mathbf{t})$ for simplicity and consistent notations. First, we note that under the time-vertex framework, the GP prior $\mathcal{GP}(0, k)$ is a JWSS GRP. Recall that a stochastic process f on $\mathcal{V} \times \mathcal{T}$ is said to be JWSS if its covariance operator commutes with the shift operator $\mathbf{S} := \mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$ on $L^2(\mathcal{J})$ (cf . Definition 3.2), where $\mathbf{A}_{\mathcal{H}}$ is the shift operator on $L^2(\mathcal{T})$. Consider the case where $\mathcal{T} = \{1, \ldots, T\}$, and $\mathbf{K}_{\mathcal{T}} := (k_{\mathcal{T}}(i, j)) \in \mathbb{R}^{T \times T}$ is a symmetric positive-definite circulant matrix. Then the covariance operator of $\mathcal{GP}(0, k)$ is $\mathbf{C}_f = \mathbf{K}_G \otimes \mathbf{K}_{\mathcal{T}}$. Let

$\mathbf{A}_{\mathcal{H}}$ be the shift operator

$$\mathbf{A}_{\mathcal{H}}(g)(t) = g((t+1) \bmod T),$$

which models the case where the vertex observation is a discrete-time signal with T time steps. Since $\mathbf{K}_{\mathcal{T}}$ is a circulant matrix, it commutes with $\mathbf{A}_{\mathcal{H}}$. On the other hand, by the construction of the kernel k_G in (4.4), we know that $\mathbf{K}_G$ commutes with $\mathbf{A}_G$. Therefore, $\mathbf{C}_f$ commutes with the shift operator $\mathbf{S} = \mathbf{A}_G \otimes \mathbf{A}_{\mathcal{H}}$, hence $\mathcal{GP}(0, k)$ is a JWSS prior.

Example 4.1. The GP prior generalizes the Gaussian process over a graph (GPG) framework [75], which defined a GPG as a vector-valued GP whose covariance matrix takes the form

$$\begin{aligned} \operatorname{cov}(\mathbf{s}, \mathbf{t}) &= k_{\mathcal{T}}(\mathbf{s}, \mathbf{t})\mathbf{B}(a), \\ \mathbf{B}(a) &= (\mathbf{I}_N + a\mathbf{L}_G)^{-2} := (B(a)_{ij}), \end{aligned}$$

where $a > 0$ is a parameter and $\mathbf{L}_G$ is the graph Laplacian matrix. We see that this covariance structure corresponds to a GP prior in $L^2(\mathcal{J})$ with $k_G(i, j) = B(a)_{ij}$. The GPG also assumes that each observation is $(\mathbf{t}, \mathbf{x})$, where $\mathbf{x}$ is a complete graph signal, while in (4.5) we allow the observed graph signals to be incomplete. Therefore, this generalization allows us to reconstruct the generalized graph signal when the observations come from different subsets of vertices at different instances.

We next consider the posterior. The observations $\{(v_m, \mathbf{t}_m, y_m)\}$ are denoted as $\mathcal{D}_{\text{train}}$. According to Section 2.5, the MAP estimator is given by (4.6) with $\mu = \sigma^2$. Since f is a GP, (4.6) is also the posterior expectation given $\mathcal{D}_{\text{train}}$, i.e., $\hat{f}(v, \mathbf{t}) = \mathbb{E}[f(v, \mathbf{t}) \,|\, \mathcal{D}_{\text{train}}]$. The posterior variance can be calculated by

$$\operatorname{var}(f(v, \mathbf{t}) \,|\, \mathcal{D}_{\text{train}}) = k((v, \mathbf{t}), (v, \mathbf{t})) - \mathbf{k}^{\intercal}(\mathbf{K}(\mathcal{S}, \mathcal{S}) + \sigma^2 \mathbf{I}_M)^{-1}\mathbf{k}, \tag{4.12}$$

where $\mathbf{k} := (k((v, \mathbf{t}), (v_1, \mathbf{t}_1)), \ldots, k((v, \mathbf{t}), (v_m, \mathbf{t}_m)))^{\intercal}$. This observation indicates that the time-vertex signal reconstruction approach is a special case of the KRR-GGSP approach.

Example 4.2. In the time-vertex signal reconstruction problem, the observed signal $\mathbf{X}_o \in \mathbb{R}^{N \times T}$ is an incomplete and noisy observation of the original signal

$\mathbf{X}_r \in \mathbb{R}^{N\times T}$. The mask matrix is $\mathbf{\Pi}_{\mathcal{S}} \in \{0,1\}^{N\times T}$. The paper [40] formulated the graph signal reconstruction via Sobolev smoothness (GTRSS) problem as follows:

$$\begin{aligned}\hat{\mathbf{X}}_r &= \underset{\mathbf{X}\in\mathbb{R}^{N\times T}}{\arg\min} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2 + \mu_{\text{TV}} \operatorname{tr}((\mathbf{X}\mathbf{D}_h)^\intercal (\mathbf{L}+\alpha\mathbf{I})^\beta \mathbf{X}\mathbf{D}_h) \\ &= \underset{\mathbf{X}\in\mathbb{R}^{N\times T}}{\arg\min} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2 + \mu_{\text{TV}} \operatorname{vec}(\mathbf{X})^\intercal (\mathbf{D}_h\mathbf{D}_h^\intercal) \otimes (\mathbf{L}+\alpha\mathbf{I})^\beta \operatorname{vec}(\mathbf{X}),\end{aligned} \tag{4.13}$$

where $\mathbf{D}_h$ is the first order difference operator

$$\mathbf{D}_h = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{T\times(T-1)}.$$

For ease of further analysis, we slightly modify (4.13) to be

$$\hat{\mathbf{X}}_r = \underset{\mathbf{X}\in\mathbb{R}^{N\times T}}{\arg\min} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X} - \mathbf{X}_o\|_F^2 + \mu_{\text{TV}} \operatorname{vec}(\mathbf{X})^\intercal (\mathbf{D}_h\mathbf{D}_h^\intercal + \delta_o\mathbf{I}) \otimes (\mathbf{L}+\alpha\mathbf{I})^\beta \operatorname{vec}(\mathbf{X}), \tag{4.14}$$

where $\delta_o > 0$. We also assume that $\operatorname{diag}(\operatorname{vec}(\mathbf{\Pi}_{\mathcal{S}})) + (\mathbf{D}_h\mathbf{D}_h^\intercal) \otimes (\mathbf{L}+\alpha\mathbf{I})^\beta$ is full-rank. It can be shown that the solution to (4.14) can approximate that of (4.13) arbitrarily well as long as δ_o is small enough.

We consider problem (4.14) under a Bayesian setting. Let the prior of $\operatorname{vec}(\mathbf{X}_r)$ be a Gaussian random vector with zero mean and covariance $((\mathbf{D}_h^\intercal\mathbf{D}_h+\delta_o\mathbf{I})\otimes(\mathbf{L}+\alpha\mathbf{I})^\beta)^{-1}$. In other words, if we let $k_{\mathcal{T}}(s,t) = (\mathbf{D}_h\mathbf{D}_h^\intercal + \delta_o\mathbf{I})^{-1}_{s,t}$, and $\mathbf{K}_G = (\mathbf{L}+\alpha\mathbf{I})^{-\beta}$, then $\mathbf{X}_r = (X_r(v,t))$ is a GP with covariance $\operatorname{cov}(X_r(u,s), X_r(v,t)) = k_{\mathcal{T}}(s,t)k_G(u,v)$. Suppose the noise is i.i.d. with variance μ_{TV}, then the objective function in (4.14) is the log-likelihood of the posterior $p(\mathbf{X}_r|\mathbf{X}_o)$ (up to a constant):

$$\begin{aligned}&-\log(p(\mathbf{X}_r|\mathbf{X}_o)) = -(\log(p(\mathbf{X}_r,\mathbf{X}_o)) - \log(p(\mathbf{X}_o))) \\ &= -(\log(p(\mathbf{X}_o|\mathbf{X}_r)) + \log(p(\mathbf{X}_r)) - \log(p(\mathbf{X}_o))) \\ &= \frac{1}{\mu_{\text{TV}}} \|\mathbf{\Pi}_{\mathcal{S}} \odot \mathbf{X}_r - \mathbf{X}_o\|_F^2 + \operatorname{vec}(\mathbf{X}_r)^\intercal (\mathbf{D}_h\mathbf{D}_h^\intercal + \delta_o\mathbf{I}) \otimes (\mathbf{L}+\alpha\mathbf{I})^\beta \operatorname{vec}(\mathbf{X}_r) + \text{const},\end{aligned}$$

where const is a constant independent of $\mathbf{X}_r$. Therefore, the solution to this problem is the MAP of $\mathbf{X}_r$ given $\mathbf{X}_o$. According to the Bayesian interpretation in Section 2.5, this MAP estimator $\hat{\mathbf{X}}_r = (\hat{X}_r(v,t))$ is the solution (4.6) of KRR-GGSP where $k_{\mathcal{T}}(s,t) = (\mathbf{D}_h\mathbf{D}_h^\intercal + \delta_o\mathbf{I})^{-1}_{s,t}$, $s,t \in \{1,2,\dots,T\}$, $\mathbf{K}_G = (\mathbf{L}_G + \alpha\mathbf{I})^{-\beta}$, and $\mu = \mu_{\mathrm{TV}}$.

From Example 4.2, we see that the GTRSS problem can be understood as using a specific kernel in the time domain. In the following, we show that since this kernel depends on the number of discrete time steps, it is sensitive to the sampling rate.

Consider the case where $\mathcal{V}$ is a singleton and $\mathcal{T} = [a,b]$ is a closed interval, so that the signal $f : \mathcal{V} \times \mathcal{T} \to \mathbb{R}$ can be identified with a signal $f : [a,b] \to \mathbb{R}$. Without loss of generality, let $[a,b] = [0,1]$. Suppose f is evenly sampled with interval length Δ. We denote the kernel from Example 4.2 as $k_{\mathrm{GTRSS}}(s,t;\Delta) = (\mathbf{D}_h\mathbf{D}_h^\intercal + \delta_o\mathbf{I})^{-1}_{\frac{s}{\Delta},\frac{t}{\Delta}}$, where $s,t \in \{0,\Delta,2\Delta,\dots,1\}$. This leads to the problem that the prior distribution assigned to the signal relies on the sampling frequency. According to the Bayesian interpretation (cf. Section 2.5), by using this kernel, we have assumed a prior distribution on f. We now examine the cross-correlation of the prior between $f(0)$ and $f(1)$, i.e.,

$$\mathrm{corr}(f(0),f(1);\Delta) := \frac{\mathrm{cov}(f(0),f(1))}{\sqrt{\mathrm{var}(f(0))\,\mathrm{var}(f(1))}} = \frac{k_{\mathrm{GTRSS}}(0,1;\Delta)}{\sqrt{k_{\mathrm{GTRSS}}(0,0;\Delta)k_{\mathrm{GTRSS}}(1,1;\Delta)}}.$$

By calculating this quantity with different values of Δ, we find that it is highly related to the sampling frequency (see Fig. 4.1). Specifically, when the sampling frequency is large enough, the prior correlation between $f(0)$ and $f(1)$ tends to zero. Instead, if we use other kernels $\tilde{k}_{\mathcal{T}}$ which does not depend on Δ (e.g., the RBF kernel), then the prior cross-correlation $\frac{\tilde{k}_{\mathcal{T}}(a,b)}{\sqrt{\tilde{k}_{\mathcal{T}}(a,a)\tilde{k}_{\mathcal{T}}(b,b)}}$ does not depend on Δ. This accounts for the failure of GTRSS on datasets with high sampling frequency, while KRR-GGSP with RBF kernel works well (see Section 4.4.2). This is essentially because the scale parameter in GTRSS kernel relies on the sampling frequency, while that in RBF kernel does not. Therefore the RBF kernel has one more degree of freedom than GTRSS kernel. Hence by using more flexible kernels, we can expect better reconstruction results.

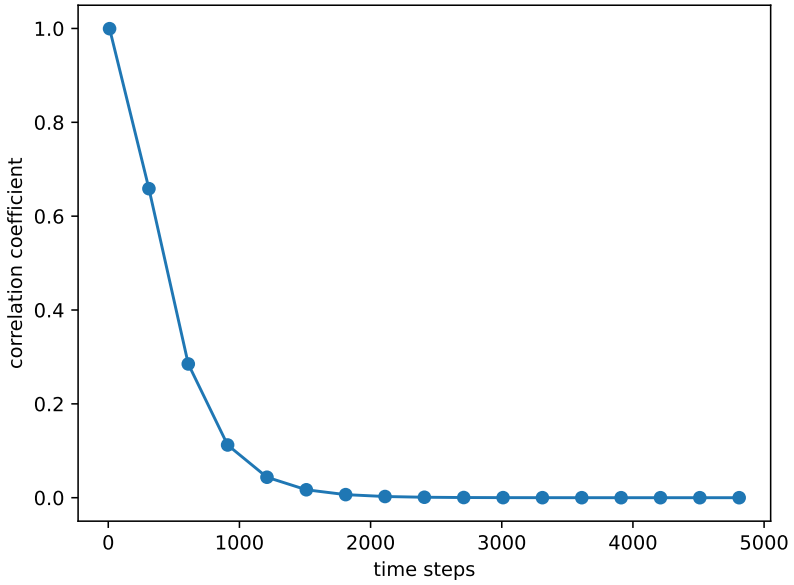


FIGURE 4.1: The prior correlation coefficients $\mathrm{corr}(f(0), f(1); \Delta)$ as a function of $\frac{1}{\Delta} + 1$ (i.e., number of time steps) with $\delta_o = 10^{-5}$.

4.2.3 Online and Distributed Implementation

We now consider the online learning problem where the data stream $\{(v_m, \mathbf{t}_m, y_m)\}$ arrives sequentially. Upon each arrival of $(v_m, \mathbf{t}_m)$, the learner provides a *distributed* prediction of $f(v_m, \mathbf{t}_m)$, i.e., the update and evaluation steps are implemented by each vertex exchanging information with its neighbors within a certain number of hops. After that, y_m is observed and the error is measured by comparing the prediction with y_m. The estimator of $f(v_m, \mathbf{t}_m)$ cannot depend on y_m, and the error is used to update the learner for the next prediction. Problem (4.5) can be adapted to this setting via RFFs[76] when $k_{\mathcal{T}}$ is a RBF kernel. Denote the columns of $\mathbf{K}_G^{\frac{1}{2}}$ by $[\mathbf{p}_1, \ldots, \mathbf{p}_N]$, and write $\mathbf{p}_v = (p_{1,v}, \ldots, p_{N,v})^\intercal$. For the kernel k, the RFF can be constructed as

$$\boldsymbol{\eta}(v, \mathbf{t}) = \mathbf{p}_v \otimes \mathbf{z}(\mathbf{t}),$$

where $\mathbf{z}(\mathbf{t}) \in \mathbb{R}^F$ is the RFF of the kernel $k_{\mathcal{T}}$, i.e., $\mathbb{E}[\mathbf{z}(\mathbf{s})^\intercal \mathbf{z}(\mathbf{t})] = k_{\mathcal{T}}(\mathbf{s}, \mathbf{t})$. By the construction of $\boldsymbol{\eta}(v, \mathbf{t})$, we have $\mathbb{E}[\boldsymbol{\eta}(u, \mathbf{s})^\intercal \boldsymbol{\eta}(v, \mathbf{t})] = k((u, \mathbf{s}), (v, \mathbf{t}))$. The reconstructed signal is then $\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \mathbf{c}^\intercal \boldsymbol{\eta}(v, \mathbf{t})$. Problem (4.5) is therefore converted

to the linear regression problem [77, (7)]:

$$\min_{\mathbf{c}\in\mathbb{R}^{NF}} q(\mathbf{c}) = \sum_{m=1}^{M} (\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)^2 + \mu \|\mathbf{c}\|_2^2 . \tag{4.15}$$

Alternatively, if we define $q_m(\mathbf{c}) := (\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)^2 + \frac{\mu}{M} \|\mathbf{c}\|_2^2$, then (4.15) turns out to be

$$\min_{\mathbf{c}\in\mathbb{R}^{NF}} q(\mathbf{c}) = \sum_{m=1}^{M} q_m(\mathbf{c}). \tag{4.16}$$

The evaluation of $\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \mathbf{c}^\intercal \boldsymbol{\eta}(v, \mathbf{t})$ can be distributed. To illustrate this, write $\mathbf{c} = (\mathbf{c}_1^\intercal, \ldots, \mathbf{c}_N^\intercal)^\intercal$ where $\mathbf{c}_n \in \mathbb{R}^F$, $n = 1, \ldots, N$. Since k_G takes the form (4.4), $\mathbf{K}_G^{\frac{1}{2}}$ can be represented as a polynomial of $\mathbf{A}_G$ of degree L_0, so that $p_{u,v} = 0$ for all $u \notin \mathcal{N}_{L_0}(v)$. Then for any input $(v, \mathbf{t})$, $\boldsymbol{\eta}(v, \mathbf{t}) = (p_{1,v}\mathbf{z}(\mathbf{t})^\intercal, \ldots, p_{N,v}\mathbf{z}(\mathbf{t})^\intercal)^\intercal$, $\hat{f}_{\mathrm{RFF}}$ is evaluated by

$$\hat{f}_{\mathrm{RFF}}(v, \mathbf{t}) = \sum_{u\in\mathcal{N}_{L_0}(v)} \mathbf{c}_u^\intercal p_{u,v} \mathbf{z}(\mathbf{t}),$$

which only requires information from $\mathcal{N}_{L_0}(v)$.

Problem (4.15) can be solved in an online and distributed way by stochastic gradient descent (SGD). To be specific, suppose the datastream is $\{(v_m, \mathbf{t}_m, y_m) : m = 1, 2, \ldots\}$. At the m-th step, we approximate ∇q with the instantaneous sample $(v_m, \mathbf{t}_m, y_m)$:

$$\nabla q_m = 2(\mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) - y_m)\boldsymbol{\eta}(v_m, \mathbf{t}_m) + 2\frac{\mu}{M}\mathbf{c}.$$

Note that $y_m - \mathbf{c}^\intercal \boldsymbol{\eta}(v_m, \mathbf{t}_m) = y_m - \hat{f}_{\mathrm{RFF}}(v_m, \mathbf{t}_m) := \hat{e}_m$ is the approximation error at the current sample point $(v_m, \mathbf{t}_m)$. We can update $\mathbf{c}$ at the m-th iteration via

$$\mathbf{c}^{(m)} = \mathbf{c}^{(m-1)} - \theta \nabla q_m = \theta_1 \mathbf{c}^{(m-1)} + \theta_2 \hat{e}_m \boldsymbol{\eta}(v_m, \mathbf{t}_m), \tag{4.17}$$

where $\theta, \theta_1, \theta_2 > 0$. Note that:

- ∇q_m is Lipschitz continuous with Lipschitz constant $\mathrm{Lip}_m = 2\|\boldsymbol{\eta}(v_m, \mathbf{t}_m)\|^2 + 2\frac{\mu}{M}$. Define $\mathrm{Lip}_{\max} = \max_m \mathrm{Lip}_m$.

- q_m is convex.
- q is 2μ-strongly convex (cf. [78, Lemma 2.12]).

According to [78, Theorem 5.7], if $\theta \in (0, \frac{1}{2\mathrm{Lip}_{\max}})$, the convergence rate of SGD is linear when $\mu > 0$. Since $\mathbf{p}_{v_m}$ only has non-zero entries in $\mathcal{N}_{L_0}(v_m)$, and $\hat{e}_m$ can be evaluated in a distributed way, we see that (4.17) is an online and distributed update. This is always achievable when $k_{\mathcal{T}}$ is a RBF kernel.

4.3 Conditional MSE of KRR-GGSP in the Bayesian framework

In this section, we consider $f \sim \mathcal{GP}(0, k)$, i.e., the Bayesian framework considered in Section 4.2.2. We derive the MSE of the estimate given by KRR-GGSP at a particular node $v_0 \in \mathcal{V}$ and time $\mathbf{t}_0 \in \mathcal{T}$, conditioned on an observation set $\{(v_m, \mathbf{t}_m, y_m) : m = 1, \ldots, M\}$. To be specific, we analyze

$$\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \{(v_m, \mathbf{t}_m, y_m)\}) = \mathbb{E}\Big[(\hat{f}(v_0, \mathbf{t}_0) - f(v_0, \mathbf{t}_0))^2 \,\Big|\, \{(v_m, \mathbf{t}_m, y_m)\}\Big] \quad (4.18)$$

under the scenario when the noise energy is unknown, and the MSE is hard to compute when $M \to \infty$ as it involves taking the inverse of the kernel matrix of the observations. We study the dependence of the MSE on the graph structure when a subset of vertices have dense observation samples ($M \to \infty$). The asymptotic MSE and its upper bound can be used as a criterion to choose an optimal sampling vertex set.

We consider the asymptotic MSE of inference for $f(v_0, \mathbf{t}_0)$, i.e., the limit of (4.18) when $M \to \infty$. Note that if we allow uniform sampling on every vertex with an ever-growing sample size, then it is known that the posterior variance will uniformly converge to 0 [79]. In order to examine the effect of leveraging information from other vertices in KRR-GGSP, we consider the case where there are no available sample points on $\{v_0\} \times \mathcal{T}$, and the value of $f(v_0, \mathbf{t})$ is to be estimated.

Mathematically, let $\mathcal{S}(v; M_0)$ be a set of M_0 samples i.i.d. from Unif $(\{v\} \times \mathcal{T})$, where $v \in \{v_0\}^{\mathsf{c}}$. The sample set $\mathcal{S}(M_0)$ is then obtained by $\mathcal{S}(M_0) = \bigcup_{v \in \{v_0\}^{\mathsf{c}}} \mathcal{S}(v; M_0)$.

This sampling scheme is illustrated in Fig. 4.2, and we call it *uniform exclusive sampling*. In practice, this scheme mimics the scene where only limited knowledge can be obtained from a certain vertex, and an inference for that is desired.

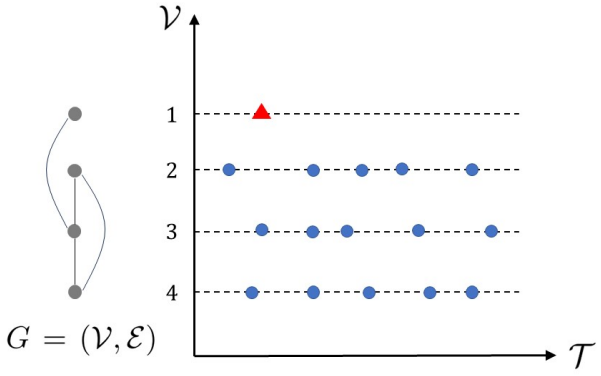


FIGURE 4.2: The uniform exclusive sampling scheme with $M_0 = 5$. The blue circles denote $\mathcal{S}(M_0)$, and the red triangle is $(v_0, \mathbf{t}_0)$.

For ease of notation, we define $\mathcal{J}_S := \{v_0\}^{\mathsf{c}} \times \mathcal{T}$. We write $\mathbf{y}(M_0)$ to represent the observations $\mathbf{y}(\mathcal{S}(M_0))$ from the sampling set $\mathcal{S}(M_0)$, and $\mathbf{z}$ to represent the restriction of f on $\mathcal{J}_S$. We analyze $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ from two aspects: first, in Theorem 4.2 we analyze the integration of $\mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0))$ over $\mathbf{t}$; then in Theorem 4.3 we provide an asymptotic upper bound for $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$.

