Biometrika (2011), **98**, 1, *pp*. 1–27

 Advance Access publication on 30 June 2011

Printed in Great Britain

Bayesian graphical models for multivariate functional data

BY H. ZHU, D. B. DUNSON

Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708 U.S.A.

hz52@stat.duke.edu dunson@stat.duke.edu

AND N. STRAWN

Department of Mathematics, Duke University, Box 90320, Durham, NC 27708 U.S.A.

nstrawn@math.duke.edu

SUMMARY

In many applications there is interest in the dependence structure in multivariate functional data. For vector data, conditional independence relationships can be inferred through allowing zeros in the precision matrix in a Gaussian graphical model. Bayesian methods can allow unknown locations of zeros relying on hyper inverse-Wishart priors for the covariance. To generalize these methods to multivariate functional data, we propose a multivariate Gaussian process with an extended block hyper-inverse Wishart prior for the covariance structure. Theoretical properties of this prior are considered. Posterior computation is performed in the frequency domain using orthogonal basis expansions, with Markov chain Monte Carlo algorithms developed with and without measurement errors. The methods are evaluated through simulation studies and are applied to Electroencephalography data.

Some key words: Functional data analysis; Gaussian process; Graphical model; Model uncertainty; Stochastic search.

1. INTRODUCTION

Although there is an increasingly rich literature on methods for functional data analysis, little consideration has been given to the multivariate case. We are interested in inference on the dependence structure in multiple functions measured for each subject. Let $\mathrm{f}_i = (f_{i1}, f_{i2}, \ldots, f_{ip})^T$ denote a vector of p random functions for subject i. We assume that the jth function f_{ij} is defined on a compact domain T_j of the real line. We will also consider error-prone measurements, where for the jth function of subject i, we have

$$y_{ij}(t) = f_{ij}(t) + \varepsilon_{ij}(t) \quad (t \in T_j;\ j = 1, \ldots, p;\ i = 1, \ldots, n), \tag{1}$$

where $\varepsilon_{ij}(t)$ is the random error process. We assume that the raw data associated with f_{ij} is recorded on the finite grid $\mathrm{t}_j = (t_{j1}, \ldots, t_{jn_j})^T$ which is common across index i.

Our goal is to estimate the conditional independence relationships in the random functional components of f_i. For example, it may be the case that f_{i1} is conditionally independent of f_{i2} given $f_{i3}, \ldots, f_{ip}$. If f_i was simply a Gaussian vector with $\mathrm{f}_i \sim N_p(\mu, \Sigma)$, then the zeros in the precision matrix Σ^{-1} would correspond to the conditional independence relationships in the components of f_i (Dempster, 1972; Lauritzen, 1996). A particular Gaussian graphical model would correspond to the zero/non-zero structure in the precision matrix, and we could potentially select an optimal graphical model or average across graphical models through Bayesian methods. More details can be found in Giudici & Green (1999); Roverato (2002); Jones et al. (2005); Scott & Carvalho (2008); Carvalho & Scott (2009); Lenkoski & Dobra (2011). Motivated by the substantial practical utility of graphical models, there is an increasingly rich literature on generalizations beyond simple vector cases to accommodate matrix-variate graphical models (Wang & West, 2009), dynamic linear models (Carvalho & West, 2007) and other complications.

Our focus is on developing graphical models and associated Bayesian methods for inferring conditional independence relationships in multivariate functional data. Although there are a number of articles focusing on Bayesian analysis of functional data that contain repeated measures of the same types of functions or functions in nested designs (Morris & Carroll, 2006; Rosen & Thompson, 2009), little consideration has been given to multivariate methods for analyzing dependence in different types of functions from a Bayesian perspective. We provide a formal definition of conditional independence of random functions, and introduce a class of Gaussian process graphical models for modeling the pairwise conditional independence of multivariate functional data.

2. Gaussian process graphical models

2·1. *Review of Gaussian graphical models*

Let $G = (V, E)$ be an undirected graph with a vertex set V and an edge set $E = \{(i, j)\}$. 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 maximum subgraph is called a *clique*. Let A, B, C be subsets of V and $V = A \cup B$, $C = A \cap B$, 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. In this case 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 to be minimal, we can iteratively decompose the graph into a sequence of *prime components*: a sequentially defined collection of subgraphs that cannot be further decomposed. If all the prime components of a connected graph are complete, the graph is said to be *decomposable*. All prime components of a decomposable graph are cliques. We will mainly study decomposable graphs in this paper.

In graphical models, one associates the graph G with a set of random variables $X = \{X_i, i \in V\}$ which has a probability measure P. For $A \subset V$, denote $X_A = \{X_k, k \in A\}$. We say that P is Markov with respect to G if for any decomposition (A, B) of G, X_A is conditionally independent of X_B given $X_{A\cap B}$, denoted by $X_A \perp X_B \mid X_{A\cap B}$. A graphical model requires that P of X is Markov with respect to the graph. In a Gaussian graphical model, X is assumed to be multivariate Gaussian, denoted by $X \sim N(\mu, \Sigma)$. Without loss of generality, we assume that μ equals the zero vector. A property that connects the multivariate Gaussian with conditional independence states that: $X_i \perp X_j \mid X_{V\setminus\{i,j\}}$ if and only if the (i, j)th component of the precision Σ^{-1} is zero. This implies that the graph G is associated with the covariance of X through the zero/nonzero patterns of Σ^{-1}. If G is decomposable, one can factorize the probability density of X according to its cliques and separators: $p(X \mid G, \Sigma) = \prod_{C\in\mathcal{C}} p(X_C \mid \Sigma_C)/\prod_{S\in\mathcal{S}} p(X_S \mid \Sigma_S)$, where $p(X_C \mid \Sigma_C)$ and $p(X_S \mid \Sigma_S)$ are the marginal density of X_C and X_S, and Σ_C and Σ_S denote the sub-covariance matrices corresponding to clique C and S, respectively. Here $\mathcal{C}$ and $\mathcal{S}$ denote the collections of clique and separator sets. A widely used conjugate prior for Σ is the hyper-inverse Wishart prior, which has density $p(\Sigma \mid G, \delta, Q) = \prod_{C\in\mathcal{C}} p(\Sigma_C \mid \delta, Q_C)/\prod_{S\in\mathcal{S}} p(\Sigma_S \mid \delta, Q_S)$, where $p(\Sigma_C \mid \delta, Q_C) \propto |\Sigma_C|^{-(\delta/2+|C|)} \exp\{-\frac{1}{2}\text{tr}(\Sigma_C^{-1}Q_C)\}$, which corresponds to the Inverse-Wishart distribution IW(δ, Q_C) as defined in Dawid (1981). This distribution always exists for $\delta > 0$ and Q_C symmetric and positive-definite. A similar form holds for $p(\Sigma_S \mid \delta, Q_S)$.

2·2. *Gaussian process graphical models for multivariate functional data*

Let $\mathrm{f} = (f_1, \ldots, f_p)^T$ be a vector of random processes defined on a probability space $(\coprod_{j=1}^p \Omega_j, \mathcal{F}, P)$, where $\coprod_{j=1}^p \Omega_j$ denotes the disjoint union of p sample spaces, with each Ω_j being a space of real functions defined on a compact set T_j. For example, Ω_j can be the function space $L^2(T_j)$, with T_j a closed set on the real line. For each $\omega \in \coprod_{j=1}^p \Omega_j$, $\mathrm{f}(\cdot, \omega)$ is a vector of p functions, with the jth function $f_j(\cdot, \omega) : T_j \to R$. For fixed $t_j^0 \in T_j$, denoting

$\mathrm{t}^0 = (t_1^0, \ldots, t_p^0)^T$, then $\mathrm{f}(\mathrm{t}^0, \cdot) = [f_1(t_1^0, \cdot), \ldots, f_p(t_p^0, \cdot)]^T$ is an $\mathcal{F}$-measurable random vector of length p.

To construct a graphical model on f, it is important to define the pairwise conditional independence of its components. We approach this through finite grid discretization. Let $\widetilde{T} = \coprod_j T_j$ be the disjoint union of $\{T_j\}$, and $\mathrm{t} = \coprod_j \mathrm{t}_j$ be a finite grid of $\widetilde{T}$ with $\mathrm{t}_j = (t_{j1}, \ldots, t_{jn_j})$. Unless stated otherwise, we assume that t contains at least one grid point on each T_j. Denote the finite dimensional discretization of f_j over t_j as $f_j(\mathrm{t}_j) = [f_j(t_{j1}), \ldots, f_j(t_{jn_j})]^T$ ($j = 1, \ldots, p$), we state the definition for a pair of elements being conditionally independent in Definition 1.

