Bayesian Graphical Models for Multivariate Functional Data

Hongxiao Zhu\textsuperscript{1}, Nate Strawn\textsuperscript{2}, and David B. Dunson\textsuperscript{3}

\textsuperscript{1} Department of Statistics, Virginia Tech, Blacksburg, VA 24061

\textsuperscript{2} Department of Mathematics, Duke University, Durham, NC 27708

\textsuperscript{3} Department of Statistical Science, Duke University, Durham, NC 27708

November 18, 2014

Author Footnote:

Zhu is Assistant Professor, Department of Statistics, Virginia Tech, Blacksburg VA 24061 (E-mail: hongxiao@vt.edu). Strawn is Visiting Assistant Professor, Department of Mathematics, Duke University, Durham NC 27708 (E-mail: nstrawn@math.duke.edu). Dunson is Arts and Sciences Professor, Department of Statistical Science, Duke University, Durham NC 27708 (E-mail: dunson@duke.edu). Zhu’s research was supported by Institute for Critical Technology and Applied Science, Virginia Tech. Dunson’s research was partly supported by the National Institute of Environmental Health Sciences, U.S.A. This work grew out of the Analysis of Object Data program at the Statistical and Mathematical Sciences Institute.
Abstract

Graphical models express conditional independence relations among variables. Although methods for vector data are well established, functional data graphical models have not yet been considered. 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 composition of likelihoods and priors according to decomposable graphs. A 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 basis-coefficient sequences, establishing existence, uniqueness and conditions for the strong Markov property and conjugacy. Stochastic search Markov chain Monte Carlo algorithms are developed for approximate inference, assessed through simulations, and applied to a study of brain activity and alcoholism.

KEYWORDS: Functional data analysis; Gaussian process; Graphical model; 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 & Lauritzen (1993) define Markov distributions (distributions with Markov property with respect to a graph) of random vectors which can be factorized according to the structure of a graph. They also introduce hyper-Markov laws serving as prior distributions in Bayesian analysis. The special case of Gaussian graphical models is well studied, 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). 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 & Green 1999; Roverato 2002; Jones et al. 2005; Scott & Carvalho 2008; Carvalho & Scott 2009).

The graphical modeling literature focuses primarily on vector-valued data in which each node corresponds to one variable. Many applications, however, involve functional data objects. 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 regions of interest). Although there is an increasingly rich literature on generalizations to accommodate matrix-variate graphical models (Wang & West 2009) and dynamic linear models (Carvalho & West 2007), the generalization to functional data case remains untouched.

In functional data analysis literature, there are a number of articles focusing on functional data with correlations induced by nested design (Morris & Carroll 2006; Rosen & Thompson 2009), but little consideration has been given to the conditional independence of random functional objects. In this paper, we focus on developing Bayesian graphical models for inferring conditional independence structures in multivariate functional data. Previous work on graphical models has only examined distributions on finite-dimensional metric spaces where many measure-theoretic issues are trivial. Since we must deal with infinite dimensional distributions, we provide a full measure-theoretic analysis of the constructions and properties. In particular, we extend Markov distributions and hyper
Markov laws from the random variable to the random process case, which facilitates a Bayesian framework for graphical modeling. We then demonstrate the special case of multivariate Gaussian processes in the space of square integrable functions. 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. In particular, we propose a hyper-inverse-Wishart-process prior for the covariance kernels of the coefficient sequences, and then demonstrate the theoretical properties of the proposed prior, such as existence, uniqueness, and conditions under which it satisfies the strong hyper Markov property and conjugacy. To perform practical posterior inference, we introduce regularity conditions which allow us to write likelihood and prior density, and design stochastic search MCMC algorithms for posterior sampling. Performance of the proposed approach is demonstrated through simulation studies and a real data analysis in brain activity and alcoholism research.

To the best of our knowledge, the proposed approach is the first work in the statistical literature on functional data graphical models. It extends the theory of Dawid & Lauritzen (1993) from multivariate data to multivariate functional data. Potentially, one can naively apply multivariate methods to functional data after performing discretization or feature extraction. However, such an approach 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.

The rest of the paper is organized as follows: Section 2 presents the proposed approach, where we first review graphical models for multivariate data in Section 2.1, then introduce Markov distribution and hyper Markov laws for functional data in Section 2.2, and present the specific case of Gaussian process graphical models in Section 2.3. In Section 3, we demonstrate the regularity conditions which facilitate an approximate posterior inference, and design the corresponding Markov Chain Monte Carlo algorithms for posterior sampling. Simulations are performed in Section 4 to evaluate
the model performance. We finally apply the proposed method to a real EEG dataset in a brain activity and alcoholism study in Section 5. Section 6 contains discussions and conclusions. Proofs are postponed to the Appendix. The online supplementary materials contain further computational details and more simulation results.

2. GRAPHICAL MODELS FOR MULTIVARIATE FUNCTIONAL DATA

2.1 Review of Graph Theory and Gaussian Graphical Models for Multivariate Data

We review graph theory and Gaussian graphical models for multivariate data following Dawid & 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; see e.g., 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 $C = \{C_1, \ldots, C_m\}$ denote the set of cliques and $S = \{S_2, \ldots, S_m\}$ 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 $X = (X_1, \ldots, X_p)^T$ obey conditional independence ac-
According to a decomposable graph $G$, the joint distribution can be factorized as

$$p(X \mid G) = \prod_{C \in C} p(X_C) \prod_{S \in S} p(X_S),$$

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

$$p(\Sigma \mid G, \delta, U) = \prod_{C \in C} p(\Sigma_C \mid \delta, U_C) \prod_{S \in S} p(\Sigma_S \mid \delta, U_S),$$

where $p(\Sigma_C \mid \delta, U_C)$ and $p(\Sigma_S \mid \delta, U_S)$ are densities of inverse-Wishart (IW) distributions. In this paper, the inverse-Wishart follows the parameterization of Dawid (1981), i.e., $\Sigma \sim \text{IW}(\delta, U)$ if and only if $\Sigma^{-1}$ has a Wishart distribution $W(\delta + p - 1, U^{-1})$.

2.2 Graphical Models for Multivariate Functional Data

Let $f = (f_1, \ldots, f_p)$ 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 $f$ is denoted by $T = \bigsqcup_{j=1}^p T_j$, where $\bigsqcup$ denotes the disjoint union defined by $\bigsqcup_{j=1}^p T_j = \bigsqcup_{j=1}^p \{ (t,j) : t \in T_j \}$. 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 $\phi_{jk}$ in the $j$th 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 $\sigma$-algebra on $L^2(T)$. For $V = \{1,2,\ldots,p\}$ and $A \subset V$, denote by $f_A$ the subset of $f$ with domain $T_A = \bigsqcup_{j \in A} T_j$. We define the conditional independence relations for components of $f$ in Definition 1.

**Definition 1** (Conditional independence). Let $A, B, C$ be subsets of $V$. Then $f_A$ is said to be conditionally independent of $f_B$ given $f_C$ under $P$, written as $f_A \perp \perp f_B \mid f_C[P]$, if for any $f_A \in D_A$, where $D_A$ is a measurable set in $L^2(T_A)$, there exists a version of the conditional probability
\( p(f_A \in D_A \mid f_B, f_C) \) which is \( B(L^2(T_C)) \) measurable, and hence one may write \( p(f_A \in D_A \mid f_B, f_C) = p(f_A \in D_A \mid f_C) \). Here \( B(L^2(T_C)) \) denotes the Borel \( \sigma \)-algebra on \( L^2(T_C) \).

We would like to use a decomposable graph \( G = (V, E) \) to describe the conditional independence relations of components in \( f \), 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 \( f \) with \( G \) by assuming that \( P \) is Markov over \( G \), as defined in Definition 2.

**Definition 2 (Markov property).** Let \( G = (V, E) \) denote a decomposable graph. A probability measure \( P \) of \( f \) is called Markov over \( G \) if for any decomposition \((A, B)\) of \( G \), \( f_A \perp \perp f_B \mid f_{A \cap B}[P] \).