Let $\mathcal{T}_0$ be a subset of $\mathcal{T}$. We consider the following integration

$$\int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \, \mathrm{d}\tau(\mathbf{t}), \tag{4.19}$$

which represents the conditional MSE of the KRR-GGSP estimator over $\mathcal{T}_0$. Intuitively, when M_0 tends to infinity, the situation can be interpreted as f on $\mathcal{J}_S$ is known and can be utilized for inference. We formally address this in the following theorem:

Theorem 4.2. *Under Assumption 4.1, the limit posterior covariance of $f(v_0, \mathbf{t})$ over $\mathcal{T}_0$ converges:*

$$\lim_{M_0 \to \infty} \int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \, \mathrm{d}\tau(\mathbf{t}) = \int_{\mathcal{T}_0} \mathrm{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z}) \, \mathrm{d}\tau(\mathbf{t}). \tag{4.20}$$

Proof. See Appendix 4.A. □

From Theorem 4.2 we know the limiting posterior variance given an infinite number of sample points. This result can also be applied when only a subset of vertices have dense samples. In that case, the R.H.S. of (4.20) becomes an asymptotic upper bound by letting $\mathbf{z}$ be the restriction of f on the vertices with dense samples. Moreover, we can get a rough idea of the behavior of $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ if we consider the following sequence of continuous functions

$$\rho_{M_0}(\alpha) := \begin{cases} \frac{1}{\alpha} \int_{B(\mathbf{t}_0,\alpha)} \operatorname{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{y}(M_0)) \,\mathrm{d}\tau(\mathbf{t}), & \alpha > 0 \\ \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)), & \alpha = 0 \end{cases}$$

where $B(\mathbf{t}_0, \alpha)$ is the open ball centered at $\mathbf{t}_0$ with measure α. Specifically, by [79, Theorem 3] we note that $\rho_{M_0}(\alpha)$ is a monotonic sequence, i.e., $\rho_{M_0}(\alpha) \leq \rho_{M_0'}(\alpha)$ if $M_0 > M_0'$. According to Theorem 4.2, the limit function of $\rho_{M_0}(\alpha)$ is

$$\rho(\alpha) = \lim_{M_0 \to \infty} \rho_{M_0}(\alpha) = \frac{1}{\alpha} \int_{B(\mathbf{t}_0,\alpha)} \operatorname{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z}) \,\mathrm{d}\tau(\mathbf{t})$$

when $\alpha > 0$, and

$$\rho(0) = \lim_{M_0 \to \infty} \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)).$$

Therefore, if we assume that the limit function of $\rho_{M_0}(\alpha)$ is continuous w.r.t. α and $\operatorname{var}(f(v_0, \mathbf{t}) \,|\, \mathbf{z})$ is continuous w.r.t. $\mathbf{t}$, then $\rho(0) = \lim_{\alpha \to 0} \rho(\alpha) = \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z})$, i.e.,

$$\lim_{M_0 \to \infty} \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)) = \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}). \tag{4.21}$$

From (4.21) we know that, although $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ is random due to the randomness of $\mathcal{S}(M_0)$, its limit $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z})$ is a deterministic quantity when $M_0 \to \infty$. In addition, it can be shown by Lemma 4.6 that

$$\begin{aligned} &\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q})) \\ &= \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}) + \operatorname{var}(\mathbb{E}[f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}] \,|\, f(\mathcal{Q})) \\ &\geq \operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{z}), \end{aligned}$$

for arbitrary finite set $\mathcal{Q} \subset \mathcal{J}_S$. Therefore, according to (4.21), $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q}))$ can always serve as an upper bound for $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$ when M_0 is large enough. Since $\mathcal{Q}$ is finite, $\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q}))$ may be numerically computed. In

contrast, We note that the quantities in (4.20) involve the pseudo-inverse of a possibly infinite-rank operator (cf . Lemma 4.8), which may be difficult to numerically compute. Consider the case when $\mathcal{Q} = \mathcal{N}_d(v_0) \times \{\mathbf{t}_0\}$ where $d \in \mathbb{N}$ is the number of neighborhood hops. Let $N_d := |\mathcal{N}_d(v_0)|$. For simplicity, we introduce the following notations:

$$\begin{aligned}
\mathbf{k}_G(v_0, \mathcal{N}_d) &:= (k_G(v_0, v))_{v \in \mathcal{V} \setminus \{v_0\}} \in \mathbb{R}^{N_d} \\
\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d) &:= (k_G(u, v))_{u,v \in \mathcal{V} \setminus \{v_0\}} \in \mathbb{R}^{N_d \times N_d} \\
l(v_0, d) &:= k_G(v_0, v_0) \\
&\quad - \mathbf{k}_G(v_0, \mathcal{N}_d)^\intercal \mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1} \mathbf{k}_G(v_0, \mathcal{N}_d),
\end{aligned}$$

so that

$$\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q})) = k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l(v_0, d).$$

To provide an explicit upper bound for (4.18), we derive an asymptotic bound with a convergence rate for the posterior variance which is locally computable.

Theorem 4.3. *Suppose $\mathcal{T}$ is a compact subset of $\mathbb{R}^D$ whose boundary set has measure zero, and $\mathbf{t}_0$ is an interior point of $\mathcal{T}$. Suppose $k_{\mathcal{T}}$ is Lipschitz continuous on $\mathcal{T}$. Let $d \in \mathbb{N}_+$. For any arbitrary $c_0 \in (0, 1)$ we have*

$$\operatorname{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0)) \leq k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l(v_0, d) + (C_1 c_0^{-1} + C_2 c_0^2) M_0^{-\frac{1}{3D+1}} + C_3 c_0 M_0^{-\frac{2}{3D+1}} \tag{4.22}$$

with probability at least

$$\left(1 - \frac{1}{2} \frac{1}{(1 - c_0)^2 C_D M_0^{\frac{1}{3D+1}}}\right)^{N_d}. \tag{4.23}$$

Proof. As the proof is technical in nature, it is provided in Section 4.C. □

We note that when $k_{\mathcal{T}}$ is RBF kernel, $k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l(v_0, d)$ only depends on the graph structure. In other words, if we are allowed to select a subset of vertices $\mathcal{V}' \subset \mathcal{V}$ to recover the signal on v_0, then it is preferred that the subgraph with vertex set $\mathcal{V}' \bigcup \{v_0\}$ has a small $l(v_0, d)$. Theorem 4.3 indicates a trade-off between the quality and confidence of the upper bound (4.22). From the proof of Theorem 4.3, when

the number of samples in a small neighborhood of every $(v, \mathbf{t}_0) \in \mathcal{N}_d(v_0) \times \{\mathbf{t}_0\}$ is larger than a threshold m_0, the conditional variance of $f(v_0, \mathbf{t}_0)$ given these samples is approximately $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, f(\mathcal{Q}))$. However, the probability that this event happens is smaller when we require more vertices to at least m_0 samples in their neighborhoods. This explains why the probability lower bound (4.23) decreases as N_d increases. On the other hand, for a fixed $(v_0, \mathbf{t}_0)$, a larger $\mathcal{N}_d(v_0)$ indicates a better asymptotic upper bound for $\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{y}(M_0))$, i.e, $l(v_0, d)$ decreases with a larger $\mathcal{N}_d(v_0)$. This is because $l(v_0, d)$ is a conditional variance of a Gaussian random variable by definition, and it is known that when we condition on a larger set of Gaussian random variables, the variance decreases [79, Lemma 9].

4.4 Numerical Experiments

In this section, we conduct experiments to illustrate the theory and methods of the KRR-GGSP approach. In the experiments, $\mathcal{T}$ is an interval, and the target signal is a function on $\mathcal{V} \times \mathcal{T}$. In the datasets, the target signal is downsampled on every vertex. We aim to reconstruct the target signal from the randomly selected samples with additive noise. We compare the following algorithms in the experiments:

1. KRR-GGSP. We reconstruct the signal using (4.5) with the tensor product kernel (4.3). We set $\mathbf{K}_G = a(\mathbf{L} - \lambda_N \mathbf{I})^2 + b\mathbf{I}$ such that

$$a(\lambda_1 - \lambda_N)^2 + b = 1, \tag{4.24}$$

   and $0 \leq b \leq 1$ is a tunable parameter. This parameter setting ensures that $1 = r(\lambda_1) \geq \cdots \geq r(\lambda_N) = b$ (cf . (4.4)). We set $k_{\mathcal{T}}$ to be the RBF kernel $k_{\mathcal{T}}(s, t) = \exp(-|s - t|^2/\beta_{\text{scale}})$, where β_{scale} is a tunable parameter.

2. Isolated KRR. We recover the signal on each vertex separately using KRR (cf . (2.30) and (4.8)). In Section 4.2, we have shown that this method is equivalent to using $\mathbf{K}_G = \mathbf{I}$ in KRR-GGSP, i.e., fixing $b = 1$ in (4.24).

3. GTRSS. We recover the signal using (4.13), where μ_{TV}, α and β are tunable parameters.

4. Graph recurrent imputation network (GRIN). We implement this method using the Spatiotemporal library [80].

5. Bandlimited- GGSP. We recover the signal using (4.11), where B', B'' and μ are tunable parameters. The eigenvalues $r(\lambda_n)$ and γ_i in (4.11) are set to be 1.

4.4.1 ECoG Dataset

We test the reconstruction performance of KRR-GGSP on an ECoG multivariate time series dataset.[2] This dataset contains measurements from 76 electrodes on an epilepsy patient during both ictal and pre-ictal periods [70]. We make use of the data from 2 ictal periods. Each period lasts 10 seconds with a sampling rate of 400 Hz. Therefore, the dataset we use is a 76×8000 matrix. We use the last 320 time steps for testing and the 160 time steps before the test set for training. We add AWGN to the dataset and randomly mask the data so that both training and test sets are incomplete and noisy. We test the recovery performances of KRR-GGSP, GTRSS, isolated KRR and GRIN on this dataset.

Except for the isolated KRR method, all other methods rely on a graph structure. To construct the graph, we first use the isolated KRR to roughly reconstruct the unknown signal values on 160 time steps in the training set, and then calculate the correlation coefficients of these recovered data. We regard two electrodes as connected if the correlation coefficients between them are larger than 0.5. We set the edge weights to be the correlation coefficients. For GRIN, the training set is used for model training and validation. Besides the small training set with 160 time steps, we also show its performance trained on all available training data from the dataset, i.e., 7680 time steps. For other methods, the training set is used for tuning parameters. The recovery performance is measured by the relative error

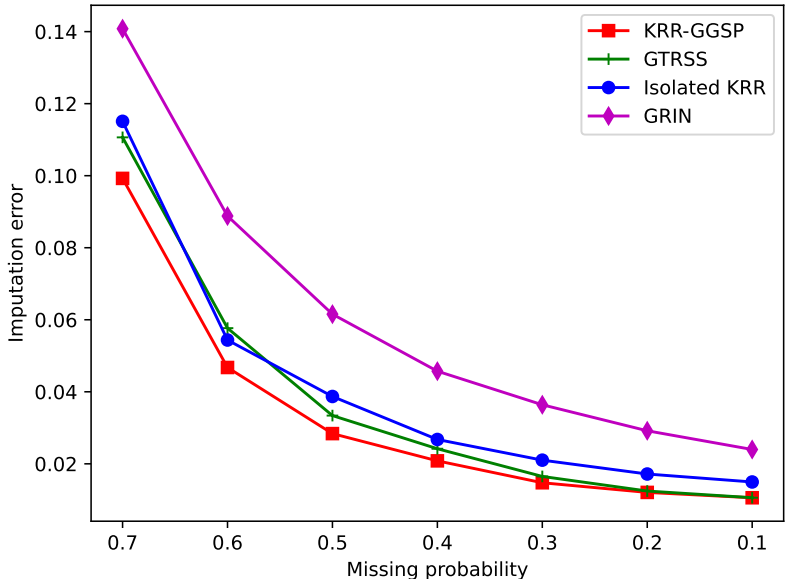
$$\frac{\mathbb{E}\left[(f(v,t)-\hat{f}(v,t))^2\right]}{\mathbb{E}[f(v,t)^2]}. \tag{4.25}$$

Similarly, we define the noise level to be

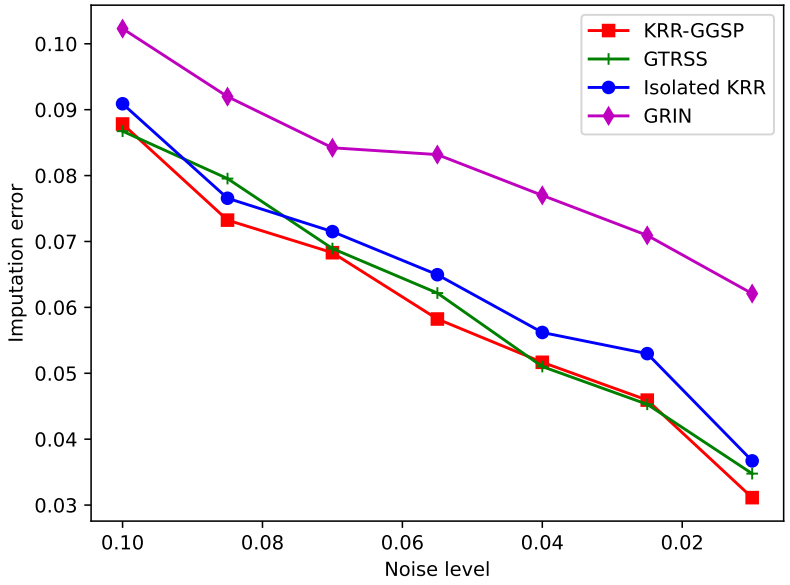
$$\frac{\mathbb{E}[\epsilon^2]}{\mathbb{E}[f(v,t)^2]}. \tag{4.26}$$

The recovery results are shown in Fig. 4.3. We observe that KRR-GGSP shows good recovery results and outperforms other methods. Since KRR-GGSP has a

[2] `https://math.bu.edu/people/kolaczyk/datasets.html`



(A) Reconstruction performances under different missing value probabilities. The noise energy of AWGN is set to be 0.01 of the signal energy.



(B) Reconstruction performances under different noise levels (cf. (4.26)). The missing value probability is set to be 0.5.

FIGURE 4.3: Comparison of different reconstruction methods on ECoG dataset. Each point in the figure is obtained by 20 repetitions.

tunable kernel in the time domain, it shows better performance than GTRSS. This effect can be better observed in Section 4.4.2. The isolated KRR method has a tunable kernel, but it is not able to take advantage of the graph structure, hence is outperformed by KRR-GGSP. Here, we show the performance of GRIN trained with 7680 time steps. We remark that the deep learning method GRIN requires a sufficiently large training set to obtain reasonable results. When the training set is

as small as 160 time steps, GRIN does not yield reasonable reconstruction results. Since the bandlimited-GGSP method does not have comparable performances with the other methods (when noise level = 0.01 and missing value probability = 0.5, its imputation error is 0.28), we do not show its performance here.

4.4.2 Intel-lab Temperature Data

We test the reconstruction performance of KRR-GGSP on the Intel lab temperature dataset illustrated in Fig. 1.1. In this experiment, we use the data from the first and second days. Since there are 86400 seconds in a day, the entire dataset we use is a 54×172800 matrix. Here we remark that since the sampling rate of each sensor is much smaller than 1 Hz and not uniform, only 1.93% of the entries are non-null. Therefore, this dataset is very sparse. We identify the temperature records outside the upper 99.92% quantile and lower 0.001% quantile as outliers and discard them. We subtract the mean value of all observed temperature records from the dataset. We treat each sensor as a vertex and construct a 5-NN graph using their locations. We use half of the first day's records for training and the second day's for testing. As in Section 4.4.1, we add AWGN to the data and assign a random mask. In this experiment, the noise energy is set to be 5% of the signal energy. We compare the methods as described in Section 4.4.1 with performance measurement (4.25).

From the result in Fig. 4.4, we observe that KRR-GGSP outperforms the isolated KRR and bandlimited-GGSP. This indicates that by utilizing infinitely many features, the reconstruction performance can be improved. On this dataset, GRIN and GTRSS fail to yield reasonable results. For example, when the observation ratio is 0.15, GTRSS has relative MSE around 0.8, and GRIN has relative MSE around 1.0. For GRIN, this is mainly due to the sparsity of the available data in the dataset. For GTRSS, this is due to the improper prior assumption on the dataset.

4.4.3 COVID-19 Case Prediction

We use the online reconstruction method in Section 4.2.3 to predict COVID-19 cases using only historical data. We use the data from The New York Times, based on reports from state and local health agencies[3]. From this dataset, we retrieve the

[3] `https://github.com/TorchSpatiotemporal/tsl`

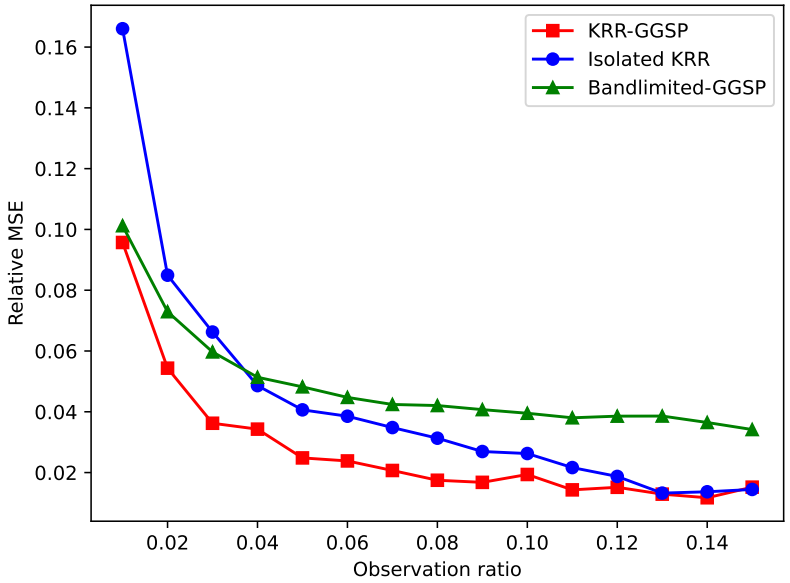


FIGURE 4.4: Reconstruction error under different proportions of samples to be used for reconstruction. Each point in the figure is obtained by 10 repetitions.

records from California's 58 counties, starting from the first day when all counties have cases reported so that there are 886 days in total. We treat each county as a vertex and connect them if they are adjacent geographically. We set the datastream and prediction rule as follows: on each date t, we randomly choose a subset of vertices $\mathcal{V}_S = \{v_1, \ldots, v_Q\} \subset \mathcal{V}$ such that the learner is assumed to have access to $\mathbf{y}(\mathcal{V}_S \times \{t\})$. Besides, for each date t, the sample points $\{(v_i, t, y(v_i, t))\}$ are observed sequentially, one datum at a time.

We compare the online KRR-GGSP with several existing online and distributed reconstruction methods. The implementation details are the following:

1. Online KRR-GGSP. For each $(v_i, t) \in \mathcal{V}_S \times \{t\}$, we first calculate the prediction $\hat{f}_{\text{RFF}}(v_i, t)$. Then we compute the error $\hat{e}_i = y(v_i, t) - \hat{f}_{\text{RFF}}(v_i, t)$, and update the predictor by (4.17). Then for each $(v_j, t) \in \mathcal{V}_S^{\mathsf{c}} \times \{t\}$, we also make predictions and compute the error, but will not update the predictor since the learner is not supposed to have access to the observations on them. We set $\mathbf{K}_G = g(\mathbf{L})^2$, where g is a polynomial of degree one such that $g(\lambda_1) = 1, g(\lambda_N) = 0.4$. We let $k_{\mathcal{T}}(s, t) = \exp(-(s-t)^2/\beta_{\text{scale}})$, where β_{scale} is an adjustable parameter. We set the dimension of $\mathbf{z}(t)$ to be 60.

2. Online isolated KRR. This is implemented by letting $\mathbf{K}_G = \mathbf{I}$ in the online KRR-GGSP method.

3. Online GTRSS. This method is a generalization of [39, (35)], by replacing $\mathbf{L}$ with $(\mathbf{L}+\alpha\mathbf{I})^{\beta}$. Let $\hat{\mathbf{f}}_t^l \in \mathbb{R}^N$ be the estimation of $\mathbf{f}_t = (y(1,t),\dots,y(N,t))^\intercal$ after observing l samples on date t. The samples are denoted by $\mathbf{y}_t^l \in \mathbb{R}^N$ such that the unobserved entries are zero. We write $\mathbf{m}_t^l$ to denote the mask after observing l samples on date t. Let $\hat{\mathbf{f}}_{t-1}$ be the estimation of $\mathbf{f}_{t-1}$ after observing all available samples on date $t-1$. Then the update rule goes as follows:

$$\begin{aligned}\hat{\mathbf{f}}_t^l = \hat{\mathbf{f}}_t^{l-1} &- \mu(\mathbf{m}_t^l \odot \hat{\mathbf{f}}_t^{l-1} - \mathbf{y}_t^l) \\ &- \mu\lambda(\mathbf{L}+\alpha\mathbf{I})^{\beta}(\hat{\mathbf{f}}_t^{l-1} - \hat{\mathbf{f}}_{t-1}).\end{aligned} \tag{4.27}$$

When the $l+1$-th sample arrives, we evaluate the error $\hat{e}_{l+1} = y(v_{l+1},t) - \hat{f}(v_{l+1},t)$, where $\hat{f}(v_{l+1},t)$ is the v_{l+1}-th entry of $\hat{\mathbf{f}}_t^l$. λ, μ, α and β are adjustable parameters in this method.

We show the best performance of the methods with different parameters in Fig. 4.5. The error measurement is (4.25). We observe that the online KRR-GGSP method outperforms other online and distributed methods. We also tested the ARMA method on each vertex, but due to the missing values, it usually fails to converge and yields unstable results. For example, when the proportion of observed vertices is 80%, the ARMA$(2,0,2)$ model fails to converge on about 29% vertices, and the prediction error on each vertex varies from 0.004 to 665×10^4.

4.A Appendix: Proof of Theorem 4.2

In order to prove Theorem 4.2, we introduce the following definitions and lemmas. The proofs of the lemmas are included in Section 4.B for completeness.

Let $\mathbf{x}_0$ be the restriction of f on $\{v_0\} \times \mathcal{T}_0$, and $\mathbf{C}_{\mathbf{x}_0|\mathbf{y}} := \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{y}(M_0))$. Note that (4.19) can be equally written as $\operatorname{tr}(\mathbf{C}_{\mathbf{x}_0|\mathbf{y}})$ (cf . (2.25)). Based on this observation, we analyze the asymptotic behavior of $\mathbf{C}_{\mathbf{x}_0|\mathbf{y}}$.

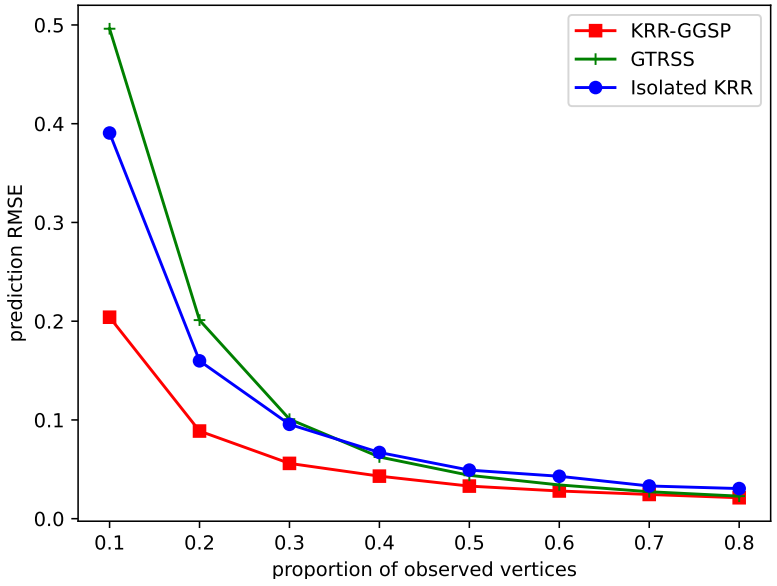


FIGURE 4.5: Prediction error under different proportions of vertices to be sampled for learning. Each point in the figure is obtained by 10 repetitions.

We compute the covariance operators $\mathbf{C}_{\mathbf{zz}}$ and $\mathbf{C}_{\mathbf{zx}_0}$ for later use:

$$\begin{aligned}
\mathbf{C}_{\mathbf{zz}} : L^2(\mathcal{J}_S) &\to L^2(\mathcal{J}_S) \\
g(\cdot) &\mapsto \int_{\mathcal{T}} \sum_{u \in \{v_0\}^{\mathrm{c}}} k_G(v,u) k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(u,\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}), \\
\mathbf{C}_{\mathbf{zx}_0} : L^2(\mathcal{T}_0) &\to L^2(\mathcal{J}_S) \\
g(\cdot) &\mapsto \int_{\mathcal{T}_0} k_G(v_0,v) k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(v_0,\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}).
\end{aligned} \tag{4.28}$$

Define the integral operators

$$\begin{aligned}
\mathbf{H} : L^2(\mathcal{T}) &\to L^2(\mathcal{T}) \\
g(\cdot) &\mapsto \int_{\mathcal{T}} k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}), \\
\mathbf{H}_0 : L^2(\mathcal{T}_0) &\to L^2(\mathcal{T}) \\
g(\cdot) &\mapsto \int_{\mathcal{T}_0} k_{\mathcal{T}}(\mathbf{t},\mathbf{s}) g(\mathbf{s}) \, \mathrm{d}\tau(\mathbf{s}).
\end{aligned}$$

Define $\mathbf{K}_{G,**}$ as the submatrix of $\mathbf{K}_G$ without the v_0-th row and the v_0-th column. Let $\mathbf{k}_{G,0*}$ be the v_0-th column of $\mathbf{K}_G$ but without the v_0-th entry. Then we have

$$\begin{aligned} \mathbf{C}_{\mathbf{zz}} &= \mathbf{K}_{G,**} \otimes \mathbf{H}, \\ \mathbf{C}_{\mathbf{z}\mathbf{x}_0} &= \mathbf{k}_{G,0*} \otimes \mathbf{H}_0. \end{aligned} \tag{4.29}$$

Lemma 4.4. *Suppose a sequence of operators $\{\mathbf{C}_n\}$ on a separable Hilbert space $\mathcal{H}$, all of which are compact, self-adjoint, positive semi-definite and trace-class. Suppose $\mathbf{J}$ is a bounded linear operator from $\mathcal{H}$ to $\mathcal{G}$, where $\mathcal{G}$ is also a separable Hilbert space. If $\lim_{n\to\infty} \operatorname{tr}(\mathbf{C}_n) = 0$, then $\lim_{n\to\infty} \operatorname{tr}(\mathbf{J}\mathbf{C}_n\mathbf{J}^*) = 0$.*

Lemma 4.5. *Suppose $\mathbf{w}_1$ is a random element in $\mathcal{H}_1$, and $\mathbf{w}_2$ is a random element in $\mathcal{H}_2$. $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable Hilbert spaces. $\mathcal{F}'$ is a sub σ-algebra of the underlying probability space. Suppose $\mathbf{w}_2 \in \mathcal{F}'$, then we have*

$$\begin{aligned} \mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'] &= \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'] \circledast \mathbf{w}_2, \\ \mathbb{E}[\mathbf{w}_2 \circledast \mathbf{w}_1 \,|\, \mathcal{F}'] &= \mathbf{w}_2 \circledast \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}']. \end{aligned}$$

Using Lemma 4.5 we can simplify the definition of conditional covariance operator as

$$\operatorname{cov}(\mathbf{w}_1, \mathbf{w}_2 \,|\, \mathcal{F}') = \mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'] - \mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'] \circledast \mathbb{E}[\mathbf{w}_2 \,|\, \mathcal{F}'].$$

Lemma 4.6. *We have*

$$\mathbf{C}_{\mathbf{x}_0|\mathbf{y}} = \mathbb{E}[\operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) \,|\, \mathbf{y}(M_0)] + \operatorname{cov}(\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] \,|\, \mathbf{y}(M_0)).$$

Lemma 4.7. *Let $\mathbf{C}_{\mathbf{z}|\mathbf{y}}$ be the conditional covariance operator of $\mathbf{z}$ given $\mathbf{y}(M_0)$. Then $\lim_{M_0\to\infty} \operatorname{tr}(\mathbf{C}_{\mathbf{z}|\mathbf{y}}) = 0$ almost surely.*

Lemma 4.8. *The conditional expectation and covariance of $\mathbf{x}_0$ given $\mathbf{z}$ are as follows:*

$$\begin{aligned} \mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] &= (\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^*\mathbf{z}, \\ \operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) &= \mathbf{C}_{\mathbf{x}_0} - \mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}^*, \end{aligned} \tag{4.30}$$

where the operator $\mathbf{C}_{\mathbf{zz}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}$ is bounded.