DEFINITION 1 (CONDITIONAL INDEPENDENCE OF RANDOM PROCESSES). *Let* f *be defined as above. We say that* f_l *and* f_m *(*$l \neq m$*) are conditionally independent given the other components of* f *if for any finite grid discretization* t*,* $f_l(\mathrm{t}_l) \perp f_m(\mathrm{t}_m) \mid \{f_j(\mathrm{t}_j), j \neq l, m\}$*, with respect to* P_{t}*, which is the probability measure induced by* P *through the finite dimensional projection. We denote such conditional independence as* $f_l \perp f_m \mid \mathrm{f}_{(-l,-m)}[P]$.

Given the definition of conditional independence, we can define graphical models for random processes, as done in Definition 2.

DEFINITION 2 (GRAPHICAL MODELS FOR RANDOM PROCESSES). *Let* f *be a vector of random processes with probability measure* P*. A graphical model for* f *associates* P *with a* p*-node undirected graph* $G = (V, E)$*, such that for any pair* $(i, j) \notin E$*,* $f_i \perp f_j \mid \mathrm{f}_{(-i,-j)}[P]$.

Depending on the nature of the functional data, P may be any measure which satisfies Definition 2. In this work, we will consider multivariate Gaussian processes as a convenient special case. Assume that f is a multivariate Gaussian process, denoted by $\mathrm{f} \sim \mathrm{MGP}(\mathrm{f}_0, \mathcal{Q})$, where f_0 is the expected value and $\mathcal{Q} = \{q_{l,m}\}$ is a class of covariance kernels that reflect the within-function and across-function covariances. In particular, $q_{l,m} : T_l \times T_m \to R$, such that

for $s \in T_l$ and $t \in T_m$, $\text{cov}(f_l(s), f_m(t)) = q_{l,m}(s,t)$. With a slight abuse of notation, we use the same notation f to denote the long vector formed by concatenating the discretized functions $\{f_j(\mathrm{t}_j), j = 1, \ldots, p\}$, and use Q to denote the corresponding covariance matrix. The matrix Q is a block-wise matrix of size $(\sum_j n_j) \times (\sum_j n_j)$, with the (i,j)th block Q_{ij} formed by evaluating the corresponding covariance kernel over discrete grid $\mathrm{t}_l \times \mathrm{t}_m$. The diagonal blocks of Q describe the within-function covariances and the off-diagonal blocks describe the between-function covariances. When necessary, we will also use $Q_{\mathrm{t}\times\mathrm{t}}$ to emphasize the grid on which Q is evaluated. We assume that $\mathcal{Q}$ is positive definite, which means that for a finite grid $\mathrm{t} = \coprod_{j\in I} \mathrm{t}_j (I \subseteq \{1, 2, \ldots, p\})$, Q is a symmetric and positive definite matrix. Under the Gaussian process assumption, the distribution of the vector f is multivariate Gaussian: $\mathrm{f} \sim \mathrm{N}(\mathrm{f}_0, Q)$. Proposition 1 connects the pairwise conditional independence with the covariance $\mathcal{Q}$. A sketch of the proof can be found in the Appendix.

PROPOSITION 1. *Assume that* $\mathrm{f} \sim$ *MGP*$(\mathrm{f}_0, \mathcal{Q})$. *Then for any* $l \neq m$, $f_l \perp f_m \mid \mathrm{f}_{(-l,-m)}$ *if and only if for any finite grid* t, *the* (l,m)*th block of* $K = Q^{-1}$ *is a zero matrix, where* Q *is the covariance of the vectorized* f, *and the* (l,m)*th block of* K *is the block that corresponds to* $f_l(\mathrm{t}_l)$ *and* $f_m(\mathrm{t}_m)$.

The above result indicates that graphical models can be learned through the zero/nonzero structure of the precision matrix K. The multivariate Gaussian process assumption also enables factorization of the joint density of f based on the decomposition of G, as stated in Proposition 2.

PROPOSITION 2. *Assume that* $\mathrm{f} \sim$ *MGP*$(\mathrm{f}_0, \mathcal{Q})$ *is associated with a decomposable graph* G, *which can be decomposed into cliques in* $\mathcal{C}$ *using separators in* $\mathcal{S}$. *Then for any finite grid* t, *we can factorize the density of the vectorized* f *as*

$$p(\mathrm{f} \mid \mathrm{f}_0, Q, G) = \frac{\prod_{C\in\mathcal{C}} p(\mathrm{f}_C \mid \mathrm{f}_{0,C}, Q_C)}{\prod_{S\in\mathcal{S}} p(\mathrm{f}_S \mid \mathrm{f}_{0,S}, Q_S)}, \tag{2}$$

where f_C, $\mathrm{f}_{0,C}$ *are sub-vectors of* f, f_0 *corresponding to functions in clique* C, *and* Q_C *is the diagonal block of* Q *corresponding to the clique* C. *Similar explanations hold for* f_S *and* Q_S. *Here* $p(\mathrm{f}_C \mid \mathrm{f}_{0,C}, Q_C)$ *and* $p(\mathrm{f}_S \mid \mathrm{f}_{0,S}, Q_S)$ *are multivariate Gaussian densities.*

The proof can be found in the Appendix. The factorization in (2) brings tremendous computational convenience. Conditional on G, the distribution of f can be computed based on the marginal distributions corresponding to the cliques and separators, which are of much lower dimensions. With $\mathrm{f} \sim \mathrm{MGP}(\mathrm{f}_0, \mathcal{Q})$ and the factorization form in (2), we now propose a prior for $\mathcal{Q}$ which extends the hyper inverse Wishart. Given δ a positive integer and a known collection of covariance kernels $\mathcal{U} = \{u_{l,m}\}$ that is also positive definite, define a prior probability measure for $\mathcal{Q}$ conditional on G such that for any t, the discretized covariance matrix Q has density

$$p(Q \mid G) = \frac{\prod_{C \in \mathcal{C}} p(Q_C \mid \delta, U_C)}{\prod_{S \in \mathcal{S}} p(Q_S \mid \delta, U_S)}, \tag{3}$$

where $p(Q_C \mid \delta, U_C)$ is the density of $\mathrm{IW}(\delta, U_C)$ and U_C is the diagonal block of the covariance matrix U corresponding to functions in clique C. Similar explanations hold for Q_S and U_S. We call such a probability measure of $\mathcal{Q}$ block hyper-inverse Wishart, denoted as $(\mathcal{Q} \mid G, \delta, \mathcal{U}) \sim \mathrm{BHIW}(G, \delta, \mathcal{U})$. When discretization is applied, we also denote $(Q|G, \delta, U) \sim \mathrm{BHIW}(G, \delta, U)$. As this prior is defined through finite discretization, one question that need to be answered is whether this measure is well defined, or whether the measure exists. This is stated in Theorem 1. The proof and some definitions therein are put in the Appendix.

THEOREM 1. *Suppose that* δ *is a positive integer and that* $T_j \subset R$ *is compact for* $j = 1, \ldots, p$. *Further suppose that the collection* $\mathcal{Q} = \{q_{lm}\}$ *contains kernels such that its projection* $Q_{\mathrm{t}\times\mathrm{t}}$ *is symmetric positive definite for any finite* $\mathrm{t} \subset \widetilde{T}$. *In addition, assume that* $\mathcal{Q}$ *is associated with a p-node decomposable graph* G. *Then there exists a unique probability measure* $\mu_{\widetilde{T}\times\widetilde{T}}$ *on* $(R_{\widetilde{T}\times\widetilde{T}}, \mathcal{B}_{\widetilde{T}\times\widetilde{T}})$ *satisfying:*

i. for any finite $\mathrm{t} \subset \widetilde{T}$ *such that* $\mathrm{t} = \coprod_j \mathrm{t}_j$ *and each* t_j *is nonempty, we have* $(\pi_{\mathrm{t}\times\mathrm{t}})_* \mu_{\widetilde{T}\times\widetilde{T}} = \mu_{\mathrm{t}\times\mathrm{t}}$*, where* $\mu_{\mathrm{t}\times\mathrm{t}}$ *is the law of BHIW*$(G, \delta, U_{\mathrm{t}\times\mathrm{t}})$ *as defined in (3);*

ii. if $B = \{(\alpha_i, \beta_i)\}_{i=1}^n \subset \widetilde{T} \times \widetilde{T}$*, and* $\{\alpha_i\}_{i=1}^n \cup \{\beta_i\}_{i=1}^n \subset \mathrm{g}$*, where* g *is a grid that contains at least one grid point on each* T_j*, then* $(\pi_B)_* \mu_{\widetilde{T}\times\widetilde{T}} = \mu_B$*, where* $\mu_B = (\pi_{B \leftarrow \mathrm{g}\times\mathrm{g}})_* \mu_{\mathrm{g}\times\mathrm{g}}$.