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

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

With Lemma 1, we can construct a joint probability measure for \( f \) 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 pairwisely consistent probability measures for \( \{f_{C_i}, i = 1, \ldots, m\} \). We construct a Markov probability measure \( P \) over \( G \) through the following recursive procedure

\[
\begin{align*}
P_{C_1} &= M_{C_1}, \\
P_{H_{i+1}} &= P_{H_i} \star M_{C_{i+1}}, \quad i = 1, \ldots, m - 1. 
\end{align*}
\]

One can show that the probability measure constructed this way is the unique Markov distribution.
over \( G \), and the proof follows that of Theorem 2.6 in Dawid & Lauritzen (1993). We call the constructed probability measure 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 & Lauritzen (1993) to define hyper Markov laws and use them as prior laws for \( P_G \). A prior law \( \mathcal{L} \) of \( P_G \) is called hyper Markov over \( G \) if for any decomposition \((A, B)\) of \( G \),

\[
(P_G)_A \perp \perp (P_G)_B \mid (P_G)_{A \cap B}[\mathcal{L}],
\]

where \((P_G)_A\) take 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 & Lauritzen (1993) applies to our random process setup.

**Proposition 1.** The theory of hyper Markov laws over undirected decomposable graphs, as described in Dawid & Lauritzen (1993, Section 3), 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 \( \{\mathcal{L}_C, C \in \mathcal{C}\} \) in a similar fashion as (1) - (2). Denote by \( \mathcal{L}_G \) the constructed hyper Markov prior for \( P_G \) and by \( \Pi \) a prior distribution for the graph \( G \). A Bayesian graphical model for the collection of random processes \( \mathbf{f} \) can be described as

\[
\mathbf{f} \sim P_G; \quad P_G \sim \mathcal{L}_G; \quad G \sim \Pi.
\]

(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 \( \Pi \) 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
\[ \sum_{i,j=1}^{p} \sum_{k,l=1}^{\infty} c_{ik} c_{jl} \int_{T_j} \int_{T_i} k_{ij}(s,t) \phi_{ik}(s) \phi_{jl}(t) dsdt \geq 0 \]
for any square summable sequence \( \{c_{ik}, i = 1, \ldots, p, k = 1, \ldots, \infty\} \); trace class means that
\[ \sum_{j=1}^{p} \sum_{l=1}^{\infty} \int_{T_j} \int_{T_i} k_{jj}(s,t) \phi_{jl}(s) \phi_{jl}(t) dsdt < \infty. \]

Then \( f_0 \) and \( K \) uniquely determine a Gaussian process on \( L^2(T) \) (Prato 2006, Section 1.5), which we call multivariate Gaussian process, and write \( \text{MGP}(f_0, K) \). The definition of multivariate Gaussian process implies that for \( A \subset V, f_A \sim \text{MGP}(f_0_A, K_A) \) where \( K_A = \{k_{ij}, i, j \in A\} \). Furthermore, on a sequence of cliques \( C = \{C_1, \ldots, C_m\} \), the marginal Gaussian process measures for \( \{f_C, C \in C\} \) are automatically consistent because they are induced from the same joint distribution. Therefore, we can construct a Markov distribution for \( f \) over \( G \) through procedure (1) - (2). We denote the resulting distribution of \( f \) by \( \text{MGP}_G(f_0, K_C) \), where \( K_C = \{k_{ij}, i, j \in C, C \in \mathcal{C}\} \). It is clear from this construction that the distribution \( \text{MGP}_G \) is Markov over \( G \) whereas \( \text{MGP} \) 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 \( L^2(T_j) \), and \( f_j(t) = \sum_{k=1}^{\infty} c_{jk} \phi_{jk}(t) \) 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 \( \ell^2_j = \{c_{jk} : \sum_{k=1}^{\infty} c_{jk}^2 < \infty\} \). Denote \( \ell^2 = \prod_{j=1}^{p} \ell^2_j \). Since \( \ell^2_j \) and \( L^2(T_j) \) are isometrically isomorphic for each \( j \), there is a natural identification between the probability measures defined on \( \ell^2 \) and \( L^2(T_j) \); therefore we can construct statistical models on \( \ell^2 \) without loss of generality. Let \( c = (c_1, \ldots, c_p) \) denote the coefficient sequence of \( f \). Then \( f \sim \text{MGP}(f_0, K) \) corresponds to \( c \sim \text{dMGP}(c_0, Q) \), where \( \text{dMGP} \) denotes the infinite dimensional discrete multivariate Gaussian processes, \( c_0 \) is the coefficient sequence of \( f_0 \) and \( Q = \{q_{ij}(\cdot, \cdot), i, j \in V\} \). 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, \( f \sim \text{MGP}_G(f_0, K_C) \) corresponds to \( c \sim \text{dMGP}_G(c_0, Q_C) \) where \( Q_C = \{q_{ij}(\cdot, \cdot), i, j \in C, C \in \mathcal{C}\} \). The
collection $Q$ 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, $K$ relates to $Q$ through equation $k_{ij}(s,t) = \sum_{k,l=1}^{\infty} q_{ij}(k,l) \phi_{ik}(s) \phi_{jl}(t)$. Denote by $P^c$ and $P^f$ the probability measures of $c$ and $f$ respectively. Then $f_A \perp \perp f_B | f_C [P^f]$ implies $c_A \perp \perp c_B | c_C [P^c]$ and vice versa. Thus the distribution $dMGP_G(c_0, Q_C)$ of $c$ is again Markov.

Assume that $c \sim dMGP_G(c_0, Q_C)$. The parameters involved in this distribution include $c_0$ and $Q_C$. In this study, we assume that $c_0$ is fixed (e.g., a zero sequence) so that the distribution of $c$ is uniquely determined by $Q_C$. 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 $Q_C$, it is equivalent to construct a hyper Markov law for $Q_C$. Let $U = \{u_{ij} : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}, i, j \in V\}$ denote a collection of kernel functions where $\mathbb{N}$ is the set of positive integers and $\mathbb{R}$ is the set of real numbers. We say that $U$ is self-adjoint (symmetric, in particular), positive semidefinite, trace class, or has finite rank if the operator from $\ell^2(\mathbb{N})^V \rightarrow \ell^2(\mathbb{N})^V$ defined by

$$(c_i)_{i \in V} \mapsto \left(\sum_{j \in V} u_{ij} c_j\right)_{i \in V}$$

is (respectively) self-adjoint, positive semidefinite, trace class, or has finite rank. Given a positive integer $\delta$ and the collection $U$ which is symmetric, positive semidefinite, and trace class, we construct a hyper-inverse-Wishart-process (HIWP) prior for $Q_C$ following Theorem 1.

**Theorem 1.** Assume that $c \sim dMGP_G(c_0, Q_C)$. Suppose that $\delta$ is a positive integer, and $U$ is a collection of kernels that is symmetric, positive semidefinite and trace class. Then there exists a sequence of pairwisely consistent inverse-Wishart processes determined by $\delta$ and $U_C = \{u_{ij}, i, j \in C\}, C \in C$, based on which one can construct a unique hyper Markov law for $Q_C$, which we call a hyper-inverse-Wishart-process, and write $Q_C \sim \text{HIWP}_G(\delta, U_C)$, where $U_C = \{u_{ij}, i, j \in C, C \in C\}$.

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

$$c \sim dMGP_G(c_0, Q_C), \quad Q_C \sim \text{HIWP}_G(\delta, U_C), \quad G \sim \Pi.$$  \hspace{1cm} (4)
It is of interest to investigate the properties of the HIWP prior and the corresponding posterior distribution. As shown in Dawid & 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), \(Q_B|A \perp \perp Q_A\), where \(Q_B|A\) denotes the conditional distribution (i.e., conditional covariance) of \(c_B\) given \(c_A\). After a careful investigation, we have only found a single, relatively strong condition under which the HIWP satisfies the strong hyper Markov property.

**Problem 1.** Whether the HIWP distribution in model (4) satisfies the strong hyper Markov property when the collection of kernels \(U\) does not have finite rank remains a challenging open problem.