Proof of Theorem 4.2. We can rewrite $\mathbf{C}_{\mathbf{x}_0|\mathbf{y}}$ as follows:

$$\begin{aligned}\mathbf{C}_{\mathbf{x}_0|\mathbf{y}} &= \mathbb{E}[\mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) \,|\, \mathbf{y}(M_0)] + \mathrm{cov}(\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] \,|\, \mathbf{y}(M_0)) \\ &= \mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) + \mathrm{cov}\big(\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{z}\mathbf{z}}^{\dagger}\mathbf{z} \,\big|\, \mathbf{y}(M_0)\big) \\ &= \mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) + \mathbf{C}_{\mathbf{z}\mathbf{z}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{z}|\mathbf{y}}(\mathbf{C}_{\mathbf{z}\mathbf{z}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^*. \end{aligned} \tag{4.31}$$

The first equality holds by Lemma 4.6. The second equality holds by the fact that $\mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z})$ is deterministic. By taking trace and limit on (4.31) we have

$$\lim_{M_0\to\infty} \mathrm{tr}(\mathbf{C}_{\mathbf{x}_0|\mathbf{y}} - \mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z})) = \lim_{M_0\to\infty} \mathrm{tr}(\mathbf{C}_{\mathbf{z}\mathbf{z}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}}\mathbf{C}_{\mathbf{z}|\mathbf{y}}(\mathbf{C}_{\mathbf{z}\mathbf{z}}^{\dagger}\mathbf{C}_{\mathbf{x}_0\mathbf{z}})^*).$$

From Lemma 4.4 and Lemma 4.7 we know that the R.H.S. tends to zero. By writing $\mathrm{cov}(\mathbf{x}_0 \,|\, \mathbf{z})$ as (4.30) and using (2.28) we conclude the proof. □

4.B Appendix: Proof of Lemmas for Theorem 4.2

In this section, we prove the lemmas for the proof of Theorem 4.2.

Proof of Lemma 4.4. Let $\{\mathbf{h}_i^{(n)} : i = 1, 2, \ldots\}$ be the orthonormal basis of $\mathbf{C}_n$. Let $\{\lambda_i(\mathbf{C}_n) : i = 1, 2, \ldots\}$ be the corresponding eigenvalues. By definition of trace we have

$$\mathrm{tr}(\mathbf{J}\mathbf{C}_n\mathbf{J}^*) = \sum_{i=1}^{\infty} \langle \mathbf{J}\mathbf{C}_n\mathbf{J}^*\mathbf{h}_i^{(n)}, \mathbf{h}_i^{(n)} \rangle = \sum_{i=1}^{\infty} \langle \mathbf{C}_n\mathbf{J}^*\mathbf{h}_i^{(n)}, \mathbf{J}^*\mathbf{h}_i^{(n)} \rangle. \tag{4.32}$$

We now compute each term in (4.32). Assume that

$$\mathbf{J}^*\mathbf{h}_i^{(n)} = \sum_{j=1}^{\infty} \alpha_{ij}^{(n)} \mathbf{h}_j^{(n)}.$$

then we have

$$\begin{aligned}\langle \mathbf{C}_n\mathbf{J}^*\mathbf{h}_i^{(n)}, \mathbf{J}^*\mathbf{h}_i^{(n)} \rangle &= \langle \sum_{j=1}^{\infty} \alpha_{ij}^{(n)} \lambda_j(\mathbf{C}_n)\mathbf{h}_j^{(n)}, \sum_{j=1}^{\infty} \alpha_{ij}^{(n)} \mathbf{h}_j^{(n)} \rangle \\ &= \sum_{j=1}^{\infty} (\alpha_{ij}^{(n)})^2 \lambda_j(\mathbf{C}_n).\end{aligned}$$

Substituting this result into (4.32) we have

$$\operatorname{tr}(\mathbf{J}\mathbf{C}_n\mathbf{J}^*) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(\alpha_{ij}^{(n)})^2\lambda_j(\mathbf{C}_n) = \sum_{j=1}^{\infty}\lambda_j(\mathbf{C}_n)\sum_{i=1}^{\infty}(\alpha_{ij}^{(n)})^2.$$

Notice that $\mathbf{J}\mathbf{h}_j^{(n)} = \sum_{i=1}^{\infty}\alpha_{ij}^{(n)}\mathbf{h}_i^{(n)}$, and $\left\|\mathbf{J}\mathbf{h}_j^{(n)}\right\|^2 = \sum_{i=1}^{\infty}(\alpha_{ij}^{(n)})^2 \leq \|\mathbf{J}\|^2$. Therefore,

$$\operatorname{tr}(\mathbf{J}\mathbf{C}_n\mathbf{J}^*) \leq \|\mathbf{J}\|^2\sum_{j=1}^{\infty}\lambda_j(\mathbf{C}_n) = \|\mathbf{J}\|^2\operatorname{tr}(\mathbf{C}_n) \to 0.$$

□

Proof of Lemma 4.5. For any $\mathbf{h}_1 \in \mathcal{H}_1$ and $\mathbf{h}_2 \in \mathcal{H}_2$ we have

$$\begin{aligned}\langle\mathbb{E}[\mathbf{w}_1 \circledast \mathbf{w}_2 \,|\, \mathcal{F}'](\mathbf{h}_2), \mathbf{h}_1\rangle &= \mathbb{E}[\langle\mathbf{w}_2, \mathbf{h}_2\rangle\langle\mathbf{w}_1, \mathbf{h}_1\rangle \,|\, \mathcal{F}'] \\ &= \langle\mathbf{w}_2, \mathbf{h}_2\rangle\mathbb{E}[\langle\mathbf{w}_1, \mathbf{h}_1\rangle \,|\, \mathcal{F}'] \\ &= \langle\mathbf{w}_2, \mathbf{h}_2\rangle\langle\mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'], \mathbf{h}_1\rangle.\end{aligned}$$

The first and third equality are obtained by (2.23). The second equality is due to the fact that $\mathbf{w}_2 \in \mathcal{F}'$. On the other hand, by definition we have

$$\langle\mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'] \circledast \mathbf{w}_2(\mathbf{h}_2), \mathbf{h}_1\rangle = \langle\mathbf{w}_2, \mathbf{h}_2\rangle\langle\mathbb{E}[\mathbf{w}_1 \,|\, \mathcal{F}'], \mathbf{h}_1\rangle,$$

which concludes the Proof of the first equation. The second equation can be proved by a similar argument. □

Proof of Lemma 4.6. To prove this equality, we mainly make use of the fact that $\mathbf{y}(M_0) \in \sigma(\mathbf{z}, \{\epsilon_m\}, \mathcal{S}(M_0))$. We write $\sigma(\mathbf{z}, \{\epsilon_m\}, \mathcal{S}(M_0))$ as $\mathcal{F}_0$ for simplicity. Notice that $\operatorname{cov}(\mathbf{x}_0 \,|\, \mathbf{z}) = \operatorname{cov}(\mathbf{x}_0 \,|\, \mathcal{F}_0)$ and $\mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{z}] = \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0]$ since both $\{\epsilon_m\}$ and $\mathcal{S}(M_0)$ are jointly independent of the GP f. The first term in the R.H.S. of Lemma 4.6 can be computed as follows:

$$\begin{aligned}&\mathbb{E}[\operatorname{cov}(\mathbf{x}_0 \,|\, \mathcal{F}_0) \,|\, \mathbf{y}(M_0)] \\ &= \mathbb{E}[\mathbb{E}[(\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0]) \circledast (\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0]) \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] \\ &= \mathbb{E}[(\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0]) \circledast (\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0]) \,|\, \mathbf{y}(M_0)]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbf{x}_0 \circledast \mathbf{x}_0 \,|\, \mathbf{y}(M_0)] - \mathbb{E}[\mathbf{x}_0 \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] \\
&- \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbf{x}_0 \,|\, \mathbf{y}(M_0)] \\
&+ \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)]. \qquad (4.33)
\end{aligned}
$$

The second equality is derived by the fact that $\mathbf{y}(M_0) \in \mathcal{F}_0$. We further use this fact and Lemma 4.5 to calculate the second and third term in (4.33):

$$
\begin{aligned}
&\mathbb{E}[\mathbf{x}_0 \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)], \\
&\mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbf{x}_0 \,|\, \mathbf{y}(M_0)] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] \\
&= \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)].
\end{aligned}
$$

Substituting this result into (4.33), we obtain

$$
\mathbb{E}[\mathrm{cov}(\mathbf{x}_0 \,|\, \mathcal{F}_0) \,|\, \mathcal{F}_0)\mathbf{y}(M_0)] = \mathbb{E}[\mathbf{x}_0 \circledast \mathbf{x}_0 \,|\, \mathbf{y}(M_0)] - \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)]. \qquad (4.34)
$$

Using a similar argument as above, we have

$$
\mathrm{cov}(\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)) = \mathbb{E}[\mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathcal{F}_0] \,|\, \mathbf{y}(M_0)] - \mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{y}(M_0)] \circledast \mathbb{E}[\mathbf{x}_0 \,|\, \mathbf{y}(M_0)]. \qquad (4.35)
$$

By combining (4.34) and (4.35), we obtain the conclusion in Lemma 4.6. □

Proof of Lemma 4.7. According to [79, Theorem 3],

$$\sup_{(v,\mathbf{t})\in\mathcal{J}_S} \operatorname{var}(f(v,\mathbf{t})\,|\,\mathbf{y}(M_0)) \to 0$$

almost surely monotonically, hence

$$\operatorname{tr}(\mathbf{C}_{\mathbf{z}|\mathbf{y}}) = \int_{\mathcal{J}_S} \operatorname{var}(f(v,\mathbf{t})\,|\,\mathbf{y}(M_0))\,\mathrm{d}\zeta(v,\mathbf{t}) \to 0, \text{a. s.}.$$

□

Proof of Lemma 4.8. We first prove that $(\mathbf{x}_0, \mathbf{z})$ meets the compatible condition in [55, Section 4.2], i.e., $\operatorname{im}(\mathbf{C}_{\mathbf{z}\mathbf{x}_0}) \subset \operatorname{im}(\mathbf{C}_{\mathbf{z}\mathbf{z}})$. From (4.29) and the fact that $\mathbf{K}_G$ is full-rank we know that $\operatorname{im}(\mathbf{C}_{\mathbf{z}\mathbf{x}_0}) = \operatorname{span}\{\mathbf{k}_{G,0*}\} \otimes \operatorname{im}(\mathbf{H}_0)$ and $\operatorname{im}(\mathbf{C}_{\mathbf{z}\mathbf{z}}) = \mathbb{R}^{T-1} \otimes \operatorname{im}(\mathbf{H})$. Hence it suffices to prove that $\operatorname{im}(\mathbf{H}_0) \in \operatorname{im}(\mathbf{H})$, which can be shown by definition of $\mathbf{H}$ and $\mathbf{H}_0$.

Second, since f is a GP, according to Theorem 2.1, $(\mathbf{x}_0, \mathbf{z})$ is a Gaussian random element on $(L^2(\mathcal{J}_S \cup (\{v_0\} \times \mathcal{T}_0)), \mathcal{B})$. Then we obtain (4.30) by using [55, Theorem 4.8] and [55, Section 6]. □

4.C Appendix: Proof of Theorem 4.3

Proof. We prove this theorem in two steps: first, we prove that with high probability, there are enough sample points in a small neighborhood of $\mathcal{Q}$. Then, we prove that since the neighborhood is small, we can asymptotically upper bound $\operatorname{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{y}(M_0))$ by $\operatorname{var}(f(v_0,\mathbf{t}_0)\,|\,f(\mathcal{Q}))$.

Consider a neighborhood of $\mathbf{t}_0$: $B(\mathbf{t}_0,\delta) \subset \mathcal{T}$. Let $C_D = \dfrac{\tau(B(\mathbf{t}_0,1))}{\tau(\mathcal{T})}$, then $\tau(B(\mathbf{t}_0,\delta)) = \delta^D \tau(\mathcal{T}) C_D$. Let $\mathcal{S}(v,M_0,\delta) = \mathcal{S}(v;M_0) \bigcap (\{v\} \times B(\mathbf{t}_0,\delta))$ be the sample points that fall into $B(\mathbf{t}_0,\delta)$ on vertex v. For a sufficiently small δ, we have $C_D\delta^D < 1$. Then on each vertex v, $|\mathcal{S}(v,M_0,\delta)|$ is a Binomial random variable $n_0 \overset{\text{i.i.d.}}{\sim} \operatorname{Binom}(M_0, C_D\delta^D)$.

Consider an arbitrary $c_0 \in (0,1)$. We evaluate the probability that there are at least $c_0\mathbb{E}[n_0]$ sample points inside $B(\mathbf{t}_0,\delta)$ on vertex v. By Chebyshev's inequality we have

$$\mathbb{P}\big(\big|n_0 - M_0C_D\delta^D\big| \geq M_0C_D\delta^D(1-c_0)\big) \leq \frac{1-C_D\delta^D}{(1-c_0)^2M_0C_D\delta^D}$$

Besides, due to the symmetry of the binomial distribution, we have

$$\begin{aligned}
&\mathbb{P}\big(n_0 \leq c_0M_0C_D\delta^D\big)\\
&= \mathbb{P}\big(n_0 - M_0C_D\delta^D \leq -M_0C_D\delta^D(1-c_0)\big)\\
&= \frac{1}{2}\mathbb{P}\big(\big|n_0 - M_0C_D\delta^D\big| \geq M_0C_D\delta^D(1-c_0)\big).
\end{aligned}$$

Therefore, the number of samples in $B(\mathbf{t}_0,\delta)$ can be lower bounded by

$$\mathbb{P}\big(n_0 > c_0M_0C_D\delta^D\big) \geq 1 - \frac{1}{2}\frac{1-C_D\delta^D}{(1-c_0)^2M_0C_D\delta^D}.$$

For ease of notation, we use m_0 to denote $c_0M_0C_D\delta^D$ in the proof. Since the samples are obtained independently on each vertex, the probability that every vertex v in $\mathcal{N}_d$ has more than m_0 sampled instances in $B(\mathbf{t}_0,\delta)$ can be lower bounded by

$$\begin{aligned}
&\mathbb{P}(|\mathcal{S}(v,M_0,\delta)| > m_0, \forall\, v \in \mathcal{N}_d)\\
&= (\mathbb{P}(n_0 > m_0))^{N_d} \geq \left(1 - \frac{1}{2}\frac{1-C_D\delta^D}{(1-c_0)^2M_0C_D\delta^D}\right)^{N_d}.
\end{aligned} \tag{4.36}$$

In the sequel, we will work on this event. We will prove that with more than m_0 samples in $B(\mathbf{t}_0,\delta)$ on each $v \in \mathcal{N}_d$, we are able to upper bound $\mathrm{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{y}(M_0))$ by $\mathrm{var}(f(v_0,\mathbf{t}_0)\,|\,f(\mathcal{Q}))$. Let $\mathcal{S}(v,M_0,\delta,m_0) \subset \mathcal{S}(v,M_0,\delta)$ be any subset with cardinality m_0. Let $\mathcal{S}(v,M_0,\delta,m_0) = \{v\} \times \mathbf{t}^{(v)}$, where $\mathbf{t}^{(v)} = \{\mathbf{t}_1^{(v)},\ldots,\mathbf{t}_{m_0}^{(v)}\}$. We write $\mathbf{Y}(M_0,\delta) := (y(v,\mathbf{t}_j^{(v)}))_{v\in\mathcal{N}_d(v_0),j\in[m_0]} \in \mathbb{R}^{N_d\times m_0}$. According to [79, Lemma 9], the posterior variance can be bounded by

$$\mathrm{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{y}(M_0)) \leq \mathrm{var}(f(v_0,\mathbf{t}_0)\,|\,\mathbf{Y}(M_0,\delta)) = k_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}_0)k_G(v_0,v_0) - \mathbf{k}^\intercal\mathbf{B}^{-1}\mathbf{k}, \tag{4.37}$$

where $\mathbf{k}$ and $\mathbf{B}$ are calculated as in (4.12):

$$\mathbf{k}_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}^{(v)}) = (k_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}_1^{(v)}),\ldots,k_{\mathcal{T}}(\mathbf{t}_0,\mathbf{t}_{m_0}^{(v)}))^\intercal \in \mathbb{R}^{m_0},$$

$$\mathbf{K}_{\mathcal{T}}(\mathbf{t}^{(u)}, \mathbf{t}^{(v)}) = (k_{\mathcal{T}}(\mathbf{t}_i^{(u)}, \mathbf{t}_j^{(v)}))_{i,j\in[m_0]} \in \mathbb{R}^{m_0\times m_0},$$

$$\begin{aligned}\mathbf{k} &= (k_G(v_0, v)\mathbf{k}_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}^{(v)})^\intercal)^\intercal_{v\in\mathcal{N}_d(v_0)} \in \mathbb{R}^{N_d m_0},\\ \mathbf{B} &= (k_G(u, v)\mathbf{K}_{\mathcal{T}}(\mathbf{t}^{(u)}, \mathbf{t}^{(v)}))_{u,v\in\mathcal{N}_d(v_0)} + \sigma^2\mathbf{I}_{N_d m_0} \in \mathbb{R}^{N_d m_0\times N_d m_0}.\end{aligned}$$

Intuitively, when δ is small enough, the points in $\mathbf{t}^{(v)}$ will be close to $\mathbf{t}_0$, thus all $\mathbf{t}_i^{(v)}$ in $\mathbf{k}$ and $\mathbf{B}$ can be replaced by $\mathbf{t}_0$. Following this idea, we define

$$\begin{aligned}\boldsymbol{\kappa} &= k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)\mathbf{k}_G(v_0, \mathcal{N}_d) \otimes \mathbf{1}_{m_0}\\ \mathbf{B}' &= k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d) \otimes \mathbf{1}_{m_0}\mathbf{1}_{m_0}^\intercal + \sigma^2\mathbf{I}_{N_d m_0}.\end{aligned}$$

We aim to approximate (4.37) by replacing $\mathbf{k}$ with $\boldsymbol{\kappa}$ and $\mathbf{B}$ with $\mathbf{B}'$.

$$\begin{aligned}&|\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{Y}(M_0, \delta)) - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)l_1|\\ &= |k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)\mathbf{k}_G(v_0, \mathcal{N}_d)^\intercal\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1}\mathbf{k}_G(v_0, \mathcal{N}_d) - \mathbf{k}^\intercal\mathbf{B}^{-1}\mathbf{k}|\\ &\le |k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)\mathbf{k}_G(v_0, \mathcal{N}_d)^\intercal\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1}\mathbf{k}_G(v_0, \mathcal{N}_d) - \boldsymbol{\kappa}^\intercal\mathbf{B}'^{-1}\boldsymbol{\kappa}| + \left|\boldsymbol{\kappa}^\intercal\mathbf{B}'^{-1}\boldsymbol{\kappa} - \mathbf{k}^\intercal\mathbf{B}^{-1}\mathbf{k}\right|\end{aligned} \tag{4.38}$$

we are going to treat the two terms in (4.38) respectively. We denote the first term as (i) and the second term as (ii). To this end, we first need to calculate $\mathbf{B}'^{-1}$. By respectively calculating the eigenvalues and eigenvectors of $\mathbf{B}'$ on $\mathbb{R}^{N_d} \otimes \mathrm{span}\{\mathbf{1}_{m_0}\}$ and $\mathbb{R}^{N_d} \otimes \mathrm{span}\{\mathbf{1}_{m_0}\}^\perp$, it can be shown that

$$\mathbf{B}'^{-1} = (k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d) + \sigma^2\mathbf{I}_{N_d})^{-1} \otimes \frac{1}{m_0}\mathbf{1}_{m_0}\mathbf{1}_{m_0}^\intercal + \frac{1}{\sigma^2}\mathbf{I}_{N_d} \otimes (\mathbf{I}_{m_0} - \frac{1}{m_0}\mathbf{1}_{m_0}\mathbf{1}_{m_0}^\intercal).$$

To simplify the notation, we define the matrix

$$\mathbf{Q}_G = (k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d) + \sigma^2\mathbf{I}_{N_d})^{-1},$$

so that

$$\boldsymbol{\kappa}^\intercal\mathbf{B}'^{-1}\boldsymbol{\kappa} = k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)^2 m_0\mathbf{k}_G(v_0, \mathcal{N}_d)^\intercal\mathbf{Q}_G\mathbf{k}_G(v_0, \mathcal{N}_d).$$

$$\begin{aligned}(i) &= |\mathbf{k}_G(v_0, \mathcal{N}_d)^\intercal(k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1} - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)^2 m_0\mathbf{Q}_G)\mathbf{k}_G(v_0, \mathcal{N}_d)|\\ &\le |k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)|\,\|\mathbf{k}_G(v_0, \mathcal{N}_d)\|_2^2 \cdot \left\|\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1} - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0\mathbf{Q}_G\right\|_2.\end{aligned} \tag{4.39}$$

Notice that $\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)$ and $\mathbf{Q}_G$ has the same set of eigenvectors. Specifically, if $\boldsymbol{\psi}$ is $\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)$'s eigenvector associated with eigenvalue α, then it is $\mathbf{Q}_G$'s eigenvector associated with eigenvalue $(k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0\alpha + \sigma^2)^{-1}$. Let $\sigma_{\min} > 0$ be the minimum eigenvalue of $\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)$. Using this relationship we can derive a bound for the norm of the matrix difference

$$\begin{aligned}\left\|\mathbf{K}_G(\mathcal{N}_d, \mathcal{N}_d)^{-1} - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0\mathbf{Q}_G\right\|_2 &= \frac{1}{\sigma_{\min}} - \frac{1}{\sigma_{\min} + \sigma^2/(k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0)} \\ &< \frac{1}{\sigma_{\min}^2}\frac{\sigma^2}{k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)m_0}.\end{aligned}$$

By substituting this result into (4.39), and noticing that the $k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0)$ and $\mathbf{k}_G(v_0, \mathcal{N}_d)$ do not depend on m_0, we have

$$(i) < \frac{C_1}{m_0}, \tag{4.40}$$

where C_1 is a constant which only depends on d.