Theorem 2 shows that the block hyper-inverse Wishart is the conjugate prior for $\mathcal{Q}$ when f is a multivariate Gaussian process associated with a decomposable graph. The proof is in the Appendix.

THEOREM 2. *Assume that* $\mathrm{f}_i \sim$ *MGP*$(\mathrm{f}_0, \mathcal{Q})(i = 1, \ldots, n)$ *are independent and identically distributed and the distribution is associated with a decomposable graph* G*. With the prior* $(\mathcal{Q} \mid G) \sim$ *BHIW*$(G, \delta, \mathcal{U})$*, the conditional posterior for* $\mathcal{Q}$ *is* $(\mathcal{Q} \mid \{\mathrm{f}_i\}, G) \sim$ *BHIW*$(G, \widetilde{\delta}, \widetilde{\mathcal{U}})$*, where* $\widetilde{\delta} = \delta + n$ *and* $\widetilde{\mathcal{U}}$ *is such that for any finite grid discretization* t*,* $\widetilde{U} = U + \sum_{i=1}^n (\mathrm{f}_i - \mathrm{f}_0)(\mathrm{f}_i - \mathrm{f}_0)^T$*. Furthermore, on fixed* t*, one can integrate out the covariance matrix* Q *to get*

$$p(\{\mathrm{f}_i\} \mid \mathrm{f}_0, G) = \int p(\{\mathrm{f}_i\} \mid \mathrm{f}_0, Q, G) p(Q \mid G) dQ = (2\pi)^{-\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^p n_{ij}} \frac{h(\delta, U)}{h(\widetilde{\delta}, \widetilde{U})}, \quad (4)$$

where

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

Here $\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{i=0}^{d-1} \Gamma(a - i/2)$*, and* d_c*,* d_s *are the size of* U_C *and* U_S*, respectively.*

Based on Theorem 2, we can build a Gaussian process graphical model for estimating the conditional independence relationships between functional components. For any finite discretization t, the posterior distribution can be written as

$$p(G \mid \{\mathrm{f}_i\}, \mathrm{f}_0, \delta, U) \propto p(\{\mathrm{f}_i\} \mid \mathrm{f}_0, \delta, U, G) p(G), \quad (5)$$

where $p(\{\mathrm{f}_i\} \mid \mathrm{f}_0, \delta, U, G)$ takes the form as in (4) and $p(G)$ is the prior for G. There are different choices for $p(G)$. One used in Giudici & Green (1999) is the uniform prior $p(G) = 1/d$, where d

is the total number of decomposable graphs with p vertices, which does not need to be computed in implementation. Jones et al. (2005) used independent Bernoulli for each pair of edges with inclusion probability $r = 2/(p-1)$, which favors sparser graphs (Giudici, 1996).

2·3. *Multivariate functional data subject to measurement errors*

In model (5), the data associated with $\{\mathrm{f}_i\}$ are assumed to be directly observed. In reality, it is common that functional data are subject to measurement errors or noise. An ad hoc remedy is to perform smoothing/denoising in preprocessing steps. Alternatively, one can rely on model (1), in which case $\{y_{ij}\}$ are the functions observed and $\{f_{ij}\}$ and $\{\varepsilon_{ij}\}$ are unknown. We focus on the simple case in which ε_{ij} is white noise: $\varepsilon_{ij}(t) \sim N(0, \sigma_j^2)$ independently of t. Over a finite grid t, we can concatenate $\{y_{ij}\}_{j=1}^p$ to form a long vector $\mathrm{y}_i = \{y_{i1}(\mathrm{t}_1), \ldots, y_{ip}(\mathrm{t}_p)\}^T$, where $y_{ij}(\mathrm{t}_j) = \{y_{ij}(t_{ij1}), \ldots, y_{ij}(t_{ijn_j})\}$. We similarly concatenate $\{f_{ij}\}_{j=1}^p$ to get f_i and $\{\varepsilon_{ij}\}_{j=1}^p$ to get ϵ_i. The discretized model (1) can therefore be written as $\mathrm{y}_i = \mathrm{f}_i + \epsilon_i$, where $\epsilon_i \sim N(0, \Lambda)$, with $\Lambda = \mathrm{diag}(\sigma_1^2 1_{n1}^T, \ldots, \sigma_p^2 1_{n_p}^T)$, and $(\mathrm{y}_i \mid \mathrm{f}_i, \Lambda) \sim N(\mathrm{f}_i, \Lambda)$, with $\{\mathrm{f}_i\}$ unknown. Assuming a conditional prior for f_i as in (2), and a block hyper-inverse Wishart prior for the covariance of f_i, we obtain the joint posterior as follows:

$$p(\{\mathrm{f}_i\}, Q, G \mid \{\mathrm{y}_i\}) \propto \prod_{i=1}^{n} p(\mathrm{y}_i \mid \mathrm{f}_i, \Lambda) p(\mathrm{f}_i \mid \mathrm{f}_0, Q, G) p(Q \mid G) p(G), \tag{6}$$

where $p(\mathrm{f}_i \mid \mathrm{f}_0, Q, G)$ takes the form of (2) and $p(Q \mid G)$ takes the form of (3). From (6), we can either integrate out Q or $\{\mathrm{f}_i\}$ to obtain the marginalized posterior distribution.

3. Model fitting through orthogonal basis expansion

As in other Gaussian process-based models, posterior computation in Gaussian process graphical models can encounter substantial bottlenecks. Computing the inverse of the covariance matrix of f_C involves $O(n_C^3)$ operations, where $n_C = \sum_{j \in C} n_j$. As the grid becomes finer, correlations

between nearby points of a smooth function increase, leading to covariance matrices that are nearly rank deficient. In addition, due to the curse of dimensionality (Hughes, 1968), as the number of grid points increases, the samples needed for accurately estimating a graph grow exponentially (Miller et al., 2010). There is a rich literature on methods for reducing bottlenecks in Gaussian process computation, ranging from kernel convolutions (Higdon, 2002) to sparse covariance approximation (Furrer et al., 2006). In functional data analysis, basis expansions are widely adopted, such as principal component analysis (Yao et al., 2005) and wavelets (Morris & Carroll, 2006). Here, we focus on orthonormal basis expansions.

3·1. *Orthonormal basis expansion and model fitting in the transformed space*

One can represent a function in a separable Hilbert space by linear combinations of orthonormal bases, which provides a convenient way for decorrelation and dimensional reduction. Assume that f_{ij} takes values in $L^2[T_j]$ which has a complete orthonormal basis $\{\phi_{jk}\}_{k=1}^{\infty}$. We can represent f_{ij} as $f_{ij}(t) = \sum_{k=1}^{\infty} c_{ijk}\phi_{jk}(t)$, where $c_{ijk} = \int_{T_j} f_{ij}(t)\phi_{jk}(t)dt$, and $E\{\sum_k c_{ijk}^2\} = E\{||f_{ij}||^2\} < \infty$. The space $L^2(T_j)$ and $l^2 = \text{span}\{c_{ijk}, k = 1, \ldots, \infty\}$ are isometrically isomorphic. If f_{ij} is a Gaussian process, then $\{c_{ijk}\}$ are also Gaussian. In practice, the orthonormal basis can be chosen as a known basis such as Fourier or wavelets. They can also be chosen as eigenfunctions of the covariance operator and can be estimated from the data. In the latter case, the coefficients $\{c_{ijk}\}_{k=1}^{\infty}$ are called functional principal component scores of f_{ij}. Various estimating methods have been proposed in Ramsay & Silverman (1997); Yao et al. (2005). In our implementation, we approximate f_{ij} by truncating the linear combination at $K_j < \infty$, and fit the model in the transformed space which is also called *frequency domain*. Approximating f_{ij} using orthonormal basis can significantly reduce the dimension. Furthermore, the coefficients of f_{ij} are either independent or nearly independent. Therefore setting parameters for the block

hyper-inverse Wishart prior in the frequency domain is easier. We discuss the details for fitting both models (5) and (6).

