Journal of Machine Learning Research 17 (2016) 1-27 Submitted 4/16; Revised 9/16; Published 10/16

Bayesian Graphical Models for Multivariate Functional Data

Hongxiao Zhu HONGXIAO@VT.EDU
Department of Statistics
Virginia Tech, 250 Drillfield Drive (MC 0439)
Blacksburg, VA 24061, USA

Nate Strawn NATE.STRAWN@GEORGETOWN.EDU
Department of Mathematics and Statistics
Georgetown University
Washington D.C. 20057, USA

David B. Dunson DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham NC 27708, USA

Editor: Jie Peng

Abstract

Graphical models express conditional independence relationships among variables. Although methods for vector-valued data are well established, functional data graphical models remain underdeveloped. By functional data, we refer to data that are realizations of random functions varying over a continuum (e.g., images, signals). We introduce a notion of conditional independence between random functions, and construct a framework for Bayesian inference of undirected, decomposable graphs in the multivariate functional data context. This framework is based on extending Markov distributions and hyper Markov laws from random variables to random processes, providing a principled alternative to naive application of multivariate methods to discretized functional data. Markov properties facilitate the composition of likelihoods and priors according to the decomposition of a graph. Our focus is on Gaussian process graphical models using orthogonal basis expansions. We propose a hyper-inverse-Wishart-process prior for the covariance kernels of the infinite coefficient sequences of the basis expansion, and establish its existence and uniqueness. We also prove the strong hyper Markov property and the conjugacy of this prior under a finite rank condition of the prior kernel parameter. Stochastic search Markov chain Monte Carlo algorithms are developed for posterior inference, assessed through simulations, and applied to a study of brain activity and alcoholism.

Keywords: graphical model, functional data analysis, gaussian process, model uncertainty, stochastic search

1. Introduction

Graphical models provide a powerful tool for describing conditional independence structures between random variables. In the multivariate data case, Dawid and Lauritzen (1993) defined Markov distributions (distributions with Markov property over a graph) of random vectors which can be factorized according to the structure of a graph. They also introduced

hyper-Markov laws serving as prior distributions in Bayesian analysis. The special case of Gaussian graphical models, in which a multivariate Gaussian distribution is assumed and the graph structure corresponds to the zero pattern of the precision matrix (Dempster, 1972; Lauritzen, 1996), is well studied. Computational algorithms, such as Markov chain Monte Carlo (MCMC) and stochastic search, are developed to estimate the graph based on the conjugate hyper-inverse-Wishart prior and its extensions (Giudici and Green, 1999; Roverato, 2002; Jones et al., 2005; Scott and Carvalho, 2008; Carvalho and Scott, 2009).

In the frequentist literature, notable works on graphical models include the graphical LASSO (Yuan and Lin, 2007; Friedman et al., 2008; Mazumder and Hastie, 2012a,b) and the neighborhood selection approach (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010). The graphical LASSO induces sparse estimation of the precision matrix of the Gaussian likelihood through l_1 regularization. The neighborhood selection approach relies on estimating the neighborhood of each node separately by regressing each variable on all the remaining variables, sparsifying with l_1 regularization, and then stitching the neighborhoods together to form the global graph estimate. Various extensions, computational methods, and theoretical properties have been developed in these frameworks (Lam and Fan, 2009; Höfling and Tibshirani, 2009; Cai et al., 2011; Witten et al., 2011; Yang et al., 2012; Mazumder and Hastie, 2012a,b; Anandkumar et al., 2012; Loh and Wainwright, 2013).

The graphical modeling literature focuses primarily on vector-valued data with each node corresponding to one variable. Many applications, however, involve *functional data*—data that are realizations of random functions varying over a continuum such as a time interval or a spatial domain. Common types of functional data include signals, images, and many emerging high-throughput digital measurements. The dependence structure of functional data is of interest in a wide range of applications. For example, in neuroimaging, we are often interested in the dependence network across brain regions, where data from each region are of functional form (e.g., EEG/ERP signals, MRI/fMRI regions). In bioinformatics, we often need to model gene networks based on time-course gene expression data (Ma et al., 2006), treating each time-course as a continuous process. In epigenetics, it is of interest to study how cells are differentiated into organs (cell lineage and differentiation) by exploring the dependence structure of genome-wide methylation levels across different cell types, and for each cell type, the methylation level can be considered as a function of the genomic locations.

Although there is increasingly rich literature on generalizations to accommodate matrix-variate graphical models (Wang and West, 2009), time varying graphical models (Zhou et al., 2010; Kolar and Xing, 2011), and dynamic linear models (Carvalho and West, 2007), the generalization to functional data has not received much attention in the literature. In recent work, Qiao et al. (2015) extended the graphical LASSO of Yuan and Lin (2007) to the functional data case. They estimate the graph by maximizing a penalized log-Gaussian likelihood constructed through truncated basis expansion, and prove the consistency of the estimated edges. In this paper, we propose Bayesian graphical models for functional data following a fundamentally different approach. In particular, we construct the graphical model directly in the space of infinite dimensional random functions through establishing the Markov distributions and hyper Markov laws for random processes, and propose a Bayesian framework that generally holds for all random processes. We then demonstrate the special case of a multivariate Gaussian process in the space of square integrable func-

tions. Through representing the random functions with orthogonal basis expansions, we transform functional data from the function space to the isometrically isomorphic space of basis coefficients, where Markov distributions and hyper Markov laws can be conveniently constructed. We further propose a hyper-inverse-Wishart-process prior for the covariance kernels of the coefficient sequences, and study theoretical properties of the proposed prior such as existence and uniqueness. We also establish the strong hyper Markov property and conjugacy of this prior under a finite rank condition for the prior kernel parameter, which implies that the covariance kernel of the coefficient sequences is a priori finite dimensional. To perform posterior inference, we introduce a regularity condition which allows us to write the likelihood and prior density and design stochastic search MCMC algorithms for posterior sampling. Performance of the proposed approach is demonstrated through simulation studies and analysis of brain activity and alcoholism data.

To our knowledge, the proposed approach is the first considering functional data graphical models from a Bayesian perspective. It extends the theory of Dawid and Lauritzen (1993) from multivariate data to multivariate functional data. Most existing graphical model approaches often naively apply multivariate methods to functional data after performing discretization or feature extraction. Such approaches may not take full advantage of the fact that data arise from a function and can lack reasonable limiting behavior. Our graphical model framework guarantees proper theoretical behavior as well as computational convenience.

2. Graphical Models for Multivariate Functional Data

In this section, we first review graphical models for multivariate data in Section 2.1, then introduce graphical models for multivariate functional data in Section 2.2, and finally present the specific case of Gaussian process graphical models in Section 2.3.

2.1 Review of Graph Theory and Gaussian Graphical Models

We follow Dawid and Lauritzen (1993), Lauritzen (1996), and Jones et al. (2005). Let $G = (V, E)$ denote an undirected graph with a vertex set V and a set of edge pairs $E = \{(i, j)\}$. Each vertex corresponds to one variable. Two variables a and b are conditionally independent if and only if $(a, b) \notin E$. A graph or a subgraph is *complete* if all possible pairs of vertices are joined by edges. A complete subgraph is *maximal* if it is not contained within another complete subgraph. A maximal subgraph is called a *clique*. If A, B, C are subsets of V with $V = A \cup B$, $C = A \cap B$, then C is said to separate A from B if every path from a vertex in A to a vertex in B goes through C. C is called a *separator* and the pair (A, B) forms a decomposition of G. The separator is *minimal* if it does not contain a proper subgraph which also separates A from B. While keeping the separators minimal, we can iteratively decompose a graph into a sequence of *prime components* – a sequentially defined collection of subgraphs that cannot be further decomposed (Jones et al., 2005). If all the prime components of a connected graph are complete, the graph is called *decomposable*. All the prime components of a decomposable graph are cliques. Iteratively decomposing a decomposable graph G produces a *perfectly ordered* sequence of cliques and separators $(C_1, S_2, C_2, \ldots, S_m, C_m)$ such that $S_i = H_{i-1} \cap C_i$ and $H_{i-1} = C_1 \cup \cdots \cup C_{i-1}$. Let $\mathcal{C} = \{C_1, \ldots, C_m\}$ denote the set of cliques and $\mathcal{S} = \{S_2, \ldots, S_m\}$ denote the set of

separators. The perfect ordering means that for every $i = 2, \ldots, m$, there is a $j < i$ with $S_i \subset C_j$ (Lauritzen, 1996, page 15).

If the components of a random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ obey conditional independence according to a decomposable graph G, the joint density can be factorized as

$$p(\mathbf{X} \mid G) = \frac{\prod_{C \in \mathcal{C}} p(\mathbf{X}_C)}{\prod_{S \in \mathcal{S}} p(\mathbf{X}_S)},$$

where $\mathbf{X}_A = \{X_i, i \in A\}$. If $\mathbf{X}$ is Gaussian with zero mean and precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, then X_i is conditionally independent of X_j given $\mathbf{X}_{V \setminus \{i,j\}}$, denoted by $X_i \perp\!\!\!\perp X_j \mid \mathbf{X}_{V \setminus \{i,j\}}$, if and only if the (i,j)th element of $\mathbf{\Omega}$ is zero. In this case $p(\mathbf{X} \mid G)$ is uniquely determined by marginal covariances $\{\mathbf{\Sigma}_C, \mathbf{\Sigma}_S, C \in \mathcal{C}, S \in \mathcal{S}\}$, which are sub-diagonal blocks of $\mathbf{\Sigma}$ according to the clique and separator sets. For a given G, a convenient conjugate prior for $\mathbf{\Sigma}$ is hyper-inverse-Wishart (HIW) with density

$$p(\mathbf{\Sigma} \mid G, \delta, \mathbf{U}) = \frac{\prod_{C \in \mathcal{C}} p(\mathbf{\Sigma}_C \mid \delta, \mathbf{U}_C)}{\prod_{S \in \mathcal{S}} p(\mathbf{\Sigma}_S \mid \delta, \mathbf{U}_S)},$$

where $p(\mathbf{\Sigma}_C \mid \delta, \mathbf{U}_C)$ and $p(\mathbf{\Sigma}_S \mid \delta, \mathbf{U}_S)$ are densities of inverse-Wishart (IW) distributions. In this paper, the inverse-Wishart follows the parameterization of Dawid (1981), i.e., $\mathbf{\Sigma} \sim \mathrm{IW}(\delta, \mathbf{U})$ if and only if $\mathbf{\Sigma}^{-1}$ has a Wishart distribution $\mathrm{W}(\delta + p - 1, \mathbf{U}^{-1})$, where $\delta > 0$ and $\mathbf{\Sigma}$ is a p by p matrix.

2.2 Graphical Models for Multivariate Functional Data

Let $\mathbf{f} = \{f_j\}_{j=1}^p$ denote a collection of random processes where each component f_j is in $L^2(T_j)$ and each T_j is a closed subset of the real line. The domain of $\mathbf{f}$ is denoted by $T = \bigsqcup_{j=1}^p T_j$, where $\bigsqcup$ denotes the disjoint union defined by $\bigsqcup_{j=1}^p T_j = \bigcup_{j=1}^p \{(t,j) : t \in T_j\}$. For each j, let $\{\phi_{jk}\}_{k=1}^\infty$ denote an orthonormal basis of $L^2(T_j)$. The extended basis functions $\psi_{jk} = (0, \ldots, 0, \phi_{jk}, 0, \ldots, 0)$, with ϕ_{jk} in the jth component and 0 functions elsewhere for $j = 1, \ldots, p$ and $k = 1, \ldots, \infty$, form an orthonormal basis of $L^2(T)$. Let $(L^2(T), \mathcal{B}(L^2(T)), P)$ be a probability space, where $\mathcal{B}(L^2(T))$ is the Borel σ-algebra on $L^2(T)$. For $V = \{1, 2, \ldots, p\}$ and $A \subset V$, denote by $\mathbf{f}_A$ the subset of $\mathbf{f}$ with domain $T_A = \bigsqcup_{j \in A} T_j$. We define the conditional independence relationships for components of $\mathbf{f}$ in Definition 1.

Definition 1 *Let A, B, and C be subsets of V. Then $\mathbf{f}_A$ is conditionally independent of $\mathbf{f}_B$ given $\mathbf{f}_C$ under P, written as $\mathbf{f}_A \perp\!\!\!\perp \mathbf{f}_B \mid \mathbf{f}_C[P]$, if for any $\mathbf{f}_A \in D_A$, where D_A is a measurable set in $L^2(T_A)$, there exists a version of the conditional probability $P(\mathbf{f}_A \in D_A \mid \mathbf{f}_B, \mathbf{f}_C)$ which is $\mathcal{B}(L^2(T_C))$ measurable, and hence one may write $P(\mathbf{f}_A \in D_A \mid \mathbf{f}_B, \mathbf{f}_C) = P(\mathbf{f}_A \in D_A \mid \mathbf{f}_C)$. Here, $\mathcal{B}(L^2(T_C))$ denotes the Borel σ-algebra on $L^2(T_C)$. Note that this implies $P(\mathbf{f}_A \in D_A, \mathbf{f}_B \in D_B \mid \mathbf{f}_C) = P(\mathbf{f}_A \in D_A \mid \mathbf{f}_C)\, P(\mathbf{f}_B \in D_B \mid \mathbf{f}_C)$.*

We would like to use a decomposable graph $G = (V, E)$ to describe the conditional independence relationships of components in $\mathbf{f}$, whereby a Bayesian framework can be constructed and G can be inferred through posterior inference. To this end, we link the probability measure P of $\mathbf{f}$ with G by assuming that P is *Markov* over G, as defined in Definition 2.