A theoretical discussion of the challenges in Problem 1 is included in the Appendix. In the following proposition, we show that the \(\text{HIWP}_G\) prior constructed in Theorem 1 is strong hyper Markov if the collection of kernels \(U\) has finite rank.

**Proposition 2.** Suppose that the collection of kernels \(U\) has finite rank. Then the hyper-inverse-Wishart-process prior constructed in Theorem 1 satisfies the strong hyper Markov property. That is, if \(Q_C \sim \text{HIWP}_G(\delta, U_C)\), then for any decomposition \((A, B)\) of \(G\), \(Q_B|A \perp \perp Q_A\), where \(Q_B|A\) denotes the conditional distribution (e.g., conditional covariance) of \(c_B\) given \(c_A\).

The strong hyper Markov property of \(\text{HIWP}_G\) ensures that the joint posterior of \(Q_C\) (conditional on \(G\)) can be constructed from the marginal posterior of \(Q_C\) (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 \(c_i \sim \text{dMGP}_G(c_0, Q_C), i = 1, \ldots, n\) are independent and identically distributed. Further assume that the prior of \(Q_C\) is \(\text{HIWP}_G(\delta, U_C)\) where the collection of kernels \(U\) has finite rank. Then the conditional posterior of \(Q_C\) given \(\{c_i\}\) and \(G\) is \(\text{HIWP}_G(\tilde{\delta}, \tilde{U}_C)\), where \(\tilde{\delta} = \delta + n, \tilde{U}_C = \{\tilde{u}_{ij}, i, j \in C, C \subseteq C\}\) and \(\tilde{u}_{ij} = u_{ij} + \sum_{i=1}^{n}(c_i - c_{0i}) \otimes (c_j - c_{0j})\). Here \(\otimes\) denotes the
outer product. Furthermore, the marginal distribution of \( \{ c_i \} \) given \( \{ G, c_0, \delta, \bar{U}_C \} \) is again Markov over \( G \).

Theorem 2 implies that when \( U_C \) is assumed to have finite rank, the HIWP\(_G(\delta, U_C)\) prior is a conjugate prior for \( Q_C \) in the dMGP\(_G(c_0, Q_C)\) likelihood. Note that here the likelihood, the prior and the posterior are all conditional on \( G \). In practice, functional data can only be collected in finite dimension (through e.g., measuring values on a grid). Model (4) and results in Theorem 2 provide the theoretical foundation for practical Bayesian inference under reasonable regularity conditions, as discussed in Section 3.

3. APPROXIMATE 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. The conditional independence relations of random components in \( f = (f_1, \ldots, f_p) \) may be violated after these functions are projected onto a finite dimensional subspace through, for example, discretization. That is, \( f_A \perp \perp f_B \mid f_C \) does not imply that \( f_A \perp \perp f_B \mid P(f_C) \) for a projection operator \( P \). These facts make practical inference of graphical models in functional data challenging. Essentially, estimating conditional independence structure of infinite dimensional random processes based on finite number of measurement points is an inverse problem, therefore requires regularization. Müller & 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 to perform approximate posterior inference based on two regularization conditions.

**Condition 1:** Supposing that the functional data are observed discretely on a grid \( t = \bigsqcup t_j \), one can approximate the underlying infinite dimensional object \( f \) by \( \tilde{f} \) through interpolation (e.g., kriging (Chilés & Delfiner 1999)) or regularized basis representation (by using e.g., spline smoothing, data-driven eigenbasis estimation). Measurements are assumed to be sufficiently dense so that
\( \mathbf{f}_A \perp \perp \mathbf{f}_B \mid \mathbf{f}_C \) if and only if \( \mathbf{f}_A \perp \perp \tilde{\mathbf{f}}_C \), for \( A, B, C \subset V \).

**Condition 2:** When projecting \( \tilde{\mathbf{f}} \) from \( L^2(T) \) onto the \( \ell^2 \) space using orthogonal basis expansion, there exists a suitable set of finite integers \( M = (m_1, \ldots, m_p) \) so that the conditional independence relations hold after truncating the sequences at \( M \), i.e. \( \mathbf{c}_A \perp \mathbf{c}_B \mid \mathbf{c}_C \) if and only if \( \mathbf{c}_A \perp \mathbf{c}_B \mid c^M_C \).

Condition 1 requires that the discretely observed functional data capture sufficient information about the conditional independence structure so that the theory established in Section 2 still applies after replacing the true random functions by the smoothed approximations. Condition 2 requires that the coefficient sequences in \( \mathbf{c} = (c_1, \ldots, c_p) \) decay to zero fast so that the tails of the sequences have negligible influence on the conditional independence structure. This condition enables us to write out the density functions of the Markov distributions and hyper Markov laws so that posterior sampling can be practically implemented. The above regularization conditions are necessary to facilitate direct application of our theory to real data analysis. They are also reasonable in that they allow the conditional independence relations to be invariant up to small perturbations.

If the distribution of the functional data is supported on a finite dimensional subspace of continuous functions over \( T \), then Condition 1 will hold for a suitable choice of \( t \) and Condition 2 follows since the representation of the original functions in \( L^2(T) \) shall form a finite dimensional subspace in \( \ell^2(\mathbb{N}) \), and hence only a finite collection of coefficients suffices to identify the remaining coefficients. Since the continuous functions over \( T \) are dense in \( L^2(T) \), this is an appropriately large class of functions. In particular, high-frequency oscillations are considered noise in most applications and therefore little information will be lost by assuming that the functions of interest are band-limited.

3.1 Approximate Inference Under Regularization Conditions

Under the above regularity conditions, we can explicitly write the density function for truncated processes \( \mathbf{c}^M \). An MCMC algorithm can be designed for the approximate posterior inference of the underlying graph \( G \). In particular, with truncation, the density function of \( \mathbf{c}^M \) is

\[
p(\mathbf{c}^M \mid \mathbf{c}_0^M, Q_C, G) = \frac{\prod_{C \in \mathcal{C}} p(\mathbf{c}_C^M \mid \mathbf{c}_0^M_G, Q_C)}{\prod_{S \in \mathcal{S}} p(\mathbf{c}_S^M \mid \mathbf{c}_0^M_S, Q_S)},
\]
where $Q_c$ 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 $Q_C, Q_S$ are submatrices of $Q_c$ corresponding to clique $C$, separator $S$, respectively. The HIWP$_G$ prior of $Q_c$ induces a finite dimensional hyper inverse-Wishart prior with density

$$p(Q_C \mid G) = \frac{\prod_{C \in C} p(Q_C \mid \delta, U_C)}{\prod_{S \in S} p(Q_S \mid \delta, U_S)},$$

where $p(Q_C \mid \delta, U_C)$ is the density of inverse-Wishart defined in Dawid (1981), $U_C$ is a submatrix of $U_C$ corresponding to clique $C$, and $U_C$ is a block-wise matrix formed by $\{u_{ij}\}$ similarly as $Q_C$ formed by $\{q_{ij}\}$. The $p(Q_S \mid \delta, U_S)$ component in denominator is defined similarly. Based on (5) and (6), and assume that $\{c_i, i = 1, \ldots, N\}$ is a random sample of $c$, one can further integrate out $Q_c$ to get the marginal density

$$p(\{c_i^M\} \mid c_0^M, G) = (2\pi)^{-\frac{N}{2}(\sum_m m_i)} h(\delta, U_C) h(\tilde{\delta}, \tilde{U}_C),$$

where

$$h(\delta, U_C) = \frac{\prod_{C \in C} |\frac{1}{2} U_C|^\frac{1}{2} \Gamma_{d_c}^{-1}\{\frac{1}{2}(\delta + d_c - 1)\}}{\prod_{S \in S} |\frac{1}{2} U_S|^\frac{1}{2} \Gamma_{d_s}^{-1}\{\frac{1}{2}(\delta + d_s - 1)\}},$$

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

$$p(G \mid \{c_i^M\}, c_0^M) \propto p(\{c_i^M\} \mid c_0^M, G) p(G),$$

where $p(G)$ is the density function of the prior distribution $G \sim \Pi$, which is a finite dimensional discrete distribution supported on all decomposable graphs with $p$ nodes. Giudici & 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 pair of edges, which favors sparser graphs (Giudici 1996). The following MCMC algorithm describes the steps to generate posterior samples based on (8).