For the case of no measurement error, denote $\mathrm{c}_{ij} = (c_{ij1}, \ldots, c_{ijK_j})$, which contains the basis coefficients of f_{ij} truncated as K_j. Let $\mathrm{c}_i = (\mathrm{c}_{i1}, \ldots, \mathrm{c}_{ip})^T$, which are the concatenated coefficients of f_i. We denote the covariance matrix of c_i as Q^*. Then Q^* is a $\sum_j K_j \times \sum_j K_j$ matrix with the (l, m)th block $Q^*_{l,m}$. Simple algebra reveals that the $Q^*_{l,m}$ relates to $Q_{l,m}$ through the following equation: $Q_{l,m} = \Psi_l(\mathrm{t}_l) Q^*_{l,m} \Psi_m^T(\mathrm{t}_m)$, where $\Psi_l(\mathrm{t}_l) = (\phi_{l1}(\mathrm{t}_l), \ldots, \phi_{lK_l}(\mathrm{t}_l))$, and each $\phi_{lk}(\mathrm{t}_l)$ is a vector formed by evaluating $\phi_{lk}(t)$ on the grid t_l. We can write the likelihood of c_i conditional on a decomposable graph G as in (2) by replacing f_i, f_0 by c_i, c_0 respectively and replacing Q by Q^*. We similarly adopt a block hyper-inverse Wishart prior for Q^* as in (3): $(Q^* \mid G) \sim \mathrm{BHIW}(G, \delta, U^*)$. The U^* is the prior scale matrix. The posterior of the graph can therefore be written as $p(G \mid \{\mathrm{c}_i\}) \propto p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, G) p(G)$, where $p(\{\mathrm{c}_i\} \mid \mathrm{c}_0, G)$ takes a similar form to (4) except replacing f_i, f_0 by c_i, c_0, replacing n_j by K_j, and replacing U by U^*. Details on determining δ, $\{K_j\}$ and U^* are available in the supplementary material.

For the case with measurement error, we assume that both f_{ij} and ϵ_{ij} are Gaussian processes taking values in $L^2(T_j)$. Orthonormal basis expansions transform model (1) to $d_{ijk} = c_{ijk} + e_{ijk}$ where $d_{ijk} = \int_{T_j} y_{ij}(t)\phi_{jk}(t)dt$ and $e_{ijk} = \int_{T_j} \varepsilon_{ij}(t)\phi_{jk}(t)dt$. Concatenating the coefficients to vector forms, we obtain the model: $\mathrm{d}_i = \mathrm{c}_i + \mathrm{e}_i$. The white noise assumption of ε_{ij} implies that $\mathrm{e}_i \sim N(0, \Lambda^*)$, with $\Lambda^* = \mathrm{diag}(s_1^2 1_{K_1}^T, \ldots, s_p^2 1_{K_p}^T)$. Under this setup the likelihood is $\mathrm{d}_i \sim N(\mathrm{c}_i, \Lambda^*)$. The prior for c_i takes a similar form as in (2), except replacing $\mathrm{f}_i, \mathrm{f}_0, Q$ by $\mathrm{c}_i, \mathrm{c}_0, Q^*$ respectively. A block hyper-inverse Wishart prior is assumed for Q^* in a similar fashion as in the case of no measurement error. Because c_i and e_i are both Gaussian and they are assumed to be mutually independent, we have $\mathrm{cov}(\mathrm{d}_i) = Q^* + \Lambda^*$. The diagonal of Q^* is not identifiable from that of Λ^* without extra constraints. To avoid this problem, we treat Λ^* as a fixed model

parameter, whose quantity can be pre-determined through estimating $\{\sigma_j^2\}$ using local smoothing and the approximation $s_j^2 \approx \sigma_j^2 T_j/(n_j - 1)$. The joint posterior can thus be written as

$$p(\{\mathrm{c}_i\}, Q^*, G \mid \{\mathrm{d}_i\}) \propto \prod_{i=1}^{n} p(\mathrm{d}_i \mid \mathrm{c}_i, \Lambda^*) p(\mathrm{c}_i \mid Q^*, G) p(Q^* \mid G) p(G). \tag{7}$$

3·2. *Algorithms*

We use Markov chain Monte Carlo algorithms to obtain posterior samples. One widely used sampling scheme in Gaussian graphical model is the Metropolis-Hastings sampler using local proposals, where in each iteration a new graph is proposed by randomly adding or deleting one edge while preserving decomposability. The chain can mix slowly when the posterior distribution is either multi-modal or spiky. There are many options for improving mixing. We rely on the small-world sampler of Guan et al. (2006) and Guan & Krone (2007), which uses a mixture of a local and heavier-tailed proposal in Metropolis-Hastings, leading to much faster convergence than for local proposals. We briefly describe the steps in Algorithm 1 and Algorithm 2, which correspond to models with and without measurement errors, respectively. More details on computation can be found in the supplementary material.

Algorithm 1

Step 0 Set an initial decomposable graph G_0 and the prior parameters c_0, δ, U^*.

Step 1 With probability $1 - p$, propose $\widetilde{G}$ by randomly adding or deleting an edge, each with probability 0.5, within the space of decomposable graphs. Accept the new $\widetilde{G}$ with the computed acceptance probability. With probability p, propose $\widetilde{G}$ from a discrete uniform distribution supported on the set of all decomposable graphs, and accept the proposal with the computed acceptance probability.

Algorithm 2

Step 0 Set initial values for $\{\mathrm{c}_i\}$ and set parameters δ, U^* and Λ.

Step 1 Conditional on $\{\mathrm{c}_i\}$, update $G \sim p(G \mid \{\mathrm{c}_i\})$ following the small-world sampler as described in Step 1 of Algorithm 1, where $p(G \mid \{\mathrm{c}_i\}) \propto \int p(\{\mathrm{c}_i\}, Q^*, G \mid \{\mathrm{d}_i\}) dQ^*$ computed based on the joint posterior in (7).

Step 2 Conditional on G, update $Q^* \sim p(Q^* \mid \{\mathrm{c}_i\}, G)$, which is a block hyper-inverse Wishart. See details for sampling in Carvalho & West (2007); Wang & Carvalho (2010).

Step 3 Conditional on G and Q^*, update $\mathrm{c}_i \sim N(\mu_i, V)$, where $V = (\Lambda^{-1} + (Q^*)^{-1})^{-1}$, and $\mu_i = V(\Lambda^{-1}\mathrm{d}_i + (Q^*)^{-1}\mathrm{c}_0)$.

One can further improve mixing using small-world sampling with tempering (Guan and Stephens, unpublished) which, rather than using the uniform distribution as the heavy tailed proposal, uses a tempered posterior distribution. We have used this technique in our simulation 1 and real data application.

4. Simulation study

Two simulation studies are conducted to assess the performance of the proposed Gaussian process graphical models for multivariate functional data. Simulation 1 corresponds to the situation where the data are smooth and Simulation 2 corresponds to the case with measurement errors.

4·1. *Simulation 1*

We set the number of functional components to $p = 6$. All functions are generated on the domain $[0, 1]$ using functional principal component analysis. In particular, eigenfunctions are chosen to be the first K_j Fourier bases, with a small perturbation to the phase for each j, $j = 1, \ldots, p$. The truncation parameters $\{K_j\}$ vary from 3 to 7. The true eigenvalues are sampled from Gamma distributions with exponentially decaying mean values. The conditional indepen-

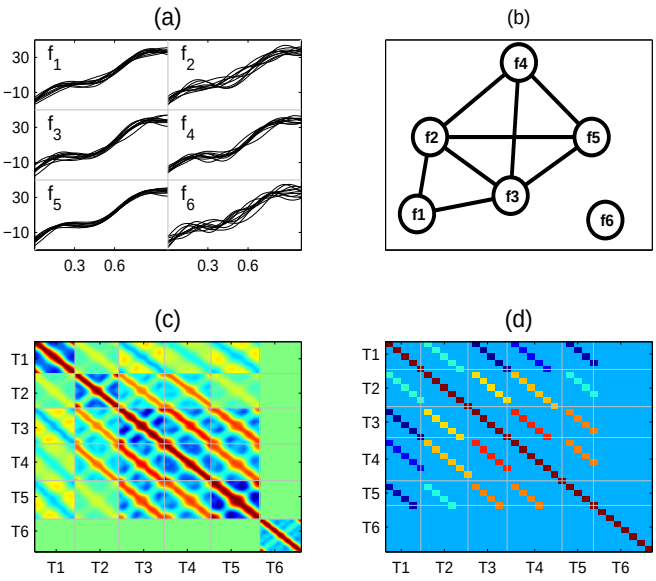

Fig. 1. Plots of Simulation 1. (a) The first 10 samples of each functional component. (b) The true underlying graph. (c) The image plot of the true time domain correlation matrix. (d) The corresponding true frequency domain correlation matrix.

dence structure of the data is determined by a $p \times p$ correlation matrix R_0^* whose inverse has zero patterns corresponding to the graph shown in Figure 1(b). The principal component scores are generated from the multivariate normal with zero mean and an expanded covariance matrix $Q^* = ZR^*Z$, which is of size $\sum_j K_j$. Here R^* is a block-wise correlation matrix that has a diagonal form in each block, expanded based on R_0^*. An image plot of R^* is shown in Figure 1(d), with its time domain counterpart shown in (c). Multivariate functional data are finally obtained through linear combinations of eigenfunctions. A common mean function is added to each curve. The first 10 samples of each functional component are plotted in Figure 1(a). The data contain $n = 200$ independent samples. Each sample contains six curves measured on six different grid designs. All six designs are equally spaced and are common across the samples.