Definition 2 *Let $G = (V, E)$ denote a decomposable graph. A probability measure P of $\mathbf{f}$ is called Markov over G if for any decomposition (A, B) of G, $\mathbf{f}_A \perp\!\!\!\perp \mathbf{f}_B \mid \mathbf{f}_{A\cap B}[P]$.*

Given a decomposable graph G, a probability measure of $\mathbf{f}$ with Markov property may be constructed. To enable the construction, we first state Lemma 1, which generalizes Lemma 2.5 of Dawid and Lauritzen (1993) from the random variable to the random process case.

Lemma 1 *Let $\mathbf{f} = (f_1, \ldots, f_p)$ be a collection of random processes in $L^2(T)$. For subsets $A, B \subset V = \{1, \ldots, p\}$ with $A \cap B \neq \varnothing$, suppose that P_1 and P_2 are probability measures of $\mathbf{f}_A$ and $\mathbf{f}_B$, respectively. If P_1 and P_2 are consistent, meaning that they induce the same measure for $\mathbf{f}_{A\cap B}$, then there exists a unique probability measure P for $\mathbf{f}_{A\cup B}$ such that (i) $P_A = P_1$, (ii) $P_B = P_2$, and (iii) $\mathbf{f}_A \perp\!\!\!\perp \mathbf{f}_B \mid \mathbf{f}_{A\cap B}[P]$. The measure P is called a Markov combination of P_1 and P_2, denoted as $P = P_1 \star P_2$.*

We provide a proof of Lemma 1 through construction in Appendix B. The main idea is to first construct the conditional probability $P_1\{\cdot \mid \pi_{A\cap B}(\mathbf{f}_A)\}$ from P_1, where $\pi_{A\cap B} : L^2(T_A) \to L^2(T_{A\cap B})$ is a projection map and $T_A = \bigsqcup_{j\in A} T_j$. We then define $P\{\cdot \mid \pi_B(\mathbf{f})\}$ based upon $P_1\{\cdot \mid \pi_{A\cap B}(\mathbf{f}_A)\}$ using disintegration theory (Chang and Pollard, 1997), and finally construct the joint measure P that satisfies conditions (i)–(iii). With Lemma 1, we can construct a joint probability measure for $\mathbf{f}$ that is Markov over G. The construction is based on the perfectly ordered decomposition $(C_1, S_2, C_2, \ldots, S_m, C_m)$ of G with $S_i = H_{i-1} \cap C_i$ and $H_{i-1} = C_1 \cup \cdots \cup C_{i-1}$. Let $\{M_{C_i}, i = 1, \ldots, m\}$ be a sequence of pairwise consistent probability measures for $\{\mathbf{f}_{C_i}, i = 1, \ldots, m\}$. We construct a Markov probability measure P over G through the following recursive procedure

$$P_{C_1} = M_{C_1}, \tag{1}$$

$$P_{H_{i+1}} = P_{H_i} \star M_{C_{i+1}}, \quad i = 1, \ldots, m-1. \tag{2}$$

One can show that the probability measure constructed this way is the unique Markov probability measure over G with marginals $\{M_{C_i}\}$, and the proof follows that of Theorem 2.6 in Dawid and Lauritzen (1993). We call the probability distribution induced by the probability measure constructed above the *Markov distribution* of $\mathbf{f}$ over G.

Denote the Markov distribution of $\mathbf{f}$ constructed in (1) - (2) by P_G, and denote the space of all Markov distributions over G by $\mathcal{M}(G)$. A prior law for P_G is then supported on $\mathcal{M}(G)$. We follow Dawid and Lauritzen (1993) to define hyper Markov laws and use them as prior laws for P_G. A prior law $\mathfrak{L}$ of P_G is called *hyper Markov* over G if for any decomposition (A, B) of G, $(P_G)_A \perp\!\!\!\perp (P_G)_B \mid (P_G)_{A\cap B}[\mathfrak{L}]$, where $(P_G)_A$ takes values in $\mathcal{M}(G_A)$ which is the space of all Markov distributions over subgraph G_A. Here, we have assumed that G is collapsible onto A, therefore $\phi \in \mathcal{M}(G_A)$ if and only if $\phi = (P_G)_A$ for some $(P_G) \in \mathcal{M}(G)$. The following Proposition 1 states that the theory of hyper Markov laws of Dawid and Lauritzen (1993) applies to our random process setup.

Proposition 1 *The theory of hyper Markov laws over undirected decomposable graphs, as described in Section 3 of Dawid and Lauritzen (1993), holds for random processes.*

According to the theory of hyper Markov laws, one can construct a prior law for P_G using a sequence of consistent marginal laws $\{\mathfrak{L}_C, C \in \mathcal{C}\}$ in a similar fashion as (1) - (2).

Denote by $\mathfrak{L}_G$ the constructed hyper Markov prior for P_G and by Π a prior distribution for the graph G. A Bayesian graphical model for the collection of random processes $\mathbf{f}$ can be described as

$$\mathbf{f} \sim P_G; \quad P_G \sim \mathfrak{L}_G; \quad G \sim \Pi. \tag{3}$$

As we have yet to specify a concrete example for the probability measure P_G, the above Bayesian framework remains abstract at the moment. In Section 2.3, we construct P_G using Gaussian processes and propose a hyper-inverse-Wishart-process law as the prior for P_G. The prior distribution Π is supported on the finite dimensional space of decomposable graphs with p nodes.

2.3 Gaussian Process Graphical Models for Multivariate Functional Data

Let $\mathbf{f}_0 = (f_{01}, \ldots, f_{0p})$ be an element in $L^2(T)$. Denote by $\mathcal{K} = \{k_{ij} : T_i \times T_j \to \mathbb{R}\}$ a collection of covariance kernels such that $\text{cov}\{f_i(s), f_j(t)\} = k_{ij}(s,t), s \in T_i, t \in T_j$. We assume that $\mathcal{K}$ is positive semidefinite and trace class. Positive semidefinite means that

$$\sum_{i,j=1}^{p} \sum_{k,l=1}^{\infty} c_{ik} c_{jl} \int_{T_j} \int_{T_i} k_{ij}(s,t) \phi_{ik}(s) \phi_{jl}(t) ds dt \geq 0$$

for any square summable sequence $\{c_{ik},\, i = 1, \ldots, p,\, k = 1, \ldots, \infty\}$; trace class means that

$$\sum_{j=1}^{p} \sum_{l=1}^{\infty} \int_{T_j} \int_{T_i} k_{jj}(s,t) \phi_{jl}(s) \phi_{jl}(t) ds dt < \infty.$$

Then $\mathbf{f}_0$ and $\mathcal{K}$ uniquely determine a Gaussian process on $L^2(T)$ (Prato, 2006), which we call a multivariate Gaussian process, and write $\text{MGP}(\mathbf{f}_0, \mathcal{K})$. The definition of multivariate Gaussian process implies that for $A \subset V$, $\mathbf{f}_A \sim \text{MGP}(\mathbf{f}_{0A}, \mathcal{K}_A)$ where $\mathcal{K}_A = \{k_{ij}, i, j \in A\}$. Furthermore, on a sequence of cliques $\mathcal{C} = \{C_1, \ldots, C_m\}$, the marginal Gaussian process measures for $\{\mathbf{f}_C, C \in \mathcal{C}\}$ are automatically consistent because they are induced from the same joint distribution. Therefore, we can construct a Markov distribution for $\mathbf{f}$ over G through procedure (1) - (2). We denote the resulting distribution of $\mathbf{f}$ by $\text{MGP}_G(\mathbf{f}_0, \mathcal{K}_\mathcal{C})$, where $\mathcal{K}_\mathcal{C} = \{k_{ij} : i, j \in C, C \in \mathcal{C}\}$. It is clear from this construction that the distribution MGP_G is Markov over G whereas MGP is not.

For the convenience of both theoretical analysis and computation, we represent elements in $L^2(T)$ using orthonormal basis expansions and construct a Bayesian graphical model in the dual space of basis coefficients. Let $\{\phi_{jk}\}_{k=1}^{\infty}$ denote an orthonormal basis of $L^2(T_j)$. For example, $\{\phi_{jk}\}_{k=1}^{\infty}$ could be a wavelet basis. We have the representation $f_j(t) = \sum_{k=1}^{\infty} c_{jk} \phi_{jk}(t)$ where $c_{jk} = \langle f_j, \phi_{jk} \rangle = \int_{T_j} f_j(t) \phi_{jk}(t) dt$. The coefficient sequence $c_j = \{c_{jk}, k = 1, \ldots, \infty\}$ lies in the space of square-summable sequences, denoted by $\ell_j^2 = \left\{ c_{jk} : \sum_{k=1}^{\infty} c_{jk}^2 < \infty \right\}$. Denote $\ell^2 = \prod_{j=1}^{p} \ell_j^2$. Since ℓ_j^2 and $L^2(T_j)$ are isometrically isomorphic for each j, once an orthonormal basis of $L^2(T)$ has been chosen, we have an identification between the Borel probability measures defined on ℓ^2 and $L^2(T)$; therefore we can construct statistical models on ℓ^2 without loss of generality. Let $\mathbf{c} = (c_1, \ldots, c_p)$ denote

the coefficient sequence of $\mathbf{f}$. Then $\mathbf{f} \sim \text{MGP}(\mathbf{f}_0, \mathcal{K})$ corresponds to $\mathbf{c} \sim \text{dMGP}(\mathbf{c}_0, \mathcal{Q})$, where dMGP denotes the infinite dimensional discrete multivariate Gaussian processes, $\mathbf{c}_0$ is the coefficient sequence of $\mathbf{f}_0$ and $\mathcal{Q} = \{q_{ij}(\cdot,\cdot), i, j \in V\}$. Here, q_{ij} is the covariance kernel so that $\text{cov}(c_{ik}, c_{jl}) = q_{ij}(k,l)$ for $k, l \in \{1, 2, 3, \ldots\}$. Similarly, $\mathbf{f} \sim \text{MGP}_G(\mathbf{f}_0, \mathcal{K}_\mathcal{C})$ corresponds to $\mathbf{c} \sim \text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_\mathcal{C})$ where $\mathcal{Q}_\mathcal{C} = \{q_{ij}(\cdot,\cdot), i, j \in C, C \in \mathcal{C}\}$. The collection $\mathcal{Q}$ is also positive semidefinite and trace class, so that $\sum_{i,j=1}^{p} \sum_{k,l=1}^{\infty} c_{ik} c_{jl} q_{ij}(k,l) \geq 0$ for any square summable sequence $\{c_{ik}, i = 1, \ldots, p, k = 1, \ldots, \infty\}$, and $\sum_{j=1}^{p} \sum_{k=1}^{\infty} q_{jj}(k,k) < \infty$. Furthermore, $\mathcal{K}$ relates to $\mathcal{Q}$ through equation $k_{ij}(s,t) = \sum_{k,l=1}^{\infty} q_{ij}(k,l)\phi_{ik}(s)\phi_{jl}(t)$. Denote by $P^{\mathbf{c}}$ and $P^{\mathbf{f}}$ the probability measures of $\mathbf{c}$ and $\mathbf{f}$ respectively, then $\mathbf{f}_A \perp\!\!\!\perp \mathbf{f}_B \mid \mathbf{f}_C[P^{\mathbf{f}}]$ implies $\mathbf{c}_A \perp\!\!\!\perp \mathbf{c}_B \mid \mathbf{c}_C[P^{\mathbf{c}}]$ and vice versa. Thus, the distribution $\text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_\mathcal{C})$ of $\mathbf{c}$ is again Markov.

Assume that $\mathbf{c} \sim \text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_\mathcal{C})$. The parameters involved in this distribution include $\mathbf{c}_0$ and $\mathcal{Q}_\mathcal{C}$. In this study, we assume that $\mathbf{c}_0$ is fixed (e.g., a zero sequence) so that the distribution of $\mathbf{c}$ is uniquely determined by $\mathcal{Q}_\mathcal{C}$. As indicated in Section 2.2, we would like to construct a hyper Markov law for the dMGP_G distribution. Since dMGP_G is uniquely determined by $\mathcal{Q}_\mathcal{C}$, it is equivalent to construct a hyper Markov law for $\mathcal{Q}_\mathcal{C}$. Given a positive integer δ and a collection $\mathcal{U} = \{u_{ij} : \mathbb{N} \times \mathbb{N} \to \mathbb{R}, i, j \in V\}$ which is symmetric, positive semidefinite, and trace class, we construct a hyper-inverse-Wishart-process (HIWP) prior for $\mathcal{Q}_\mathcal{C}$ following Theorem 1.