**Algorithm 1**

14
Step 0. Set an initial decomposable graph $G$ and set the prior parameters $c_0$, $\delta$, $U_C$.

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

$$\alpha = \min \left\{ 1, \frac{p(\tilde{G} \mid \{c_i\}, c_0^M) p(G \mid \tilde{G})}{p(G \mid \{c_i\}, c_0^M) p(\tilde{G} \mid G)} \right\}.$$ 

Detailed derivations are available in supplementary materials.

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 current graph, and the “global” move is achieved through the discrete uniform proposal. Guan et al. (2006) and Guan & 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 Approximate Inference for Noisy Functional Data

The theory in Section 2 and the approximate inference in Section 3.1 relies on the assumption that the distribution of $f$ (and $c$) 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 for the coefficient process:

$$d_{ijk} = c_{ijk} + e_{ijk}, \quad i = 1, \ldots, N, \quad j = 1 \ldots, p, \quad k = 1, \ldots, \infty,$$

where $\{c_{ijk}\}$ and $\{e_{ijk}\}$ are mutually independent with Gaussian distributions. This induces an additive model in the $L^2(T)$ space: $y_{ij} = f_{ij} + \varepsilon_{ij}$, where $\{y_{ij}\}$ are the functional data observed, $\{f_{ij}\}$ are the underlying true functions and $\{\varepsilon_{ij}\}$ are residuals. We assume that $e_{ijk} \sim N(0, s_j^2)$ which corresponds to white noise for $\varepsilon_{ij}$. After concatenating the $p$ coefficient sequences to vector
forms, we obtain the model $d_i = c_i + e_i$, where $d_i = (d_{i1}, \ldots, d_{ip})$, $d_{ij} = (d_{ij1}, d_{ij2}, \ldots)$, and $c_i, e_i$ follow similar forms.

After truncation at $M$, $e_i^M \sim N(0, \Lambda)$ and $\Lambda = \text{diag}(s_1^2T_{m1}, \ldots, s_p^2T_{mp})$. Notice that here $\text{cov}(d_i^M) = Q_C + \Lambda$, thus the diagonals of $Q_C$ and $\Lambda$ can not be separately identifiable. Therefore, we treat $\Lambda$ as a fixed model parameter, whose quantity can be predetermined through the approximation: $s_j^2 \approx \hat{\sigma}_j^2/|t_j|/(|t_j| - 1)$, where $\hat{\sigma}_j^2$ is the estimated variance of $\varepsilon_{ij}$ using local smoothing, $|T_j|$ is the width of interval $T_j$, and $|t_j|$ is the number of grid points in $T_j$. Applying a prior for $c_i^M$ in the form of (5) (conditional on $G$) and the HIWP prior for the covariance matrix $Q_C$ in the form of (6), we obtain the density function for the joint posterior:

$$p(\{c_i^M\}, Q_C, G | \{d_i^M\}) \propto \prod_{i=1}^n p(d_i^M | c_i^M, \Lambda) p(c_i^M | c_0^M, Q_C, G) p(Q_C | G) p(G).$$

From (10), we can integrate out $Q_C$ to obtain the marginal posterior distribution of $\{c_i^M\}$ and $G$.

The MCMC algorithm for generating posterior samples based on (10) is listed in Algorithm 2.

**Algorithm 2**

**Step 0.** Set initial values for $\{c_i^M\}$, $G$ and set the model parameters $\delta, c_0^M, U$ and $\Lambda$.

**Step 1.** Conditional on $\{c_i^M\}$, update $G \sim p(G | \{c_i^M\}, c_0^M)$ using the small-world sampler as described in Step 1 of Algorithm 1, where $p(G | \{c_i^M\}, c_0^M)$ is computed based on (10).

**Step 2.** Given $G$, update $Q_C \sim p(Q_C | \{c_i^M\}, G)$, which takes the same form as (6) except that $\delta$ and $U$ are replaced by $\tilde{\delta}$ and $\tilde{U}$ respectively using the formulae in Theorem 2.

**Step 3.** Conditional on $G$ and $Q_C$, update $c_i^M \sim N(\mu_i, V)$, where $V = (\Lambda^{-1} + Q_C^{-1})^{-1}$ and $\mu_i = V(\Lambda^{-1}d_i^M + Q_C^{-1}c_0^M)$.

### 3.3 Other Practical Computational Issues

To calculate the coefficient sequences $\{c_i\}$ from the functional observations $\{f_i\}$ requires the selection of an orthonormal basis $\{\phi_{jk} : j = 1, \ldots, p, k = 1, \ldots, \infty\}$. If a known set of basis is chosen (e.g., Fourier basis), 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 \( \{f_i\} \), 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 & Silverman (2005) or the principal component analysis through 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 \( U_C \) 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 \( \{c_i\} \) is needed to facilitate practical posterior inference. Although the regularization condition 2 guarantees the existence of a finite truncation, to theoretically verify that a particular truncation is suitable is difficult due to the lack of knowledge about the characteristics of the underlying random processes. Nevertheless, one may still determine the truncation parameter empirically by measuring the sensitivity of the posterior distribution to the change of truncations through sensitivity analysis (Saltelli et al. 2000). A truncation is suitable if adding more components does not cause large variation on the posterior distribution. If one has a clear understanding of the frequency range of the underlying processes, a convenient empirical solution is to pre-determine the truncation parameters using objective approximation criteria, following Rice & Silverman (1991) or Yao et al. (2005). This includes cross-validation (Rice & Silverman 1991), 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

Two simulation studies were conducted to assess the performance of the approximate 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).

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 $R_0$, with the inverse $R_0^{-1}$ containing a zero pattern corresponding to the graph in Figure 1 (a). We then generate principal component scores from a multivariate normal with zero mean and a block-wise covariance matrix $Q = ZRZ$, which has dimension $\sum_{j=1}^p m_j$. Here $R$ is a block-wise correlation matrix that has a diagonal form in each block. In particular, the $(i, j)th$ block of $R$, denoted by $R_{ij}$, satisfies that $R_{ij} = (R_0)_{i,j}I$ where $I$ is a rectangular identity matrix with size $m_i \times m_j$. An image plot of $R$ is shown in Figure 1(d), with its time-domain counterpart (the correlation of $f$ evaluated on $t$) 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. 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 $\{c_i\}$ using the PACE algorithm of Yao et al. (2005) and determine the truncation parameter $\{m_j\}$ using the FVE criterion with 90% threshold, resulting in $\{m_j\}$ values around 5. We apply Algorithm 1 and set $\delta = 5$ and $U = \hat{Z}\hat{R}\hat{Z}$, where $\hat{Z} = \text{diag}(\hat{\lambda}_{jk}^{1/2}, k = 1, \ldots, m_j, j = 1, \ldots, p)$, $\{\hat{\lambda}_{jk}\}$ are the estimated eigenvalues and $\hat{R}$ is set to be the identity matrix. 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 result 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 & 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 third method assumes matrix data, so we take the first five principal component scores and stack them up to form a $6 \times 5$ matrix for each sample. In the third 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 compared 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. When the true graph is unknown, the tuning procedure can be time-consuming. The matrix-normal graphical model is much slower to implement, perhaps due to the numerical approximation of marginal density in the MCMC algorithm.

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 edge pairs, 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 estimation. This suggests that for problems involving more nodes (>100), we can use these methods to obtain an initial estimate before applying our approach. Compared with others, the matrix-normal method tends to under-estimate the number of edges, and the resulting mis-estimation rate is relatively high.