As the generated multivariate functional data are smooth, we apply Algorithm 1. Functional principal component analysis is firstly applied to estimate the eigenfunctions, eigenvalues and

the corresponding FPC scores. This reduces the data from dimension 610 with 93-111 grids per function to 22 with 3-5 principal components per function. For the prior parameters, the degrees of freedom δ is set as 3, and the scale matrix is set to be $U^* = \widehat{Z}\widehat{R}\widehat{Z}$, where $\widehat{Z} = \text{diag}\{\widehat{\lambda}_{jk}^{1/2}, k = 1, \ldots, K_j, j = 1, \ldots, p\}$ and $\{\widehat{\lambda}_{jk}\}$ are the estimated eigenvalues. The $\widehat{R}$ is set to be identity. The truncation parameters $\{K_j\}$ are determined by controlling the fraction of variables explained to be above $0 \cdot 9$. The proportion of uniform proposal of the small-world sampler is set to be $0 \cdot 2$. A total of $5,000$ MCMC iterations are conducted. Starting from the empty graph, the chain reaches the true underlying graph in around 500 iterations. We have also tried running multiple chains with varying initial graphs. All chains result in the same posterior mode which is the true graph.

We compared the performance of our approach with three other methods: the Gaussian graphical model of Jones et al. (2005) using Metropolis-Hastings with local proposal, the graphical lasso of Friedman et al. (2008), and the matrix normal graphical model of Wang & West (2009). As the former two methods both assume vector data, we reduce the multivariate functional data to multivariate data by replacing each function by its first principal component score. The third method assumes matrix-variate data. We take the first five principal component scores for each function and stack them row by row to form a 6×5 matrix for each sample. The matrix normal graphical model provides graph estimates both across the rows and across the columns of the data matrix. In our case, only the graph across the rows is of interest.

Some summary statistics are listed on the top panel of Table 1. The running times were recorded on a laptop with Intel(R) Core(TM) i5 CPU, M430 @2.27 GHZ processor and 4GB RAM. As for computing speed, the graphical lasso algorithm seems to be the fastest since it does not require posterior sampling. However, as a frequentist method, it requires pre-specification of the tuning parameter which typically relies on cross-validation. Here we only reported the result using a tuning parameter that produces the best estimation, which gives this method an unfair

advantage in both running time and the prediction accuracy. The mis-estimation rate reported is the proportion of missed or over-estimated edges, which is averaged across all posterior samples if using Bayesian methods. The sensitivity is the proportion of missed edges among the true edge pairs. The specificity is the proportion of over-estimated edges among the true non-edge pairs. The matrix normal graphical model is slow compared with all other methods because it involves marginal density approximations and the code is developed in Matlab whereas the other two are in C++. The top panel of Table 1 shows that the proposed Algorithm 1 provides the smallest mis-estimation rate. Furthermore, although only using the first principal component per function, the Gaussian graphical model and graphical lasso perform reasonably well. This suggests that for more complex problems, we can use either of these methods to get an initial estimate of the graph. The matrix normal graphical model tends to under-estimate the number of edges, and the mis-estimation rate is relatively high.

4·2. *Simulation 2*

In this simulation, white noise was added to the functional data generated in simulation 1, with variances generated from a gamma distribution with mean 2·5 and variance 0·25, resulting in signal-to-noise ratio around 9, where the signal-to-noise ratio is defined by $f_{ij}(t)/\text{var}\{\varepsilon_{ij}(t)\}$ and is averaged across grids and samples. Algorithm 2 is applied. The eigenbasis is firstly estimated based on the pre-smoothed curves obtained from a locally weighted least squares smoother. Functional principal component scores are then obtained by projecting raw data on the estimated eigenbasis. The Λ matrix is determined using the noise variances estimated in the pre-smoothing step. Other parameters are set similarly as in Algorithm 1. The method is also compared with the same three methods as well as the Algorithm 1. The summary statistics were reported on the bottom panel of table 1. Similar patterns are observed as in Simulation 1. Using the posterior samples of $\{\mathrm{c}_i\}$, we can estimate the underlying function $f_{ij}(t)$ and their time

Table 1. *Summary statistics and comparison with three other methods*

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 (n=200)	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 (n=200)	FDGM-N	3 - 5	112·27	8·00	1	0·00	1·0	1·0
	FDGM-S	3 - 5	36	7·75	4	0·02	0·97	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

Sec: sensitivity; Spec: specificity; FDGM-S: functional data graphical model–smooth case, based on Algorithm 1; FDGM-N: functional data graphical model for noisy data, based on Algorithm 2; GGM-MH: Gaussian graphical model; gLasso: graphical lasso; MNGM: matrix normal graphical model.

domain correlations. Related plots can be found in the supplementary material. An additional comparison was performed by applying Algorithm 1 to the pre-smoothed noisy data. The results show that the Algorithm 1 performs almost as well as Algorithm 2, with a mis-estimation rate 0·02.

5. Real data analysis

We demonstrate the performance of the proposed method using EEG data measured through multiple electrodes. Data were obtained from a study examining EEG correlates of genetic predisposition to alcoholism. Measurements were obtained from multiple electrodes placed on subject's scalps which catch EEG signals at 256 Hz for one second. There were 122 subjects, 77

of which are in the alcoholism group with the remaining 45 in the control group. Each subject completed 120 trials. During each trial, the subject was exposed to either a single stimulus or two stimuli, which were a single picture or a pair of pictures shown on a computer monitor. Data were provided by Henri Begleiter at the Neurodynamics Laboratory, State University of New York Health Center at Brooklyn, and is publicly available on UC Irvine Machine Learning Repository. We select 10 electrodes, namely F1, F2, C3, C4, P3, P4, O1, O2, TP7, TP8 according to 10-20 *system*, American EEG Society 1990. We aim to estimate the conditional independence structure between the corresponding EEG signals, which may reflect the correlations between different regions of the brain and hence can be useful for studying brain connectivities (Zheng & Rajapakse, 2004).

Before applying the proposed Gaussian process graphical models, several preprocessing steps were performed. The measurement points at the beginning and the end of each trial are more susceptible to experimental artifacts. Therefore, we truncate the signals at both ends by removing the points according to the first 20 ms and the last 200 ms. This gives 200 points for each EEG trajectory. Additionally, the data show non-Gaussian behavior both globally and regionally. A simple preprocessing step was conducted to remove the extreme outliers. A curve is detected as an outlier if it has measurements going beyond quantile bounds of [0·005,0·997]. These preprocessing procedures result in 10-variate functional data with sample size 2182 for the alcoholism group and 2189 for the control group. In this study we neither differentiate the trials associated with different stimulus, nor differentiate them according to different subjects.

Based on the preprocessed data, we apply the functional principal component analysis to estimate their eigenfunctions and the corresponding FPC scores. By retaining 70% of the total variability, we reduce the dimension of the alcoholic group data from 2000 with 200 points per curve to 51 with $4-6$ scores per curve, and reduce the dimension of the control group data to

865 50 with $4-6$ per curve. The initial graph was estimated via Gaussian graphical model using
866 the first principal component scores. Algorithm 1 was applied to both datasets. It took around
867 50 minutes to complete $30,000$ iterations on the same machine reported in the simulation study.
868 A burn-in period of $20,000$ is removed in the posterior inference. For both datasets, the pos-
869 terior distributions have very spiky shapes which brings challenges to the mixing. We used the
870 small-world sampler with tempering to improve mixing. The convergence of the Algorithm 1
871 was assessed through running multiple chains with varying initial graphs. Due to the spiky shape
872 of the posterior, the chains tend to settle near a local mode around the maximum a posteriori
873 value. The posterior frequency for the mode is above 80% for both datasets.