We find the upper bound for (ii) by triangle inequality:

$$\begin{aligned}(ii) &\leq \left|\boldsymbol{\kappa}^\intercal\mathbf{B}'^{-1}\boldsymbol{\kappa} - \boldsymbol{\kappa}^\intercal\mathbf{B}^{-1}\boldsymbol{\kappa}\right| + \left|\boldsymbol{\kappa}^\intercal\mathbf{B}^{-1}\boldsymbol{\kappa} - \mathbf{k}^\intercal\mathbf{B}^{-1}\boldsymbol{\kappa}\right| + \left|\mathbf{k}^\intercal\mathbf{B}^{-1}\boldsymbol{\kappa} - \mathbf{k}^\intercal\mathbf{B}^{-1}\mathbf{k}\right| \\ &\leq \|\boldsymbol{\kappa}\|_2^2 \left\|\mathbf{B}'^{-1} - \mathbf{B}^{-1}\right\|_2 + \|\boldsymbol{\kappa} - \mathbf{k}\|_2 \left\|\mathbf{B}^{-1}\right\|_2 \|\boldsymbol{\kappa}\|_2 + \|\boldsymbol{\kappa} - \mathbf{k}\|_2 \left\|\mathbf{B}^{-1}\right\|_2 \|\mathbf{k}\|_2 .\end{aligned} \tag{4.41}$$

Then it suffices to find bounds for the norms of vectors and matrices in (4.41). By definition, we have

$$\|\boldsymbol{\kappa}\|_2 = k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) \|\mathbf{k}_G(v_0, \mathcal{N}_d)\|_2 m_0^{\frac{1}{2}}.$$

Since $k_{\mathcal{T}}$ is continuous and $\mathcal{T}$ is compact, $k_{\mathcal{T}}$ can achieve its maximum, denoted as $M(k_{\mathcal{T}}) := \max_{(\mathbf{s},\mathbf{t})}\{k_{\mathcal{T}}(\mathbf{s}, \mathbf{t})\}$. Besides, $k_{\mathcal{T}}$ is Lipschitz continuous with Lipschitz constant $L(k_{\mathcal{T}})$. Then we have

$$\begin{aligned}\|\mathbf{k}\|_2 &\leq M(k_{\mathcal{T}}) \|\mathbf{k}_G(v_0, \mathcal{N}_d)\|_2 m_0^{\frac{1}{2}}, \\ \|\boldsymbol{\kappa} - \mathbf{k}\|_2 &\leq \|\mathbf{k}_G(v_0, \mathcal{N}_d)\|_2 L(k_{\mathcal{T}})\delta m_0^{\frac{1}{2}} \\ \|\mathbf{B} - \mathbf{B}'\|_2 &\leq \|\mathbf{B} - \mathbf{B}'\|_\infty \leq \sqrt{2} \|\mathbf{k}_G(v_0, \mathcal{N}_d)\|_1 L(k_{\mathcal{T}})\delta m_0.\end{aligned}$$

By definition we know that $\mathbf{B} \succeq \sigma^2 \mathbf{I}_{N_d^1 m_0}$, so $\mathbf{B}^{-1} \preceq \frac{1}{\sigma^2}\mathbf{I}_{N_d^1 m_0}$, $\|\mathbf{B}^{-1}\|_2 \leq \frac{1}{\sigma^2}$. Using the same argument we have $\|\mathbf{B}'^{-1}\| \leq \frac{1}{\sigma^2}$.

$$\begin{aligned}\left\|\mathbf{B}'^{-1} - \mathbf{B}^{-1}\right\|_2 &= \left\|\mathbf{B}^{-1}(\mathbf{B} - \mathbf{B}')\mathbf{B}'^{-1}\right\|_2 \\ &= \left\|\mathbf{B}^{-1}\right\|_2 \|\mathbf{B} - \mathbf{B}'\|_2 \left\|\mathbf{B}'^{-1}\right\|_2 \\ &\leq \sqrt{2}\sigma^{-4} \left\|\mathbf{k}_G(v_0, \mathcal{N}_d)\right\|_1 L(k_{\mathcal{T}})\delta m_0.\end{aligned}$$

Combining all the bounds on vectors and matrices' norms with (4.41) we obtain that

$$(ii) \leq C_2 \delta m_0^2 + C_3 \delta m_0, \tag{4.42}$$

where C_2 and C_3 are constants only depend on d. By combining (4.40), (4.42) with (4.38) we obtain that

$$\left|\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{Y}(M_0, \delta)) - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l_1\right| \leq \frac{C_1}{m_0} + C_2 \delta m_0^2 + C_3 \delta m_0.$$

Now we are to examine the asymptotic case when $M_0 \to \infty$. Recall that $m_0 = c_0 M_0 C_D \delta^D$. If we let $\delta = M_0^{-\beta}$ where $\beta > 0$, we have

$$\begin{aligned}&\left|\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{Y}(M_0, \delta)) - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l_1\right| \\ &\leq C_1 c_0^{-1} M_0^{\beta D - 1} + C_2 c_0^2 M_0^{2-(2D+1)\beta} + C_3 c_0 M_0^{1-(D+1)\beta},\end{aligned} \tag{4.43}$$

where the constant C_D is absorbed by C_1, C_2 and C_3. By requiring the powers of M_0 to be negative, β should be in the range $(\frac{2}{2D+1}, \frac{1}{D})$. By adjusting β in this range, the best rate is achieved when $\hat{\beta} = \frac{3}{3D+1}$. By substituting this into (4.43), we obtain

$$\left|\mathrm{var}(f(v_0, \mathbf{t}_0) \,|\, \mathbf{Y}(M_0, \delta)) - k_{\mathcal{T}}(\mathbf{t}_0, \mathbf{t}_0) l_1\right| \leq (C_1 c_0^{-1} + C_2 c_0^2) M_0^{-\frac{1}{3D+1}} + C_3 c_0 M_0^{-\frac{2}{3D+1}}. \tag{4.44}$$

On the other hand, by substituting $\delta = M_0^{-\hat{\beta}}$ into (4.36), we have

$$\begin{aligned}\mathbb{P}(|\mathcal{S}(v, M_0, \delta)| > m_0, \forall\, v \in \mathcal{N}_d(v_0)) &\geq \left(1 - \frac{1}{2}\frac{1 - C_D\delta^D}{(1-c_0)^2 M_0 C_D \delta^D}\right)^{N_d} \\ &\geq \left(1 - \frac{1}{2}\frac{1}{(1-c_0)^2 C_D M_0^{\frac{1}{3D+1}}}\right)^{N_d}.\end{aligned} \tag{4.45}$$

Finally, by combining (4.37), (4.44) and (4.45), we conclude the proof. □

Chapter 5

Multiple Hypothesis Testing over the Joint Domain

5.1 Statistical Model

5.1.1 Problem formulation

In Chapters 3 and 4 we have introduced the methods for estimating random generalized graph signals. In this chapter, we address the hypothesis testing problem for random generalized graph signals. Our solution to this problem leverages the joint Fourier basis $\{\boldsymbol{\phi}_k \otimes \psi_l\}$, and includes estimation of bandlimited generalized graph signals.

Suppose we have a hypothesis testing problem $H(u, \mathbf{s})$ on every point $(u, \mathbf{s}) \in \mathcal{V} \times \mathcal{T}$. As in previous chapters, we write $\mathcal{V} \times \mathcal{T}$ as $\mathcal{J}$. We define the *null region* $\mathcal{J}_0 := \{(u, \mathbf{s}) \in \mathcal{J} : H(u, \mathbf{s}) \text{ is null}\}$ and the *alternative region* $\mathcal{J}_1 := \mathcal{J} \backslash \mathcal{J}_0$.

Note that the inference of $\mathcal{J}_0$ and $\mathcal{J}_1$ may contain an infinite number of tests. In practice, we only have access to p-values from a finite subset of $\mathcal{J}$, denoted by $\mathcal{S} = \{(v_m, \mathbf{t}_m) : m = 1, \ldots, M\}$, where $\mathcal{S}$ is a random finite subset of $\mathcal{J}$. The corresponding (random) set of p-values is denoted as $\{p_m : m = 1, \ldots, M\}$. In this chapter, we consider the hypothesis tests on $\mathcal{S}$. Let $\mathbf{p} := (p_1, \ldots, p_m) \in \mathbb{R}^M$. Define $\{0, 1\}^{\mathcal{S}}$ as the set of functions on $\mathcal{S}$ that take values in $\{0, 1\}$. A *detection strategy*

is a mapping

$$h : \mathbb{R}^M \to \{0,1\}^{\mathcal{S}}$$
$$\mathbf{p} \mapsto h(\mathbf{p}),$$

i.e., given a vector of p-values $\mathbf{p}$, $h(\mathbf{p})$ is a function that maps each sample point in $\mathcal{S}$ to either 0 or 1. We can then define $h(\mathbf{p})^{-1}\{a\}$ as the set of sample points in $\mathcal{S}$ that are mapped to a by $h(\mathbf{p})$, for $a = 0, 1$. Correspondingly, we define $\mathcal{J}_{0,M} := \mathcal{J}_0 \bigcap \mathcal{S}$, $\mathcal{J}_{1,M} := \mathcal{J}_1 \bigcap \mathcal{S}$, $\widehat{\mathcal{J}}_{0,M} := h(\mathbf{p})^{-1}\{0\} \bigcap \mathcal{S}$ and $\widehat{\mathcal{J}}_{1,M} := h(\mathbf{p})^{-1}\{1\} \bigcap \mathcal{S}$, where $\widehat{\mathcal{J}}_{0,M}$ and $\widehat{\mathcal{J}}_{1,M}$ are the detected null and alternative regions.

The *FDR* and the *power* of h are defined as

$$\mathrm{FDR}(h; M) = \mathbb{E}\left[\frac{\left|\widehat{\mathcal{J}}_{1,M} \bigcap \mathcal{J}_{0,M}\right|}{\max(\left|\widehat{\mathcal{J}}_{1,M}\right|, 1)}\right], \tag{5.1}$$

$$\mathrm{pow}(h; M) = \mathbb{E}\left[\frac{\left|\widehat{\mathcal{J}}_{1,M} \bigcap \mathcal{J}_{1,M}\right|}{\max(|\mathcal{J}_{1,M}|, 1)}\right]. \tag{5.2}$$

Our goal is to design a strategy h such that $\mathrm{pow}(h; M)$ is as large as possible while $\mathrm{FDR}(h; M)$ is controlled by a prescribed level α. In MHT, one widely used approach is the *empirical-Bayesian approach* based on the two-groups model assumption [41]. This model assumes that, first, the p-values have different distributions f_0 and f_1 under null and alternative hypotheses. Second, the hypotheses being alternative is random with probability π_0. Under this model, it is proved that if we use $p \in \mathcal{Q}$ as a rejection rule, then the Bayes FDR $\mathbb{P}(\text{hypothesis is null} \mid p \in \mathcal{Q})$ dominates the FDR. Therefore, in order to make as many discoveries as possible with FDR control, it is proposed to choose the largest set of hypotheses whose average lfdr value does not exceed the nominal significance level. Since this model assumes that f_0, f_1 and π_0 are the same for all hypotheses, and the p-values are independent, it describes the situation where the p-values are exchangeable among different tests. However, this exchangeability assumption may not always be appropriate in practice. For example, in [30], the authors considered the situation where each hypothesis is associated with a vertex in a graph. They proposed to allow π_0 to vary across vertices to raise the detection power. In this chapter, we consider a more general

and flexible framework, where f_0, f_1 and π_0 can be inhomogeneous among different vertices and time instances. We assume the hierarchical model depicted in Fig. 5.1.

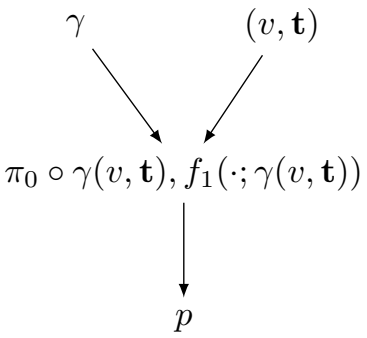


FIGURE 5.1: The scheme of the Bayesian model for MHT in GGSP.

In this model, γ is a stochastic process on $\mathcal{J}$ with $\gamma(u, \mathbf{s}) \in \mathcal{Z}$ for all $(u, \mathbf{s}) \in \mathcal{J}$ and a given space $\mathcal{Z}$. A sample point $(v, \mathbf{t})$ is drawn from a positive probability measure ρ on $\mathcal{J}$, independently with the process γ. Let π_0 be a continuous function from $\mathcal{Z}$ to $[0, 1]$. Given γ and $(v, \mathbf{t})$, we obtain a probability $\pi_0 \circ \gamma(v, \mathbf{t}) \in [0, 1]$. We assume that two steps generate the p-value p: first, generate θ as

$$\theta = \begin{cases} 0 & \text{w.p. } \pi_0 \circ \gamma(v, \mathbf{t}), \\ 1 & \text{w.p. } 1 - \pi_0 \circ \gamma(v, \mathbf{t}). \end{cases}$$

Then, under null and alternative hypotheses, we assume that the p-value p is generated independently by

$$\begin{aligned} p \mid \{\theta = 0\} &\sim f_0(\cdot; (v, \mathbf{t})), \\ p \mid \{\theta = 1\} &\sim f_1(\cdot; \gamma(v, \mathbf{t})), \end{aligned}$$

where f_1 is from a parametric family of pdfs $\mathcal{P}^1 := \{f_1(\cdot; \zeta) : \zeta \in \mathcal{Z}\}$ on $[0, 1]$ with parameter ζ.

On the other hand, we assume that $f_0(\cdot; (u, \mathbf{s}))$ is known and deterministic for all $(u, \mathbf{s}) \in \mathcal{J}$. Suppose the p-value p is derived from a test statistic w that is a continuous random variable with cumulative distribution function (cdf) F_w under the null hypothesis. When the hypothesis test is a one-sided right-tail test, $p = 1 - F_w(w)$ has distribution Unif $(0, 1)$. Similarly, it can be shown that when $H(v, \mathbf{t})$ is a one-sided left-tail test, $p \sim$ Unif $(0, 1)$ holds true. When none of the above situations

apply and $f_0(\cdot;(v,\mathbf{t}))$ is unknown, then $f_0(\cdot;(v,\mathbf{t}))$ may be estimated from data beforehand by controlling the null hypotheses to be true. For the M i.i.d. samples $(v_m,\mathbf{t}_m)$ in $\mathcal{S}$, we obtain M independent samples $(\theta_m)_{m=1}^M$, and p-values $(p_m)_{m=1}^M$.

We note that this model generalizes the two-group model in the MHT literature. In traditional MHT, each θ_m is modeled as a Bernoulli random variable Bern $(1-\pi_0)$ where π_0 is a fixed value. The distributions f_0 and f_1 are also assumed to be the same for all p-values. This model assumes that the hypothesis tests are exchangeable. However, in practice, the exchangeable assumption may not be appropriate when auxiliary information is available. For example, in [45], the authors consider the situation where the hypotheses' relative chances of being null are known. In other words, we know that π_0 are different among different tests, and we also know their order. The authors of [45] thus model π_0 as a non-decreasing function of m. Similarly, by parametrizing f_1 and π_0 with $\gamma(v,\mathbf{t})$, we indicate that both the probability of hypotheses being null and the distribution of alternative p-values can be inhomogeneous over $\mathcal{J}$ (see Example 5.1 and Fig. 5.2 for illustration).

Under our model, given γ, the conditional joint pdf of the p-value p, the indicator θ, and the sample point $(v,\mathbf{t})$ is

$$f_{p,\theta,(v,\mathbf{t})}(q,\vartheta,(u,\mathbf{s})\mid\gamma)=\Big(\pi_0\circ\gamma(u,\mathbf{s})\mathbb{I}\{\vartheta=0\}f_0(q;(u,\mathbf{s}))+(1-\pi_0\circ\gamma(u,\mathbf{s}))\mathbb{I}\{\vartheta=1\}f_1(q;\gamma(u,\mathbf{s}))\Big)\rho(u,\mathbf{s}),$$

where $q\in(0,1]$, $\vartheta\in\{0,1\}$, and $(u,\mathbf{s})\in\mathcal{J}$. In the rest of this chapter, we will use q, ϑ and $(u,\mathbf{s})$ as the corresponding variables in the pdfs for the random variables p, θ and $(v,\mathbf{t})$. From $f_{p,\theta,(v,\mathbf{t})}$, we derive the conditional pdf of p and $(v,\mathbf{t})$ as

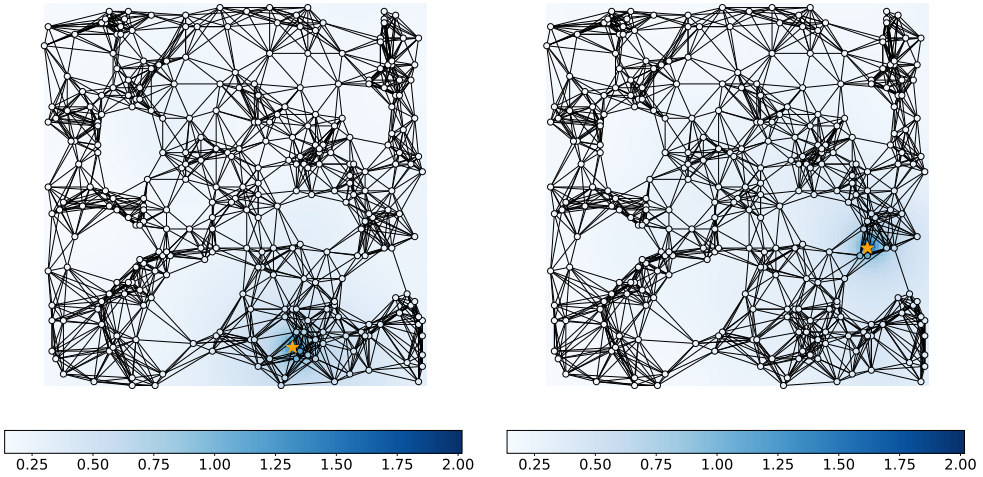
$$f_{p,(v,\mathbf{t})}(q,(u,\mathbf{s})\mid\gamma)=\Big(\pi_0\circ\gamma(u,\mathbf{s})f_0(q;(u,\mathbf{s}))+(1-\pi_0\circ\gamma(u,\mathbf{s}))f_1(q;\gamma(u,\mathbf{s}))\Big)\rho(u,\mathbf{s}),\tag{5.3}$$

and the conditional pdf of p given γ and $(v,\mathbf{t})$ as

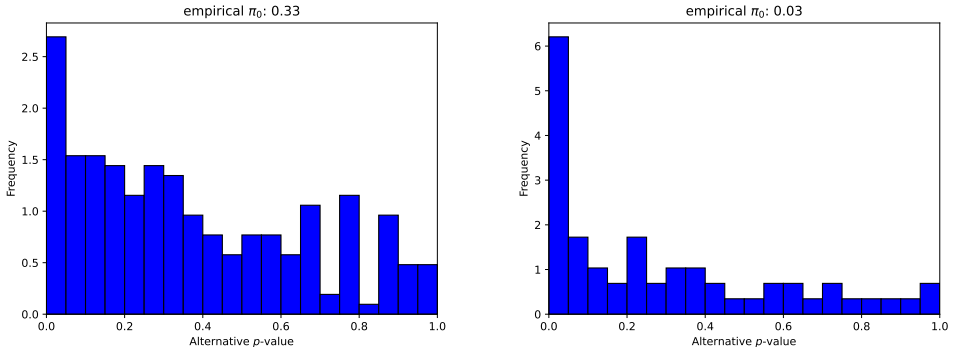
$$f_{\text{mix}}(q\mid\gamma(u,\mathbf{s}))=\pi_0\circ\gamma(u,\mathbf{s})f_0(q;(u,\mathbf{s}))+(1-\pi_0\circ\gamma(u,\mathbf{s}))f_1(q;\gamma(u,\mathbf{s})).\tag{5.4}$$

In other words, we parameterize the distribution of p-values on $\mathcal{J}$ by a *random generalized graph signal* γ. This model combines the information from the joint

domain $\mathcal{J}$ via γ and the two-group model widely adopted in the MHT literature [41, 44, 45]. We illustrate our model in the following example.



(A) The generalized graph signal $\gamma(u, \mathbf{s}) = (x_{\text{sf}}(u, \mathbf{s}, i))_{i=1,2}$. Here, we show the snapshot of $\gamma(u, \mathbf{s})$ at a fixed time instance $\mathbf{s}$. For clarity of visualization, the first image shows $x_{\text{sf}}(u, \mathbf{s}, 1)^{1/2}$, and the second shows $x_{\text{sf}}(u, \mathbf{s}, 2)^{1/2}$. The transmitters' positions are highlighted as stars in orange.



(B) The left figure shows the empirical π_0 and empirical histogram corresponding to f_1 when $\|\gamma(u, \mathbf{s})\|_2 \in [0.2, 0.35)$. The right figure shows these quantities when $\|\gamma(u, \mathbf{s})\|_2 \in [0.35, 0.5)$.

FIGURE 5.2: Illustration of $\gamma(u, \mathbf{s})$ in Example 5.1 and how the proportion of null hypotheses and the empirical distribution of p-values from alternatives vary with $\gamma(u, \mathbf{s})$.

Example 5.1. Consider a radio signal emitted by multiple transmitters in a 2-dimensional (2D) region. It is monitored by a sensor network G. We first introduce the signal propagation model from a single transmitter. Denote the location of the transmitter as $\mathbf{c} = (c_x, c_y)$. Let the signal magnitude (i.e., the absolute value of the signal) being transmitted by this transmitter be x_0. We denote the distance between a sensor u (with coordinate (u_x, u_y)) and the transmitter as $d(u, \mathbf{c})$. The signal value received by each sensor u from this single transmitter, denoted by $x(u)$,

is subject to path loss, shadow fading and fast fading [81, 82]:

$$x_{\text{sf}}(u) = Cx_0 \cdot \frac{\lambda}{4\pi d(u, \mathbf{c})} \cdot \exp(s(u_x, u_y)), \tag{5.5}$$

$$x_{\text{ff}}(u) \sim \begin{cases} \text{Rice}(\text{ratio}, x_{\text{sf}}(u)^2) & \text{if LOS path dominates,} \\ \text{Rayleigh}(x_{\text{sf}}(u)^2) & \text{otherwise.} \end{cases} \tag{5.6}$$

In (5.5), s is a two-dimensional Gaussian process, λ is the wavelength, and C is a constant. The quantity $x_{\text{sf}}(u)$ denotes the signal magnitude after path loss and shadow fading. In (5.6), we present two typical types of fast fading. In cases where there is a dominant line-of-sight (LOS) path, the received signal magnitude follows a Rician distribution where ratio represents the ratio between signal power from the LOS and the remaining multipath. Another parameter equals to the signal power $x_{\text{sf}}(u)^2$. On the other hand, when there is no single dominant signal path, the received signal magnitude follows a Rayleigh distribution where the parameter is the signal power.

When there are multiple moving transmitters, we denote the signal that node u receives from transmitter i at time $\mathbf{s}$ as $x_{\text{ff}}(u, \mathbf{s}, i)$. We assume that node u receives the strongest signal. i.e.,

$$|x(u, \mathbf{s})| = \max_i |x_{\text{ff}}(u, \mathbf{s}, i)| . \tag{5.7}$$

Finally, the signal energy measured by the node u at instance $\mathbf{s}$ is

$$y(u, \mathbf{s}) := x(u, \mathbf{s}) + e,$$

where e is AWGN.

Suppose we want to test whether the received signal energy $x(v_m, \mathbf{t}_m)$ is above the noise floor τ_0, i.e.,

$$H_0(v_m, \mathbf{t}_m) : |x(v_m, \mathbf{t}_m)| \leq \tau_0,$$
$$H_1(v_m, \mathbf{t}_m) : |x(v_m, \mathbf{t}_m)| > \tau_0,$$

then the summary statistics at $(v_m, \mathbf{t}_m)$ is $p_m = \mathbb{P}(y^2 \geq y(v_m, \mathbf{t}_m)^2 \,|\, H_0(v_m, \mathbf{t}_m))$. Here, y^2 is a chi-squared random variable whose degree of freedom is 1.

In this propagation model, the vector $(x_{\mathrm{sf}}(u, \mathbf{s}, i))_{i\geq 1}$ corresponds to the stochastic process $\gamma(u, \mathbf{s})$. Given γ and $(v_m, \mathbf{t}_m)$, the distribution of $x(v_m, \mathbf{t}_m)$ is determined by (5.6) and (5.7). Besides, According to the nature of fast fading, $x(v_m, \mathbf{t}_m)$ for different $(v_m, \mathbf{t}_m)$ are independent given γ. Hence, $(\theta_m)_{m=1}^M$ are independent with inhomogeneous probabilities on $\mathcal{J}$ determined by γ. When $H_0(v_m, \mathbf{t}_m)$ is true, the amplitude of $x(v_m, \mathbf{t}_m)$ is small if τ_0 is chosen to be small, and it can be approximated by 0. In this case, the distribution of p_m is Unif $(0, 1)$. When $H_1(v_m, \mathbf{t}_m)$ is true, the distribution of p_m only relies on the value of $\gamma(v_m, \mathbf{t}_m)$. This coincides with the hierarchical Bayesian model. We illustrate the dependence of the null probability and the alternative p-value distribution on $\gamma(v_m, \mathbf{t}_m)$ in Fig. 5.2.

Besides our model, there are other works that aim to model MHT problems over graphs or spatial areas. For example, the work [32] studies the MHT problem when all vertices have similar alternative distributions of p-values. In [30, 31], the models utilize the graph structure by respectively imposing sparsity of the π_0 differences and p-value weights among the edges. In [33, (12)], the authors assumed a finite number of alternative distributions of p-values over a spatial area. From Example 5.1, we see that these assumptions may not be flexible enough since, on the one hand, the distance $d(u, \mathbf{c}(\mathbf{s}))$ varies continuously over $\mathcal{J}$ instead of being piecewise constant on $\mathcal{J}$. On the other hand, the number of alternative p-value distributions may be the same as the number of p-values. Therefore, by parametrizing the alternative distributions with a smooth γ over $\mathcal{J}$, we obtain a finer model than the existing ones.