Theorem 1 Assume that $\mathbf{c} \sim \text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_\mathcal{C})$. Suppose that δ is a positive integer, and $\mathcal{U}$ is a collection of kernels that is symmetric, positive semidefinite and trace class. Then there exists a sequence of pairwise consistent inverse-Wishart processes determined by δ and $\mathcal{U}_C = \{u_{ij}, i, j \in C\}, C \in \mathcal{C}$, based on which one can construct a unique hyper Markov law for $\mathcal{Q}_\mathcal{C}$, which we call a hyper-inverse-Wishart-process, and write $\mathcal{Q}_\mathcal{C} \sim \text{HIWP}_G(\delta, \mathcal{U}_\mathcal{C})$, where $\mathcal{U}_\mathcal{C} = \{u_{ij}, i, j \in C, C \in \mathcal{C}\}$.

Based on Theorem 1, a Bayesian Gaussian process graphical model can be written as

$$\mathbf{c} \sim \text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_\mathcal{C}), \quad \mathcal{Q}_\mathcal{C} \sim \text{HIWP}_G(\delta, \mathcal{U}_\mathcal{C}), \quad G \sim \Pi. \tag{4}$$

It is of interest to investigate the properties of the HIWP prior and the corresponding posterior distribution. As shown in Dawid and Lauritzen (1993), one nice property of the HIW law is the strong hyper Markov property, which leads to conjugacy as well as convenient posterior computation at each clique. In case of the HIWP prior, the strong hyper Markov property is defined such that for any decomposition (A, B) of G in model (4), $\mathcal{Q}_{B|A} \perp\!\!\!\perp \mathcal{Q}_A$, where $\mathcal{Q}_{B|A}$ denotes the conditional distribution (i.e., conditional covariance) of $\mathbf{c}_B$ given $\mathbf{c}_A$. In the following proposition, we show that the HIWP_G prior constructed in Theorem 1 is strong hyper Markov when $\text{rank}(u_{ij}) < \infty$ for $i, j \in V$.

Proposition 2 *Suppose that the collection of kernels $\mathcal{U}$ satisfies that rank$(u_{ij}) < \infty$ for $i, j \in V$, then the hyper-inverse-Wishart-process prior constructed in Theorem 1 satisfies the strong hyper Markov property. That is, if $\mathcal{Q}_\mathcal{C} \sim HIWP_G(\delta, \mathcal{U}_\mathcal{C})$, then for any decomposition (A, B) of G, $\mathcal{Q}_{B|A} \perp\!\!\!\perp \mathcal{Q}_A$, where $\mathcal{Q}_{B|A}$ denotes the conditional distribution (e.g., conditional covariance) of $\mathbf{c}_B$ given $\mathbf{c}_A$.*

The finite rank condition for the prior parameters $\{u_{ij}\}$ in Proposition 2 is a relatively strong condition under which the HIWP_G satisfies the strong hyper Markov property. It

implies that the covariance kernel $\mathcal{Q}_{\mathcal{C}}$, thus the sequence $\mathbf{c}$, is a priori finite dimensional. Whether the strong hyper Markov property still holds without this condition remains a challenging open problem. In the online appendix, we have included several interesting results made through our preliminary study, which may provide useful insights into further investigations of this problem. The strong hyper Markov property of HIWP_G ensures that the joint posterior of $\mathcal{Q}_{\mathcal{C}}$ (conditional on G) can be constructed from the marginal posterior of $\mathcal{Q}_C$ (conditional on G) at each clique C, as stated in Theorem 2. Therefore one essentially transforms the Bayesian analysis to a sequence of sub-analyses at the cliques, which substantially reduces the size of the problem.

Theorem 2 *Suppose that* $\mathbf{c}_i \sim dMGP_G(\mathbf{c}_0, \mathcal{Q}_{\mathcal{C}}), i = 1, \ldots, n$ *are independent and identically distributed. Further assume that the prior of* $\mathcal{Q}_{\mathcal{C}}$ *is* $HIWP_G(\delta, \mathcal{U}_{\mathcal{C}})$ *where the collection of kernels* $\mathcal{U}$ *satisfies that* $rank(u_{ij}) < \infty$ *for* $i, j \in V$*. Then the conditional posterior of* $\mathcal{Q}_{\mathcal{C}}$ *given* $\{\mathbf{c}_i\}$ *and* G *is* $HIWP_G(\widetilde{\delta}, \widetilde{\mathcal{U}}_{\mathcal{C}})$*, where* $\widetilde{\delta} = \delta + n$*,* $\widetilde{\mathcal{U}}_{\mathcal{C}} = \{\widetilde{u}_{ij}, i, j \in C, C \in \mathcal{C}\}$ *and* $\widetilde{u}_{ij} = u_{ij} + \sum_{i=1}^{n} (\mathbf{c}_i - \mathbf{c}_{0i}) \otimes (\mathbf{c}_j - \mathbf{c}_{0j})$*. Here* $\otimes$ *denotes the outer product. Furthermore, the marginal distribution of* $\{\mathbf{c}_i\}$ *given* $\{G, \mathbf{c}_0, \delta, \widetilde{\mathcal{U}}_{\mathcal{C}}\}$ *is again Markov over* G*.*

Theorem 2 implies that when $\text{rank}(u_{ij}) < \infty$ for $i, j \in V$, the $\text{HIWP}_G(\delta, \mathcal{U}_{\mathcal{C}})$ prior is a conjugate prior for $\mathcal{Q}_{\mathcal{C}}$ in the $\text{dMGP}_G(\mathbf{c}_0, \mathcal{Q}_{\mathcal{C}})$ likelihood. Note that here the likelihood, the prior, and the posterior are all conditional on G, which makes Bayesian inference of G tractable. Model (4) and results in Theorem 2 provide the theoretical foundation for practical Bayesian inference under a reasonable regularity condition, as discussed in Section 3.

3. Bayesian Posterior Inference

Despite the fact that functional data are realizations of inherently infinite-dimensional random processes, data can only be collected at a finite number of measurement points. Essentially, estimating the conditional independence structure of infinite-dimensional random processes based on a finite number of measurement points is an inverse problem and therefore requires regularization. Müller and Yao (2008) reviewed two main approaches for regularization in functional data analysis—finite approximation through, e.g., suitably truncating the basis expansion representation and penalized likelihood. In this paper, we suggest performing posterior inference based on approximating the underlying random processes with orthogonal basis functions. In particular, we assume the following regularity condition:

Condition 1 *The functional data* $\mathbf{f}$ *are observed discretely on a dense grid* $\mathbf{t} = \bigsqcup \mathbf{t}_j$ *with* $\mathbf{t}_j = (t_{j1}, \ldots, t_{jm_j(n)})$ *and* $m_j(n) \to \infty$ *as* $n \to \infty$*. One can find* $M_j(n)$ *so that the underlying random process* f_j *can be approximated with an* M_j*-term orthogonal basis expansion* $\widehat{f}_j = \sum_{l=1}^{M_j} c_{jl}\phi_{jl}$*, with approximation error* $||f_j - \widehat{f}_j||_{L^2} = O_p(n^{-\beta})$ *with* $\beta \geq 1/2$ *for all* $j \in V$*.*

Essentially, Condition 1 requires that the discretely-measured functional data capture sufficient information about the underlying random processes, so that we can approximate each f_j with a negligible approximation error. This condition provides the consistency of the basis representation, i.e., the approximation error converges to zero with order $O_p(n^{-\beta})$

when M_j increases with the sample size n. We need such a condition in order to guarantee that the behavior of $\{f_j\}$ is not too outrageous. Certain assumptions, such as the decay rate of the eigenvalues of f_j, the smoothness property of f_j, or the characteristics of the basis functions, will determine the specific rate β (De Boor, 2001; Jansen and Oonincx, 2015). However, under our generic setup, since we prefer not to specify particular assumptions, we only require a mild range for the convergence rate. Condition 1 is a basic assumption in the functional setting, and a similar regularity condition has been adopted by Qiao et al. (2015) in a functional graphical model based on the group LASSO penalty.

3.1 Bayesian Posterior Inference under the Regularization Condition

The regularity from Condition 1 enables us to write the density functions of the Markov distributions and hyper Markov laws so that posterior inference can be practically implemented. Denoting $M = (M_1, \ldots, M_p)$, we can explicitly write the density function for the truncated process $\mathbf{c}^M = (c_1^{M_1}, \ldots, c_p^{M_p})$, and an MCMC algorithm can then be designed for the posterior inference of the underlying graph G. The density function of $\mathbf{c}^M$ is

$$p(\mathbf{c}^M \mid \mathbf{c}_0^M, \mathbf{Q}_\mathcal{C}, G) = \frac{\prod_{C\in\mathcal{C}} p(\mathbf{c}_C^M \mid \mathbf{c}_{0,C}^M, \mathbf{Q}_C)}{\prod_{S\in\mathcal{S}} p(\mathbf{c}_S^M \mid \mathbf{c}_{0,S}^M, \mathbf{Q}_S)}, \tag{5}$$

where $\mathbf{Q}_\mathcal{C}$ is a block-wise covariance matrix with the (i,j)th block formed by $\{q_{ij}(k,l), k = 1,\ldots,M_i, l = 1,\ldots,M_j\}$, and $\mathbf{Q}_C$, $\mathbf{Q}_S$ are submatrices of $\mathbf{Q}_\mathcal{C}$ corresponding to clique C and separator S, respectively. The HIWP_G prior of $\mathcal{Q}_\mathcal{C}$ induces a hyper inverse-Wishart prior with density

$$p(\mathbf{Q}_\mathcal{C} \mid G) = \frac{\prod_{C\in\mathcal{C}} p(\mathbf{Q}_C \mid \delta, \mathbf{U}_C)}{\prod_{S\in\mathcal{S}} p(\mathbf{Q}_S \mid \delta, \mathbf{U}_S)}, \tag{6}$$

where $p(\mathbf{Q}_C \mid \delta, \mathbf{U}_C)$ is the density of inverse-Wishart defined in Dawid (1981), $\mathbf{U}_C$ is a submatrix of $\mathbf{U}_\mathcal{C}$ corresponding to clique C, and $\mathbf{U}_\mathcal{C}$ is a block-wise matrix formed by $\{u_{ij}\}$ in the same way as $\mathbf{Q}_\mathcal{C}$ is formed by $\{q_{ij}\}$. The $p(\mathbf{Q}_S \mid \delta, \mathbf{U}_S)$ component in the denominator is defined similarly. Based on (5) and (6), and assuming that $\{\mathbf{c}_i, i = 1,\ldots,n\}$ is a random sample of $\mathbf{c}$, one can further integrate out $\mathbf{Q}_\mathcal{C}$ to get the marginal density

$$p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, G) = (2\pi)^{-\frac{n}{2}(\sum_i M_i)} \frac{h(\delta, \mathbf{U}_\mathcal{C})}{h(\widetilde{\delta}, \widetilde{\mathbf{U}}_\mathcal{C})}, \tag{7}$$

where

$$h(\delta, \mathbf{U}_\mathcal{C}) = \frac{\prod_{C\in\mathcal{C}} |\tfrac{1}{2}\mathbf{U}_C|^{(\frac{\delta+d_c-1}{2})} \Gamma_{d_c}^{-1}\{\tfrac{1}{2}(\delta+d_c-1)\}}{\prod_{S\in\mathcal{S}} |\tfrac{1}{2}\ \mathbf{U}_S|^{(\frac{\delta+d_s-1}{2})} \Gamma_{d_s}^{-1}\{\tfrac{1}{2}(\delta+d_s-1)\}},$$

and d_c and d_s are the dimensions of $\mathbf{U}_C$ and $\mathbf{U}_S$ respectively, and $\Gamma_b(a) = \pi^{b(b-1)/4} \prod_{i=0}^{b-1} \Gamma(a - i/2)$. The denominator $h(\widetilde{\delta}, \widetilde{\mathbf{U}}_\mathcal{C})$ in (7) is defined in the same way. Based on these results, posterior inference can be done through sampling from the posterior density

$$p(G \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M) \propto p(\{\mathbf{c}_i^M\} \mid \mathbf{c}_0^M, G)\, p(G), \tag{8}$$

where $p(G)$ is the density function corresponding to the prior distribution $G \sim \Pi$, which is a discrete distribution supported on all decomposable graphs with p nodes. Giudici and Green (1999) used the discrete uniform prior $\Pr(G = G_0) = 1/d$ for any fixed p-node decomposable graph G_0, where d is the total number of such graphs; Jones et al. (2005) used the independent Bernoulli prior with probability $2/(p-1)$ for each edge, which favors sparser graphs (Giudici, 1996). The following MCMC algorithm describes the steps to generate posterior samples based on (8).