874 In Figure 2, we plot the posterior modes for both groups and the marginal inclusion prob-
875 abilities, where the latter is defined by the frequency of having an edge for each possible
876 edge. For illustration purposes, the graphs were superimposed on a head diagram which in-
877 dicates the positioning of the 10 EEG electrodes. The bottom plot of Figure 2 indicates
878 that most of the edges in the modal graph have marginal inclusion probabilities close to 1,
879 and the colored edges are significantly different between the two groups (inclusion prob-
880 abilities are close to 1 for one group and are less than 0.2 for the other). For a 10-
881 node graph, there are 45 possible edges. A direct comparison of the posterior modes shows
882 that 31 edges are the same for the alcoholic and control group. For the alcoholic group,
883 the algorithm detected edges between pairs $(F_1, P_4), (F_1, O_2), (F_1, TP_7), (C_3, O_2), (C_4, TP_7)$
884 and (O_2, TP_7), while these pairs are estimated to be conditionally independent by
885 the control group. On the other hand, the algorithm detected edges between pairs
886 $(F_2, P_3), (F_2, P_4), (F_2, TP_8), (C_3, P_4), (C_3, O_1), (C_3, TP_8), (C_4, O_1)$ and (P_4, O_1) in the con-
887 trol group, while these pairs are found to be conditionally independent for the alcoholic group.
888 Similar results were obtained using Algorithm 2 which requires longer computation time.
889
890
891
892
893

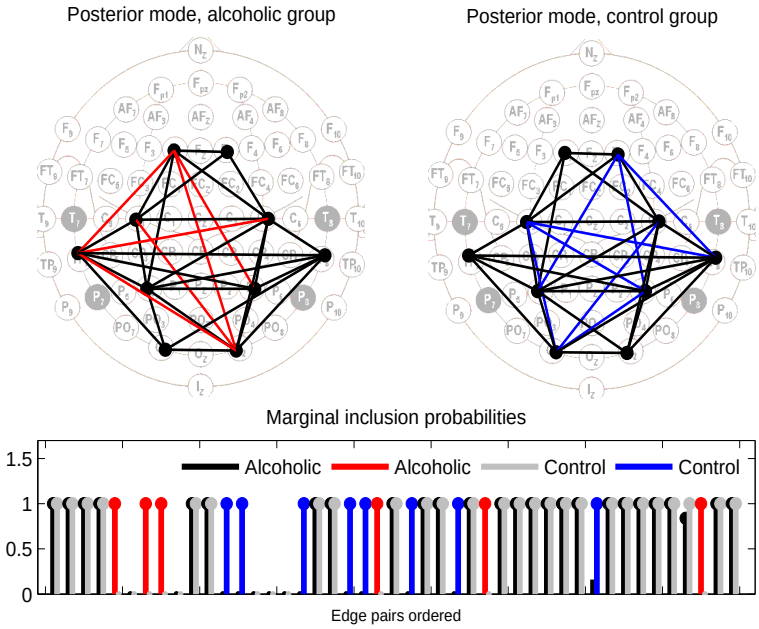

Fig. 2. Graphs at the posterior modes, and the marginal inclusion probabilities. Upper panels: black edges are common to the modal graph for both groups; red/blue edges are specific to the alcoholic/control group, respectively. Bottom panel: Solid round dots on top of vertical bars indicate edges found in the modal graphs.

6. DISCUSSION

We have focused on decomposable graphs. In case of non-decomposable graphs, the block hyper-inverse Wishart prior still applies with the exception that the collection of clique sets $\mathcal{C}$ should be replaced by the collection of prime components $\mathcal{P}$ which may contain non-complete components. Correspondingly, the densities $p(Q_C|\delta, U_C)$ should be replaced by $p(Q_P|\delta, U_P)$, with $P \in \mathcal{P}$. For non-complete prime components, $p(Q_P|\delta, U_P)$ is not the density of an inverse-Wishart anymore because the missing edges provide an extra constraint to the distribution. The distribution contains non-free elements of U_P which can be determined by the free elements. A widely used form that preserves conjugacy is the Diaconis-Ylvisaker prior (Roverato, 2002), where the most challenging part is to estimate its normalizing constant (Giudici & Castelo, 2003;

Jones et al., 2005). Extension to non-decomposable graphs is therefore more computationally intensive than the decomposable cases.

In our model fitting, the truncation parameters $\{K_j\}$ are predetermined using approximation criteria such as the fraction of variance explained. One can consider setting priors to K_j, in which case hybrid Markov chain Monte Carlo algorithms need to be used for fitting both model (5) and (6). The posterior sampling of the latter model would be much more difficult because the dimension of the parameter space changes whenever $\{K_j\}$ are updated.

ACKNOWLEDGEMENT

This work was partially supported by the National Institute of Environmental Health Sciences, U.S.A. Part of the work was done when the first author worked in the Analysis of Object Data program at Statistical and Applied Mathematical Sciences Institute, U.S.A.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes more details on algorithm 1, model parameter setup, methods for improving mixing and extra results for simulation 2.

APPENDIX

Proof of Proposition 1. From the Definition 1, we have that $f_l \perp f_m \mid \mathrm{f}_{(-l,-m)}$ if and only if $f_l(\mathrm{t}_l) \perp f_m(\mathrm{t}_m) \mid \{f_j(\mathrm{t}_j), j \neq l, m\}$ for any t, where $f_j(\mathrm{t}_j) = (f_j(t_{j1}), \ldots, f_j(t_{jn_j}))^T$. We just need to prove that the statement holds for the vectorized f with $\mathrm{f} \sim \mathrm{N}(\mathrm{f}_0, Q)$. Let $N(\mu_{lm,lm|(-l,-m)}, \Sigma_{lm,lm|(-l,-m)})$ be the conditional distribution of the blocks $\{f_l(\mathrm{t}_l), f_m(\mathrm{t}_m)\}$ conditional on the rest. We can partition $\Sigma_{lm,lm|(-l,-m)}$ into four blocks:$\{\Sigma_{i,j|(-l,-m)}, i, j = l, m\}$. Then the blockwise conditional independence hold if and only if the off-diagonal block $\Sigma_{l,m|(-l,-m)} = 0$. On the other hand, since $K = Q^{-1}$ and $K_{l,m}$ is the (l, m)th block of K, similar argument as in (C.3) in Lauritzen (1996) implies that

$K_{lm,lm} = \Sigma^{-1}_{lm,lm|(-l,-m)}$, where $K_{lm,lm}$ can also be partitioned into four blocks $\{K_{i,j}, i,j = l,m\}$. Applying the formula (B.2) of in Lauritzen (1996) to the partitioned form of $K_{lm,lm}$, we have that $\Sigma_{lm|(-l,-m)} = -(K_{l,l} - K_{l,m}K^{-1}_{m,m}K_{m,l})^{-1}K_{l,m}K^{-1}_{m,m}$, which equals zero if and only if $K_{l,m} = 0$, assuming that the matrix inverse exists. □

Proof of Proposition 2. Since $\mathrm{f} \sim \mathrm{MGP}(\mathrm{f}_0, \mathcal{Q})$, for any finite grid t, we have $(\mathrm{f} \mid \mathrm{f}_0, Q) \sim \mathrm{N}(\mathrm{f}_0, Q)$. Therefore the marginal distributions are: $(\mathrm{f}_C|\mathrm{f}_{0,C}, Q_C) \sim \mathrm{N}(\mathrm{f}_{0,C}, Q_C)$, for all $C \in \mathcal{C}$, where f_C is the subvector of f that corresponds to functions in set C and Q_C is the corresponding diagonal block of Q. We just need to show that the marginals for $\{\mathrm{f}_C, C \in \mathcal{C}\}$ are pairwise consistent, the construction procedure (3)-(4) and the Theorem 2.6 of Dawid & Lauritzen (1993) then imply that the distribution with density (2) is the unique Markov distribution over G. The pairwise consistency is obvious due to the marginalization properties of multivariate Gaussian. In particular, for any $A, B \subseteq V$, we have $(\mathrm{f}_A \mid \mathrm{f}_{0,A}, Q_A) \sim N(\mathrm{f}_{0,A}, Q_A)$, $(\mathrm{f}_B \mid \mathrm{f}_{0,B}, Q_B) \sim N(\mathrm{f}_{0,B}, Q_B)$, both yield the marginals $(\mathrm{f}_{A\cap B} \mid \mathrm{f}_{0,A\cap B}, Q_{A\cap B}) \sim N(\mathrm{f}_{0,A\cap B}, Q_{A\cap B})$. □