4.2 Simulation 2: graph estimation for noisy functional data.

We add white noise to the functional data generated in Simulation 1 to demonstrate the performance of approximate inference for noisy data. The variances of the additive white noise \( \{ \epsilon_{ij}(t) \} \) 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 \( \frac{f_{ij}(t)}{\text{var}\{\epsilon_{ij}(t)\}} \) and is averaged across the grid points and the samples. We apply model (10) and generate posterior samples using Algorithm 2. The eigenbasis and the variance of the noise are estimated simultaneously using the PACE algorithm. The principal component scores \( d_i \) are estimated by projecting the raw data on the estimated eigenbasis. The parameter \( \Lambda \) is determined using the estimated variance of the white noise and the other model parameters are set to be the same as in Simulation 1. The posterior inference 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 supplementary material.

5. ANALYSIS OF EEG DATA IN A BRAIN-ALCOHOL STUDY

We demonstrate the performance of the proposed method using the EEG data in an alcoholism study. The data were obtained from 64 electrodes placed on subjects’ scalp that catch EEG signals at 256 Hz during a one-second period. The measurements were taken from 122 subjects, with 77 in alcoholism group and 45 in 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 $\alpha$ frequency band in the range of 8–12.5 Hz, which is known to be associated with inhibitory control (Knyazev 2007). Research has shown that, alcoholic subjects demonstrate decreased inhibitory control relative to control subjects (Sher et al. 2005). Reflected on the $\alpha$-band signal, less prevalent and lower signal power are observed in alcoholics (Porjesz et al. 2005; Finn & Justus 1999). Moreover, regional asymmetric patterns have been found in alcoholics (Hayden et al. 2006); alcoholics exhibit lower left $\alpha$-band activities in anterior regions relative to right. In this study, we aim to estimate the conditional independence relations of $\alpha$-band signals from different regions of the scalp, and expect to find evidence that reflects differences in brain connectivity and asymmetric pattern between the two groups.

To best describe the regional interactions, we select 13 electrodes from five regions of the scalp. As illustrated in Figure 2 (a)-(c), these electrodes are (AF7, AF8), (F3, F4) and AFz in frontal region, (C3, C4) in central region, (P3, P4) in parietal region, (TP7, TP8) in temporal region, and (O1, O2) in occipital region. The electrodes in parentheses are in symmetric positions, and the coding follows the 10–20 system of the American EEG Society. The signals can be influenced by stimulus type, and the samples may not be independent due to multiple trials per subject. To reduce the influence of these effects, we use the subset containing samples with single stimulus, and further remove some samples so that all data are coming from non-consecutive trials. To retain sufficiently large sample size, we still allow multiple samples per subject and make an independence assumption across samples, as done in many previous studies (Mao & Li 2009). This gives 2,010 samples in alcoholic group and 2,006 samples in control group. The $\alpha$-band filtering is conducted using eegfilt function in the EEGLAB toolbox of Matlab. We then apply model (4) using coefficients of the eigenbasis expansion. The number of eigenbasis $\{m_j\}$ is determined through retaining 90% of the total variation; this results in 6–7 coefficients per $f_j$. We collect 30,000 posterior samples using Algorithm 1, in which the first 20,000 are treated as the burn-in period. The model is fitted for both the alcoholic and the control group.
The posterior results are summarized in Figure 2. The plots in (a) and (b) are the posterior modes of the alcoholic and the control group, respectively, where the gray edges are common across the two groups, and the black ones are those that differ from the other group. Comparing (a) with (b), we see that the alcoholic group contains more edges connecting the left frontal/central regions (AF7, C3) with the right temporal/parietal/occipital regions (TP8, P4, O2). The control group, on the other hand, contains more edges connecting the frontal regions (AF8, AFz, F3, F4) with the temporal (TP7), parietal (P3, P4) and occipital (O1, O2) regions. In (c) we demonstrate the edge pairs that have more than 0.5 differences between the two groups, based on the marginal inclusion probabilities. We can see that most black edges marked in (a) and (b) appear in (c).

To further compare with established results, we calculate two summary statistics for connectivity: the number of edges per node and the total number of edges, and calculate another two summary statistics for asymmetry: the number of asymmetric pairs per node and the total number of asymmetric pairs. We compare the four summary statistics across the two groups using boxplots in Figure 2 (d)–(g), and calculate the posterior probability that the alcoholic group is greater than, equal to, or less than the control group for each statistic. In (f), since symmetric positions have the same number of asymmetric pairs, we collapse the plots for F3 with F4, and P3 with P4. The posterior probability calculations show that, with probability 1, the alcoholic group has fewer edges than the control at F3; with probability 0.91, the alcoholic group has more asymmetric pairs than the control at F3/F4; with probability 0.99, the alcoholic group has higher total number of asymmetric pairs than the control. These results indicate that the alcoholic group exhibits decreased connectivity at F3, increased asymmetry at (F3, F4), and increased overall asymmetry. These observations are compatible with the findings of Hayden et al. (2006), who studied the asymmetric patterns at (F3, F4) and (P3, P4) using the analysis of variance method based on the resting-state α-band power. Furthermore, our analysis provides connectivity and asymmetric pattern of all 13 nodes simultaneously whereas Hayden et al. (2006) only focus on the four representative nodes.
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 an approximate posterior inference based on two regularization conditions, which enables posterior sampling through MCMC algorithms.

One concern is the possibility of performing exact posterior inference, i.e., inferring the graph directly from the joint posterior

$$p(G|\{c_i\}) \propto p(\{c_i\}|G)p(G)$$

based on model (4), where $p(\{c_i\}|G)$ is the marginal likelihood (with the covariance kernel $Q_C$ 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 evaluated on a finite dimensional projection of $\{c_i\}$.

In the approximate posterior inference, the approximation errors (caused by truncation) and their influence on posterior distribution may be quantified empirically. Assuming that the functional data are pre-smoothed, the approximation error can be quantified by calculating the difference of the $\ell^2$ norms between the full sequence and the truncated sequence. The influence on the posterior distribution can be quantified by tools of sensitivity analysis as noted in Section 3.3. For example, based on model (4) one can calculate the Kullback-Leibler divergence $KL(M, M') = \sum_G p(G|\{c_i^M\}) \log_2 [p(G|\{c_i^M\})/p(G|\{c_i^{M'}\})]$ as one increases $M$ to $M'$. A suitable truncation parameter is the one beyond which the $KL(M, M')$ function is stabilized. An alternative method to pre-determining the truncation parameter is to set prior for $M$ in a Bayesian hierarchical model, in which case hybrid MCMC algorithms are needed for fitting both models (4) and (10). The posterior sampling in these models would become more complicated because the dimension of the truncated sequences and the size of the covariance matrix $Q_C$ would change whenever $M$ are...
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 $Q_P$ is subject to extra constraint induced by missing edges.

Although we have applied the proposed method to relatively small graphs, extra simulations (results not reported for conciseness) have shown that our approach can be safely applied to multivariate functional data to estimate graphs with up to 50 nodes. 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).

**APPENDIX**

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 $\pi_{(\alpha, \beta)} \{ \{ x_{(l,m)} : (l, m) \in T \times T \} \} = x_{(\alpha, \beta)}$. 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 $\pi_B \{ \{ x_{(l,m)} : (l, m) \in T \times T \} \} = \{ x_{(s,t)} : (s, t) \in B \}$. 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 $\sigma$-algebra.* Let $\mathcal{B}_{(\alpha, \beta)}$ be a $\sigma$-algebra on $\mathbb{R}^{(\alpha, \beta)}$. We can create a $\sigma$-algebra on $\mathbb{R}^{T \times T}$ by pulling back the $\mathcal{B}_{(\alpha, \beta)}$ using the inverse of the projection map and define $\pi^*_{{(\alpha, \beta)}}(\mathcal{B}_{(\alpha, \beta)}) = \{ \pi^{-1}_{(\alpha, \beta)}(A) : A \in \mathcal{B}_{(\alpha, \beta)} \}$. 24
One can verify that $\pi_{(\alpha,\beta)}(B_{(\alpha,\beta)})$ is a $\sigma$-algebra. (III) Product $\sigma$-algebra. We define the product $\sigma$-algebra as $\mathcal{B}(\mathbb{R}^{T \times T}) = \prod_{(\alpha,\beta) \in T \times T} \mathcal{B}_{(\alpha,\beta)}$, where $\prod_{(\alpha,\beta) \in T \times T} \mathcal{B}_{(\alpha,\beta)} = \sigma \left( \bigcup_{(\alpha,\beta) \in T \times T} \pi^*_\alpha(B_{(\alpha,\beta)}) \right)$.