In this chapter, we assume that $\pi_0 \circ \gamma(u, \mathbf{s})$ and $f_1(\cdot; \gamma(u, \mathbf{s}))$ are identifiable from $f_{\mathrm{mix}}(\cdot \mid \gamma(u, \mathbf{s}))$ using (5.4). This can be guaranteed by the following assumptions:

Assumption 5.1.

(i) $f_0(q; (u, \mathbf{s}))$ is non-decreasing w.r.t. $q \in [0, 1]$.

(ii) $f_1(\cdot; \zeta)$ is non-increasing on $(0, 1]$ for all $\zeta \in \mathcal{Z}$.

(iii) $\min_{q\in[0,1]} f_1(q; \zeta) = 0$ for all $\zeta \in \mathcal{Z}$, i.e., with condition (ii), $f_1(1; \zeta) = 0$ for all $\zeta \in \mathcal{Z}$.

(iv) $f_0(q; (u, \mathbf{s}))$ is continuous on $[0, 1] \times \mathcal{J}$, and $f_1(q; \zeta)$ is continuous on $(0, 1] \times \mathcal{Z}$.

In Assumption 5.1, conditions (i) and (ii) indicate that p-value is more likely to be small under the alternative hypothesis and more likely to be large under the null hypothesis. The assumptions in Assumption 5.1 are commonly assumed in the MHT literature (cf . [44, Theorem 2] and [45, Section 2.1]). They ensure that $\pi_0 \circ \gamma(u, \mathbf{s})$ and $f_1(\cdot; \gamma(u, \mathbf{s}))$ are identifiable from $f_{\text{mix}}(\cdot \mid \gamma(u, \mathbf{s}))$, since according to conditions (i) to (iii), we have

$$\pi_0 \circ \gamma(u, \mathbf{s}) = \frac{f_{\text{mix}}(1 \mid \gamma(u, \mathbf{s}))}{f_0(1; (u, \mathbf{s}))}, \tag{5.8}$$

$$f_1(q; \gamma(u, \mathbf{s})) = \frac{1}{1 - \pi_0 \circ \gamma(u, \mathbf{s})} \Big(f_{\text{mix}}(q \mid \gamma(u, \mathbf{s})) - \pi_0 \circ \gamma(u, \mathbf{s}) f_0(q; (u, \mathbf{s})) \Big). \tag{5.9}$$

The lfdr is defined as

$$\text{lfdr}(q; \gamma(u, \mathbf{s})) = \frac{\pi_0 \circ \gamma(u, \mathbf{s}) f_0(p; (u, \mathbf{s}))}{f_{\text{mix}}(p \mid \gamma(u, \mathbf{s}))}. \tag{5.10}$$

The lfdr for $\{p_m, (v_m, \mathbf{t}_m)\}$ is the conditional probability $\mathbb{P}(\theta_m = 0 \,|\, \gamma(v_m, \mathbf{t}_m), p_m)$. In Section 5.2.2, we estimate $f_{\text{mix}}(\cdot \mid \gamma(v_m, \mathbf{t}_m))$, and thus $\pi_0 \circ \gamma(v_m, \mathbf{t}_m)$ and $f_1(\cdot; \gamma(v_m, \mathbf{t}_m))$.

In this chapter, we assume that the functions of random variables are always measurable or have measurable choices.

5.2 Asymptotic FDR Control Approach

In this section, we first explain the detection strategy when the true values of lfdr are known. Next, we propose a method to estimate the lfdr and present a theoretical guarantee of FDR control by using the estimated lfdr for detection.

5.2.1 Oracle solution

In this subsection, we show that, if $\pi_0 \circ \gamma(v_m, \mathbf{t}_m)$ and $f_1(\cdot; \gamma(v_m, \mathbf{t}_m))$ are known for $m \geq 0$, then the optimal detection strategy is thresholding the lfdr. To see this, we introduce two complementary definitions, marginal FDR and marginal power,

defined as

$$\mathrm{mFDR}(h;\gamma,\mathcal{S}) = \frac{\mathbb{E}\Big[\Big|\widehat{\mathcal{J}}_{1,M}\bigcap\mathcal{J}_{0,M}\Big|\,\Big|\,\gamma(\mathcal{S})\Big]}{\mathbb{E}\Big[\Big|\widehat{\mathcal{J}}_{1,M}\Big|\,\Big|\,\gamma(\mathcal{S})\Big]}, \tag{5.11}$$

$$\mathrm{mpow}(h;\gamma,\mathcal{S}) = \frac{\mathbb{E}\Big[\Big|\widehat{\mathcal{J}}_{1,M}\bigcap\mathcal{J}_{1,M}\Big|\,\Big|\,\gamma(\mathcal{S})\Big]}{\mathbb{E}[|\mathcal{J}_{1,M}|\,|\,\gamma(\mathcal{S})]}. \tag{5.12}$$

When the hypothesis tests are conducted separately, the null hypotheses are rejected if the p-value is lower than the pre-determined significance level. In this chapter, for each individual test on $(v_m, \mathbf{t}_m)$, we use the following thresholding rule:

$$h(\mathbf{p})(v_m,\mathbf{t}_m) = \begin{cases} 1 & \text{if } p_m \le z_m, \\ 0 & \text{otherwise.} \end{cases} \tag{5.13}$$

The difference is that, in (5.13), the threshold of p-value is no longer the significance level but chosen by certain criteria that we will soon introduce. Under this rule, designing the detection strategy h amounts to designing $\mathbf{z} := (z_1, \ldots, z_m)$. Since the rejection rule h is fully determined by $\mathbf{z}$, we denote $\mathrm{mFDR}(h;\gamma,\mathcal{S})$ and $\mathrm{mpow}(h;\gamma,\mathcal{S})$ by $\mathrm{mFDR}(\mathbf{z};\gamma,\mathcal{S})$ and $\mathrm{mpow}(\mathbf{z};\gamma,\mathcal{S})$ under (5.13). We thus consider the following problem:

$$\begin{aligned} &\max_{\mathbf{z}\in[0,1]^M} \mathrm{mpow}(\mathbf{z};\gamma,\mathcal{S}) \\ &\text{s.t. } \mathrm{mFDR}(\mathbf{z};\gamma,\mathcal{S}) \le \alpha. \end{aligned} \tag{5.14}$$

In Theorem 5.1, we show that the optimal solution to problem (5.14) is a level surface of lfdr. This theorem is a modification of [44, Theorem 2]. In Section 5.2.2, we show that by solving an approximate version of (5.14), we can obtain an h that yields asymptotic FDR control by a pre-determined FDR level.

Theorem 5.1. *Suppose Assumption 5.1 holds and there exists $q \in (0,1)$ and $(u,\mathbf{s}) \in \mathcal{S}$ such that $\mathrm{lfdr}(q;\gamma(u,\mathbf{s})) < \alpha$. Then the optimal solution $\mathbf{z}^* = (z_1^*, \ldots, z_M^*)$ to problem (5.14) satisfies*

$$\mathrm{lfdr}(z_m^*;\gamma(v_m,\mathbf{t}_m)) = \eta,\ \ m = 1,\ldots,M,$$

where η is a constant independent of m.

Proof. See Section 5.A for the proof. □

Different from [44, Theorem 2] and [45, Proposition 2.1], in Theorem 5.1, we do not require that $f_1(\cdot;\gamma(u,\mathbf{s}))$ is continuous on the *closed interval* $[0,1]$. This means that we allow for unbounded f_1 such as the Beta distribution. Theorem 5.1 implies that the optimal threshold for $\{p_m\}$ corresponds to a level set of lfdr. Therefore, the rejection rule becomes

$$h(\mathbf{p})(v_m,\mathbf{t}_m)=\begin{cases}1 & \text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta,\\ 0 & \text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))>\eta.\end{cases} \tag{5.15}$$

Note that in practice we usually do not have access to the ground truth $\pi_0\circ\gamma(v_m,\mathbf{t}_m)$ and $f_1(\cdot;\gamma(v_m,\mathbf{t}_m))$. Therefore, we call (5.15) an *oracle* rejection rule. In the next section, we replace (5.15) by its estimate determined from samples. Here, we first explain the choice of η under this oracle rule, and the sample-based version will then easily follow. To clarify the choice of η, we define the following quantities as in [45]:

$$d_{1,M}(\eta):=\frac{1}{M}\sum_{m=1}^{M}\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\mathbb{I}\{\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta\}, \tag{5.16}$$

$$d'_{1,M}(\eta):=\frac{1}{M}\sum_{m=1}^{M}(1-\theta_m)\mathbb{I}\{\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta\} \tag{5.17}$$

$$d_{0,M}(\eta):=\frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\{\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta\}. \tag{5.18}$$

By conditioning on p_m and noting that $\mathbb{P}(\theta_m=0\,|\,p_m,\gamma(v_m,\mathbf{t}_m))=\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))$, we see that $d_{1,M}(\eta)$ estimates the proportion of false rejections among all tests:

$$\begin{aligned}&\mathbb{E}[(1-\theta_m)\mathbb{I}\{\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta\}\,|\,\gamma,\mathcal{S}]\\&=\mathbb{E}[\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\mathbb{I}\{\text{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m))\le\eta\}\,|\,\gamma,\mathcal{S}].\end{aligned}$$

Therefore, by taking expectation over $\mathcal{S}$, we see that $\mathbb{E}[d_{1,M}(\eta)\,|\,\gamma]=\mathbb{E}\Big[\Big|\widehat{\mathcal{J}}_{1,M}\bigcap\mathcal{J}_{0,M}\Big|\,\Big|\,\gamma\Big]=$

$\mathbb{E}\big[d'_{1,M}(\eta)\,\big|\,\gamma\big]$. Besides, it can be shown that $\mathbb{E}[d_{0,M}(\eta)\,|\,\gamma] = \mathbb{E}\Big[\Big|\widehat{\mathcal{J}}_{1,M}\Big|\,\Big|\,\gamma\Big]$. Therefore, the quantity

$$r_M(\eta) := \frac{d_{1,M}(\eta)}{d_{0,M}(\eta)}$$

approximates $\mathrm{mFDR}(h;\gamma,\mathcal{S})$ in (5.11). On the other hand, note that $\mathrm{mpow}(h;\gamma,\mathcal{S})$ in (5.12) increases with η. Therefore, we choose the optimal η under the oracle rule (5.15) in the following way:

$$\eta_M := \sup\{\eta\,:\,r_M(\eta) \le \alpha\}. \tag{5.19}$$

5.2.2 Joint density estimation and testing procedure

In this subsection, we propose a method to estimate the unknown densities $f_{\mathrm{mix}}(\cdot\,|\,\gamma(v_m,\mathbf{t}_m))$ for $m = 1,\dots,M$, and then solve the MHT problem with these estimates. Estimating the unknown densities is not an easy task in general since the number of unknown densities is the same as the number of p-values. In this chapter, we take advantage of the GGSP model to largely reduce the number of unknown parameters to a constant, so that the MLE can be calculated and thus the density estimation is consistent. This consistency then ensures asymptotic FDR control.

To ensure the consistency of the MLE, we make the following assumptions:

Assumption 5.2.

(i) The signal γ is bandlimited, i.e., for all $(u,\mathbf{s}) \in \mathcal{J}$,

$$\gamma(u,\mathbf{s}) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \xi_{k_1,k_2}\cdot\phi_{k_1}(u)\psi_{k_2}(\mathbf{s}), \tag{5.20}$$

where K_1, K_2 are known positive integers, $\{\phi_k(u)\,:\,k=1,\dots,N\}$ is the graph Fourier basis and $\{\psi_k(\mathbf{s})\,:\,k\in\mathbb{N}\}$ is a set of orthonormal basis of $L^2(\mathcal{T})$. The coefficient matrix $\boldsymbol{\Xi} := (\xi_{k_1,k_2}) \in \mathbb{R}^{K_1\times K_2}$ is a random matrix and takes values in a convex and compact set $\mathcal{K} \subset \mathbb{R}^{K_1\times K_2}$. Under this assumption, we may write $\gamma(u,\mathbf{s})$ as $\gamma(u,\mathbf{s};\boldsymbol{\Xi})$ to highlight the relationship (5.20).

(ii) The function $\phi_{k_1}(u)\psi_{k_2}(\mathbf{s})$ is continuous in $(u,\mathbf{s}) \in \mathcal{J}$ for all $1 \le k_1 \le K_1, 1 \le k_2 \le K_2$.

(iii) For any distinct $\Xi \neq \Xi'$, $f_{\text{mix}}(p \mid \gamma(u, \mathbf{s}; \Xi)) \neq f_{\text{mix}}(p \mid \gamma(u, \mathbf{s}; \Xi'))$ on a set in $(0, 1] \times \mathcal{J}$ with positive measure. We denote the distribution and expectation of $p, (v, \mathbf{t})$ under Ξ by $\mathbb{P}_\Xi$ and $\mathbb{E}_\Xi$.

(iv) For all Ξ,

$$\mathbb{E}_\Xi \left[\ln \left\| \frac{f_{\text{mix}}(p \mid \gamma(v, \mathbf{t}; \Xi'))}{f_{\text{mix}}(p \mid \gamma(v, \mathbf{t}; \Xi))} \right\|_\infty \right] < \infty, \tag{5.21}$$

where the sup norm is taken w.r.t. Ξ', and the expectation is taken w.r.t. $p, (v, \mathbf{t})$.

Under Assumption 5.2, the number of variables to be estimated is $K_1 K_2$, which is independent of M. The MLE can thus be calculated by maximizing the log-likelihood function:

$$\max_{\Xi \in \mathcal{K}} \sum_{m=1}^{M} l(\Xi; p_m, (v_m, \mathbf{t}_m))$$

where

$$l(\Xi; p_m, (v_m, \mathbf{t}_m)) = \ln f_{\text{mix}}(p_m \mid \gamma(v_m, \mathbf{t}_m; \Xi)) + \ln \rho(v_m, \mathbf{t}_m).$$

Note that $\rho(v_m, \mathbf{t}_m)$ does not depend on Ξ, so the MLE $\widehat{\Xi}$ can be obtained by

$$\arg\max_{\Xi \in \mathcal{K}} \sum_{m=1}^{M} \ln f_{\text{mix}}(p_m \mid \gamma(v_m, \mathbf{t}_m; \Xi)). \tag{5.22}$$

By the consistency of MLEs, $\widehat{\Xi}$ converges to Ξ in probability as $M \to \infty$, hence

$$\widehat{\gamma}(u, \mathbf{s}) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \widehat{\xi}_{k_1, k_2} \cdot \phi_{k_1}(u) \psi_{k_2}(\mathbf{s})$$

converges to $\gamma(u, \mathbf{s})$ in probability. We formally state this in the following theorem:

Theorem 5.2. *Under Assumption 5.2, we have* $\widehat{\Xi} \xrightarrow{\text{p}} \Xi$ *as the number of samples* $M \to \infty$ *and*

$$\sup_{(u, \mathbf{s}) \in \mathcal{J}} |\widehat{\gamma}(u, \mathbf{s}) - \gamma(u, \mathbf{s})| \xrightarrow{\text{p}} 0 \tag{5.23}$$

under the probability measure conditional on $\boldsymbol{\Xi}$.

Theorem 5.2 indicates that the MLE of the parameter $\gamma(u, \mathbf{s})$ uniformly converges on $\mathcal{J}$. As we will see in Theorem 5.3, this property ensures the asymptotic control of FDR, hence justifies the usage of MLE in the inference of pdfs.

In practice, we need to choose the hyperparameters K_1 and K_2 to balance the goodness of fit and model complexity. Let $l^*_{K_1,K_2}$ be the optimal value of (5.22). We propose to use the Bayesian information criterion (BIC) for choosing these parameters:

$$\mathrm{BIC} = K_1 K_2 \ln M - 2l^*_{K_1,K_2}.$$

Using Theorem 5.2, we can estimate $f_{\mathrm{mix}}(\cdot \mid \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi}))$ by $f_{\mathrm{mix}}(\cdot \mid \gamma(v_m, \mathbf{t}_m; \widehat{\boldsymbol{\Xi}}))$. This approach is inspired by the work [44], which proposes an EM algorithm to estimate the varying prior probability of null hypotheses and alternative distributions of p-values by the generalized linear model (GLM). This EM algorithm also parametrizes the prior probability of hypotheses being null and the alternative distributions as linear combinations of feature vectors. The main difference here is that we estimate the marginal density $f_{\mathrm{mix}}(p \mid \gamma(u, \mathbf{s}))$ instead of separately estimating $\pi_0 \circ \gamma(u, \mathbf{s})$ and $f_1(p; \gamma(u, \mathbf{s}))$. After estimating $f_{\mathrm{mix}}(p \mid \gamma(u, \mathbf{s}))$, we infer $\pi_0 \circ \gamma(u, \mathbf{s})$ and $f_1(p; \gamma(u, \mathbf{s}))$ using (5.8) and (5.9):

$$\pi_0 \circ \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}}) = \frac{f_{\mathrm{mix}}(1 \mid \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}}))}{f_0(1; (u, \mathbf{s}))}, \tag{5.24}$$

$$f_1(p; \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}})) = \frac{1}{1 - \pi_0 \circ \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}})} (f_{\mathrm{mix}}(p \mid \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}})) - \pi_0 \circ \gamma(u, \mathbf{s}; \widehat{\boldsymbol{\Xi}}) f_0(p; (u, \mathbf{s}))). \tag{5.25}$$

In previous works addressing the MHT problem on graphs, the optimization problems often involve a number of parameters equal to the number of p-values (cf. [31, (5)], [30, (6)]), making it challenging to solve these high-dimensional optimization problems. In contrast, our method only requires $K_1 K_2$ parameters, which do not increase with M and are typically much smaller than M.

Once we have estimated γ, we can then estimate lfdr, $d_{1,M}(\eta)$, $d_{0,M}(\eta)$ and $r_M(\eta)$:

$$\begin{aligned}
\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) &:= \frac{\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})f_0(q;(u,\mathbf{s}))}{f_{\mathrm{mix}}(q\mid\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))},\\
\widehat{d}_{1,M}(\eta) &:= \frac{1}{M}\sum_{m=1}^{M}\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))\le\eta\},\\
\widehat{d'}_{1,M}(\eta) &:= \frac{1}{M}\sum_{m=1}^{M}(1-\theta_m)\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))\le\eta\},\\
\widehat{d}_{0,M}(\eta) &:= \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))\le\eta\},\\
\widehat{r}_M(\eta) &:= \frac{\widehat{d}_{1,M}(\eta)}{\widehat{d}_{0,M}(\eta)}.
\end{aligned}$$

Therefore, by (5.19), we design the rejection threshold z_m such that

$$\mathrm{lfdr}(z_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))=\widehat{\eta}_M,$$

where

$$\widehat{\eta}_M := \sup\{\eta\,:\,\widehat{r}_M(\eta)\le\alpha\}.$$

In the rest of this paper, we write h to denote the thresholding strategy (5.15) with $\eta=\widehat{\eta}_M$. We call this method MHT-GGSP. To achieve asymptotic FDR control, we make the following regularity assumptions:

Assumption 5.3. Assume that the following conditions hold:

(i) $f_1(q;\zeta)$ is strictly decreasing on $q\in(0,1]$ for all ζ.

(ii) $1>\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})>0$ for all $(u,\mathbf{s})\in\mathcal{J}$ and $\boldsymbol{\Xi}\in\mathcal{K}$.

(iii) $f_0(q;u,\mathbf{s})>0$ whenever $q>0$.

(iv) Let $\chi(u,\mathbf{s};\boldsymbol{\Xi}):=\lim_{q\to0^+}\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$. Then $\chi(u,\mathbf{s};\boldsymbol{\Xi})$ is always continuous on $\mathcal{J}$.

(v) There always exists $(u,\mathbf{s})\in\mathcal{J}$ and $q>0$ such that $\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))<\alpha$.

(vi) Rewriting (5.4), with $\zeta = \gamma(u, \mathbf{s})$, as

$$f'_{\text{mix}}(q \mid \zeta, (u, \mathbf{s})) = \pi_0(\zeta) f_0(p; (u, \mathbf{s})) + (1 - \pi_0(\zeta)) f_1(p; \zeta),$$

$$\text{we suppose } \frac{\partial f'_{\text{mix}}}{\partial \zeta}(p \mid \zeta, (u, \mathbf{s})) \text{ is continuous on } (0, 1] \times \mathcal{Z} \times \mathcal{J}.$$

In Assumption 5.3, condition (i) is a slightly stronger condition than condition (ii) in Assumption 5.1. Condition (ii) ensures that the Bayesian model in Section 5.1 is non-trivial, i.e., condition on any γ, θ_m is always random. Condition (iii) states that it is always possible to observe small p-values under null hypotheses, which is the motivation of MHT. Conditions (iv) and (v) assume good identifiability of alternative hypothesis when p is small enough. Combining these two conditions, we know that there exists a non-empty open set in $\mathcal{J}$ such that for any $(u, \mathbf{s})$ in this open set, there exists q such that $\text{lfdr}(q; \gamma(u, \mathbf{s}; \boldsymbol{\Xi})) < \alpha$. Since ρ is a positive measure, This implies that the probability that the assumption in Theorem 5.1 holds tends to 1 as M tends to infinity. These conditions hold true, for example, when $\lim_{q \to 0} f_1(q; (u, \mathbf{s})) = \infty$ and $f_0(q; (u, \mathbf{s}))$ is bounded on $[0, 1]$ for all $(u, \mathbf{s})$, we have $\chi(u, \mathbf{s}; \boldsymbol{\Xi}) = 0$ for all $(u, \mathbf{s})$. Condition (vi) is a smoothness assumption on f'_{mix}.

To state the result on asymptotic FDR control, we define the following quantities:

$$\begin{aligned}
d_0(\eta) &:= \mathbb{P}(\text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m); \boldsymbol{\Xi})) \leq \eta | \boldsymbol{\Xi}), \\
d_1(\eta) &:= \mathbb{E}[(1 - \theta_m) \text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \leq \eta \,|\, \boldsymbol{\Xi}], \\
d_2(\eta) &:= d_0(\eta) - d_1(\eta), \\
r(\eta) &:= \frac{d_1(\eta)}{d_0(\eta)}, \\
\kappa_0 &:= \mathbb{E}[\mathbb{I}\{\theta_m = 0\} \,|\, \boldsymbol{\Xi}].
\end{aligned}$$

Under the aforementioned assumptions, we have the following results. Their proofs are in Section 5.C.

Theorem 5.3. *Under Assumption 5.1, Assumption 5.2, and Assumption 5.3, we have*

$$\overline{\lim_{M \to \infty}} \text{FDR}(h; M) \leq \alpha.$$

Theorem 5.4. *Suppose Assumption 5.1, Assumption 5.2, and Assumption 5.3 hold. Let* $\eta_0 := \sup\{\eta : r(\eta) \leq \alpha\}$. *We have*

$$\mathrm{pow}(h; M) \xrightarrow{\mathrm{p}} \frac{d_2(\eta_0)}{1-\kappa_0} \text{ as } M \to \infty.$$

We remark here that our result is more universal than the existing results for FDR control over graphs. In [32, Proposition 3], the authors prove asymptotic FDR control by their method under the assumption of the alternative distribution of p-values being homogeneous among the vertices. When the homogeneity assumption does not strictly hold, the FDR upper bound shows inflation and relies on the deviation of the proportions of null hypotheses among different vertices [32, Theorem 2]. By making use of the domain information of $\mathcal{J}$, we allow the alternative distribution of p-values to vary in the joint domain while still maintaining FDR control. In [31, Lemma 3], the asymptotic control of FDR simultaneously depends on the norm of the incidence matrix of G and the sparsity level of p-values' weights, which restricts the user's choices of sparsity level, and making the FDR control even not accessible in some cases. In contrast, the asymptotic control of FDR always holds for our approach irrespective of K_1 and K_2.

5.3 Numerical Results

In this section, we compare different MHT methods on sensor network datasets. We compare the following methods:

1. MHT-GGSP (ours): We set $f_{\mathrm{mix}}(p \mid \gamma(u,s)) = \mathrm{sigmoid} \circ \gamma(u,s) p^{\mathrm{sigmoid}\circ\gamma(u,s)-1}$, where $\mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}$ (in the experiment, $\mathcal{T} = [-\pi, \pi]$). We use the graph Laplacian as the GSO, and ψ_k the trigonometric basis of $L^2[-\pi, \pi]$:

$$\{\frac{1}{\sqrt{2\pi}} 1[-\pi,\pi]\} \bigcup \{\frac{1}{\sqrt{\pi}}\sin(ks), \frac{1}{\sqrt{\pi}}\cos(ks) \ : \ k = 1, 2, \ldots\}.$$

2. BH [41, 42]: This method sets adaptive threshold on p-values. Suppose the p-values are ordered as $\{p_{(i)} : i = 1, \ldots, M\}$. This method rejects the p-values

less than or equal to

$$\max\left\{p_{(i)} \,:\, p_{(i)} \leq \frac{i}{M}\alpha\right\}.$$

3. lfdr-sMoM [33]: This method assumes that all the p-values can be divided into several groups such that in each group they follow the same marginal distribution f_{mix} which is a Beta mixture distribution [33, (12)]. The group assignment and the marginal distribution are inferred from the data. We implement this method using the original code [83].