Definitions: Some definitions used in Lemma 1 and Theorem 1 are listed as follows: (I) *Projection map.* Let R be the real line and T be an index set. Consider the (double indexed) Cartesian product space $R_{T\times T} = \prod_{(\alpha,\beta)\in T\times T} R_{(\alpha,\beta)}$. For a fixed point $(\alpha, \beta) \in T \times T$, we can define the projection map $\pi_{(\alpha,\beta)} : R_{T\times T} \to R_{(\alpha,\beta)}$ 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 can define the partial projection $\pi_B : R_{T\times T} \to R_B$ 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 can define the partial subprojections $\pi_{B_2\leftarrow B_1} : R_{B_1} \to R_{B_2}$, 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 $R_{(\alpha,\beta)}$. We can create a σ-algebra on $R_{T\times T}$ by pulling back the $\mathcal{B}_{(\alpha,\beta)}$ using the inverse of the projection map, i.e. 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 can then define a product σ-algebra as $\mathcal{B}_{T\times T} = \prod_{(\alpha,\beta)\in T\times T} \mathcal{B}_{(\alpha,\beta)}$, 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 can define the pushforward measure $\mu_B = (\pi_B)_* \mu_{T\times T}$ on 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 $(\pi_{B_2 \leftarrow B_1})_* \mu_{B_1} = \mu_{B_2}$.

LEMMA 1. *Suppose that δ is a positive integer and that $T \subset R$ is compact. Further suppose that $u : T \times T \to R$ is a random symmetric positive definite kernel. Then there exists a unique probability measure $\mu_{T\times T}$ on $(R_{T\times T}, \mathcal{B}_{T\times T})$ satisfying*

i. for any finite $\mathrm{t} \subset T$, we have $(\pi_{\mathrm{t}\times\mathrm{t}})_* \mu_{T\times T} = \mu_{\mathrm{t}\times\mathrm{t}}$, where $\mu_{\mathrm{t}\times\mathrm{t}}$ is the law of IW$(\delta, U_{\mathrm{t}\times\mathrm{t}})$, and $(U_{\mathrm{t}\times\mathrm{t}})_{ij} = u(t_i, t_j)$;

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

Proof of Lemma 1. Denote $Q_{\mathrm{t}\times\mathrm{t}}$ the random matrix formed by evaluating u on $\mathrm{t} \times \mathrm{t}$. We will prove using Theorem 2.4.3 of Tao (2011) as follows: (1) First, we verify that the μ_B are compatible for all finite $B \subset T \times T$. There are two successive cases we shall consider. Case 1: Suppose $\mathrm{t}_2 \subset \mathrm{t}_1$ are two finite subsets of T, then $Q_{\mathrm{t}_2\times\mathrm{t}_2}$ is the submatrix of $Q_{\mathrm{t}_1\times\mathrm{t}_1}$ obtained by deleting the rows and columns with indices in $\mathrm{t}_1 \setminus \mathrm{t}_2$. If $Q_{\mathrm{t}_1\times\mathrm{t}_1}$ has law $\mu_{\mathrm{t}_1\times\mathrm{t}_1} = \mathrm{IW}(\delta, U_{\mathrm{t}_1\times\mathrm{t}_1})$, then $Q_{\mathrm{t}_2\times\mathrm{t}_2}$ has law $\mathrm{IW}(\delta, U_{\mathrm{t}_2\times\mathrm{t}_2})$ due to the consistency property of the inverse Wishart distribution. Consequently, $(\pi_{\mathrm{t}_2\times\mathrm{t}_2\leftarrow\mathrm{t}_1\times\mathrm{t}_1})_* \mu_{\mathrm{t}_1\times\mathrm{t}_1} = \mu_{\mathrm{t}_2\times\mathrm{t}_2}$. Case 2: Let $B_1 = \{(\alpha_i, \beta_i)\}_{i=1}^n \subset T \times T$ 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,

$$\begin{aligned}(\pi_{B_2\leftarrow B_1})_* \mu_{B_1} &= (\pi_{B_2\leftarrow B_1})_* (\pi_{B_1\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_* \mu_{\mathrm{g}_1\times\mathrm{g}_1} = (\pi_{B_2\leftarrow B_1} \circ \pi_{B_1\leftarrow \mathrm{g}_1\times\mathrm{g}_1})_* \mu_{\mathrm{g}_1\times\mathrm{g}_1} \\ &= (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2} \circ \pi_{\mathrm{g}_2\times\mathrm{g}_2\leftarrow\mathrm{g}_1\times\mathrm{g}_1})_* \mu_{\mathrm{g}_1\times\mathrm{g}_1} = (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2})_* (\pi_{\mathrm{g}_2\times\mathrm{g}_2\leftarrow\mathrm{g}_1\times\mathrm{g}_1})_* \mu_{\mathrm{g}_1\times\mathrm{g}_1} \\ &= (\pi_{B_2\leftarrow \mathrm{g}_2\times\mathrm{g}_2})_* \mu_{\mathrm{g}_2\times\mathrm{g}_2} = \mu_{B_2}, \end{aligned} \tag{A1}$$

where the second to last equality holds because of our demonstration in Case 1.

(2) Second, we claim that the finite dimensional measure $\mu_{\mathrm{t}\times\mathrm{t}} = \mathrm{IW}(\delta, U_{\mathrm{t}} \times \mathrm{t})$ is an inner regular probability measure on the product σ-algebra $\mathcal{B}_{\mathrm{t}\times\mathrm{t}}$. We will show that $\mu_{\mathrm{t}\times\mathrm{t}}$ is a finite Borel measure on a Polish space, which then implies that $\mu_{\mathrm{t}\times\mathrm{t}}$ is regular, hence inner regular by Lemma 26.2 of Bauer (2001). This is done through (i)-(iii) as follows: (i) For finite t, $Q_{\mathrm{t}\times\mathrm{t}}$ takes values in the space of symmetric and positive definite matrices, denoted by $\Omega_{|\mathrm{t}|}$ where $|\mathrm{t}|$ denotes the size of t. Since the subset of symmetric matrices is closed in $R_{\mathrm{t}\times\mathrm{t}}$, it is Polish. Furthermore, the space of symmetric positive definite matrices is an open convex cone in the space of symmetric matrices, hence it is Polish as well. Therefore the space $\Omega_{|\mathrm{t}|}$ is Polish. (ii) Since $\mu_{\mathrm{t}\times\mathrm{t}}$ (the law of $Q_{\mathrm{t}\times\mathrm{t}} \sim IW(\delta, U_{\mathrm{t}\times\mathrm{t}})$) has an almost everywhere continuous density function, $\mu_{\mathrm{t}\times\mathrm{t}}$ is a measure defined by Lebesgue integration against an almost everywhere continuous function. Therefore $\mu_{\mathrm{t}\times\mathrm{t}}$ is Borel on $\Omega_{|\mathrm{t}|}$. As $\Omega_{|\mathrm{t}|} \subset R_{\mathrm{t}\times\mathrm{t}}$, we may extend the measure $\mu_{\mathrm{t}\times\mathrm{t}}$ from $\Omega_{|\mathrm{t}|}$ to $R_{\mathrm{t}\times\mathrm{t}}$ via the Carathéodory theorem. In particular, define $\widetilde{\mu}_{\mathrm{t}\times\mathrm{t}}(A) = \mu_{\mathrm{t}\times\mathrm{t}}(A \cap \Omega_{|\mathrm{t}|})$ for $A \in \mathcal{B}(R_{\mathrm{t}\times\mathrm{t}})$. With extension, $\mu_{\mathrm{t}\times\mathrm{t}}$ is Borel on $R_{\mathrm{t}\times\mathrm{t}}$, and the σ-algebra associated is $\mathcal{B}(R_{\mathrm{t}\times\mathrm{t}}) = \mathcal{B}_{\mathrm{t}\times\mathrm{t}} = \prod_{(\alpha,\beta)\in T\times T} \mathcal{B}_{(\alpha,\beta)}$. (iii) The measure $\mu_{\mathrm{t}\times\mathrm{t}}$ 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 $\mu_{T\times T}$ on the product σ-algebra $\mathcal{B}_{T\times T}$ that satisfies (i) and (ii). □