(IV) Pushforward measure. Given a measure $\mu_{T \times T}$ on the product $\sigma$-algebra, and a subset $B$ of $T \times T$, we define the pushforward 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 $A \in \mathcal{B}_{\mathbb{R}^B}$, where $\mathcal{B}_{\mathbb{R}^B} = \prod_{(\alpha,\beta) \in \mathbb{R}^B} \mathcal{B}_{(\alpha,\beta)}$. (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 $\mu_{B_1}$ and $\mu_{B_2}$ are said to obey compatibility relation if $(\pi_{B_2 \cap B_1})_*\mu_{B_1} = \mu_{B_2}$.

Proof of Lemma 1

This proof involves some measure-theoretic arguments. The essential idea is to use disintegration theory (Chang & Pollard 1997) to first construct the conditional probability measure $P_1 \{ \cdot \mid \pi_{A \cap B}(f_A) \}$ on $\mathcal{B}(L^2(T_A))$, extend this to $P \{ \cdot \mid \pi_B(f) \}$ 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 $P_1 \{ \cdot \mid \pi_{A \cap B}(f_A) \}$ on $\mathcal{B}(L^2(T_A))$ satisfying: (a.1) $P_1(\mathcal{X} \mid f_{A \cap B}) = P_1(\mathcal{X} \cap [L^2(T_{A \setminus B}) \times \{\pi_{A \cap B}(f_A)\}]) \mid \pi_{A \cap B}(f_A)$ for all $\mathcal{X} \in \mathcal{B}(L^2(T_A))$, (b.1) the map $f_{A \cap B} \mapsto (P_1)_t f_{A \cap B} H : = \int H(f_A) dP_1(f_A) \mid f_{A \cap B}$ is measurable for all nonnegative measurable $H : L^2(T_A) \to \mathbb{R}$, and (c.1) $P_1 H = ((\pi_{A \cap B})_* P_1)(P_1)_t f_{A \cap B} H$ 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 $P \{ \cdot \mid \pi_B(f) \}$ by setting $P \{ A \mid \pi_B(f) \} = P_1 \{ \pi_A(A \cap [L^2(T_{A \setminus B}) \times \{\pi_B(f)\}]) \mid \pi_{A \cap B}(f) \}$. Note that this is well defined for all measurable $A \in \mathcal{B}(L^2(T_{A \cup B}))$ since the sections $\pi_A(A \cap [L^2(T_{A \setminus B}) \times \{\pi_B(f)\}])$ are always measurable, and also that (a) $P \{ A \mid \pi_B(f) \} = P \{ A \cap [L^2(T_{A \setminus B}) \times \{\pi_B(f)\}] \mid \pi_B(f) \}$ 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) $f_B \mapsto P_{T_B} H$ 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 $f_B$ 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 $P_{f_B}H_n \to P_{f_B}H$ in an increasing manner. Since this holds for each $f_B$, we conclude that $P_{f_B}H$ 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 $P_{f_B}1_{X \times Y} = P_1(\mathcal{X} \mid f_{A \cap B})1_Y(f_{B \setminus A})$ 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 $1_{X \times Y} \in \mathcal{M}$. 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 $PH = P_{f_B}P_{f_B}H$. By construction, we have that $P1_{L^2(T_{A \setminus B}) \times Y} = P_2P_1(L^2(T_{A \setminus B}) \times \{f_{A \cap B}\} \mid f_{A \cap B})1_Y(f_B) = P_2(Y)$ and $P1_{X \times L^2(T_{B \setminus A})} = P_2P_1(\mathcal{X} \mid f_{A \cap B}) = ((\pi_{A \cap B})_*P_2)P_1(\mathcal{X} \mid f_{A \cap B}) = ((\pi_{A \cap B})_*P_1)P_1(\mathcal{X} \mid f_{A \cap B}) = P_1(\mathcal{X})$. Thus, we also have that $PH = P_{P_{f_B}H} = (P_{f_B})_*P\pi_B(f_B)H$ for all measurable $H$, and this is the final property establishing that $P(\cdot \mid f_B)$ is a disintegration of $P$ with respect to the map $\pi_B$. By the disintegration theorem, this disintegration is a version of the regular conditional probability of $f_A$ given $f_B$. Since this version only depends upon $f_{A \cap B}$, we conclude that (iii) holds. Finally, we note that any other measure satisfying these properties must agree with the measure we have constructed on $\pi$-system, and therefore the uniqueness of $P$ immediately follows.

Proof of Proposition 1

The Properties 1 - 4 in Dawid & 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 & Lauritzen (1993) continue to hold. Properties 1 - 4 and results in Appendix A imply that results in B1- B7 of Dawid & Lauritzen (1993) continue to hold when $P$ is a Markov distribution constructed in Lemma 1. Theorem 2.6 and Corollary 2.7 of Dawid & 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 & Lauritzen (1993) hold.

**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 \( u : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R} \) 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 \( \mu \) on \((\mathbb{R}^{N \times N}, \mathcal{B}(\mathbb{R}^{N \times N}))\) satisfying

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

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

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

Proof of Lemma 2

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 \( \mu_B \) 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 \( Q_{I_2 \times I_2} \) is the sub-matrix of \( Q_{I_1 \times I_1} \) obtained by deleting the rows and columns with indices in \( I_1 \setminus I_2 \). If \( Q_{I_1 \times I_1} \) has law \( \mu_{I_1 \times I_1} = \text{IW}(\delta, \mathbf{U}_{I_1 \times I_1}) \), then \( Q_{I_2 \times I_2} \) has law \( \text{IW}(\delta, \mathbf{U}_{I_2 \times I_2}) \) due to the consistency property of the inverse-Wishart distribution (Dawid & 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 \( B_2 = \{\tilde{\alpha}_i, \tilde{\beta}_i\}_{i=1}^{m} \subset B_1 \). Set \( g_1 = \{\alpha_i\}_{i=1}^{n} \cup \{\beta_i\}_{i=1}^{n} \) and \( g_2 = \{\tilde{\alpha}_i\}_{i=1}^{m} \cup \{\tilde{\beta}_i\}_{i=1}^{m} \) so that \( g_2 \times g_2 \subset g_1 \times g_1 \). It is clear that \( \pi_{B_2 \leftarrow B_1} \circ \pi_{B_1 \leftarrow g_1 \times g_1} = \pi_{B_2 \leftarrow g_1 \times g_1} = \pi_{B_2 \leftarrow g_2 \times g_2} \circ \pi_{g_2 \times g_2 \leftarrow g_1 \times g_1} \).

Thus,

\[
(\pi_{B_2 \leftarrow B_1})_{*} \mu_{B_1} = (\pi_{B_2 \leftarrow B_1})_{*} (\pi_{B_1 \leftarrow g_1 \times g_1})_{*} \mu_{g_1 \times g_1} = (\pi_{B_2 \leftarrow B_1} \circ \pi_{B_1 \leftarrow g_1 \times g_1})_{*} \mu_{g_1 \times g_1}
\]

\[
= (\pi_{B_2 \leftarrow g_2 \times g_2} \circ \pi_{g_2 \times g_2 \leftarrow g_1 \times g_1})_{*} \mu_{g_1 \times g_1} = (\pi_{B_2 \leftarrow g_2 \times g_2})_{*} (\pi_{g_2 \times g_2 \leftarrow g_1 \times g_1})_{*} \mu_{g_1 \times g_1}
\]

\[
= (\pi_{B_2 \leftarrow g_2 \times g_2})_{*} \mu_{g_2 \times g_2} = \mu_{B_2}.
\]
where the second to last equality holds because of our demonstration in Case 1. (2) Second, we claim that the finite dimensional measure $\mu_{I \times I} = IW(\delta, U_{I \times I})$ is an inner regular probability measure on the product $\sigma$-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 $I$, $Q_{I \times I}$ 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 $Q_{I \times I} \sim IW(\delta, U_{I \times I})$, 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 $\sigma$-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 $\mu$ on the product $\sigma$-algebra $\mathcal{B}(\mathbb{R}^{N \times N})$ that satisfies (i) and (ii).