4. Proportion-matching [32]: This method assumes that f_1 is homogeneous over the graph, while the prior probability of being null is different on each vertex. By adjusting the FDR control levels on different vertices, and applying BH method on each vertex, it is expected to match the performance of using global BH method on all p-values from all vertices.

5. FDR-smoothing [30]: This method uses z-values. It assumes that f_1 is homogeneous over the graph, while π_0 is different on each vertex. This method estimates f_1 and π_0 by combining the negative log-likelihood function with the penalty term being the l_1-smoothness of π_0. The graph is constructed by the time-vertex approach, i.e., the product of the graph with the cyclic graph. We implement this method using the original code [84].

6. SABHA [31]: This method first reweighs the p-values and then apply the BH method on the weighted p-values. The weights are understood as π_0, and the reciprocals are assumed to come from a feasible set. In this experiment, the feasible set is the graph signals that have l_1 norm less than a predetermined threshold. The weights are obtained by solving the optimization problem [31, (5)]. The graph is constructed by the time-vertex approach. We implement this method using the original code [85].

7. AdaPT [44]: This method takes an iterative strategy that masks most of the p-values in the beginning. At each iteration it reveals a certain amount of p-values according to a threshold, estimates the FDR and updates the threshold. It stops when the FDR estimate is lower than the nominal FDR level, and rejects the p-values below the threshold. The threshold of the p-values can be updated using any method as long as it decreases with iterations. The

paper [44] proposed an EM algorithm to estimate the lfdr and uses it as the threshold. We use its R package [86] to implement this method.

5.3.1 Signal detection in communication sensor network

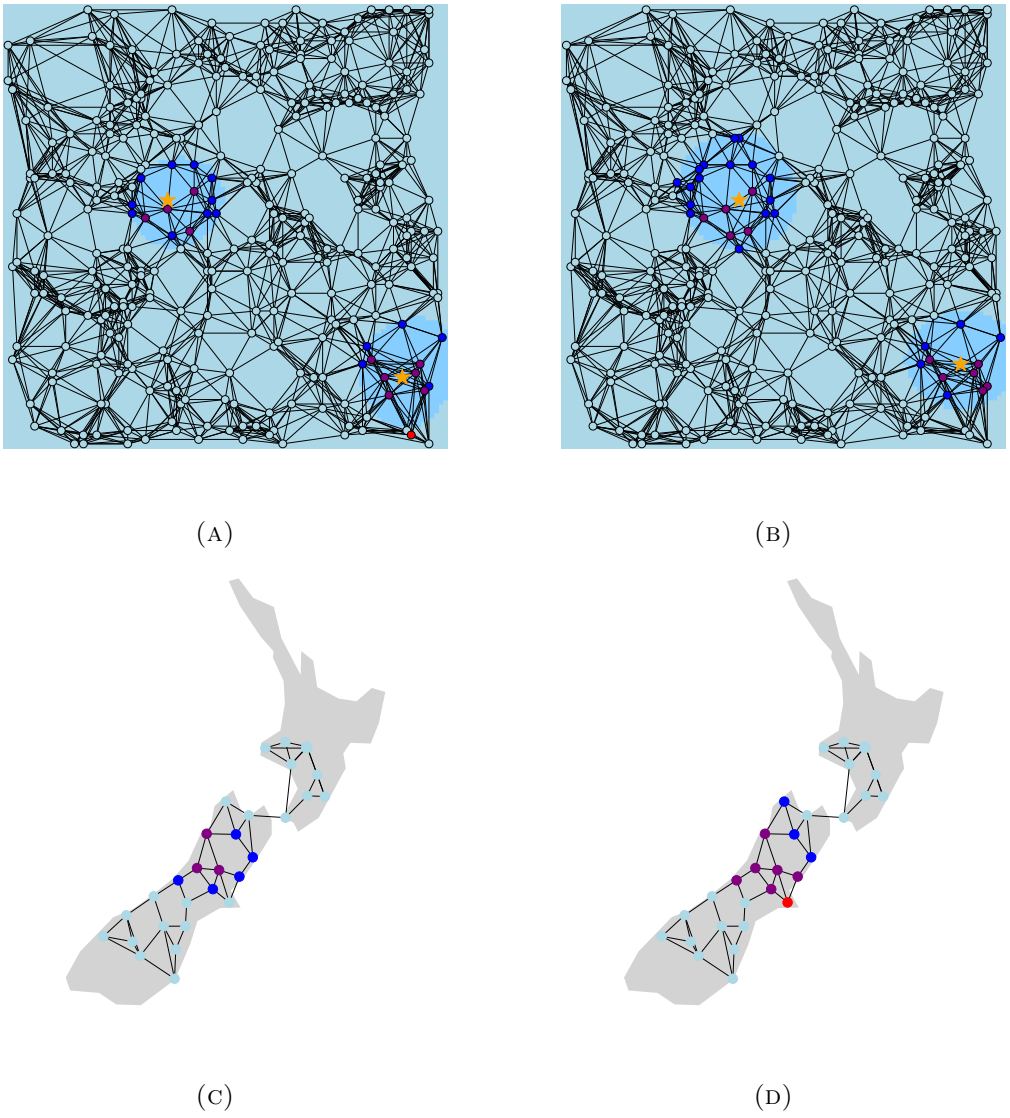
Similar to Example 5.1, we consider a 2D area where two wireless transmitters are performing random walks on a 100×100 grid (cf. Figs. 5.3a and 5.3b). Receivers are randomly placed on 300 points of the grid. We model the receiver sensor network as a 10-NN graph according to their coordinates, and $\mathcal{T} = [-\pi, \pi]$. For each receiver, the received signal is affected by path loss, shadow fading and fast fading. Besides, the observation on each receiver has AWGN. We suppose the sample set $\mathcal{S}$ is given by $\mathcal{V} \times \{-\pi + \frac{j}{T} \cdot 2\pi \ : \ j = 0, \ldots, T\}$. In this experiment, we set $T = 9$. We are interested in determining whether each sensor, at each time instance, has received a signal above the noise floor from at least one transmitter. The data is generated using the source code in [33] modified such that the transmitters are performing random walks, and the noise energy of AWGN is raised from 1 to 1.5, yielding a more challenging task to identify the hypotheses. The proportion of null hypotheses is approximately 10%.

5.3.2 Seismic signal detection in sensor network

We consider the seismic signal detection task in a sensor network (cf . Figs. 5.3c and 5.3d). The seismic event occurred at 17:32:40, 22 December 2020 UTC in New Zealand, with the origin latitude and longitude (-42.14, 171.94). The event ID is "2020p964137" in the GeoNet dataset.[1] We focus on the 32 stations that are either within a radius of 4 from the origin or recorded this event. The sensor network is constructed as a 3-NN graph based on the stations' latitudes and longitudes. We are interested in a time window of 60 seconds before and 60 seconds after the event.

For each second, each station tests the presence of a seismic signal using the z-detector [87]. To compute the signal energy, we first calculate the average of the squared values of the observed waveform in a one-second time window, and then take the logarithm. We denote this quantity as $\log \mathrm{STA}$. To compute the mean and standard deviation of $\log \mathrm{STA}$, we use the period from 7200 seconds to 60 seconds

[1] https://www.geonet.org.nz/data/access/FDSN



(A) (B)

(C) (D)

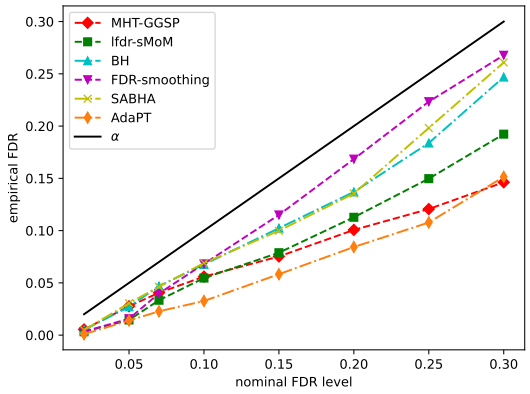
FIGURE 5.3: Example of detection results by MHT-GGSP with nominal FDR level 0.2. Figures 5.3a and 5.3b are detection results on the communication network for two consecutive instances. In the background, light blue denotes the null region, while a deeper color denotes the alternative region. On the graph, light blue represents correctly identified nulls, and purple represents correctly rejected alternatives. Deep blue represents undetected alternatives, and red represents incorrectly rejected nulls. We use orange stars to highlight the transmitters' locations. Figures 5.3c and 5.3d are detection results on the seismic dataset for two consecutive instances.

before the event time as the long-term history. Finally, the z-detector for each second is computed by normalizing $\log \mathrm{STA}$ using the mean and standard deviation. Under the null hypothesis (no seismic signal present), the z-detector is assumed to follow a standard normal distribution. Therefore, a one-sided z-test is conducted to detect the signal.

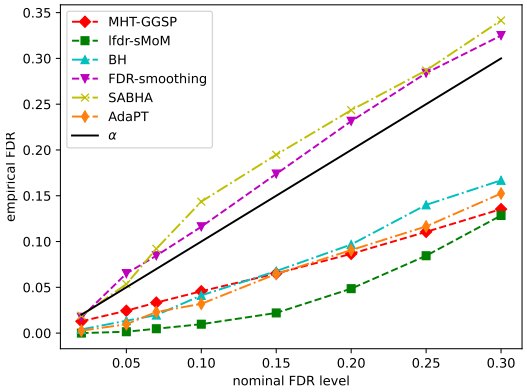
In this experiment, we add AWGN with a noise energy that is 900 times the background noise. The proportion of null hypotheses is 12.8%.

5.3.3 Performance Analysis

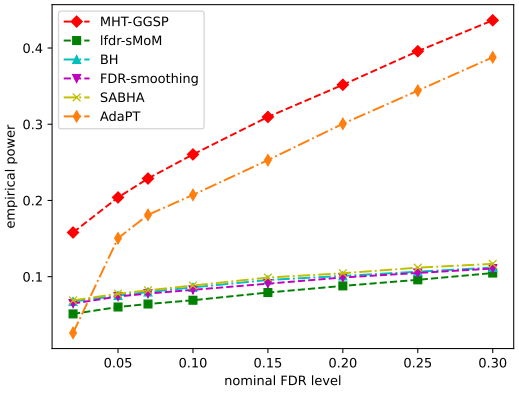
The performance is shown in Fig. 5.4. From the results, we see that the FDR-



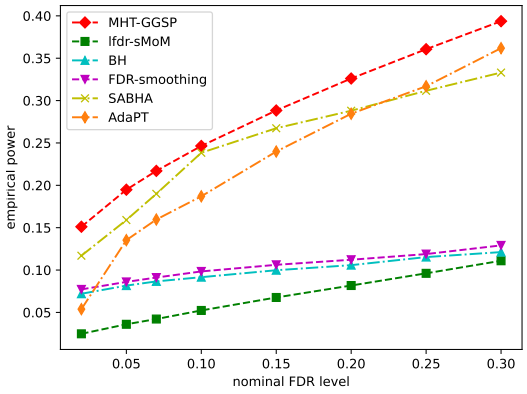
(A) Empirical FDR on communication network data



(B) Empirical FDR on seismic data



(C) Empirical power on communication network data



(D) Empirical power on seismic data

FIGURE 5.4: FDR and detection power under different target FDR levels. Each point is obtained by 20 repetitions.

smoothing and SABHA methods show significant FDR inflation. Besides, the empirical FDR by proportion-matching method is far above the nominal level on both datasets (greater than 0.5 when $\alpha \leq 0.3$), hence we do not show its performance here. For proportion-matching and FDR-smoothing, the FDR inflation is because they assume the same distribution of p-values under the alternative distribution. This assumption does not hold in general since the signal energy is different for different vertex and time, hence $f_1(\cdot;(v_m,t_m))$ will be different for different (v_m,t_m).

Furthermore, note that the proportion-matching method aims to approximate the BH procedure, whose power is not as competitive as most of the baselines. The SABHA method has specific requirements on the Rademacher complexity of the feasible set of reciprocal of the weights, which may not be achievable, thus it may also observe FDR inflation. The methods utilizing the auxiliary information appear to have better power than the classical BH method. Since the signal energy varies smoothly over vertex and time, the MHT-GGSP model is more reasonable and thus observes the best power.

5.A Appendix: Proof of Theorem 5.1

We adopt a similar procedure to the proof of [44, Theorem 2]. Specifically, we first rewrite problem (5.14) to a convex problem, and then utilize Karush–Kuhn–Tucker (KKT) condition to prove the final result. For any $(u, \mathbf{s}) \in \mathcal{J}$, we use $F_0(\cdot; (u, \mathbf{s}))$ and $F_1(\cdot; \gamma(u, \mathbf{s}))$ to denote the cdfs of $f_0(\cdot; (u, \mathbf{s}))$ and $f_1(\cdot; \gamma(u, \mathbf{s}))$, respectively. First we simplify problem (5.14) by introducing the following notations:

$$a_0(\mathbf{z}) := \sum_{m=1}^{M} \mathbb{E}[\mathbb{I}\{p_m \le z_m, \theta_m = 0\} \,|\, \gamma(\mathcal{S})] = \sum_{m=1}^{M} F_0(z_m; (v_m, \mathbf{t}_m))\pi_0 \circ \gamma(v_m, \mathbf{t}_m),$$
$$a_1(\mathbf{z}) := \sum_{m=1}^{M} \mathbb{E}[\mathbb{I}\{p_m \le z_m, \theta_m = 1\} \,|\, \gamma(\mathcal{S})] = \sum_{m=1}^{M} F_1(z_m; \gamma(v_m, \mathbf{t}_m))(1 - \pi_0 \circ \gamma(v_m, \mathbf{t}_m)).$$

Note that the denominator of (5.12) does not depend on h. Therefore, problem (5.14) is equivalent to

$$\begin{aligned} &\max_{\mathbf{s}\in[0,1]^M} a_1(\mathbf{z}) \\ \text{s. t. } &\frac{a_0(\mathbf{z})}{a_0(\mathbf{z}) + a_1(\mathbf{z})} \le \alpha. \end{aligned}$$

This can be further simplified as

$$\begin{aligned} &\min_{\mathbf{s}\in[0,1]^M} -a_1(\mathbf{z}) \\ \text{s. t. } &-\alpha a_1(\mathbf{z}) + (1-\alpha)a_0(\mathbf{z}) \le 0. \end{aligned} \tag{5.26}$$

From the monotonicity of $f_0(\cdot;(v_m,\mathbf{t}_m))$ and $f_1(\cdot;\gamma(v_m,\mathbf{t}_m))$, we know that $-a_1(\mathbf{z})$ and $-\alpha a_1(\mathbf{z})+(1-\alpha)a_0(\mathbf{z})$ are convex in $\mathbf{z}$. Hence (5.26) is a convex optimization problem. Besides, since the feasible region is compact and $F_0(\cdot;(v_m,\mathbf{t}_m))$, $F_1(\cdot;\gamma(v_m,\mathbf{t}_m))$ are continuous, we know that problem (5.26) has a global optimal solution. Next, we verify Slater's condition and then apply KKT condition. Slater's condition requires that there exists a $\mathbf{z}=(z_1,\ldots,z_M)$ such that $z_m\in(0,1)$ for all $m=1,\ldots,M$, and $-\alpha a_1(\mathbf{z})+(1-\alpha)a_0(\mathbf{z})<0$. To find such $\mathbf{z}$, let $g(\mathbf{z}):=-\alpha a_1(\mathbf{z})+(1-\alpha)a_0(\mathbf{z})$. According to the assumption, we suppose that there exists $q_0\in(0,1)$ such that $\mathrm{lfdr}(q_0;\gamma(v_1,\mathbf{t}_1))<\alpha$. We define the constant

$$c_1=-\alpha f_1(q_0;\gamma(v_1,\mathbf{t}_1))(1-\pi_0\circ\gamma(v_1,\mathbf{t}_1))+(1-\alpha)f_0(q_0;(v_1,\mathbf{t}_1))\pi_0\circ\gamma(v_1,\mathbf{t}_1)$$

which is negative by assumption. According to condition (iv) in Assumption 5.1, we suppose $f_0(q;(u,\mathbf{s}))$ is uniformly bounded by c_2 on $[0,1]\times\mathcal{J}$. Given arbitrary $\epsilon>0$, let $\mathbf{z}\in[0,1]^M$ such that:

$$\begin{aligned}0<z_m\le\min(q_0,\frac{\epsilon}{2(M-1)c_2}),\forall\, m=2,\ldots,M,\\ -\frac{\epsilon}{c_1}<z_1\le q_0.\end{aligned}$$

Note that g is differentiable on the interior of $[0,1]^M$. Therefore, using mean value theorem [88, Corollary 10.2.9], we know that there exists a $\mathbf{z}'$ such that

$$\begin{aligned}&g(\mathbf{z})-g(\mathbf{0})\\&=\nabla g(\mathbf{z}')^\intercal\mathbf{z}\\&=(-\alpha f_1(z_1';\gamma(v_1,\mathbf{t}_1))(1-\pi_0\circ\gamma(v_1,\mathbf{t}_1))+(1-\alpha)f_0(z_1';(v_1,\mathbf{t}_1))\pi_0\circ\gamma(v_1,\mathbf{t}_1))z_1+\\&\sum_{m=2}^{M}(-\alpha f_1(z_m';\gamma(v_m,\mathbf{t}_m))(1-\pi_0\circ\gamma(v_m,\mathbf{t}_m))+(1-\alpha)f_0(z_m';(v_m,\mathbf{t}_m))\pi_0\circ\gamma(v_m,\mathbf{t}_m))z_m\\&\le(-\alpha f_1(q_0;\gamma(v_1,\mathbf{t}_1))(1-\pi_0\circ\gamma(v_1,\mathbf{t}_1))+(1-\alpha)f_0(q_0;(v_1,\mathbf{t}_1))\pi_0\circ\gamma(v_1,\mathbf{t}_1))z_1+\\&\sum_{m=2}^{M}(-\alpha f_1(q_0;\gamma(v_m,\mathbf{t}_m))(1-\pi_0\circ\gamma(v_m,\mathbf{t}_m))+(1-\alpha)f_0(q_0;(v_m,\mathbf{t}_m))\pi_0\circ\gamma(v_m,\mathbf{t}_m))z_m\\&\le c_1z_1+\sum_{m=2}^{M}c_2z_m<-\frac{\epsilon}{2}<0.\end{aligned}$$

Here $\mathbf{z}'=\mu'\mathbf{z}$, $\mu'\in(0,1)$. Notice that $g(\mathbf{0})=0$, and $\mathbf{z}$ is in the interior of $[0,1]^M$, hence Slater's condition is satisfied. In this case, $\mathbf{z}^*$ is the optimal solution if and

only if it satisfies the KKT conditions (cf. [89, pp.244]), one of which is that the gradient of Lagrangian $L(\mathbf{z}, \mu) = -a_1(\mathbf{z}) + \mu g(\mathbf{z})$ is zero. Therefore, we have

$$\begin{aligned}\frac{\partial}{\partial z_m} L(\mathbf{z}^*, \mu^*) = &-f_1(z_m^*; \gamma(v_m, \mathbf{t}_m))(1 - \pi_0 \circ \gamma(v_m, \mathbf{t}_m)) \\ &+ \mu(-\alpha f_1(z_m^*; \gamma(v_m, \mathbf{t}_m))(1 - \pi_0 \circ \gamma(v_m, \mathbf{t}_m)) \\ &+ (1-\alpha) f_0(z_m^*; (v_m, \mathbf{t}_m))\pi_0 \circ \gamma(v_m, \mathbf{t}_m)) \\ &= 0.\end{aligned}$$

Rearranging the terms we obtain

$$\mathrm{lfdr}(z_m^*; \gamma(v_m, \mathbf{t}_m)) = \frac{1 + \mu\alpha}{1 + \mu},$$

which completes the proof.

5.B Appendix: Proof of Theorem 5.2

This theorem follows directly from the consistency of MLE. Given $\boldsymbol{\Xi}$, from (5.3), the pdf of $p_m, (v_m, \mathbf{t}_m)$ can be written as

$$\begin{aligned}f_{p,(v,\mathbf{t})}(q, (u, \mathbf{s}) \mid \gamma) = &\pi_0\Big(\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \xi_{k_1,k_2} \cdot \phi_{k_1}(u)\psi_{k_2}(\mathbf{s})\Big) f_0(q; (u, \mathbf{s}))\rho(u, \mathbf{s}) \\ + \Big(1 - \pi_0\Big(\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} &\xi_{k_1,k_2} \cdot \phi_{k_1}(u)\psi_{k_2}(\mathbf{s})\Big)\Big) f_1\Big(q; \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \xi_{k_1,k_2} \cdot \phi_{k_1}(u)\psi_{k_2}(\mathbf{s})\Big)\rho(u, \mathbf{s}).\end{aligned}$$

For every q and $(u, \mathbf{s})$, this pdf is continuous w.r.t. $\boldsymbol{\Xi}$. By condition (iii) in Assumption 5.2, we know that different values of $\boldsymbol{\Xi}$ lead to different joint distributions of $p_m, (v_m, \mathbf{t}_m)$. According to [90, Theorem 9.9] and Lemma 5.5, $\widehat{\boldsymbol{\Xi}} \xrightarrow{\mathrm{p}} \boldsymbol{\Xi}$ given γ. By condition (ii) in Assumption 5.2 and the compactness of $\mathcal{J}$, suppose $\{|\phi_{k_1}(u)\psi_{k_2}(\mathbf{s})|\}$ are uniformly bounded by b. We have

$$\begin{aligned}|\widehat{\gamma}(u, \mathbf{s}) - \gamma(u, \mathbf{s})| &\le \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \left|\widehat{\xi}_{k_1,k_2} - \xi_{k_1,k_2}\right| \cdot |\phi_{k_1}(u)\psi_{k_2}(\mathbf{s})| \\ &\le b \cdot \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \left|\widehat{\xi}_{k_1,k_2} - \xi_{k_1,k_2}\right|.\end{aligned}$$

Note that the right-hand side does not depend on $(u, \mathbf{s})$ and converges to zero in probability given $\boldsymbol{\Xi}$. This concludes the proof.

5.C Appendix: Proof of Theorems 5.3 and 5.4

To prove Theorems 5.3 and 5.4, we first prove the following lemmas. First, Lemma 5.5 allows us to prove results under probability measure conditioned on γ. Second, Lemma 5.6, Lemma 5.7, Lemma 5.8 and Lemma 5.9 indicate that under Assumption 5.1, Assumption 5.2, and Assumption 5.3, the conditions [45, C1-C3] hold. Then Theorems 5.3 and 5.4 can be obtained by using [45, Theorem 2.3, Theorem 3.5] and Lemma 5.5. We denote the set of all sample paths of $\gamma(u, \mathbf{s})$ as Γ. Suppose Γ is a measure space with σ-algebra $\mathcal{F}_\Gamma$, then the random γ induces a probability measure $\mathbb{P}_\gamma$ on $(\Gamma, \mathcal{F}_\Gamma)$.

Lemma 5.5. *Let ι_m represent $(q_m, (u_m, \mathbf{s}_m), \vartheta_m)$ for simplicity. Define $\widetilde{\mathcal{J}}_M := ([0,1] \times \mathcal{J} \times \{0,1\})^M$. Let $\mathcal{F}_M$ denote the σ-algebra of the space $\widetilde{\mathcal{J}}_M \times \Gamma$. Define*

$$\begin{aligned} \mu : \Gamma \times \mathcal{F}_M &\to [0,1] \\ (\gamma, A) &\mapsto \int_{\widetilde{\mathcal{J}}_M} \mathbb{I}\{((\iota_m)_{m=1}^M, \gamma) \in A\} \prod_{m=1}^M f_{p,\theta,(v,\mathbf{t})}(\iota_m \mid \gamma) \, \mathrm{d}\iota_m. \end{aligned}$$

Then $\mu(\gamma(\omega), A)$ is a regular conditional distribution (r.c.d.) of $((p_m, (v_m, \mathbf{t}_m), \theta_m)_{m=1}^M, \gamma)$ given γ [91, Section 4.1.3].

Proof. We prove this lemma by definition of r.c.d. As a stochastic process, γ can be written as $\gamma(\omega, (u, \mathbf{s}))$. In this lemma, we abuse the notation to write $\gamma(\omega, \cdot)$ as $\gamma(\omega)$. First, we show that for each $A \in \mathcal{F}_M$, $\mu(\gamma(\omega), A)$ is a version of $\mathbb{P}\big(((p_m, (v_m, \mathbf{t}_m), \theta_m)_{m=1}^M, \gamma) \in A \,\big|\, \gamma\big)$. Let $F \in \mathcal{F}_\Gamma$. Then we have

$$\begin{aligned} &\mathbb{E}[\mu(\gamma(\omega), A)\mathbb{I}\{\gamma(\omega) \in F\}] \\ &= \int_\Omega \int_{\widetilde{\mathcal{J}}_M} \mathbb{I}\{((\iota_m)_{m=1}^M, \gamma(\omega)) \in A\} \prod_{m=1}^M f_{p,\theta,(v,\mathbf{t})}(\iota_m \mid \gamma(\omega)) \, \mathrm{d}\iota_m \mathbb{I}\{\gamma(\omega) \in F\} \, \mathrm{d}\mathbb{P}(\omega) \\ &= \int_\Gamma \int_{\widetilde{\mathcal{J}}_M} \mathbb{I}\{((\iota_m)_{m=1}^M, \gamma) \in A\} \prod_{m=1}^M f_{p,\theta,(v,\mathbf{t})}(\iota_m \mid \gamma) \, \mathrm{d}\iota_m \mathbb{I}\{\gamma \in F\} \, \mathrm{d}\mathbb{P}_\gamma(\gamma) \\ &= \mathbb{E}\big[\mathbb{I}\{((p_m, (v_m, \mathbf{t}_m), \theta_m)_{m=1}^M, \gamma) \in A\}\mathbb{I}\{\gamma \in F\}\big]. \end{aligned}$$

Second, it can be shown that for each $\gamma \in \Gamma$, $\mu(\gamma, \cdot)$ is a probability measure on $\widetilde{\mathcal{J}}_M \times \Gamma$, hence $\mu(\gamma, A)$ is a r.c.d. □

From Lemma 5.5 we know that when proving results under probability measure conditioned on γ (or $\boldsymbol{\Xi}$), we can regard γ (or $\boldsymbol{\Xi}$) as being fixed.