Proof of Theorem 1. We will apply Theorem 2.4.3 of Tao (2011) in a similar fashion as in Lemma 1. Details are listed below:

1. For each clique $C \in \mathcal{C}$, the Inverse Wishart distribution of Q_C extends to a unique probability measure on $\mathcal{B}_{\widetilde{T}_C\times\widetilde{T}_C}$ with $\widetilde{T}_C = \coprod_{j\in C} T_j$. This can be shown following the similar arguments as in the proof of Lemma 1, except replacing T in the proof by $\widetilde{T}_C$. Similarly, the extension also holds for the Inverse Wishart measure of Q_S for $S \in \mathcal{S}$. So the compatibility and regularity conditions hold for the marginal distributions at $C \in \mathcal{C}$ and $S \in \mathcal{S}$.
2. The Inverse Wishart defined for $\{Q_C, C \in \mathcal{C}\}$ admits consistency between pairs of blocks. Specifically, if $B \subseteq C_l \cap C_m$, then the distribution for Q_{C_l} and Q_{C_m} both yield the same marginal distribution for

the submatrix Q_B, which is $IW(\delta, U_B)$. This is again implied by Lemma 7.4 (i) of Dawid & Lauritzen (1993). Therefore, a Markov distribution in form of (3) can be constructed recursively following the procedure in (3)-(4) of Dawid & Lauritzen (1993) and this distribution is unique for discretization t by Theorem 2.6 of Dawid & Lauritzen (1993). This construction procedure remains the same as long as the Graph G is fixed.

3. We now show the compatibility conditions. There are two cases. Case 1. For finite grids $\mathrm{t}_2 \subset \mathrm{t}_1 \subset \widetilde{T}$, assume that both t_2 and t_2 contain at least one grid point on each T_j. The blockwise matrix $Q_{\mathrm{t}_2\times\mathrm{t}_2}$ is a submatrix of $Q_{\mathrm{t}_1\times\mathrm{t}_1}$, both with the same $p \times p$ block structure. To find the measure $\mu_{\mathrm{t}_2\times\mathrm{t}_2}$, it is sufficient to find the distributions of $\{Q_{\mathrm{t}_2\times\mathrm{t}_2,C}, Q_{\mathrm{t}_2\times\mathrm{t}_2,S}, C \in \mathcal{C}, S \in \mathcal{S}\}$, which are Inverse Wishart and are consistent with the Inverse Wishart used to construct $\mu_{\mathrm{t}_1\times\mathrm{t}_1}$ based on the argument in 1. The construction procedure described in 2 can then be applied to build the distribution of $Q_{\mathrm{t}_2\times\mathrm{t}_2}$, which gives that $Q_{\mathrm{t}_2\times\mathrm{t}_2} \sim \text{BHIW}(G, \delta, U_{\mathrm{t}_2\times\mathrm{t}_2})$. Therefore $(\pi_{\mathrm{t}_2\times\mathrm{t}_2\leftarrow\mathrm{t}_1\times\mathrm{t}_1})_*\mu_{\mathrm{t}_1\times\mathrm{t}_1} = \mu_{\mathrm{t}^2\times\mathrm{t}^2}$. Case 2. Now consider an arbitrary set $B_1 = \{(\alpha_j, \beta_j)\}_{j=1}^n$ with $(\alpha_j, \beta_j) \in \widetilde{T} \times \widetilde{T}$. Let $B_2 = \{(\widetilde{\alpha}_j, \widetilde{\beta}_j)\}_{j=1}^m$ be a subset of B_1. We would like to show that $(\pi_{B_2\leftarrow B_1})_*\mu_{B_1} = \mu_{B_2}$. Let g_2 be a grid of $\widetilde{T}$ which has at least one grid point on each T_j and such that $\{\widetilde{\alpha}_j\} \cup \{\widetilde{\beta}_j\} \subset \mathrm{g}_2$. Let $\mathrm{g}_1 = \mathrm{g}_2 \cup \{\alpha_j\} \cup \{\beta_j\}$, then $B_2 \subset \mathrm{g}_2 \times \mathrm{g}_2 \subset \mathrm{g}_1 \times \mathrm{g}_1$ and $B_2 \subset B_1 \subset \mathrm{g}_1 \times \mathrm{g}_1$. We can then use the similar argument as in (A1) to show the compatibility.
4. We now show that the regularity condition holds. Since the density of (3) is continuous almost everywhere, the corresponding law is then a measure in the form of Lebesgue integration against a continuous function. Therefore the block hyper-inverse Wishart measure of Q is a Borel measure on the space of block-wise symmetric positive definite matrices denoted by $\Omega_{|\mathrm{t}|}$, which is again Polish. Therefore the probability measure of Q is regular. We can also extend the measure from $\Omega_{|\mathrm{t}|}$ to $R_{\mathrm{t}\times\mathrm{t}}$ using Carathéodory theorem. Therefore the measure of Q is an inner regular probability measure on the product σ-algebra $\mathcal{B}_{\mathrm{t}\times\mathrm{t}}$.

From 1-4, both the compatibility and the inner regular conditions hold. Hence there exists a unique probability measure on the product σ-algebra $\mathcal{B}_{\widetilde{T}\times\widetilde{T}}$ that satisfies (i) and (ii).

Proof of Theorem 2. For a fixed clique C, $Q_C \sim \text{IW}(\delta, U_C)$ is the conjugate prior for the distribution $\mathrm{f}_{i,C}|\mathrm{f}_{0,C} \sim N(\mathrm{f}_{0,C}, Q_C)$. The conjugacy also holds for Q_S. Direct computation based on the likelihood in (2) and prior in (3) gives the posterior density corresponds to BHIW$(G, \widetilde{\delta}, \widetilde{\mathcal{U}})$. In fact, applying (iv) of Lemma 7.4 and Proposition 3.16 of Dawid & Lauritzen (1993) we see that BHIW is *strong hyper Markov*. Corollary 5.5 of Dawid & Lauritzen (1993) then implies that a strong hyper Markov prior induces a unique posterior distribution which is also strong Markov, and is specified by the clique-marginal laws. This also proves that the posterior is BHIW$(G, \widetilde{\delta}, \widetilde{\mathcal{U}})$. Furthermore, by Proposition 5.6 of Dawid & Lauritzen (1993), with a strong hyper Markov prior, we can integrate out Q to get the marginal distribution of $\{\mathrm{f}_i\}$ which is again a Markov law. Simple derivations give the formula in (4) □

REFERENCES

CARVALHO, C. M. & SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.

CARVALHO, C. M. & WEST, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Anal.* **2**, 69–98.

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

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

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

FURRER, R., GENTON, M. G. & NYCHKA, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15**, 502–523.

GIUDICI, P. (1996). Learning in graphical Gaussian models. *Bayesian Statistics 5* , 621–628.

GIUDICI, P. & CASTELO, R. (2003). Improving Markov chain Monte Carlo model search for data mining. *Machine Learning* **50**, 127–158.

GIUDICI, P. & GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 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.

HIGDON, D. (2002). *Quantitative methods for current environmental issues*, chap. Space and space-time modeling using process convolutions. Springer Verlag, pp. 37 – 56.

HUGHES, G. (1968). On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on* **14**, 55 – 63.

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**, 388–400.

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

LENKOSKI, A. & DOBRA, A. (2011). Computational aspects related to inference in gaussian graphical models with the g-wishart prior. *J. Comput. Graph. Statist.* **20**, 140–157.

MILLER, F., VANDOME, A. & MCBREWSTER, J. (2010). *Curse of Dimensionality*. VDM Publishing House Ltd.

MORRIS, J. S. & CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Statist. Soc. B* **68**, 179–199.

RAMSAY, J. O. & SILVERMAN, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.

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.

SCOTT, J. G. & CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist.* **17**, 790–808.

TAO, T. (2011). *An Introduction to Measure Theory*. Graduate Studies in Mathematics. American Mathematical Society.

WANG, H. & CARVALHO, C. M. (2010). Simulations of hyper-inverse Wishart distributions for non-decomposible graphs. *Electron. J. Statist.* **4**, 1470–1475.

WANG, H. & WEST, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96**, 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.

ZHENG, X. & RAJAPAKSE, J. C. (2004). Graphical models for brain connectivity from functional imaging data. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 1.

[*Received January* 2011. *Revised June* 2011]