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

First, we show each row of $Qx$ is finite almost surely hence is well-defined for all $x \in \ell^2(N)$. It
is sufficient to show that $E[|{(Qx)_i}|] < \infty$. 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 $E \sum_j q_{ij}^2 = \sum_j E q_{ij}^2 \leq C \sum_j u_{ii}u_{jj} = Cu_{ii} \sum_j u_{jj}$, where $C$ is the maximum of the above constants. Thus $E[|{(Qx)_i}|] \leq \|x\|\sqrt{E \sum_j q_{ij}^2} < \infty$. Because there are only countably many rows, we have that $Qx$ is finite almost surely for all rows simultaneously. Consequently, we have that $Qx$ is well-defined for all $x \in \ell^2(N)$. Now we show that $Qx \in \ell^2(N)$ almost surely. By similar considerations, let $\mathbf{q}_i = (Qx)_i$, then $E(\sum_i \|\mathbf{q}_i\|^2) \leq C (\sum_i u_{ii})^2 < \infty$ and $\|Qx\|^2 \leq C\|x\|^2 \sum_i \|q_i\|^2$; this implies that $\|Qx\| < \infty$ almost surely hence $Qx \in \ell^2(N)$ almost surely, and it also implies that the operator norm $\|Q\|_{op}$ is finite almost surely.

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

Proof of Theorem 1

Based on Lemma 2, we can define a sequence of inverse-Wishart process prior for $Q_C$, denoted by $Q_C \sim \text{IWP}(\delta, U_C), C \in C$. These sequences are pairwise consistent due to the consistency of inverse-Wishart processes and the fact that $U_C$ is a common collection of kernels. Therefore, we can construct a unique hyper Markov law for $Q_C$ following procedure (12) - (13) of Dawid & Lauritzen (1993). And Theorem 3.9 of Dawid & Lauritzen (1993) guarantees that the constructed hyper Markov law is unique.
Discussion of Problem 1

We first introduce the notation used in this discussion. Given $y \in \ell^2(N)$, we let $y^*$ denote the bounded linear operator from $\ell^2(N)$ to $\mathbb{R}$ satisfying $y^*x = \langle y, x \rangle$, where $\langle \cdot, \cdot \rangle$ is the usual inner product on the Hilbert space $\ell^2(N)$. For a bounded, self-adjoint, positive semidefinite operator $A$ from $\ell^2(N)$ to $\ell^2(N)$, there is a unique bounded, symmetric, positive semidefinite operator $A^{1/2}$ that satisfies $A^{1/2}A^{1/2} = A$. We let $A^{-1}$ denote the restricted inverse of $A$, or the unbounded linear operator defined on the image of $A$ satisfying $Ay = 0$ if and only if $A^{-1}y = 0$. $A^{-1}$ is the orthogonal projection onto the orthogonal complement of the kernel of $A$, and $AA^{-1}$ is a densely-defined operator which can be extended to the orthogonal projection onto the orthogonal complement of the kernel of $A$.

To get to the heart of the issue, we introduce a simplification of Problem 1 as follows. Suppose $\sigma_0^2 > 0$ and that $\Sigma$ is a trace class symmetric positive-definite operator on $\ell^2(N)$. With a slight abuse of notation, we may identify $\Sigma$ with an infinite diagonal array with diagonal entries $\{\sigma_1^2, \sigma_2^2, \ldots\}$. We let $\Sigma_N = \text{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, $\tilde{\Sigma}_N = \text{diag}(\sigma_0^2, \sigma_1^2, \ldots, \sigma_N^2)$, and set $\tilde{\Sigma} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \Sigma \end{pmatrix}$. We assume that the matrix $\tilde{S}_N = \begin{pmatrix} s_0 & s_N^* \\ s_N & S_N \end{pmatrix}$ has the IW distribution with parameters $\delta$ and $\tilde{\Sigma}_N$, and assume that $(x, y_N)^T \sim N(0, \tilde{S}_N)$. Let $(x, y)$ denote the infinite dimensional counterparts of $(x, y_N)$ and let $S, s, \tilde{S}$ denote the infinite dimensional counterpart of $S_N, s_N, \tilde{S}_N$, respectively. Showing the strong Markov property of HIWP simplifies to showing that the conditional distribution $\pi(x|y)$ is independent of the marginal $\pi(y)$ under the prior law $\tilde{S} \sim IW(\delta, \tilde{\Sigma})$.

From the perspective of Baker (1973), an $\tilde{S}$ drawn from $IW(\delta, \tilde{\Sigma})$ is trace class almost surely, and $s^*$ in the block representation $\tilde{S} = \begin{pmatrix} s_0 & s_N^* \\ s_N & S_N \end{pmatrix}$ has the form $\sqrt{s_0}u^*S^{1/2}$ where $\|u\| \leq 1$ and $Pu = u$ when $P$ is the orthogonal projection onto the orthogonal complement of the kernel of $S$. For $y$ drawn from a Gaussian process with mean 0 and covariance operator $S$, we have that $S^{-1/2}y$
is an isonormal Gaussian process supported on the orthogonal complement of the kernel of \( S \). Since 
\[ \|u\| = \sqrt{\sum u_i^2} < \infty, \]
we have that the inner product of \( u \) with this isonormal Gaussian process has the distribution 
\( N(0, \sum u_i^2) \). Thus, we make the (potentially non-associative) identification 
\[ s^* S^{-1} y = \sqrt{s_0 u^*} (S^{-1/2} y), \]
which is \( \sqrt{s_0} \) times a random variable distributed as 
\( N(0, \sum u_i^2) \). Now, consider drawing \( y \) from \( GP(0, S) \) and then \( x \) (conditional on \( y \)) from 
\( N(\sqrt{s_0 u^*} (S^{-1/2} y), s_0 (1 - \|u\|^2)) \). By the above analysis, it is clear that \((x, y)\) drawn in this manner is defined almost surely.

It is also a Gaussian process, and therefore is characterized by its mean and covariance operator. By the tower property, Fubini-Tonelli for \( \sigma \)-finite measures, and linearity, we have 
\[ E(x) = E\{E(x|y)\} = E\{\sqrt{s_0} u^* (S^{-1/2} y)\} = 0, \]
\[ E(xy_k) = E\{E(xy_k|y)\} = E\{y_k \sqrt{s_0} u^* (S^{-1/2} y)\} = \sqrt{s_0} u^* (S^{-1} E\{y_k y\}) = \sqrt{s_0} u^* S^{-1/2} S e_k = s^* e_k, \]
where \( e_k \) is the \( k \)th member of the standard orthonormal basis of \( \ell^2(\mathbb{N}) \), and 
\[ E(x^2) = E\{E(x^2|y)\} - E(x|y)^2 + E(x|y)^2 = E\{s_0 (1 - \|u\|^2) + s_0 (u^* (S^{-1/2} y))^2\} = s_0 (1 - \|u\|^2) + s_0 \|u\|^2 = s_0. \]
Since it is clear that \( E(y) = 0 \) and \( E(yy^*) = S \), we have verified that, for \((x, y)\) drawn from a Gaussian with mean 0 and covariance \( \tilde{S} \), the conditional distribution \( \pi(x|y) \) is 
\( N(\sqrt{s_0} u^* S^{-1/2} y, s_0 (1 - \|u\|^2)) \) (which is defined \( \pi(y) \)-almost surely). Thus, if the Gaussian measure \( \pi(x, y) \) is the Gaussian distribution associated with \( \tilde{S} \) drawn from the infinite dimensional Wishart, the conditional distribution \( \pi(x|y) \) is independent from the marginal \( \pi(y) \) if and only if 
\[ s_0 (1 - \|u\|^2) \]
and 
\[ \sqrt{s_0} (u, S^{-1/2} \cdot) \] are independent of \( S \).