Algorithm 1

Step 0. *Set an initial decomposable graph G and set the prior parameters $\boldsymbol{c}_0$, δ, and $\boldsymbol{U}_{\mathcal{C}}$.*

Step 1. *With probability $1-q$, propose $\widetilde{G}$ by randomly adding or deleting an edge from G (each with probability 0.5) within the space of decomposable graphs; with probability q, propose $\widetilde{G}$ from a discrete uniform distribution supported on the set of all decomposable graphs. Accept the new $\widetilde{G}$ with probability*

$$\alpha = \min\left\{1, \frac{p(\widetilde{G} \mid \{\boldsymbol{c}_i^M\}, \boldsymbol{c}_0^M)\ p(G \mid \widetilde{G})}{p(G \mid \{\boldsymbol{c}_i^M\}, \boldsymbol{c}_0^M)\ p(\widetilde{G} \mid G)}\right\}.$$

Repeat Step 1 for a large number of iterations until convergence is achieved.

Detailed derivations are available in the online appendix. The above algorithm is a Metropolis-Hastings sampler with a mixture of local and heavier-tailed proposals, also called a *small-world sampler*. The "local" move involves randomly adding or deleting one edge based on the current graph, and the "global" move is achieved through the discrete uniform proposal. Guan et al. (2006) and Guan and Krone (2007) have shown that the small-world sampler leads to much faster convergence especially when the posterior distribution is either multi-modal or spiky.

3.2 Bayesian Posterior Inference for Noisy Functional Data

The theory in Section 2 and the posterior inference in Section 3.1 relies on the assumption that the distribution of $\mathbf{f}$ (and $\mathbf{c}$) is Markov over G. In many situations, it is more desirable to make such an assumption in a hierarchical model. For example, when functional data are subject to measurement error, one might wish to incorporate an additive error term and consider the following model

$$y_{ijt} = f_{ij}(t) + \varepsilon_{ijt}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p, \quad t \in \mathbf{t}_j, \tag{9}$$

where $\{y_{ijt}, t \in \mathbf{t}_j\}$ are noisy observations measured on a dense grid $\mathbf{t}_j = (t_{j1}, \ldots, t_{jm_j})$, $\{f_{ij}\}$ are the underlying true functions, and $\{\varepsilon_{ijt}, t \in \mathbf{t}_j\}$ are measurement errors. We assume that $\{f_{ij}\}$ and $\{\varepsilon_{ijt}, t \in \mathbf{t}_j\}$ are mutually independent of each other. The inference of model (9) involves both smoothing (i.e., estimating f_{ij}) and estimation of the underlying graph G. We achieve these goals simultaneously through fitting a Bayesian hierarchical model.

In particular, we assume that $\{f_{ij}\}$ are Gaussian processes in $L^2(T_j)$, and denote $\{c_{ijk}\}$ their basis coefficients corresponding an orthonormal basis $\{\phi_{jk}\}_{k=1}^{\infty}$. With this representation, model (9) has the form $y_{ijt} = \sum_{k=1}^{\infty} c_{ijk}\phi_{jk}(t) + \varepsilon_{ijt}$, and $\{c_{ijk}\}$ is a discrete Gaussian

process. We further assume that the measurement error $\boldsymbol{\varepsilon}_{ij} = \{\varepsilon_{ijt}, t \in \mathbf{t}_j\}$ is Gaussian white noise with variance σ_j^2, i.e., $\varepsilon_{ijt} \sim N(0, \sigma_j^2)$ independently across all t for $t \in \mathbf{t}_j$. Truncating at the M_jth basis element, we can reparameterize the model as

$$y_{ijt} = \sum_{k=1}^{M_j} c_{ijk}\phi_{jk}(t) + \widetilde{\varepsilon}_{ijt}, \quad t \in \mathbf{t}_j \tag{10}$$

where $\widetilde{\varepsilon}_{ijt}$ is a new residual term that consists of the approximation error of the truncated series (e.g., $\sum_{k=M_j+1}^{\infty} c_{ijk}\phi_{jk}(t)$) and the measurement error. If we concatenate the noisy observations to form a vector

$$\mathbf{y}_i = (y_{i1t_{11}}, \ldots, y_{i1t_{1m_1}}, \ldots, y_{ipt_{p1}}, \ldots, y_{ipt_{pm_p}})^T$$

and denote $\mathbf{c}_i^M$ the vector formed by the basis coefficients $\{c_{ijk}, j = 1, \ldots, p, k = 1, \ldots, M_j\}$, then model (10) can be written as $\mathbf{y}_i = \boldsymbol{\Phi}\mathbf{c}_i^M + \widetilde{\boldsymbol{\varepsilon}}_i$, where $\boldsymbol{\Phi} = \text{diag}\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p\}$ is a $\sum_j m_j$ by $\sum_j M_j$ block-diagonal matrix with the jth diagonal block containing $\boldsymbol{\phi}_j = [\phi_{j1}(\mathbf{t}_j), \ldots, \phi_{jM_j}(\mathbf{t}_j)]$, and $\widetilde{\boldsymbol{\varepsilon}}_i$ denote the concatenated vector of the new residual terms. We assume that $\widetilde{\boldsymbol{\varepsilon}}_i \sim N(0, \boldsymbol{\Lambda})$ where $\boldsymbol{\Lambda} = \text{diag}(s_1^2 1_{m_1}^T, \ldots, s_p^2 1_{m_p}^T)$. Notice that if $\mathbf{Q}_{\mathcal{C}} = \text{cov}(\mathbf{c}_i^M)$, then $\text{cov}(\mathbf{y}_i^M) = \boldsymbol{\Phi}\mathbf{Q}_{\mathcal{C}}\boldsymbol{\Phi}^T + \boldsymbol{\Lambda}$. The diagonals of $\boldsymbol{\Phi}\mathbf{Q}_{\mathcal{C}}\boldsymbol{\Phi}^T$ and $\boldsymbol{\Lambda}$ can not be separately identifiable. Therefore, we treat $\boldsymbol{\Lambda}$ as a fixed model parameter, whose quantity can be pre-determined by the approximation $s_j^2 \approx \widehat{\sigma}_j^2$, where $\widehat{\sigma}_j^2$ is the estimation of σ_j^2 using local smoothing on $\{\varepsilon_{ijt}\}$.

Applying a prior for $\mathbf{c}_i^M$ in the form of (5) (conditional on G) and the HIWP$_G$ prior for the covariance matrix $\mathcal{Q}_{\mathcal{C}}$ in the form of (6), we obtain the density function for the joint posterior

$$p(\{\mathbf{c}_i^M\}, \mathbf{Q}_{\mathcal{C}}, G \mid \{\mathbf{y}_i\}) \propto \prod_{i=1}^{n} p(\mathbf{y}_i \mid \mathbf{c}_i^M, \boldsymbol{\Lambda})\, p(\mathbf{c}_i^M \mid \mathbf{c}_0^M, \mathbf{Q}_{\mathcal{C}}, G)\, p(\mathbf{Q}_{\mathcal{C}} \mid G)\, p(G). \tag{11}$$

From (11), we can integrate out $\mathbf{Q}_{\mathcal{C}}$ to obtain the marginal posterior distribution of $\{\mathbf{c}_i^M\}$ and G. The MCMC algorithm for generating posterior samples based on (11) is listed in Algorithm 2.

Algorithm 2

Step 0 *Set initial values for $\{\mathbf{c}_i^M\}$, G and set the model parameters δ, $\mathbf{c}_0^M$, $\mathbf{U}$ and $\boldsymbol{\Lambda}$.*

Step 1 *Conditional on $\{\mathbf{c}_i^M\}$, update $G \sim p(G \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)$ using the small-world sampler as described in Step 1 of Algorithm 1, where $p(G \mid \{\mathbf{c}_i^M\}, \mathbf{c}_0^M)$ is computed based on (11).*

Step 2 *Given G, update $\mathbf{Q}_{\mathcal{C}} \sim p(\mathbf{Q}_{\mathcal{C}} \mid \{\mathbf{c}_i^M\}, G)$, which takes the same form as (6) except that δ and $\mathbf{U}$ are replaced by $\widetilde{\delta}$ and $\widetilde{\mathbf{U}}$ respectively using the formulae in Theorem 2.*

Step 3 *Conditional on G and $\mathbf{Q}_{\mathcal{C}}$, update $\mathbf{c}_i^M \sim N(\boldsymbol{\mu}_i, \mathbf{V})$, where $\mathbf{V} = (\boldsymbol{\Phi}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi} + \mathbf{Q}_{\mathcal{C}}^{-1})^{-1}$ and $\boldsymbol{\mu}_i = \mathbf{V}(\boldsymbol{\Phi}^T\boldsymbol{\Lambda}^{-1}\mathbf{y}_i + \mathbf{Q}_{\mathcal{C}}^{-1}\mathbf{c}_0^M)$.*

Repeat Step 1 $\sim$ 3 for a large number of iterations until convergence is achieved.

3.3 Other Practical Computational Issues

Calculating the coefficient sequences $\{\mathbf{c}_i\}$ from the functional observations $\{\mathbf{f}_i\}$ requires the selection of an orthonormal basis $\{\phi_{jk}, j = 1, \ldots, p, k = 1, \ldots, \infty\}$. If a known basis is chosen (e.g., Fourier), the coefficient sequences can be estimated by $c_{ijk} = \langle f_{ij}, \phi_{jk} \rangle$ using numerical integration. Another convenient choice is the eigenbasis of the autocovariance operators of $\{\mathbf{f}_i\}$, in which case the coefficient sequences are called functional principal component (FPC) scores. The corresponding basis representation is called Karhunen-Loève expansion. The eigenbasis can be estimated using the method of Ramsay and Silverman (2005) or the Principal Analysis by Conditional Expectation (PACE) algorithm of Yao et al. (2005). Owing to the rapid decay of the eigenvalues, the eigenbasis provides a more parsimonious and efficient representation compared with other bases. Furthermore, the FPC scores within a curve are mutually uncorrelated, so one may set the prior parameter $\mathbf{U}_{\mathcal{C}}$ to be a matrix with blocks of diagonal sub-matrices, or simply a diagonal matrix.

In addition to the estimation of coefficient sequences, a suitable truncation of the infinite sequences $\{\mathbf{c}_i\}$ is needed to facilitate practical posterior inference. We suggest to pre-determine the truncation parameters using approximation criteria, following Rice and Silverman (1991) or Yao et al. (2005). This includes cross-validation (Rice and Silverman, 1991), applying the pseudo Akaike information criterion (Yao et al., 2005), or controlling the fraction-of-variance-explained (FVE) in the FPC analysis (Lei et al., 2014).

4. Simulation Study

Three simulation studies were conducted to assess the performance of posterior inference using the Gaussian process graphical models outlined in Section 2.3 and Section 3. Simulation 1 corresponds to the smooth functional data case (without measurement error), and Simulation 2 corresponds to the noisy data case when measurement error is considered. Both simulations are based on a true underlying graph with 6 nodes, demonstrated in Figure 1 (a). In simulation 3, we show the performance of the proposed Bayesian inference in a $p > n$ case, with the number of nodes $p = 60$ and the sample size $n = 50$.

4.1 Simulation 1: Graph Estimation for Smooth Functional Data

Multivariate functional data are generated on the domain $[0, 1]$ using Fourier basis with the number of basis functions $\{M_j\}_{j=1}^p$ varying from 3 to 7. The true eigenvalues are generated from Gamma distributions and are subject to exponential decay. The conditional independence structure is determined by a $p \times p$ correlation matrix $\mathbf{R}_0$, with the inverse $\mathbf{R}_0^{-1}$ containing a zero pattern corresponding to the graph in Figure 1 (a). We then generate principal component scores from a multivariate normal distribution with zero mean and a block-wise covariance matrix $\mathbf{Q} = \mathbf{ZRZ}$, which has dimension $\sum_{j=1}^p M_j$. Here $\mathbf{R}$ is a block-wise correlation matrix that has a diagonal form in each block. In particular, the $(i, j)th$ block of $\mathbf{R}$, denoted by $\mathbf{R}_{ij}$, satisfies that $\mathbf{R}_{ij} = (\mathbf{R}_0)_{i,j}\mathbf{I}$ where $\mathbf{I}$ is a rectangular identity matrix with size $M_i \times M_j$. An image plot of $\mathbf{R}$ is shown in Figure 1(d), with its data-domain counterpart (the correlation of $\mathbf{f}$ evaluated on a grid $\mathbf{t}$) shown in Figure 1(c). The multivariate functional data are finally generated through linearly combining the eigenbasis using the principal component scores. A common mean function is added to each curve.