Lemma 5.6. *Suppose Assumption 5.1, Assumption 5.2, Assumption 5.3 hold. Under the probability measure conditioned on $\boldsymbol{\Xi}$, we have*

$$\begin{aligned} d_{0,M}(\eta) &\xrightarrow{\text{p}} d_0(\eta), \\ d_{1,M}(\eta) &\xrightarrow{\text{p}} d_1(\eta), \\ d'_{1,M}(\eta) &\xrightarrow{\text{p}} d_1(\eta), \end{aligned}$$

where $d_0(\eta)$ and $d_1(\eta)$ are continuous w.r.t. η.

Proof. Note that (5.16) to (5.18) are summations of i.i.d. random variables (conditioned on $\boldsymbol{\Xi}$, same as below), and each random variable takes value in $[0, 1]$, thus has finite expectation and variance. Then by weak law of large numbers (WLLN), we know that $d_{0,M}(\eta) \xrightarrow{\text{p}} \mathbb{E}[\mathbb{I}\{\text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} \,|\, \boldsymbol{\Xi}]$, which equals $d_0(\eta)$. Besides, since $\mathbb{E}[d_{1,M}(\eta) \,|\, \boldsymbol{\Xi}] = \mathbb{E}\big[d'_{1,M}(\eta) \,\big|\, \boldsymbol{\Xi}\big]$, we know that the limits of $d_{1,M}(\eta)$ and $d'_{1,M}(\eta)$ are the same, both equal to

$$\mathbb{E}[(1 - \theta_m)\mathbb{I}\{\text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} \,|\, \boldsymbol{\Xi}],$$

which then equals $d_1(\eta)$.

Next, we prove the continuity of $d_0(\eta)$ and $d_1(\eta)$. For $d_0(\eta)$, note that it is a cdf, so it is right continuous [92, Theorem 1.5.1]. To verify that it is left continuous, we calculate the limit

$$\begin{aligned} &\lim_{\eta' \to \eta^-} |d_0(\eta) - d_0(\eta')| \\ &= \lim_{\eta' \to \eta^-} \mathbb{E}[\mathbb{I}\{\eta' < \text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} \,|\, \boldsymbol{\Xi}] \\ &= \mathbb{E}\left[\lim_{\eta' \to \eta^-} \mathbb{I}\{\eta' < \text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} \,\middle|\, \boldsymbol{\Xi}\right] \\ &= \mathbb{E}[\mathbb{I}\{\text{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) = \eta\} \,|\, \boldsymbol{\Xi}] \\ &= \int_{\mathcal{J}} \int_{(0,1]} \mathbb{I}\{\text{lfdr}(q; \gamma(u, \mathbf{s}; \boldsymbol{\Xi})) = \eta\} f_{\text{mix}}(q \mid \gamma(u, \mathbf{s}; \boldsymbol{\Xi})) \rho(u, \mathbf{s}) \, \mathrm{d}q \, \mathrm{d}(u, \mathbf{s}), \end{aligned}$$

where the second equality is derived by dominated convergence theorem (DCT) [69, Theorem 1.13]. By condition (i) in Assumption 5.3, $\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$ strictly increases in q. Therefore, for every fixed $(u,\mathbf{s})$, $|\{\mathrm{lfdr}(q;\gamma((u,\mathbf{s});\boldsymbol{\Xi}))=\eta\}|\leq 1$. Since $f_{\mathrm{mix}}(q\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$ is continuous in p, we know that the last line in the above equation equals zero, hence $\lim_{\eta'\to\eta^-}|d_0(\eta)-d_0(\eta')|=0$, i.e., $d_0(\eta)$ is continuous on $[0,1]$. The continuity of $d_1(\eta)$ can be proved in a similar way. □

Lemma 5.7. *Let $\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$ be the cdf of $\mathrm{lfdr}(p;\gamma(v,\mathbf{t};\boldsymbol{\Xi}))$ given $(v,\mathbf{t})$ and $\theta=0$. Specifically, it can be calculated as*

$$\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi})):=\int_{(0,1]}\mathbb{I}\{\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\leq\eta\}f_0(q;(u,\mathbf{s}))\,\mathrm{d}q.$$

Then the following quantities are equal:

$$\begin{aligned}
\eta'_{\infty,1}&:=\inf\{\eta\,:\,d_1(\eta)>0\},\\
\eta'_{\infty,2}&:=\inf\{\eta\,:\,\rho(\{(u,\mathbf{s})\,:\,\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))>0\})>0\},\\
\eta'_{\infty,3}&:=\min_{(u,\mathbf{s})\in\mathcal{J}}\chi(u,\mathbf{s};\boldsymbol{\Xi}).
\end{aligned}$$

We denote $\eta'_\infty:=\eta'_{\infty,1}=\eta'_{\infty,2}=\eta'_{\infty,3}$.

Proof. We prove the inequalities $\eta'_{\infty,2}\leq\eta'_{\infty,1}$, $\eta'_{\infty,3}\leq\eta'_{\infty,2}$ and $\eta'_{\infty,1}\leq\eta'_{\infty,3}$. Note that $d_1(\eta)$ can be calculated by

$$\begin{aligned}
d_1(\eta)&=\int_{\mathcal{J}}\int_{(0,1]}\sum_{\vartheta=0}^{1}\mathbb{I}(\mathrm{lfdr}(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\leq\eta,\vartheta=0)f_{p,\theta,(v,\mathbf{t})}(q,\vartheta,(u,\mathbf{s})\mid\gamma)\,\mathrm{d}q\,\mathrm{d}(u,\mathbf{s})\\
&=\int_{\mathcal{J}}\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\rho(u,\mathbf{s})\,\mathrm{d}(u,\mathbf{s}).
\end{aligned}\tag{5.27}$$

For any η such that $d_1(\eta)>0$, by (5.27) and condition (ii) in Assumption 5.3 we know that

$$\rho(\{(u,\mathbf{s})\,:\,\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))>0\})>0,\tag{5.28}$$

hence $\eta'_{\infty,2}\leq\eta'_{\infty,1}$. Next, we want to prove that for any η satisfying (5.28), $\eta\geq\min_{(u,\mathbf{s})\in\mathcal{J}}\chi(u,\mathbf{s};\boldsymbol{\Xi})$. We prove this claim by contradiction. Suppose there exists $\eta'_{\infty,2}\leq\eta<\min_{(u,\mathbf{s})\in\mathcal{J}}\chi(u,\mathbf{s};\boldsymbol{\Xi})$. Then for any $(u,\mathbf{s})\in\mathcal{J}$, η is below the range of

$\mathrm{lfdr}(\cdot;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$, hence $\tilde{F}_0(\eta;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))=0$, which contradicts with (5.28). Hence $\eta \geq \min_{(u,\mathbf{s})\in\mathcal{J}} \chi(u,\mathbf{s};\boldsymbol{\Xi})$, implying that $\eta'_{\infty,3} \leq \eta'_{\infty,2}$.

Finally, we prove $\eta'_{\infty,1} \leq \eta'_{\infty,3}$ by contradiction. Suppose $\eta'_{\infty,3} < \eta'_{\infty,1}$. Then there exists $\eta'_{\infty,4} \in (\eta'_{\infty,3}, \eta'_{\infty,1})$. Since $\eta'_{\infty,4} < \eta'_{\infty,1}$, we know that $d_1(\eta'_{\infty,4}) = 0$. On the other hand, since $\eta'_{\infty,4} > \eta'_{\infty,3}$, and $\chi(u,\mathbf{s};\boldsymbol{\Xi})$ is continuous (cf . condition (iv) in Assumption 5.3), we know that $\{(u,\mathbf{s}) : \chi(u,\mathbf{s};\boldsymbol{\Xi}) < \eta'_{\infty,4}\}$ is a non-empty open set. Since ρ is a positive measure, we know that this set has a positive measure w.r.t. ρ. Besides, note that $\chi(u,\mathbf{s};\boldsymbol{\Xi}) < \eta'_{\infty,4}$ implies $\tilde{F}_0(\eta'_{\infty,4};\gamma(u,\mathbf{s};\boldsymbol{\Xi})) > 0$, therefore

$$\rho(\{(u,\mathbf{s}) : \tilde{F}_0(\eta'_{\infty,4};\gamma(u,\mathbf{s};\boldsymbol{\Xi})) > 0\}) > 0,$$

and

$$d_1(\eta'_{\infty,4}) \geq \int_{\mathcal{J}} \mathbb{I}\{(u,\mathbf{s}) : \tilde{F}_0(\eta'_{\infty,4};\gamma(u,\mathbf{s};\boldsymbol{\Xi})) > 0\}\tilde{F}_0(\eta'_{\infty,4};\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$$
$$\min_{(u',s')\in\mathcal{J}} \pi_0 \circ \gamma(u',s';\boldsymbol{\Xi})\, \mathrm{d}\rho(u,\mathbf{s}) > 0,$$

which leads to a contradiction with $d_1(\eta'_{\infty,4}) = 0$. This completes the proof. □

Lemma 5.8. *Under Assumption 5.1 and Assumption 5.3, there exists $\eta_\infty \in (0,1]$ such that $r(\eta_\infty) < \alpha$.*

Proof. According to Lemma 5.7, for any $\eta > \eta'_\infty$, we have $d_0(\eta) \geq d_1(\eta) > 0$ and

$$\frac{d_1(\eta)}{d_0(\eta)} = \frac{\mathbb{E}[\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi})) \leq \eta\}]}{\mathbb{E}[\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi})) \leq \eta\}]} \leq \eta.$$

Therefore, it suffices to prove that $\eta'_\infty < \alpha$. By Lemma 5.7, $\eta'_\infty = \min_{(u,\mathbf{s})\in\mathcal{J}} \chi(u,\mathbf{s};\boldsymbol{\Xi})$. According to condition (v) in Assumption 5.3, there exists a $(u,\mathbf{s})$ such that $\chi(u,\mathbf{s};\boldsymbol{\Xi}) < \alpha$. This completes the proof. □

Lemma 5.9. *Under Assumption 5.1, Assumption 5.2 and Assumption 5.3, we have*

$$\frac{1}{M}\sum_{m=1}^{M} \left|\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}})) - \mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\right| \xrightarrow{\mathrm{p}} 0,$$

under the probability measure conditioned on $\boldsymbol{\Xi}$.

Proof. First we observe that

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M}\left|\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\widehat{\boldsymbol{\Xi}}))-\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\right| \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\sup_{(u,\mathbf{s})\in\mathcal{J}}\left|\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))-\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|.
\end{aligned}$$

By Markov's inequality, it suffices to prove that the conditional expectation of the right-hand side tends to zero, i.e.,

$$\mathbb{E}\left[\sup_{(u,\mathbf{s})\in\mathcal{J}}\left|\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))-\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|\,\middle|\,\boldsymbol{\Xi}\right]\to 0. \tag{5.29}$$

To this end, we find an upper bound for $\left|\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))-\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|$:

$$\begin{aligned}
&\left|\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))-\mathrm{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right| \\
&=\left|\frac{\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))}-\frac{\pi_0\circ\gamma(u,\mathbf{s},\boldsymbol{\Xi})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))}\right| \\
&\leq\left|\frac{\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))}-\frac{\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))}\right| \\
&+\left|\frac{\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))}-\frac{\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})f_0(p_m;(u,\mathbf{s}))}{f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))}\right|.
\end{aligned}$$

We denote the first and second terms on the right-hand side as T_1 and T_2. In the rest of this proof, we aim to find upper bounds for them. Since f_0 is continuous on $[0,1]\times\mathcal{J}$, we assume that it is upper bounded by b_1. To find a lower bound for $f_{\mathrm{mix}}(p_m\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))$, using conditions (ii) and (iii) in Assumption 5.3, we have

$$\begin{aligned}
f_{\mathrm{mix}}(q\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))&\geq\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})f_0(q;(u,\mathbf{s})) \\
&\geq\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})f_0(\frac{1}{2};(u,\mathbf{s}))>0,\forall\, q>\frac{1}{2}, \\
f_{\mathrm{mix}}(q\mid\gamma(u,\mathbf{s};\boldsymbol{\Xi}))&\geq(1-\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi}))f_1(q;\gamma(u,\mathbf{s};\boldsymbol{\Xi})) \\
&\geq(1-\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi}))f_1(\frac{1}{2};\gamma(u,\mathbf{s};\boldsymbol{\Xi}))>0,\forall\, q\leq\frac{1}{2}.
\end{aligned}$$

Let

$$b_2 := \min\left(\min_{(u,\mathbf{s})\in\mathcal{J}} \pi_0 \circ \gamma(u,\mathbf{s};\boldsymbol{\Xi}) f_0(\frac{1}{2};(u,\mathbf{s})), \min_{(u,\mathbf{s})\in\mathcal{J}} (1-\pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})) f_1(\frac{1}{2};\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right) > 0,$$

then $f_{\text{mix}}(q \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi})) \geq b_2$ for all $q \in (0,1]$ and $(u,\mathbf{s}) \in \mathcal{J}$. Then we have

$$T_1 \leq \min\left(b_1 \left| \frac{1}{f_{\text{mix}}(p_m \mid \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}))} - \frac{1}{f_{\text{mix}}(p_m \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}))} \right|, 1 + \frac{b_1}{b_2}\right) := \min(T_1', b_3),$$

thus for any $\delta \in (0,1)$, we have

$$T_1 \leq T_1' \mathbb{I}\{p_m \geq \delta\} + b_3 \mathbb{I}\{p_m \in (0,\delta)\}.$$

We first consider the event $p_m \geq \delta$. Since $\gamma(u,\mathbf{s};\boldsymbol{\Xi})$ is continuous on $(u,\mathbf{s},\boldsymbol{\Xi})$ and $\mathcal{J}\times\mathcal{K}$ is compact, the image $\gamma(\mathcal{J};\mathcal{K}) := \{\gamma(u,\mathbf{s};\boldsymbol{\Xi}) : (u,\mathbf{s}) \in \mathcal{J}, \boldsymbol{\Xi} \in \mathcal{K}\}$ is compact. According to condition (vi) in Assumption 5.3, $\dfrac{\partial f'_{\text{mix}}}{\partial \zeta}$ is continuous on $[\delta,1]\times\gamma(\mathcal{J};\mathcal{K})\times\mathcal{J}$, so there exists $b_4(\delta) > 0$ so that $\left|\dfrac{\partial f'_{\text{mix}}}{\partial\zeta}(q \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}),(u,\mathbf{s}))\right| \leq b_4(\delta)$ for any $q \in [\delta,1]$, $(u,\mathbf{s}) \in \mathcal{J}$ and $\boldsymbol{\Xi} \in \mathcal{K}$. Then there exists $\boldsymbol{\Xi}' = c\boldsymbol{\Xi} + (1-c)\widehat{\boldsymbol{\Xi}}$, $c \in (0,1)$ such that on the event $p_m \geq \delta$, we have

$$\begin{aligned} T_1' &= b_1 \frac{1}{f_{\text{mix}}(p_m \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}'))^2} \left|\frac{\partial f'_{\text{mix}}}{\partial\zeta}(p_m \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}'),(u,\mathbf{s}))\right| \left|\gamma(u,\mathbf{s};\boldsymbol{\Xi}) - \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})\right| \\ &\leq \frac{b_1}{b_2^2} b_4(\delta) \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\boldsymbol{\Xi}) - \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})\right|. \end{aligned}$$

Hence T_1 can be upper bounded as

$$T_1 \leq \frac{b_1}{b_2^2} b_4(\delta) \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\boldsymbol{\Xi}) - \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})\right| + b_3 \mathbb{I}\{p_m \in (0,\delta)\}. \tag{5.30}$$

Next, we find an upper bound for T_2. Define $b_5 := \min_{(u,\mathbf{s})\in\mathcal{J}} f_0(1;(u,\mathbf{s})) > 0$ and

$$T_2' := \left|f_{\text{mix}}(1 \mid \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) - f_{\text{mix}}(1 \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|.$$

$$\begin{aligned} T_2 &= \frac{f_0(p_m;(u,\mathbf{s}))}{f_{\text{mix}}(p_m \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}))} \left|\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})\right| \\ &\leq \frac{b_1}{b_2} \left|\pi_0\circ\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \pi_0\circ\gamma(u,\mathbf{s};\boldsymbol{\Xi})\right| \end{aligned}$$

$$\leq \frac{b_1}{b_2}\frac{1}{f_0(1;(u,\mathbf{s}))}\left|f_{\text{mix}}(1 \mid \gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) - f_{\text{mix}}(1 \mid \gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|.$$
$$\leq \frac{b_1}{b_2 b_5} T_2'.$$

Using a similar argument as before, there exists $\boldsymbol{\Xi}''$ such that

$$T_2' = \left|\frac{\partial f'_{\text{mix}}}{\partial \zeta}(1 \mid \gamma(u,\mathbf{s},\boldsymbol{\Xi}''),(u,\mathbf{s}))\right| \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right|$$
$$\leq b_4(\delta) \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right|,$$

hence

$$T_2 \leq \frac{b_1 b_4(\delta)}{b_2 b_5} \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right|. \tag{5.31}$$

Combining (5.30) and (5.31) and redefining the constants, we obtain

$$\sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\text{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) - \text{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right|$$
$$\leq b_6(\delta) \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right| + b_7 \mathbb{I}\{p_m \in (0,\delta)\}. \tag{5.32}$$

By Theorem 5.2, $\sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right| \xrightarrow{\text{p}} 0$ (condition on $\boldsymbol{\Xi}$). For any $\epsilon_1, \epsilon_2 > 0$, we first choose $\delta > 0$ such that $\mathbb{P}(p_m \in (0,\delta) \,|\, \boldsymbol{\Xi}) < \frac{\epsilon_2}{2}$. Then when M is large enough, we have

$$\mathbb{P}\left(b_6(\delta) \sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}}) - \gamma(u,\mathbf{s};\boldsymbol{\Xi})\right| \geq \epsilon_1 \,\middle|\, \boldsymbol{\Xi}\right) < \frac{\epsilon_2}{2}.$$

Combining these results, we obtain

$$\mathbb{P}\left(\sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\text{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) - \text{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right| \geq 2\epsilon_1 \,\middle|\, \boldsymbol{\Xi}\right) < \epsilon_2,$$

i.e., $\sup_{(u,\mathbf{s})\in\mathcal{J}} \left|\text{lfdr}(p_m;\gamma(u,\mathbf{s};\widehat{\boldsymbol{\Xi}})) - \text{lfdr}(p_m;\gamma(u,\mathbf{s};\boldsymbol{\Xi}))\right| \xrightarrow{\text{p}} 0$ conditioned on $\boldsymbol{\Xi}$. Note that this is a sequence of uniformly bounded random variables, hence uniformly integrable. Then by applying [90, Theorem 8.16], we obtain (5.29). This completes the proof. □

Proof of Theorem 5.3. We first prove the result under the probability measure conditioned on $\boldsymbol{\Xi}$. The result then follows by taking expectation over $\boldsymbol{\Xi}$ in the last step. According to Lemma 5.5, Lemma 5.6, Lemma 5.8, Lemma 5.9 and [45, Theorem 3.2], we know that[2]

$$\overline{\lim_{M\to\infty}}\,\mathbb{E}\left[\frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\,\middle|\,\boldsymbol{\Xi}\right]\le\alpha. \tag{5.33}$$

The result (5.33) is based on the true values of lfdr. We use this result to provide an upper bound for $\mathrm{FDR}(h;M)=\mathbb{E}\left[\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right]$:

$$\begin{aligned}
&\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}-\frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\\
\le&\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}-\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\\
+&\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}-\frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\\
\le&\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\frac{1}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\left|\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})-\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})\right|\\
+&\frac{1}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\left|\widehat{d}'_{1,M}(\widehat{\eta}_M)-d'_{1,M}(\widehat{\eta}_M)\right|\\
\le&\frac{1}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\\
\cdot&\left(\left|\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})-\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})\right|+\left|\widehat{d}'_{1,M}(\widehat{\eta}_M)-d'_{1,M}(\widehat{\eta}_M)\right|\right)
\end{aligned}$$

When $\widehat{\eta}_M\ge\eta_\infty$, $d_{0,M}(\widehat{\eta}_M)\ge d_{0,M}(\eta_\infty)$. Combining [45, (26)] and [45, Lemma 8.1], we know that $\sup_{\eta\in[0,1]}|d_{0,M}(\eta)-d_0(\eta)|\xrightarrow{\mathrm{p}}0$ and $\sup_{\eta\ge\eta_\infty}\left|d_{0,M}(\eta)-\widehat{d}_{0,M}(\eta)\right|\xrightarrow{\mathrm{p}}0$ (conditioned on $\boldsymbol{\Xi}$, the same below), hence

$$\max(d_{0,M}(\eta_\infty),\frac{1}{M})\xrightarrow{\mathrm{p}}\max(d_0(\eta_\infty),0)=d_0(\eta_\infty)>0.$$

[2]Note that the definition of FDR_m in [45] is different from the definition (5.1) in this chapter.