Our current recipe for constructing the full conditional \( \pi(x|y) \) first employs construction of the full joint measure \( \pi(x, y) \) and the marginal \( \pi(y) \) via Kolmogorov’s extension theorem, and then we invoke the disintegration theorem to obtain \( \pi(x|y) \). Using Kolmogorov’s extension invokes knowledge of all finite marginal distributions, and it is not straightforward to show that \( \pi(x|y) \perp \pi(y) \) under the prior law \( IW(\delta, \Sigma) \) from the finite dimensional counterparts. However, our above equivalence indicates that it may be possible to find distributions for 
\[ s_0 (1 - \|u\|^2) \]
and 
\[ \sqrt{s_0} (u, S^{-1/2} \cdot) \]
so that we may define \( \pi(x|y) \) independently from \( \pi(y) \), but we must then verify that drawing these independently induces a distribution on the joint \( \pi(x, y) \) that coincides with our construction \( IW(\delta, \Sigma) \) via Kolmogorov’s extension theorem. The independence of the finite-dimensional counter-
parts may play a role in the construction of such distributions. On the other hand, the discontinuity of the linear functional \( \langle u, S^{-1/2} \cdot \rangle \) implies that we cannot use simple continuity arguments to infer \( \langle u, S^{-1/2}y \rangle \) from \( \langle u_N, S_N^{-1/2}y_N \rangle \). We intend to investigate these issues very carefully in the future.

Proof of Proposition 2

Note that an operator drawn from a hyper-inverse-Wishart process with a finite-rank parameter \( \Sigma \) will be finite-rank almost surely. This follows by simply considering the covariance in the appropriate basis. By again considering the distribution in the appropriate basis, we may reduce to the finite dimensional case, in which case the proof is well-known.

Proof of Theorem 2

By the result of Proposition 1, the HIWP_G prior is a strong hyper Markov law. So by Corollary 5.5 of Dawid & Lauritzen (1993), the posterior law of \( Q_C \) 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: \( c_{i,C} \sim dMGP(c_{0,C}, Q_C) \) with prior \( Q_C \sim IWP(\delta, U_C) \) for each \( Q_C \), and use them to construct the posterior law of \( Q_C \) following (12) - (13) of Dawid & Lauritzen (1993). As before, choosing the appropriate basis for representation reduces this to the finite-dimensional case which is well-known. Finally, by Proposition 5.6 of Dawid & Lauritzen (1993), the marginal distribution of \( \{c_i\} \) given \( G, c_0, \delta, \tilde{U}_C \) is again Markov over \( G \).

REFERENCES

Baker, C. R. (1973), “Joint measures and cross-covariance operators,” *Transactions of the American Mathematical Society*, 186, 273–289.

Bauer, H. (2001), *Measure and integration theory*, De Gruyter studies in mathematics W. de Gruyter.

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

Carvalho, C. M., & West, M. (2007), “Dynamic matrix-variate graphical models,” *Bayesian Anal.*, 2(1), 69–98.
Chang, J. T., & Pollard, D. (1997), “Conditioning as disintegration,” *Statistica Neerlandica*, 51(3), 287–317.

Chilès, J. P., & Delfiner, P. (1999), *Geostatistics: modeling spatial uncertainty*, New York: John Wiley & Sons.

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

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

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

Dempster, A. P. (1972), “Covariance selection,” *Biometrics*, 28, 157–175.

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

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

Giudici, P. (1996), “Learning in graphical Gaussian models,” *Bayesian Statistics 5*, pp. 621–628.

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

Guan, Y., Fleissner, R., Joyce, P., & Krone, S. M. (2006), “Markov Chain Monte Carlo in Small Worlds,” *Stat. Comput.*, 16, 193–202.

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

Hayden, E. P., Wiegand, R. E., Meyer, E. T., Bauer, L. O., O’Connor, S. J., Nurnberger, J. I., Chorlian, D. B., Porjesz, B., & Begleiter, H. (2006), “Patterns of Regional Brain Activity in Alcohol-Dependent Subjects,” *Alcohol. Clin. Exp. Res.*, 30(12), 1986 –1991.

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

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

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

Lei, E., Yao, F., Heckman, N., & Meyer, K. (2014), “Functional Data Model for Genetically Related Individuals with Application to Cow Growth,” *Journal of Computational and Graphical Statistics*, .

Li, S., Gu, M., Wu, C. J., & Xia, J. (2012), “New Efficient and Robust HSS Cholesky Factorization of SPD Matrices,” *SIAM J. Matrix Analysis Applications*, pp. 886–904.
Mao, W., & Li, Y. (2009), “Scalp Topographic distribution of cognitive electroencephalogram power in the alcoholic subjects,” *Journal of shanghai university*, 15(5), 445 – 451.

Morris, J. S., & Carroll, R. J. (2006), “Wavelet-based functional mixed models,” *J. R. Statist. Soc. B*, 68, 179–199.

Müller, H. G., & Yao, F. (2008), “Functional additive models,” *J. Am. Statist. Assoc.*, 103, 1534–1544.

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

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

Ramsay, J. O., & Silverman, B. W. (2005), *Functional Data Analysis, Section Edition*, New York: Springer.

Rice, J. A., & Silverman, B. W. (1991), “Estimating the Mean and Covariance Structure Non-parametrically When the Data Are Curves,” *Journal of the Royal Statistical Society, Series B*, 53, 233–243.

Rosen, O., & Thompson, W. K. (2009), “A Bayesian regression model for multivariate functional data,” *Comput. Statist. Data Anal.*, 53, 3773–3786.

Roverato, A. (2002), “Hyper Inverse Wishart distribution for Non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models,” *Scand. J. Stat.*, 29, 391–411.

Saltelli, A., Chan, K., & Scott, E. M., eds (2000), *Sensitivity Analysis*, New York: John Wiley & Sons, Ltd.

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

Sher, K. J., Grekin, E., & Williams, N. A. (2005), “The development of alcohol use disorders,” *Annu. Rev. Clin. Psychol.*, 1, 493–523.

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

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

Yao, F., Müller, H. G., & Wang, J. L. (2005), “Functional data analysis for sparse longitudinal data,” *J. Am. Statist. Assoc.*, 100, 577–590.
Table 1: FPC: functional principal component; 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.

| Data Type | Method    | # of FPC per curve | Run-time of 5000 Iter. (sec) | Mean # of edges | # of Unique graphs visited | Mis-estimation rate of edges | Sen | Spec |
|-----------|-----------|--------------------|-----------------------------|-----------------|----------------------------|-------------------------------|-----|------|
| Smooth    | FDGM-S    | 3 - 5              | 38                          | 7.66            | 3                          | 0.02                          | 0.96| 1.0  |
| (n=200)   | GGM-MH    | 1                  | 0.15                        | 9.55            | 63                         | 0.10                          | 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  |
| (n=200)   | GGM-MH    | 1                  | 0.39                        | 9.62            | 59                         | 0.11                          | 0.77|       |
|           | gLasso    | 1                  | -                           | -               | -                          | 0.13                          | -   | -    |
|           | MNGM      | 5                  | 4086.38                     | 6.33            | 18                         | 0.26                          | 0.65| 0.85 |
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 underlying time-domain correlation matrix. (d) The underlying correlation matrix in the transformed space.
Figure 2: Summary of posterior inference: the posterior modes of alcoholic group (a) and control group (b), the edges with >0.5 difference in marginal inclusion probabilities (c), the boxplots of connectivity measures: the number of edges per node (d) and the total number of edges (e), the boxplots of asymmetry measures: the number of asymmetric pairs per node (f) and the total number of asymmetric pairs (g). In (c), solid lines indicate edges with higher frequency in alcoholic group and dashed lines indicate edges with higher frequency in the control group. In (d)–(g), the alcoholic group is abbreviated as “al”, and the control group is abbreviated as “ct”.

37