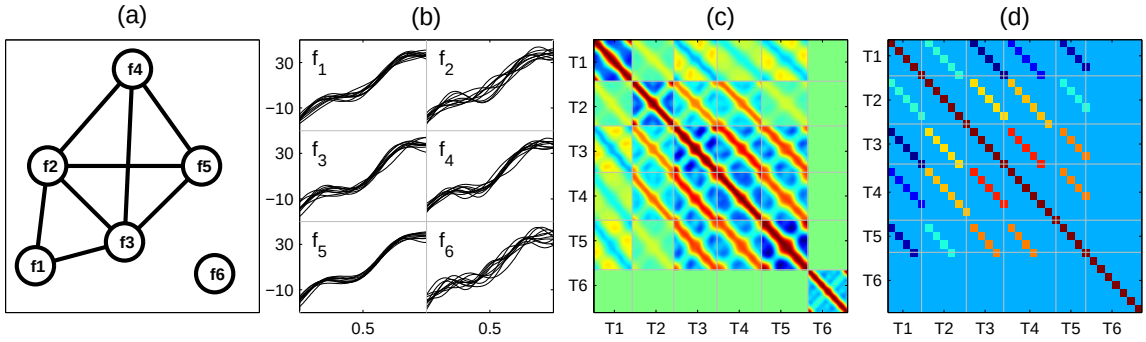

Figure 1: Plots of Simulation 1. (a) The true underlying graph; (b) The first 10 samples of $\{f_{ij}, j = 1, \ldots, 6\}$; (c) The image plot of the underlying data-domain correlation matrix; (d) The image plot of the underlying correlation matrix $\mathbf{R}$.

The generated data contain $n = 200$ independent samples, and each sample contains six curves measured on six different grids. We display the first 10 samples in Figure 1(b).

Based on the data generated above, we estimate the principal component scores $\{\mathbf{c}_i\}$ using the PACE algorithm of Yao et al. (2005) and determine the truncation parameter $\{M_j\}$ using the FVE criterion with a 90% threshold, resulting in $\{M_j\}$ values around 5. We apply Algorithm 1 and set $\delta = 5$ and $\mathbf{U} = \widehat{\mathbf{Z}}\widehat{\mathbf{R}}\widehat{\mathbf{Z}}$, where $\widehat{\mathbf{Z}} = \text{diag}\{\widehat{\lambda}_{jk}^{1/2}, k = 1, \ldots, M_j, j = 1, \ldots, p\}$, $\{\widehat{\lambda}_{jk}\}$ are the estimated eigenvalues and $\widehat{\mathbf{R}}$ is set to be the identity marix. A total of $5,000$ MCMC iterations are performed. Starting from the empty graph, the chain reaches the true underlying graph in around 500 iterations. We have also tried implementing Algorithm 1 with different initial graphs; all implementations resulted in the same posterior mode at the true underlying graph.

We compare the performance of our approach with three other related methods: the Gaussian graphical model of Jones et al. (2005) based on Metropolis-Hastings (GGM-MH), the graphical LASSO (GLASSO) of Friedman et al. (2008), and the matrix-normal graphical model (MNGM) of Wang and West (2009). As both GGM-MH and GLASSO assume that each node is associated with one variable, we reduce the dimension of the functional data by retaining only the first principal component score. The MNGM method assumes matrix data, so we take the first five principal component scores and stack them up to form a 6×5 matrix for each sample. In the MNGM method, graph estimates across the rows and columns are obtained simultaneously, and only that across the rows is of interest to us.

The simulation results are demonstrated in the top panel of Table 1. Summary statistics, such as running-time, mis-estimation rate, sensitivity and specificity are calculated for each method. The running-time was obtained using a laptop with Intel(R) Core(TM) i5 CPU, M430 with 2.27 GHZ processor and 4GB RAM. The comparison of running-time shows that the GLASSO method is the fastest. This is because GLASSO does not require posterior sampling. However, GLASSO relies on a penalized optimization approach which requires determination of the tuning parameter. In this simulation, we have selected the tuning parameter that results in the lowest mis-estimation rate with respect to the underlying true

graph. When the true graph is unknown, the tuning procedure can be time-consuming. The MNGM is much slower to implement, perhaps due to the numerical approximation of the marginal density in the MCMC algorithm.

Data	Method	nFPC	Time	nEdge	nUnique	MisR	Sen	Spec
Smooth	FDGM-S	3 - 5	38	7.66	3	0.02	0.96	1.0
	GGM-MH	1	0.15	9.55	63	0.10	1.0	0.78
	GLASSO	1	-	-	-	0.13	-	-
	MNGM	5	4067.73	5.83	36	0.21	0.66	0.93
Noisy	FDGM-N	3 - 5	64	7.86	5	0.01	0.98	1.0
	GGM-MH	1	0.39	9.62	59	0.11	1.0	0.77
	GLASSO	1	-	-	-	0.13	-	-
	MNGM	5	4086.38	6.33	18	0.26	0.65	0.85

Table 1: Summary statistics of simulation 1 and 2. nFPC: number of FPCs used to approximate each curve; Time: running time (in seconds) based on 5000 MCMC iterations; nEdge: total number of edges of the graph averaged across all posterior samples; nUnique: number of unique graphs visited after the burnin period; MisR: mean mis-estimation rate with respect to the true graph; Sen: sensitivity; Spec: specificity; FDGM-S: the proposed functional data graphical model for smooth data, based on Algorithm 1; FDGM-N: the proposed functional data graphical model for noisy data, based on Algorithm 2; GGM-MH: Gaussian graphical model; GLASSO: graphical LASSO; MNGM: matrix-normal graphical model.

In Table 1, the mis-estimation rate is defined as the proportion of mis-estimated edges, obtained by averaging across all posterior samples. The sensitivity is the proportion of missed edges among the true edges, and the specificity is the proportion of over-estimated edges among the true non-edge pairs. The top panel of Table 1 shows that the proposed functional data graphical model provides the smallest mis-estimation rate as well as the highest sensitivity and specificity. We also observe that, although relying on excessive dimension reduction, the Gaussian graphical model and the GLASSO still provide reasonably good estimates. This suggests that for problems involving more nodes (>50), we can use these methods to obtain an initial estimate before applying our approach.

4.2 Simulation 2: Graph Estimation for Noisy Functional Data

We add Gaussian white noise to the functional data generated in Simulation 1 to demonstrate the performance of posterior inference for noisy data. The variances of the additive Gaussian white noise $\{\varepsilon_{ijt}, t \in \mathbf{t}_j\}$ are generated from a gamma distribution with mean 2.5 and variance 0.25, resulting in a signal-to-noise ratio around 9, where the signal-to-noise ratio is defined by $f_{ij}(t)/\text{var}\{\varepsilon_{ijt}\}$ and is averaged across the grid points and the samples. We apply model (11) and generate posterior samples using Algorithm 2. The eigenbasis and the variance of the noise are estimated simultaneously using the PACE algorithm. The parameter $\mathbf{\Lambda}$ is determined using the estimated variance of the Gaussian white noise, and the other model parameters are set to be the same as in Simulation 1. The posterior infer-

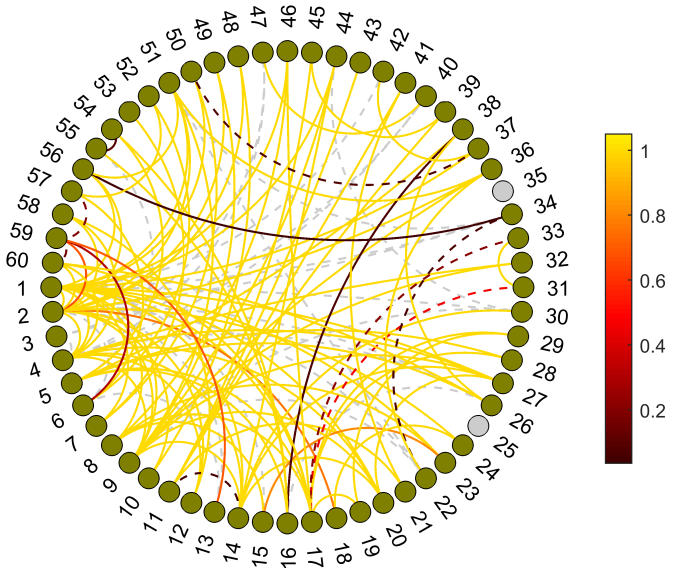

Figure 2: Plot of Simulation 3. The estimated graph based on the marginal inclusion probability for each edge.

ence results are compared with the other three methods in the bottom panel of Table 1. Similar patterns are observed as in Simulation 1. In particular, the proposed functional data graphical model shows a clear advantage in accurately estimating the graph. Estimates of the functions $\{f_{ij}\}$ and their time-domain correlations are provided in the online appendix.

4.3 Simulation 3: Graph Estimation When p is Greater than n

To further investigate the performance of the proposed approach when the number of nodes p is greater than the sample size n, we design another simulation study with $p = 60$ and $n = 55$. The true graph contains 60 nodes, among which 2 are singletons and 58 are connected with edges. The total number of edges in the true graph is 121. Smooth functional data are simulated following the procedure described in Section 4.1. With the simulated data, we apply the PACE algorithm to estimate $\{\mathbf{c}_i\}$ and determine the truncation parameters using the FVE criterion with a 95% threshod. We then apply Algorithm 1 and set prior parameters δ and $\mathbf{U}$ following Simulation 1. Posterior samples of the graph are obtained for $30,000$ MCMC iterations after removing $10,000$ burn-in samples.

The posterior inference results are summarized in a circular graph plot in Figure 2, where we show an estimated graph by thresholding the marginal inclusion probability for each edge—the proportion that each edge is included in the posterior samples—to be greater than 0.03. In Figure 2, the colors indicate the levels of the marginal inclusion probabilities, the colored dashed lines indicate edges that are mistakenly estimated, and the gray dashed lines indicate edges that are missed. This gives 105 estimated edges, among which 98 are correctly estimated, and 7 are mistakenly estimated. Additionally, 23 edges in the true graph are missed. We have also calculated the summary statistics similarly as in previous

simulations, resulting in mean mis-estimation rate 0.02, sensitivity 0.77, and specificity 0.99. Extra simulation runs show that the sensitivity level is improved when we increase the sample size n.

5. Analysis of Event-related Potential Data in an Alcoholism Study

We apply the proposed method to event-related potential data from an alcoholism study. Data were initially obtained from 64 electrodes placed on subjects' scalps that captured EEG signals at 256 Hz during a one-second period. The measurements were taken from 122 subjects, of which 77 belonged to the alcoholism group and 45 to the control group. Each subject completed 120 trials. During each trial, the subject was exposed to either a single stimulus (a single picture) or two stimuli (a pair of pictures) shown on a computer monitor. We band-pass filtered the EEG signals to extract the α frequency band in the range of 8–12.5 Hz. The filtering was performed by applying the `eegfilt` function in the EEGLAB toolbox of Matlab. The α-band signal is known to be associated with inhibitory control (Knyazev, 2007). Research has shown that, relative to control subjects, alcoholic subjects demonstrate unstable or poor rhythm and lower signal power in the α-band signal (Porjesz et al., 2005; Finn and Justus, 1999), indicating decreased inhibitory control (Sher et al., 2005). Moreover, regional asymmetric patterns have been found in alcoholics—alcoholics exhibit lower left α-band activities in anterior regions relative to right (Hayden et al., 2006). In this study, we aim to estimate the conditional independence relationships of α-band signals from different locations of the scalp, and expect to find evidence that reflects differences in brain connectivity and asymmetric pattern between the two groups.

Since multiple trials were measured over time for each subject, the EEG measurements may not be treated as independent due to the time dependence of the trials. Furthermore, since the measurements were taken under different stimuli, the signals could be influenced by different stimulus effects. To remove the potential dependence between the measurements and the influence of different stimulus types, for each subject, we averaged the band-filtered EEG signals across all trials under the single stimulus, resulting in one *Event-related potential* (ERP) curve per electrode per subject. ERP is a type of electrophysiological signal generated by averaging EEG segments recorded under repeated applications of a stimulus, with the averaging serving to reduce biological noise levels and enhance the stimulus evoked neurological signal (Brandeis and Lehmann, 1986; Bressler, 2002). Based on the preprocessed ERP curves, we further removed subjects with missing nodes, and balanced the sample size across the two groups, producing multivariate functional data with $n = 44$ and $p = 64$ for both the alcoholic and the control group. We applied model (4) using coefficients of the eigenbasis expansion. The number of eigenbasis $\{M_j\}$ was determined through retaining 90% of the total variation; this resulted in 4–7 coefficients per f_j. We collected $30,000$ posterior samples using Algorithm 1, in which the first $10,000$ were treated as the burn-in period. The model was fitted for both the alcoholic and the control group, and convergence of the MCMC was justified by running multiple chains starting with various initial values.