Since the function $\max(a, \frac{1}{M})$ is Lipschitz continuous in a with Lipschitz constant 1, we have

$$\begin{aligned}\left|\max(d_{0,M}(\widehat{\eta}_M), \frac{1}{M}) - \max(\widehat{d}_{0,M}(\widehat{\eta}_M), \frac{1}{M})\right| &\le \left|d_{0,M}(\widehat{\eta}_M) - \widehat{d}_{0,M}(\widehat{\eta}_M)\right| \\ &\le \sup_{\eta \ge \eta_\infty} \left|d_{0,M}(\eta) - \widehat{d}_{0,M}(\eta)\right| \xrightarrow{\mathrm{p}} 0.\end{aligned}$$

Besides, notice that

$$\begin{aligned}&\left|\widehat{d'}_{1,M}(\eta) - d'_{1,M}(\eta)\right| \\ &\le \frac{1}{M}\sum_{m=1}^{M} \left|\mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} - \mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \widehat{\boldsymbol{\Xi}})) \le \eta\}\right| (1 - \theta_m) \\ &\le \frac{1}{M}\sum_{m=1}^{M} \left|\mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} - \mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \widehat{\boldsymbol{\Xi}})) \le \eta\}\right|.\end{aligned}$$

In the proof of [45, Lemma 8.4], we know that

$$\sup_{\eta \ge \eta_\infty} \frac{1}{M}\sum_{m=1}^{M} \left|\mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} - \mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \widehat{\boldsymbol{\Xi}})) \le \eta\}\right| \xrightarrow{\mathrm{p}} 0.$$

Therefore, $\left|\widehat{d'}_{1,M}(\widehat{\eta}_M) - d'_{1,M}(\widehat{\eta}_M)\right| \le \sup_{\eta \ge \eta_\infty} \left|\widehat{d'}_{1,M}(\eta) - d'_{1,M}(\eta)\right| \xrightarrow{\mathrm{p}} 0$. Combining the above results, we obtain

$$\left|\frac{\widehat{d'}_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M), \frac{1}{M})} - \frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M), \frac{1}{M})}\right| \mathbb{I}\{\widehat{\eta}_M \ge \eta_\infty\} \xrightarrow{\mathrm{p}} 0. \tag{5.34}$$

When $\widehat{\eta}_M < \eta_\infty$,

$$\left|\frac{\widehat{d'}_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M), \frac{1}{M})} - \frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M), \frac{1}{M})}\right| \mathbb{I}\{\widehat{\eta}_M < \eta_\infty\} \le 2\mathbb{I}\{\widehat{\eta}_M < \eta_\infty\}.$$

According to the proof of [45, Theorem 2.3], we know that $\mathbb{P}(\widehat{\eta}_M \ge \eta_\infty \,|\, \boldsymbol{\Xi}) \to 1$, hence the R.H.S. tends to 0 in conditional probability. Combining this with (5.34), we obtain that

$$\left|\frac{\widehat{d'}_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M), \frac{1}{M})} - \frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M), \frac{1}{M})}\right| \xrightarrow{\mathrm{p}} 0.$$

This sequence of random variables is uniformly bounded by 2, hence is uniformly

integrable. By [90, Theorem 8.16] we know that its expectation tends to 0. Finally, by the existing result (5.33), the result follows by

$$\begin{aligned}
&\overline{\lim_{M\to\infty}}\ \mathbb{E}\left[\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\middle|\boldsymbol{\Xi}\right]\\
&\le \overline{\lim_{M\to\infty}}\ \mathbb{E}\left[\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}-\frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\middle|\boldsymbol{\Xi}\right]\\
&+\mathbb{E}\left[\left|\frac{d'_{1,M}(\widehat{\eta}_M)}{\max(d_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\middle|\boldsymbol{\Xi}\right]\\
&\le \alpha.
\end{aligned}$$

Further applying reverse Fatou's lemma, we obtain

$$\overline{\lim_{M\to\infty}}\ \mathbb{E}\left[\mathbb{E}\left[\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\middle|\boldsymbol{\Xi}\right]\right]\le\mathbb{E}\left[\overline{\lim_{M\to\infty}}\ \mathbb{E}\left[\left|\frac{\widehat{d}'_{1,M}(\widehat{\eta}_M)}{\max(\widehat{d}_{0,M}(\widehat{\eta}_M),\frac{1}{M})}\right|\middle|\boldsymbol{\Xi}\right]\right]\le\alpha,$$

which concludes the proof. □

Proof of Theorem 5.4. We first prove the result under the probability measure conditioned on $\boldsymbol{\Xi}$. The result then follows by taking expectation over $\boldsymbol{\Xi}$. We first prove that $r(\eta)$ strictly increases in $\eta\in(\eta'_\infty,1)$. Let $\eta_1,\eta_2\in(\eta'_\infty,1)$ satisfy $\eta_1<\eta_2$. Then

$$\begin{aligned}
r(\eta_2)-r(\eta_1)&=\frac{d_1(\eta_2)}{d_0(\eta_2)}-\frac{d_1(\eta_1)}{d_0(\eta_1)}\\
&=\frac{d_1(\eta_2)d_0(\eta_1)-d_1(\eta_1)d_0(\eta_1)}{d_0(\eta_1)d_0(\eta_2)}+\frac{d_1(\eta_1)d_0(\eta_1)-d_1(\eta_1)d_0(\eta_2)}{d_0(\eta_1)d_0(\eta_2)}.
\end{aligned}$$

The numerator can be rewritten as

$$\begin{aligned}
&d_1(\eta_2)d_0(\eta_1)-d_1(\eta_1)d_0(\eta_1)+d_1(\eta_1)d_0(\eta_1)-d_1(\eta_1)d_0(\eta_2)\\
&=d_0(\eta_1)(d_1(\eta_2)-d_1(\eta_1))-d_1(\eta_1)(d_0(\eta_2)-d_0(\eta_1))\\
&=\mathbb{E}[\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_1\}\,|\,\boldsymbol{\Xi}]\\
&\cdot\mathbb{E}[\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\mathbb{I}\{\eta_1<\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_2\}\,|\,\boldsymbol{\Xi}]\\
&-\mathbb{E}[\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_1\}\,|\,\boldsymbol{\Xi}]\\
&\cdot\mathbb{E}[\mathbb{I}\{\eta_1<\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_2\}\,|\,\boldsymbol{\Xi}]\\
&>\eta_1\mathbb{E}[\mathbb{I}\{\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_1\}\,|\,\boldsymbol{\Xi}]\mathbb{E}[\mathbb{I}\{\eta_1<\mathrm{lfdr}(p_m;\gamma(v_m,\mathbf{t}_m;\boldsymbol{\Xi}))\le\eta_2\}\,|\,\boldsymbol{\Xi}]
\end{aligned}$$

$$- \eta_1 \mathbb{E}[\mathbb{I}\{\mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta_1\} \,|\, \boldsymbol{\Xi}] \mathbb{E}[\mathbb{I}\{\eta_1 < \mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta_2\} \,|\, \boldsymbol{\Xi}] = 0.$$

Since $\eta_1, \eta_2 \in (\eta'_\infty, 1)$, $d_0(\eta_2) \ge d_0(\eta_1) > 0$, the denominator $d_0(\eta_1)d_0(\eta_2)$ is positive, i.e, $r(\eta)$ strictly increases in $\eta \in (\eta'_\infty, 1)$. Therefore, due to the continuity of d_1 and d_0, we know that for any small $\epsilon > 0$, $R(\eta_0 - \epsilon) < \alpha$ holds. By WLLN, we know that $\frac{1}{M}\sum_{m=1}^{M}(1-\theta_m) \xrightarrow{\mathrm{p}} \kappa_0$, and $\frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\{\theta_m = 1, \mathrm{lfdr}(p_m; \gamma(v_m, \mathbf{t}_m; \boldsymbol{\Xi})) \le \eta\} \xrightarrow{\mathrm{p}} d_2(\eta)$, both conditioned on $\boldsymbol{\Xi}$. Hence by applying [45, Theorem 3.5], we obtain the desired result when conditioned on $\boldsymbol{\Xi}$. The result under unconditional probability measure can be obtained by taking expectation over $\boldsymbol{\Xi}$. □

Chapter 6

Conclusion and Future Work

In this thesis, we have established a statistical model for stochastic generalized graph signal, and derived estimation and testing techniques for it.

In Chapter 3, we have introduced the concepts of a GRP and JWSS processes. These concepts generalize the existing stationarity models for GSP to the broader GGSP setting. Concrete cases such as the time-vertex model, continuous-time graph signal, and multichannel graph signal are encompassed in this framework. The stationarity models in the graph and Hilbert space domains are related in the sense that JWSS implies wide-sense stationarity in the respective domains. The explicit and approximate forms of Wiener filters for denoising and signal completion are also derived. The implementation of the GRP framework is illustrated via several numerical experiments. In the case of a finite-dimensional feature space, the shift operator in the Hilbert space can be learned from data without prior knowledge. When dealing with finite samples of GRPs taking values in a Hilbert space that may be infinite-dimensional, a variational EM algorithm is proposed to simultaneously estimate the PSD and noise energy, allowing the recovery of continuous signals with finite and noisy observations.

An alternative approach to inference for graph signals is machine learning data-driven methods like graph neural networks (GNNs). For example, graph auto-encoder (GAE) [93] has been used to denoise graph signals. In general, the training complexity of GRP is lower than GNN and a GRP model has fewer parameters than a GNN. For GRP, it suffices to determine the filter coefficients and bandwidth, while a GNN's weights depend on the number of layers, and the number of channels in

different layers. In addition, the Wiener filter for GRP has an explicit solution (3.15), while a GNN depends on training data to optimize its weights. The Wiener filters we have proposed are linear transformations, while for a GNN, the test complexity depends on the number of hidden layers, feature dimension and connection settings. Finally, it is easier to interpret GRP as it is a statistical model. On the other hand, GNNs can learn nonlinear relationships and do not require the assumption of JWSS. The pros and cons of statistical parametric models versus data-driven models are expounded in [94, 95]. Comparison of GRP and GNN in different applications is an interesting future research direction.

In Chapter 4, we have devised a signal reconstruction approach for GGSP, yielding a predictor that can be computed in a distributed fashion. We interpreted this approach from both deterministic and Bayesian aspects and cast it as an extension of existing frameworks. In the former case where the signal is a deterministic function, we showed that the approach imposes smoothness on the reconstructed signal. In the latter case, the signal is regarded as a GP, and we analyzed its moments. By utilizing RFF, the reconstruction approach can be implemented online, and the evaluation is still distributed.

We provided statistical analysis on the predictor. Under the uniform exclusive sampling scheme, we derived the limit of the posterior variance and provided a numerically computable upper bound for it. We verified the KRR-GGSP approach by numerical experiments. By testing KRR-GGSP against existing methods on real datasets, we validated that introducing the graph structure and the product kernel improves reconstruction performance.

In Chapter 5, we have proposed a novel method for conducting network multiple hypothesis tests on the joint domain of the vertex set and the measure space of each vertex signal. Our approach models the distributions of p-values over the joint domain as parametrized by a random generalized graph signal. This allows for the possibility of inhomogeneity over the joint domain. By utilizing this estimator in conjunction with the lfdr-based approach, we are able to control the FDR at a specified nominal level in the asymptotic setting. This provides a powerful tool for accurately identifying significant hypotheses in complex datasets with an underlying graph structure. This approach has the potential to be extended for online and distributed implementation over the underlying network.

In this thesis, we have studied statistical signal processing model and techniques on a static graph structure with a dynamic signal. The first possible future research topic is to study the estimation, hypothesis testing and change-point detection approaches on dynamic graph structures. In practice the dynamic graph structure is usually not provided. Therefore, a reliable approach that jointly estimates the signal and dynamic graph is expected. Given the estimated signal and graphs, the change-point detection approach should utilize both signal and graph structures. This study may be useful for processing dynamic point clouds and biomedical signals. Second, as discussed in Chapter 5, we have imposed certain conditions on the p-value distribution. If these conditions are not satisfied, or deviate significantly from the ground truth, the proposed approach may face challenges in controlling the FDR. Therefore, future work could study the MHT problem on $\mathcal{V} \times \mathcal{T}$ by relaxing these assumptions on the p-value distribution. This will lead to more flexible and universal detection approaches. Additionally, it is also important to alleviate the computational cost when solving the MLE problem. An online MHT method on graph will be desired when the hypotheses are tested in a streaming way, which is usually the case. Finally, it is valuable to model and analyze the inputs, outputs and intermediate features produced by a GNN, as they are also graph signals. Due to the randomness of these signals, it is essential to develop uncertainty quantification methods for them. For example, by incorporating the uncertainty of model prediction error, we may obtain a confidence set that contains the ground truth with a high probability. Specifically, studying the dynamic cases with an expanding graph and an online GNN will be of practical interest, as it closely resemble the real-world cases such as growing users and commodities in a commercial network.

List of Author's Publications[1]

Journal Articles

- **X. Jian**, W. P. Tay, and Y. C. Eldar, "Kernel based reconstruction for generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 72, pp. 2308–2322, Apr. 2024.
- **X. Jian** and W. P. Tay, "Wide-sense stationarity in generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 70, pp. 3414–3428, 2022.

Conference Proceedings

- P. Zhang, **X. Jian**, F. Ji, W. P. Tay and B. Wen, "Spectral Convergence of Simplicial Complex Signals," in *Proc. IEEE Int. Symp. on Inform. Theory*, Athens, Greece, Jul. 2024.
- S. Wang, R. She, Q. Kang, **X. Jian**, K. Zhao, Y. Song, and W. P. Tay, "DistilVPR: Cross-Modal knowledge distillation for visual place recognition," in *Proc. AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024.
- R. She, S. Wang, Q. Kang, K. Zhao, Y. Song, W. P. Tay, T. Geng, **X. Jian**, "PosDiffNet: Positional Neural Diffusion for Point Cloud Registration in a Large Field of View with Perturbations," in *Proc. AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024.
- **X. Jian** and W. P. Tay, "Kernel Ridge Regression for Generalized Graph Signal Processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Rhodes Island, Greece, 2023.

[1]The superscript * indicates joint first authors

- **X. Jian** and W. P. Tay, "Wide-Sense Stationarity and Spectral Estimation for Generalized Graph Signal," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Singapore, May. 2022.

Book Chapters/Monographs

- **X. Jian**, F. Ji, and W. P. Tay, "Generalizing graph signal processing: High dimensional spaces, models and structures," *Foundations and Trends in Signal Processing*, vol. 17, no. 3, pp. 209–290, 2023.

Preprints

- **X. Jian**, M. Gölz, F. Ji, W. P. Tay, and A. M. Zoubir, "A Graph Signal Processing Perspective of Network Multiple Hypothesis Testing with False Discovery Rate Control," *arXiv preprint arXiv:2408.03142*, 2024.

Bibliography

[1] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013. 1, 6, 14, 55

[2] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[3] G. Leus, A. G. Marques, J. M. Moura, A. Ortega, and D. I. Shuman, "Graph signal processing: History, development, impact, and outlook," *IEEE Signal Process. Mag.*, vol. 40, no. 4, pp. 49–60, Jun. 2023.

[4] X. Jian, F. Ji, and W. P. Tay, "Generalizing graph signal processing: High dimensional spaces, models and structures," *Foundations and Trends in Signal Processing*, vol. 17, no. 3, pp. 209–290, 2023. 1

[5] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, Sep. 2017. 1

[6] E. Pavez, B. Girault, A. Ortega, and P. A. Chou, "Two channel filter banks on arbitrary graphs with positive semi definite variation operators," *IEEE Trans. Signal Process.*, vol. 71, pp. 917–932, Mar. 2023.

[7] E. Isufi, F. Gama, D. I. Shuman, and S. Segarra, "Graph filters for signal processing and machine learning on graphs," *IEEE Trans. Signal Process.*, pp. 1–32, 2024. 1

[8] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3775–3789, Mar. 2016. 1

[9] Y. Tanaka, Y. C. Eldar, A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, Oct. 2020. 1

[10] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 260–291, Mar. 2016. 1

[11] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 764–778, 2017. 1, 8, 75, 76

[12] A. Kroizer, T. Routtenberg, and Y. C. Eldar, "Bayesian estimation of graph signals," *IEEE Trans. Signal Process.*, vol. 70, no. 5, pp. 2207–2223, Mar. 2022. 1

[13] W. Huang, T. A. Bolton, J. D. Medaglia, D. S. Bassett, A. Ribeiro, and D. Van De Ville, "A graph signal processing perspective on functional brain imaging," *Proc. IEEE*, vol. 106, no. 5, pp. 868–885, Mar. 2018. 1

[14] J. D. Medaglia, W. Huang, E. A. Karuza, A. Kelkar, S. L. Thompson-Schill, A. Ribeiro, and D. S. Bassett, "Functional alignment with anatomical networks is associated with cognitive flexibility," *Nature human behaviour*, vol. 2, no. 2, pp. 156–164, 2018.

[15] W. Huang, L. Goldsberry, N. F. Wymbs, S. T. Grafton, D. S. Bassett, and A. Ribeiro, "Graph frequency analysis of brain signals," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 7, pp. 1189–1203, Aug. 2016. 1

[16] G. Cheung, E. Magli, Y. Tanaka, and M. K. Ng, "Graph spectral image processing," *Proc. IEEE*, vol. 106, no. 5, pp. 907–930, May 2018. 1

[17] A. C. Yağan and M. T. Özgen, "Spectral graph based vertex-frequency wiener filtering for image and graph signal denoising," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 226–240, 2020. 1

[18] W. Huang, A. G. Marques, and A. R. Ribeiro, "Rating prediction via graph signal processing," *IEEE Trans. Signal Process.*, vol. 66, no. 19, pp. 5066–5081, 2018. 1

[19] F. Grassi, A. Loukas, N. Perraudin, and B. Ricaud, "A time-vertex signal processing framework: Scalable processing and meaningful representations for time-series on graphs," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 817–829, Nov. 2018. 1, 7, 43

[20] A. Loukas and D. Foucard, "Frequency analysis of time-varying graph signals," in *Proc. IEEE Global Conf. on Signal and Information Processing*, Washington, DC, USA, Dec. 2016. 1

[21] J. Yu, X. Xie, H. Feng, and B. Hu, "On critical sampling of time-vertex graph signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Ottawa, Canada, Nov. 2019. 1

[22] F. Ji and W. P. Tay, "A Hilbert space theory of generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6188–6203, Nov. 2019. 1, 3, 4, 7, 8, 19, 37, 38, 78, 79

[23] B. Girault, "Stationary graph signals using an isometric graph translation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Nice, France, Aug./Sep. 2015. 3, 7, 8, 41

[24] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 22, pp. 5911–5926, Aug. 2017. 41, 42, 53

[25] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, Jul. 2017. 3, 7, 41, 48

[26] N. Perraudin, A. Loukas, F. Grassi, and P. Vandergheynst, "Towards stationary time-vertex signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, New Orleans, US, Mar. 2017. 3, 7, 45, 48

[27] A. Loukas and N. Perraudin, "Stationary time-vertex signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, no. 1, pp. 1–19, Aug. 2019. 3, 4, 7, 8, 43, 45, 48, 53, 55, 57

[28] X. Jian and W. P. Tay, "Wide-sense stationarity in generalized graph signal processing," *IEEE Trans. Signal Process.*, vol. 70, pp. 3414–3428, 2022. 4

[29] J. Hara, Y. Tanaka, and Y. C. Eldar, "Graph signal sampling under stochastic priors," *IEEE Trans. Signal Process.*, vol. 71, pp. 1421–1434, Apr. 2023. 4

[30] R. A. P. Wesley Tansey, Oluwasanmi Koyejo and J. G. Scott, "False discovery rate smoothing," *J. Amer. Statist. Assoc.*, vol. 113, no. 523, pp. 1156–1171, 2018. 5, 6, 9, 108, 113, 119, 123

[31] A. Li and R. F. Barber, "Multiple testing with the structure-adaptive benjamini–hochberg algorithm," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 81, no. 1, pp. 45–74, Feb. 2019. 5, 6, 9, 113, 119, 122, 123

[32] M. Pournaderi and Y. Xiang, "On large-scale multiple testing over networks: An asymptotic approach," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 9, pp. 442–457, 2023. 6, 9, 113, 122, 123

[33] M. Gölz, A. M. Zoubir, and V. Koivunen, "Multiple hypothesis testing framework for spatial signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 771–787, 2022. 6, 9, 113, 123, 124

[34] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013. 6, 15, 41, 42

[35] ——, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014. 6, 15

[36] F. Ji and W. P. Tay, "Generalized graph signal processing," in *Proc. IEEE Global Conf. on Signal and Information Processing*, Anaheim, USA, Nov. 2018. 7

[37] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, Aug. 2015. 8

[38] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4845–4860, May 2016. 8

[39] K. Qiu, X. Mao, X. Shen, X. Wang, T. Li, and Y. Gu, "Time-varying graph signal reconstruction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 6, pp. 870–883, Sep. 2017. 8, 74, 95

[40] J. H. Giraldo, A. Mahmood, B. Garcia-Garcia, D. Thanou, and T. Bouwmans, "Reconstruction of time-varying graph signals via Sobolev smoothness," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 201–214, 2022. 8, 74, 82

[41] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press, 2010. 8, 108, 111, 122

[42] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. 8, 122

[43] J. D. Storey, "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 64, no. 3, pp. 479–498, Aug. 2002. 9

[44] L. Lei and W. Fithian, "AdaPT: an interactive procedure for multiple testing with side information," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 80, no. 4, pp. 649–679, 2018. 9, 111, 114, 115, 116, 119, 123, 124, 127

[45] H. Cao, J. Chen, and X. Zhang, "Optimal false discovery rate control for large scale multiple testing with auxiliary information," *Annals of Statistics*, vol. 50, no. 2, pp. 807 – 857, 2022. 9, 110, 111, 114, 116, 130, 137, 138, 140

[46] G. Strang, *Linear Algebra and its Applications*, 3rd ed. Thomson Learning, 1988. 17

[47] L. Debnath and P. Mikusinski, *Introduction to Hilbert Spaces with Applications*, 3rd ed. London, UK: Elsevier Academic Press, 2000. 18, 20, 22, 24

[48] T. Hungerford, *Algebra (Graduate Texts in Mathematics) (v. 73)*, 8th ed. Springer, 2002. 18

[49] T. Hsing and R. Eubank, *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*. John Wiley & Sons, 2015. 20, 24, 25, 26, 38, 65

[50] C. R. Baker, "Joint measures and cross-covariance operators," *Transactions of the American Mathematical Society*, vol. 186, pp. 273–289, Dec. 1973.

[51] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan, *Probability distributions on Banach spaces*. Springer Science & Business Media, 1987. 20, 27, 28, 38

[52] K. Yosida, *Functional Analysis*, 6th ed. Berlin Heidelberg: Springer-Verlag, 1980. 23

[53] J. Mikusiński, *The Bochner Integral*. Basel, Switzerland: Springer, 1978.

[54] P. Mikusiński, "Integrals with values in Banach spaces and locally convex spaces," *arXiv preprint arXiv:1403.5209*, 2014. 23

[55] I. Klebanov, B. Sprungk, and T. Sullivan, "The linear conditional expectation in Hilbert space," *Bernoulli*, vol. 27, no. 4, pp. 2267 – 2299, Nov. 2021. 27, 30, 49, 101

[56] B. S. Rajput and S. Cambanis, "Gaussian processes and Gaussian measures," *The Annals of Mathematical Statistics*, vol. 43, no. 6, pp. 1944 – 1952, 1972. 28

[57] I. Steinwart and C. Scovel, "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs," *Constructive Approximation*, vol. 35, pp. 363–417, 2012. 29, 35, 77

[58] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. New York, US: Springer Science & Business Media, 2011. 32, 33, 75

[59] B. Scholkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2002. 32

[60] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *J. Machine Learning Research*, vol. 7, no. 95, pp. 2651–2667, 2006. 33

[61] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, US: Springer, 2009. 33

[62] E. De Vito, V. Umanità, and S. Villa, "An extension of Mercer theorem to matrix-valued measurable kernels," *Applied and Computational Harmonic Analysis*, vol. 34, no. 3, pp. 339–351, May 2013. 37, 67, 68

[63] N. N. Vakhania and N. P. Kandelaki, "Random vectors with values in complex Hilbert spaces," *Theory Prob. and its Applications*, vol. 41, no. 1, pp. 116—-131, Feb. 1995. 38

[64] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer, 1998. 40

[65] B. Girault, P. Gonçalves, and E. Fleury, "Translation on graphs: An isometric shift operator," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2416–2420, Oct. 2015. 41

[66] A. Sandryhaila and J. M. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Aug. 2014. 41

[67] R. Hammack, W. Imrich, and S. Klavžar, *Handbook of Product Graphs*, 2nd ed. Boca Raton, US: CRC Press, 2011. 42

[68] S. Barik, R. B. Bapat, and S. Pati, "On the Laplacian spectra of product graphs," *Applicable Analysis and Discrete Mathematics*, vol. 9, no. 1, pp. 39–58, Apr. 2015. 42

[69] E. M. Stein and R. Shakarchi, *Real Analysis*. Princeton, US: Princeton, 2005. 43, 132

[70] M. A. Kramer, E. D. Kolaczyk, and H. E. Kirsch, "Emergent network topology at seizure onset in humans," *Epilepsy Research*, vol. 79, no. 2, pp. 173–186, May 2008. 54, 91

[71] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Dec. 2008. 61

[72] G. Wahba, "Convergence rates of certain approximate solutions to Fredholm integral equations of the first kind," *Journal of Approximation Theory*, vol. 7, no. 2, pp. 167–185, 1973. 78

[73] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005. 79

[74] X. Jian and W. P. Tay, "Kernel ridge regression for generalized graph signal processing," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Rhodes Island, Greece, Jun. 2023. 79, 80

[75] A. Venkitaraman, S. Chatterjee, and P. Händel, "Gaussian processes over graphs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020. 81

[76] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., vol. 20. Curran Associates, Inc., 2007. 84

[77] Z. Li, J.-F. Ton, D. Oglic, and D. Sejdinovic, "Towards a unified analysis of random Fourier features," *J. Machine Learning Research*, vol. 22, no. 1, p. 4887–4937, Jul. 2021. 85

[78] G. Garrigos and R. M. Gower, "Handbook of convergence theorems for (stochastic) gradient methods," *arXiv preprint arXiv:2301.11235*, 2023. 86

[79] P. Koepernik and F. Pfaff, "Consistency of Gaussian process regression in metric spaces," *J. Machine Learning Research*, vol. 22, no. 244, pp. 1–27, 2021. 86, 88, 90, 101, 102

[80] A. Cini and I. Marisca, "Torch Spatiotemporal," Mar. 2022. [Online]. Available: https://github.com/TorchSpatiotemporal/tsl 90

[81] X. Cai and G. Giannakis, "A two-dimensional channel simulation model for shadowing processes," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6, pp. 1558–1567, Nov. 2003. 112

[82] W. C. Jakes and D. C. Cox, *Microwave Mobile Communications*. Wiley-IEEE Press, 1994. 112

[83] https://github.com/mgoelz95/lfdr-sMoM. 123

[84] https://github.com/tansey/smoothfdr. 123

[85] https://rinafb.github.io/research/. 123

[86] https://cran.r-project.org/web/packages/adaptMT/index.html. 124

[87] J. Berger and R. L. Sax, "Seismic detectors: the state of the art," SSR-R, Technical Report 80-4588, 1980. 124

[88] T. Tao, *Analysis I*, 3rd ed. Springer, 2016. 128

[89] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004. 129

[90] R. W. Keener, *Theoretical Statistics: Topics for a Core Course*. New York: Springer, 2010. 129, 136, 139

[91] R. Durrett, *Probability: Theory and Examples*, 5th ed. Cambridge University Press, 2019. 130

[92] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*, 8th ed. Boston: Pearson, 2019. 131

[93] T. H. Do, D. Minh Nguyen, and N. Deligiannis, "Graph auto-encoder for graph signal denoising," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020. 141

[94] L. Breiman, "Statistical modeling: the two cultures (with comments and a rejoinder by the author)," *Statistical Science*, vol. 16, no. 3, pp. 199 – 231, 2001. 142

[95] K. M. Bzdok D, Altman N, "Statistics versus machine learning," *Nature methods*, vol. 15, Apr. 2018. 142