The posterior results are summarized in Figure 3. The plots in (a) and (b) show the marginal inclusion probabilities for edges in the alcoholic and the control group respectively, where the edge color indicates the proportion that each edge is included in the posterior

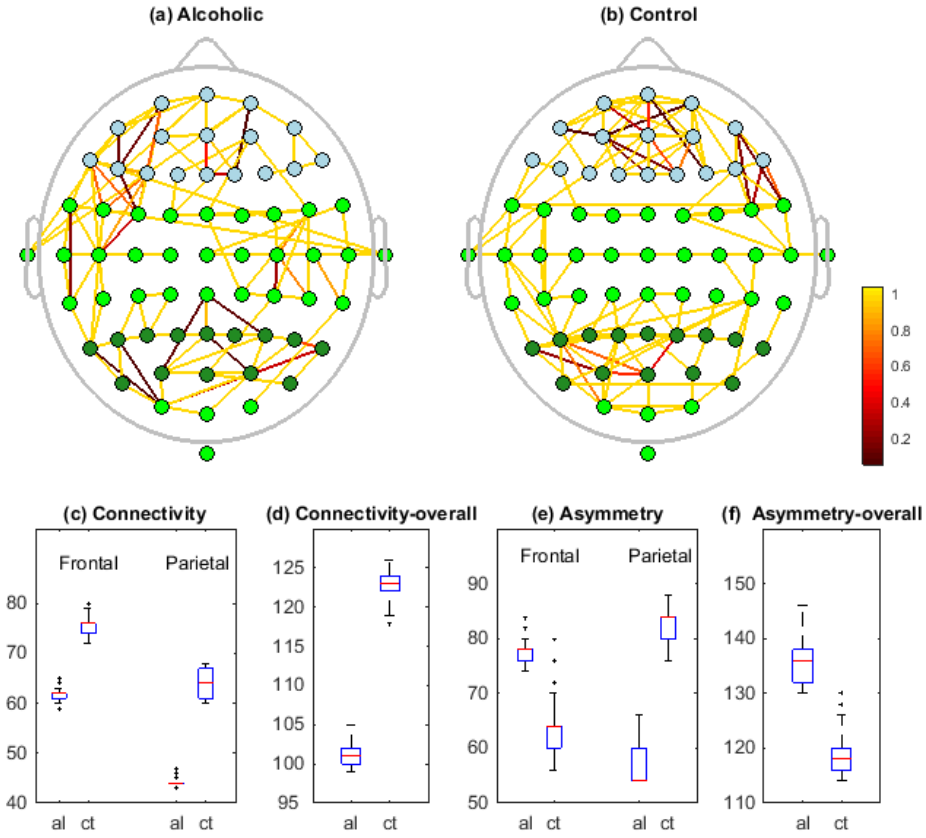

Figure 3: Summary of posterior inference: the marginal inclusion probabilities for edges in the alcoholic group (a) and the control group (b); the boxplots of connectivity measures: the number of edges connecting with nodes in the frontal and parietal regions (c), and the overall total number of edges (d); the boxplots of asymmetry measures: the number of asymmetric edges for nodes in the frontal and the parietal regions (e), and the overall total number of asymmetric edges (f). In (a) and (b), the edge color indicates the magnitude of the posterior inclusion probability. In (c)–(f), the alcoholic group is abbreviated as "al", and the control group is abbreviated as "ct".

samples. To distinguish different regions, we used light blue to highlight nodes in the frontal region, used dark green to highlight nodes in the parietal region, and used green to indicate nodes in the central and occipital regions. Comparing (a) with (b), we see that the alcoholic group contains more edges connecting the left frontal-central, right central, and right parietal regions than the control group. The control group, on the other hand, contains more edges connecting the middle and right frontal regions, as well as the left parietal region than the alcoholic group.

To further compare with established results, we calculated two summary statistics for connectivity: the number of edges connected with nodes in a specific region, and the overall total number of edges. We also calculated two additional summary statistics for asymmetry: the number of asymmetric edges for all nodes in a specific region, and the overall total number of asymmetric edges. We summarized these summary statistics across the two groups using boxplots in Figure 3 (c)–(f), and calculated the posterior probability that the alcoholic group is greater than, equal to, or less than the control group for each statistic. Results show that, with probability ≈ 1, the alcoholic group has fewer edges than the control group in the frontal and the parietal region, and has fewer overall total number of edges; with probability 0.95, the alcoholic group has more asymmetric edges than the control group in the frontal region; and with probability ≈ 1, the alcoholic group has higher overall total number of asymmetric edges than the control group. These results indicate that the alcoholic group exhibits decreased regional and overall connectivity, increased asymmetry in the frontal region, and increased overall asymmetry. These observations are consistent with the findings of Hayden et al. (2006), who studied the asymmetric patterns at two frontal electrodes (F3, F4) and two parietal electrodes (P3, P4) using the analysis of variance method based on the resting-state α-band power. In comparison, our analysis provides connectivity and asymmetric pattern of all 64 electrodes simultaneously whereas Hayden et al. (2006) only focuses on the four representative electrodes.

6. Discussion

We have constructed a theoretical framework for graphical models of multivariate functional data and proposed a HIWP prior for the special case of Gaussian process graphical models. For practical implementation, we have suggested a posterior inference approach based on a regularization condition, which enables posterior sampling through MCMC algorithms.

One concern is whether it is possible to perform exact posterior inference without the regularity condition on approximation, i.e., inferring the graph directly from the joint posterior $p(G|\{\mathbf{c}_i\}) \propto p(\{\mathbf{c}_i\}|G)p(G)$ based on model (4), where $p(\{\mathbf{c}_i\}|G)$ is the marginal likelihood (with the covariance kernel $\mathcal{Q}_{\mathcal{C}}$ integrated out) and $p(G)$ is the prior distribution for G. Although the above joint posterior is theoretically well-defined according to Theorem 2, exact posterior sampling is difficult due to the fact that the density function for the marginal likelihood can only be calculated on a finite dimensional projection of $\{\mathbf{c}_i\}$.

In posterior inference, the influence of the approximation error on the posterior distribution can be quantified empirically. Assuming that the functional data are pre-smoothed, the approximation error can be quantified by calculating the difference of the ℓ^2 norms between the full sequence and the truncated sequence. The influence on the posterior distribution can be quantified by measuring the sensitivity of the posterior distribution to the change of truncation (Saltelli et al., 2000). For example, based on model (4) one may calculate the Kullback-Leibler divergence for two different truncation parameters M and M'. An alternative method for pre-determining the truncation parameter is to choose a prior for M in a Bayesian hierarchical model, in which case hybrid MCMC algorithms are needed for fitting both models (4) and (11). The posterior sampling in these models would become

more complicated because the dimension of the truncated sequences and the size of the covariance matrix $\mathbf{Q}_{\mathcal{C}}$ would change whenever M is updated.

We have demonstrated the application of the proposed approach through an ERP data set. By treating ERPs as functional data, we are estimating the systematic brain connectivity that is common across a group of subjects and a time interval. For other modeling purposes, such as estimating the individual level or dynamic brain connectivity, one could use multivariate graphical models described in Carvalho and West (2007) or Bilmes (2010).

We have focused on decomposable graphs. In case of non-decomposable graphs, the proposed HIWP prior may still apply if we replace the inverse-Wishart process prior for each clique with that for a prime component of the graph. For a non-complete prime component P, the inverse-Wishart processes prior for $\mathbf{Q}_P$ is subject to extra constraint induced by missing edges.

We have applied the proposed method to graphs of small to moderate size, with number of nodes as large as 60. To deal with larger scale problems (e.g, multivariate functional data with hundreds or thousands of functional components), more efficient large-scale computational techniques such as the fast Cholesky factorization (Li et al., 2012) can be readily combined with our MCMC algorithms. Furthermore, non-MCMC algorithms may be more computationally efficient in case of large graphs. For example, based on the posterior distribution of G in (8), a fast search algorithm may be developed to search for the maximum a posteriori (MAP) solution following ideas similar to Daumé III (2007) and Jalali et al. (2011).

Acknowledgments

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Hongxiao Zhu's research is supported by National Science Foundation (NSF-DMS 1611901). David B. Dunson's research is supported by Office of Naval Research (N00014-14-1-0245).

Appendix A. Definitions

Definitions used in the lemmas, theorems and their proofs are listed as follows: (I) *Projection map.* Let $\mathbb{R}$ be the real line and T be an index set. Consider the Cartesian product space $\mathbb{R}^{T\times T} = \prod_{(\alpha,\beta)\in T\times T}\mathbb{R}^{(\alpha,\beta)}$. For a fixed point $(\alpha,\beta)\in T\times T$, we define the projection map $\pi_{(\alpha,\beta)} : \mathbb{R}^{T\times T}\to\mathbb{R}^{(\alpha,\beta)}$ as $\pi_{(\alpha,\beta)}\left(\{x_{(l,m)} : (l,m)\in T\times T\}\right) = x_{(\alpha,\beta)}$. For a subset $B\subset T\times T$, we define the partial projection $\pi_B : \mathbb{R}^{T\times T}\to\mathbb{R}^B$ as $\pi_B\left(\{x_{(l,m)} : (l,m)\in T\times T\}\right) = \{x_{(s,t)} : (s,t)\in B\}$. More generally, for subsets B_1, B_2, such that $B_2\subset B_1\subset T\times T$, we define the partial sub-projections $\pi_{B_2\leftarrow B_1} : \mathbb{R}^{B_1}\to\mathbb{R}^{B_2}$, by $\pi_{B_2\leftarrow B_1}(\{x_{(l,m)} : (l,m)\in B_1\}) = \{x_{(s,t)} : (s,t)\in B_2\}$. (II) *The pullback of a σ-algebra.* Let $\mathcal{B}_{(\alpha,\beta)}$ be a σ-algebra on $\mathbb{R}^{(\alpha,\beta)}$. We can create a σ-algebra on $\mathbb{R}^{T\times T}$ by pulling back the $\mathcal{B}_{(\alpha,\beta)}$ using the inverse of the projection map and define $\pi^*_{(\alpha,\beta)}(\mathcal{B}_{(\alpha,\beta)}) = \{\pi^{-1}_{(\alpha,\beta)}(A) : A\in\mathcal{B}_{(\alpha,\beta)}\}$. One can verify that $\pi^*_{(\alpha,\beta)}(\mathcal{B}_{(\alpha,\beta)})$ is a σ-algebra. (III) *Product σ-algebra.* We define the product σ-algebra as $\mathcal{B}(\mathbb{R}^{T\times T}) = \prod_{(\alpha,\beta)\in T\times T}\mathcal{B}_{(\alpha,\beta)}$, where $\prod_{(\alpha,\beta)\in T\times T}\mathcal{B}_{(\alpha,\beta)} = \sigma\left(\bigcup_{(\alpha,\beta)\in T\times T}\pi^*_{(\alpha,\beta)}(\mathcal{B}_{(\alpha,\beta)})\right)$. (IV) *Pushforward measure.* Given a measure $\mu_{T\times T}$ on the product σ-algebra, and a subset B of $T\times T$, we define the push-forward measure $\mu_B = (\pi_B)_*\mu_{T\times T}$ on $\mathbb{R}^B$ as $\mu_B(A) = \mu_{T\times T}\{\pi_B^{-1}(A)\}$ for all $A\in\mathcal{B}_B$, where $\mathcal{B}_B = \prod_{(\alpha,\beta)\in B}\mathcal{B}_{(\alpha,\beta)}$. (V) *Compatibility.* Given subsets B_1, B_2 of $T\times T$ such that $B_2\subset B_1\subset T\times T$, the pushforward measures μ_{B_1} and μ_{B_2} are said to obey compatibility relation if $(\pi_{B_2\leftarrow B_1})_*\mu_{B_1} = \mu_{B_2}$.

Appendix B. Proof of Lemma 1

This proof involves some measure-theoretic arguments. The essential idea is to use disintegration theory Chang and Pollard (1997) to first construct the conditional probability measure $P_1\{\cdot \mid \pi_{A\cap B}(\mathbf{f}_A)\}$ on $\mathcal{B}(L^2(T_A))$, extend this to $P\{\cdot \mid \pi_B(\mathbf{f})\}$ on $\mathcal{B}(L^2(T_{A\cup B}))$, and finally construct the joint measure P which satisfies conditions (i)–(iii).

Denote $T_A = \bigsqcup_{j\in A}T_j$. Since P_1 is a finite Radon measure and the projection $\pi_{A\cap B} : L^2(T_A)\to L^2(T_{A\cap B})$ is measurable, we invoke the disintegration theorem to obtain measures $P_1\{\cdot \mid \pi_{A\cap B}(\mathbf{f}_A)\}$ on $\mathcal{B}(L^2(T_A))$ satisfying:

(a.1) $P_1(\mathcal{X} \mid \mathbf{f}_{A\cap B}) = P_1\left\{\mathcal{X}\cap[L^2(T_{A\setminus B})\times\{\pi_{A\cap B}(\mathbf{f}_A)\}] \mid \pi_{A\cap B}(\mathbf{f}_A)\right\}$ for all $\mathcal{X}\in\mathcal{B}(L^2(T_A))$,

(b.1) the map $\mathbf{f}_{A\cap B}\mapsto(P_1)_{\mathbf{f}_{A\cap B}}H := \int H(\mathbf{f}_A)dP_1(\mathbf{f}_A \mid \mathbf{f}_{A\cap B})$ is measurable for all non-negative measurable $H : L^2(T_A)\to\mathbb{R}$,

(c.1) $P_1H = ((\pi_{A\cap B})_*P_1)(P_1)_{\mathbf{f}_{A\cap B}}H$ for all nonnegative measurable $H : L^2(T_A)\to\mathbb{R}$, where $(\pi_{A\cap B})_*P_1$ is the push-forward measure of P_1.

Now, we define the measure $P\{\cdot \mid \pi_B(\mathbf{f})\}$ by setting $P\{\mathcal{A} \mid \pi_B(\mathbf{f})\} = P_1\{\pi_A(\mathcal{A}\cap[L^2(T_{A\setminus B})\times\{\pi_B(\mathbf{f})\}]) \mid \pi_{A\cap B}(\mathbf{f})\}$. Note that this is well defined for all measurable $\mathcal{A}\in\mathcal{B}(L^2(T_{A\cup B}))$ since the sections $\pi_A(\mathcal{A}\cap[L^2(T_{A\setminus B})\times\{\pi_B(\mathbf{f})\}])$ are always measurable, and also that (a) $P\{\mathcal{A} \mid \pi_B(\mathbf{f})\} = P\{\mathcal{A}\cap[L^2(T_{A\setminus B})\times\{\pi_B(\mathbf{f})\}] \mid \pi_B(\mathbf{f})\}$ holds by construction. Now, let $\mathcal{M}$ denote the set of measurable functions from $L^2(T_{A\cup B})$ to $\mathbb{R}$ satisfying (b) $\mathbf{f}_B\longmapsto P_{\mathbf{f}_B}H$ is a measurable function on $L^2(T_B)$. We shall argue that $\mathcal{M}$ is a monotone class. First,

suppose H_n is a sequence of positive measurable functions in $\mathcal{M}$ increasing pointwise to a bounded measurable function H. For each fixed $\mathbf{f}_B$ in $L^2(T_B)$, we then have that H_n is a sequence of positive measurable functions increasing pointwise to H, and hence the monotone convergence theorem implies $P_{\mathbf{f}_B}H_n \longrightarrow P_{\mathbf{f}_B}H$ in an increasing manner. Since this holds for each $\mathbf{f}_B$, we conclude that $P_{\mathbf{f}_B}H$ is the point-wise increasing limit of measurable functions on $L^2(T_B)$, and hence it is measurable. Moreover, it is simple to see that $P_{\mathbf{f}_B}\mathbf{1}_{\mathcal{X}\times\mathcal{Y}} = P_1(\mathcal{X} \mid \mathbf{f}_{A\cap B})\mathbf{1}_{\mathcal{Y}}(\mathbf{f}_{B\setminus A})$ is a measurable function on $L^2(T_B)$ for all $\mathcal{X} \in \mathcal{B}(L^2(T_A))$ and $\mathcal{Y} \in \mathcal{B}(L^2(T_{B\setminus A}))$, and hence $\mathbf{1}_{\mathcal{X}\times\mathcal{Y}} \in \mathcal{M}$. By the Monotone Class Theorem, we then have that all bounded measurable functions on $L^2(T_{A\cup B})$ satisfy (b), and hence it will hold for all positive measurable functions on $L^2(T_{A\cup B})$. Since (b) is satisfied for all positive measurable functions, we may define the measure $PH = P_2P_{\mathbf{f}_B}H$. By construction, we have that $P\mathbf{1}_{L^2(T_{A\setminus B})\times\mathcal{Y}} = P_2P_1(L^2(T_{A\setminus B}) \times \{\mathbf{f}_{A\cap B}\} \mid \mathbf{f}_{A\cap B})\mathbf{1}_{\mathcal{Y}}(\mathbf{f}_B) = P_2(\mathcal{Y})$ and $P\mathbf{1}_{\mathcal{X}\times L^2(T_{B\setminus A})} = P_2P_1(\mathcal{X} \mid \mathbf{f}_{A\cap B}) = ((\pi_{A\cap B})_*P_2)P_1(\mathcal{X} \mid \mathbf{f}_{A\cap B}) = ((\pi_{A\cap B})_*P_1)P_1(\mathcal{X} \mid \mathbf{f}_{A\cap B}) = P_1(\mathcal{X})$. Thus, we also have that $PH = P_2P_{\mathbf{f}_B}H = ((\pi_B)_*P)P_{\pi_B(\mathbf{f})}H$ for all measurable H, and this is the final property establishing that $P(\cdot \mid \mathbf{f}_B)$ is a disintegration of P with respect to the map π_B. By the disintegration theorem, this disintegration is a version of the regular conditional probability of $\mathbf{f}_A$ given $\mathbf{f}_B$. Since this version only depends upon $\mathbf{f}_{A\cap B}$, we conclude that (iii) holds. Finally, we note that any other measure satisfying these properties must agree with the measure we have constructed on π-system, and therefore the uniqueness of P immediately follows. ■

Appendix C. Proof of Proposition 1

Proof. The Properties 1 - 4 in Dawid and Lauritzen (1993) are treated as axioms; they are universal properties thus also hold when X, Y, Z are random processes. Since the graph G is undirected and decomposable, the results on graphical theory in Appendix A of Dawid and Lauritzen (1993) continue to hold. Properties 1 - 4 and results in Appendix A imply that results in B1- B7 of Dawid and Lauritzen (1993) continue to hold when P is a Markov distribution constructed in Lemma 1. Theorem 2.6 and Corollary 2.7 of Dawid and Lauritzen (1993) are also implied. These results, combined with the definition of marginal distribution defined by pushforward measure and the definition of conditional probability measure based on disintegration theory, prove that Lemmas 3.1, 3.3, Theorems 3.9 - 3.10 as well as Propositions 3.11, 3.13, 3.15, 3.16, 3.18 from Dawid and Lauritzen (1993) hold. ■

Appendix D. Lemma 2 and Proof

Lemma 2 Let $\mathbb{N}$ be the set of positive integers and I an arbitrary finite subset of it. Suppose that $\delta > 4$ is a positive integer and that $u : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ is a symmetric positive semidefinite and trace class kernel so that the matrix $\mathbf{U}_{I\times I}$ formed by $\{u(i,j), i,j \in I\}$ is symmetric positive semidefinite. Then there exists a unique probability measure μ on $(\mathbb{R}^{\mathbb{N}\times\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}\times\mathbb{N}}))$ satisfying

i. $(\pi_{I\times I})_*\mu = \mu_{I\times I}$, where $\mu_{I\times I}$ is the law of $\mathrm{IW}(\delta, \mathbf{U}_{I\times I})$ defined in Dawid (1981);

ii. if $B = \{(\alpha_i, \beta_i)\}_{i=1}^n \subset \mathbb{N}\times\mathbb{N}$ and $\mathrm{g} = \{\alpha_i\}_{i=1}^n \cup \{\beta_i\}_{i=1}^n$, then $(\pi_B)_*\mu = \mu_B$, where $\mu_B = (\pi_{B\leftarrow \mathrm{g}\times\mathrm{g}})_*\mu_{\mathrm{g}\times\mathrm{g}}$.

Setting $\mu = \text{IWP}(\delta, \mathbf{U})$ so that $(\mathbf{U})_{ij} = u(i,j)$, we further have that if $\mathbf{Q} \sim \text{IWP}(\delta, \mathbf{U})$ and $\delta > 4$, the countably infinite array $\mathbf{Q}$ is a positive semidefinite trace class operator on $\ell^2(\mathbb{N})$ almost surely.

Proof. Let $\mathbf{U}_{I\times I}$ be a matrix with the law $\mu_{I\times I}$. We will prove following Tao (2011, Theorem 2.4.3) as follows: (1) we verify the compatibility of μ_B for all finite $B \subset \mathbb{N}\times\mathbb{N}$. There are two successive cases we shall consider. Case 1: Suppose $I_2 \subset I_1$ are two finite subsets of $\mathbb{N}$, then $\mathbf{Q}_{I_2\times I_2}$ is the sub-matrix of $\mathbf{Q}_{I_1\times I_1}$ obtained by deleting the rows and columns with indices in $I_1 \setminus I_2$. If $\mathbf{Q}_{I_1\times I_1}$ has law $\mu_{I_1\times I_1} = \text{IW}(\delta, \mathbf{U}_{I_1\times I_1})$, then $\mathbf{Q}_{I_2\times I_2}$ has law $\text{IW}(\delta, \mathbf{U}_{I_2\times I_2})$ due to the consistency property of the inverse-Wishart distribution (Dawid and Lauritzen, 1993, Lemma 7.4). Consequently, $(\pi_{I_2\times I_2 \leftarrow I_1\times I_1})_*\mu_{I_1\times I_1} = \mu_{I_2\times I_2}$. Case 2: Let $B_1 = \{(\alpha_i, \beta_i)\}_{i=1}^n \subset \mathbb{N}\times\mathbb{N}$ and suppose $B_2 = \{(\widetilde{\alpha}_i, \widetilde{\beta}_i)\}_{i=1}^m \subset B_1$. Set $\mathrm{g}_1 = \{\alpha_i\}_{i=1}^n \cup \{\beta_i\}_{i=1}^n$ and $\mathrm{g}_2 = \{\widetilde{\alpha}_i\}_{i=1}^m \cup \{\widetilde{\beta}_i\}_{i=1}^m$ so that $\mathrm{g}_2\times\mathrm{g}_2 \subset \mathrm{g}_1\times\mathrm{g}_1$. It is clear that $\pi_{B_2\leftarrow B_1} \circ \pi_{B_1\leftarrow \mathrm{g}_1\times\mathrm{g}_1} = \pi_{B_2\leftarrow \mathrm{g}_1\times\mathrm{g}_1} = \pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2} \circ \pi_{\mathrm{g}_2\times\mathrm{g}_2\leftarrow \mathrm{g}_1\times\mathrm{g}_1}$. Thus,

$$\begin{aligned}(\pi_{B_2\leftarrow B_1})_*\mu_{B_1} &= (\pi_{B_2\leftarrow B_1})_*(\pi_{B_1\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_*\mu_{\mathrm{g}_1\times\mathrm{g}_1} = (\pi_{B_2\leftarrow B_1}\circ\pi_{B_1\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_*\mu_{\mathrm{g}_1\times\mathrm{g}_1}\\ &= (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2}\circ\pi_{\mathrm{g}_2\times\mathrm{g}_2\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_*\mu_{\mathrm{g}_1\times\mathrm{g}_1} = (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2})_*(\pi_{\mathrm{g}_2\times\mathrm{g}_2\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_*\mu_{\mathrm{g}_1\times\mathrm{g}_1}\\ &= (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2})_*\mu_{\mathrm{g}_2\times\mathrm{g}_2} = \mu_{B_2},\end{aligned}$$

where the second to last equality holds because of our demonstration in Case 1. (2) Second, we claim that the finite dimensional measure $\mu_{I\times I} = \text{IW}(\delta, \mathbf{U}_{I\times I})$ is an inner regular probability measure on the product σ-algebra $\mathcal{B}_{I\times I}$. We will show that $\mu_{I\times I}$ is a finite Borel measure on a Polish space, which then implies that $\mu_{I\times I}$ is regular, hence inner regular by Bauer (2001, Lemma 26.2). This is done through (a)–(c) as follows: (a) For finite I, $\mathbf{Q}_{I\times I}$ takes values in the space of symmetric and positive semidefinite matrices, denoted by $\Psi_{|I|}$ where $|N|$ denotes the number of elements in I. Since the subset of symmetric matrices is closed in $\mathbb{R}^{I\times I}$, it is Polish. Furthermore, the space of symmetric positive semidefinite matrices is an open convex cone in the space of symmetric matrices, hence it is Polish as well. Therefore the space $\Psi_{|I|}$ is Polish. (b) Since $\mu_{I\times I}$, the law of $\mathbf{Q}_{I\times I} \sim \text{IW}(\delta, U_{I\times I})$, has an almost everywhere continuous density function, $\mu_{I\times I}$ is a measure defined by Lebesgue integration against an almost everywhere continuous function. Therefore $\mu_{I\times I}$ is Borel on $\Psi_{|I|}$. As $\Psi_{|I|} \subset \mathbb{R}^{I\times I}$, we may extend the measure $\mu_{I\times I}$ from $\Psi_{|I|}$ to $\mathbb{R}^{I\times I}$ via the Carathéodory theorem (Tao, 2011, Theorem 1.7.3). In particular, define $\widetilde{\mu}_{I\times I}(A) = \mu_{I\times I}(A \cap \Psi_{|I|})$ for $A \in \mathcal{B}(\mathbb{R}^{I\times I})$. With extension, $\mu_{I\times I}$ is Borel on $\mathbb{R}^{I\times I}$, and the σ-algebra associated is $\mathcal{B}(\mathbb{R}^{I\times I}) = \mathcal{B}_{I\times I} = \prod_{(\alpha,\beta)\in I\times I} \mathcal{B}_{(\alpha,\beta)}$. (c) The measure $\mu_{I\times I}$ is certainly finite since it is a probability measure.

The compatibility and regularity conditions in (1) and (2) ensure that the Kolmogorov extension theorem holds. Therefore there exists a unique probability measure μ on the product σ-algebra $\mathcal{B}(\mathbb{R}^{\mathbb{N}\times\mathbb{N}})$ that satisfies (i) and (ii).

We now prove that if $\mathbf{Q} \sim \text{IWP}(\delta, \mathbf{U})$, then the countably infinite array $\mathbf{Q}$ is a well-defined positive semidefinite trace class operator on $\ell^2(\mathbb{N})$ almost surely. First, we note that the spectral theorem ensures the existence of an orthonormal basis of $\ell^2(\mathbb{N})$ that diagonalizes U. Thus, without loss of generality, we may assume that $\mathbf{Q}$ is drawn from $\text{IWP}(\delta, \mathbf{U})$ where $\mathbf{U}$ is a diagonal positive semidefinite trace class operator on $\ell^2(\mathbb{N})$.

First, we show each row of $\mathbf{Qx}$ is finite almost surely hence is well-defined for all $\mathbf{x} \in \ell^2(\mathbb{N})$. It is sufficient to show that $E[\,|(\mathbf{Qx})_i|\,] < \infty$. We note that for arbitrary $i \neq j$,

$\begin{pmatrix} q_{ii} & q_{ij} \\ q_{ij} & q_{jj} \end{pmatrix} \sim \text{IW}\left(\delta, \begin{pmatrix} u_{ii} & 0 \\ 0 & u_{jj} \end{pmatrix}\right)$ and hence using the moments of finite dimensional inverse-Wishart, $E(q_{ii}^2) = u_{ii}^2(\delta-2)^{-1}(\delta-4)^{-1}, E(q_{ij}^2) = u_{ii}u_{jj}(\delta-1)^{-1}(\delta-2)^{-1}(\delta-4)^{-1}$, for $\delta > 4$. By Tonelli's theorem, we have that $E\sum_j q_{ij}^2 = \sum_j Eq_{ij}^2 \leq C\sum_j u_{ii}u_{jj} = Cu_{ii}\sum_j u_{jj}$, where C is the maximum of the above constants. Thus

$$E[\,|(\mathbf{Q}\mathbf{x})_i|\,] \leq \|\mathbf{x}\|\sqrt{E\sum_j q_{ij}^2} < \infty.$$

Because there are only countably many rows, we have that $\mathbf{Q}\mathbf{x}$ is finite almost surely for all rows simultaneously. Consequently, we have that $\mathbf{Q}\mathbf{x}$ is well-defined for all $\mathbf{x} \in \ell^2(\mathbb{N})$. Now we show that $\mathbf{Q}\mathbf{x} \in \ell^2(\mathbb{N})$ almost surely. By similar considerations, let $\mathbf{q}_i = (\mathbf{Q}\mathbf{x})_i$, then $E(\sum_i \|\mathbf{q}_i\|^2) \leq C\left(\sum_i u_{ii}\right)^2 < \infty$ and $\|\mathbf{Q}\mathbf{x}\|^2 \leq C\|\mathbf{x}\|^2\sum_i \|q_i\|^2$; this implies that $\|\mathbf{Q}\mathbf{x}\| < \infty$ almost surely hence $\mathbf{Q}\mathbf{x} \in \ell^2(\mathbb{N})$ almost surely, and it also implies that the operator norm $\|\mathbf{Q}\|_{op}$ is finite almost surely.

By construction, we must have that $\mathbf{Q}$ is positive semidefinite almost surely since $\langle \mathbf{Q}\mathbf{x}, \mathbf{x}\rangle = \lim_{n\to\infty}\langle \mathbf{Q}_n\mathbf{x}, \mathbf{x}\rangle \geq 0$, where $\mathbf{Q}_n$ is the restriction of $\mathbf{Q}$ to its n by n leading principal minor. Finally, $\mathbf{Q}$ is trace class almost surely since $E[\,|\text{tr}(\mathbf{Q})|\,] = \sum_i E(q_{ii}) = (\delta-2)^{-1}\sum_i u_{ii} < \infty$. ■

Appendix E. Proof of Theorem 1

Proof. Based on Lemma 2, we can define a sequence of inverse-Wishart process prior for $\mathcal{Q}_C$, denoted by $\mathcal{Q}_C \sim \text{IWP}(\delta, \mathcal{U}_C), C \in \mathcal{C}$. These sequences are pairwise consistent due to the consistency of inverse-Wishart processes and the fact that $\mathcal{U}_\mathcal{C}$ is a common collection of kernels. Therefore, we can construct a unique hyper Markov law for $\mathcal{Q}_\mathcal{C}$ following procedure (12) - (13) of Dawid and Lauritzen (1993). And Theorem 3.9 of Dawid and Lauritzen (1993) guarantees that the constructed hyper Markov law is unique. ■

Appendix F. Proof of Proposition 2

Proof. Note that an operator drawn from a hyper-inverse-Wishart process with the parameter $\mathcal{U}$ satisfies $\text{rank}(u_{ij}) < \infty$ for $i, j \in V$ will have finite-rank almost surely. This follows by noting that if $\mathcal{Q} \sim \text{HIWP}(\delta, \mathcal{U})$ and $\mathcal{W}$ is a fixed unitary transformation on ℓ^2, then $\mathcal{W}^T\mathcal{Q}\mathcal{W} \sim \text{HIWP}(\delta, \mathcal{W}^T\mathcal{U}\mathcal{W})$. Thus, choosing $\mathcal{W}$ so that the block representation $\mathcal{W}^T\mathcal{U}\mathcal{W} = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$ holds (here, U is a finite matrix and 0's represent infinite arrays of zeros), we see that the block representation $\mathcal{W}^T\mathcal{Q}\mathcal{W} = \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix}$ holds almost surely, and that $Q \sim \text{IW}(\delta, U)$. Consequently, we have reduced to the finite-dimensional setting where the result is well-known. ■

Appendix G. Proof of Theorem 2

Proof. By the result of Proposition 1, the HIWP_G prior is a strong hyper Markov law. So by Corollary 5.5 of Dawid and Lauritzen (1993), the posterior law of $\mathcal{Q}_\mathcal{C}$ is the unique

hyper Markov law specified by the marginal posterior laws at each clique. In other words, we just need to find the posterior law for the model: $\mathbf{c}_{i,C} \sim \text{dMGP}(\mathbf{c}_{0,C}, \mathcal{Q}_C)$ with prior $\mathcal{Q}_C \sim \text{IWP}(\delta, \mathcal{U}_C)$ for each $\mathcal{Q}_C$, and use them to construct the posterior law of $\mathcal{Q}_\mathcal{C}$ following (12) - (13) of Dawid and Lauritzen (1993). As in the last proof, choosing an appropriate transformation reduces this to the finite-dimensional case which is well-known. Finally, by Proposition 5.6 of Dawid and Lauritzen (1993), the marginal distribution of $\{\mathrm{c}_i\}$ given $G, \mathbf{c}_0, \delta, \widetilde{\mathcal{U}}_\mathcal{C}$ is again Markov over G. ■

Online Appendix

The online appendix contains more detailed derivations, discussions, and simulation results.

References

A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky. High-dimensional structure estimation in Ising models: local separation criterion. *Ann. Statist.*, 40(3):1346–1375, 06 2012.

H. Bauer. *Measure and Integration Theory*. De Gruyter Studies in Mathematics. W. de Gruyter, 2001.

J. Bilmes. Dynamic graphical models. *IEEE Signal Processing Magazine*, 27(6):29–42, 2010.

D. Brandeis and D. Lehmann. Event-related potentials of the brain and cognitive processes: approaches and applications. *Neuropsychologia*, pages 151–168, 1986.

S. L. Bressler. Event-related potentials. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 412–415. MIT Press, Cambridge MA, 2002.

T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):594–607, 2011.

C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

C. M. Carvalho and M. West. Dynamic matrix-variate graphical models. *Bayesian Anal.*, 2(1):69–98, 2007.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3): 287–317, 1997.

H. Daumé III. Fast search for dirichlet process mixture models. In *Eleventh International Conference on Artificial Intelligence and Statistics (AIStats)*, 2007.

A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981.

A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.*, 21(3):1272–1317, 1993.

C. De Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, Berlin, 2001.

A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

P. R. Finn and A. Justus. Reduced EEG alpha power in the male and female offspring of alcoholics. *Alcohol. Clin. Exp. Res.*, 23:256–262, 1999.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

P. Giudici. Learning in graphical Gaussian models. *Bayesian Statistics 5*, pages 621–628, 1996.

P. Giudici and P. J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

Y. Guan and S. M. Krone. Small-world MCMC and convergence to multi-modal distributions: From slow mixing to fast mixing. *Ann. Appl. Prob.*, 17:284–304, 2007.

Y. Guan, R. Fleissner, P. Joyce, and S. M. Krone. Markov chain Monte Carlo in small worlds. *Stat. Comput.*, 16:193–202, 2006.

E. P. Hayden, R. E. Wiegand, E. T. Meyer, L. O. Bauer, S. J. O'Connor, J. I. Nurnberger, D. B. Chorlian, B. Porjesz, and H. Begleiter. Patterns of regional brain activity in alcohol-dependent subjects. *Alcohol. Clin. Exp. Res.*, 30(12):1986 – 1991, 2006.

H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, 10:883–906, 2009.

A. Jalali, C. C. Johnson, and P. K. Ravikumar. On learning discrete graphical models using greedy methods. In J. Shawe-taylor, R.s. Zemel, P. Bartlett, F.c.n. Pereira, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1935–1943. 2011.

M. Jansen and P. J. Oonincx. *Second Generation Wavelets and Applications*. Springer-Verlag, London, 2015.

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.*, 20(4):388–400, 2005.

Gennady G. Knyazev. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neurosci. Biobehav. Rev.*, 31(3):377 – 395, 2007.

M. Kolar and E. Xing. On time varying undirected graphs. *J. Mach. Learn. Res.*, 15: 407–415, 2011.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278, 12 2009.

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

E. Lei, F. Yao, N. Heckman, and K. Meyer. Functional data model for genetically related individuals with application to cow growth. *J. Comp. and Graph. Stat.*, 2014.

S. Li, M. Gu, C. J. Wu, and J. Xia. New efficient and robust HSS Cholesky factorization of SPD matrices. *SIAM J. Matrix Anal. Appl.*, pages 886–904, 2012.

P-L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.*, 41(6):3022–3049, 12 2013.

P. Ma, C. I. Castillo-Davis, W. Zhong, and J. S. Liu. A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.*, 34(4):1261–1269, 2006.

R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electron. J. Statist.*, 6:2125–2149, 2012a.

R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, 13:781–794, 2012b.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

H. G. Müller and F. Yao. Functional additive models. *J. Am. Statist. Assoc.*, 103:1534–1544, 2008.

B. Porjesz, M. Rangaswamy, C. Kamarajan, K. A. Jones, A. Padmanabhapillai, and H. Begleiter. The utility of neurophysiological markers in the study of alcoholism. *Clin. Neurophysiol.*, 116(5):993 – 1018, 2005.

G. D. Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer, New York, 2006.

X. Qiao, G. James, and J. Lv. Functional graphical models. Technical report, University of Southern California, 2015.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis, Section Edition*. Springer, New York, 2005.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using l1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 06 2010.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc., Series B*, 53:233–243, 1991.

A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Stat.*, 29: 391–411, 2002.

A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity Analysis*. John Wiley & Sons, Ltd., New York, 2000.

J. G. Scott and C. M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist.*, 17(4):790–808, 2008.

K. J. Sher, E.R. Grekin, and N. A. Williams. The development of alcohol use disorders. *Annu. Rev. Clin. Psychol.*, 1:493–523, 2005.

T. Tao. *An Introduction to Measure Theory.* Graduate Studies in Mathematics. Amer. Math. Soc., 2011.

H. Wang and M. West. Bayesian analysis of matrix normal graphical models. *Biometrika*, 96(4):821–834, 2009.

D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical Lasso. *J. Comp. Graph. Stat.*, 20(4):892–900, 2011.

E. Yang, G. Allen, Z. Liu, and P. K. Ravikumar. Graphical models via generalized linear models. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1358–1366. Curran Associates, Inc., 2012.

F. Yao, H. G. Müller, and J. L. Wang. Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.*, 100:577–590, 2005.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

S. Zhou, J. D. Lafferty, and L. A. Wasserman. Time varying undirected graphs. *Mach. Learn.*, 80:295–319, 2010.