Electronic Journal of Statistics
Vol. 5 (2011) 981–1014
ISSN: 1935-7524
DOI: 10.1214/11-EJS634

Sparse covariance estimation in heterogeneous samples*

Abel Rodríguez†

Department of Applied Mathematics and Statistics
University of California, Santa Cruz, California
e-mail: abel@ams.ucsc.edu

Alex Lenkoski‡

Department of Applied Mathematics
Heidelberg University, Heidelberg, Germany
e-mail: alex.lenkoski@uni-heidelberg.de

and

Adrian Dobra§

Departments of Statistics, Biobehavioral Nursing and
Health Studies and the Center for Statistics and the Social Sciences
University of Washington, Seattle, Washington
e-mail: adobra@uw.edu

Abstract: Standard Gaussian graphical models implicitly assume that the conditional independence among variables is common to all observations in the sample. However, in practice, observations are usually collected from heterogeneous populations where such an assumption is not satisfied, leading in turn to nonlinear relationships among variables. To address such situations we explore mixtures of Gaussian graphical models; in particular, we consider both infinite mixtures and infinite hidden Markov models where the emission distributions correspond to Gaussian graphical models. Such models allow us to divide a heterogeneous population into homogenous groups, with each cluster having its own conditional independence structure. As an illustration, we study the trends in foreign exchange rate fluctuations in the pre-Euro era.

AMS 2000 subject classifications: Primary 62F15, 62H25; secondary 62H30, 62M10.
Keywords and phrases: Covariance selection, Gaussian graphical model, mixture model, Dirichlet process, hidden Markov model, nonparametric Bayes inference.

Received March 2011.

*We would like to thank Mike West for providing access to the exchange rate data set used in Section 6.

†AR gratefully acknowledges support by the National Science Foundation, grant DMS 0915272.

‡AL gratefully acknowledges support by the German Research Foundation (DFG) within the programme "Spatio-/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653.

§AD gratefully acknowledges support by the National Science Foundation, grant DMS 1120255.

Contents

1. Introduction

Problems with small sample sizes and a large number of unknown parameters represent one of the most challenging areas of current statistical research. Graphical models deal with these ill-posed problems by enforcing sparsity in the conditional dependence structure among outcomes. More specifically, given a random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$, a graphical model for X encodes the conditional independence relationships between its components through a p-vertex graph G, such that vertex i represents component X_i and the lack of an edge between nodes i and j indicates that variables i and j are conditionally independent given the rest. In particular, Gaussian graphical models (GGMs), also known as covariance selection models [13], have become extremely popular in applications ranging from genetics [61, 11] to econometrics and finance [10, 15]. Gaussian graphical models assume that the joint distribution of X follows a multivariate Gaussian distribution, and therefore conditional independence among variables can be enforced by setting to zero the appropriate off-diagonal elements of the inverse covariance (precision) matrix.

One important shortcoming of Gaussian graphical models is that they implicitly assume a linear relationship between variables. Copulas have been used in the context of graphical models to address nonlinearities. For example, [5] decompose the joint distribution of X using pairwise copulas; however, the resulting models are computationally difficult to fit, especially when p grows. An alternative to copulas is to model non-linearities through mixtures of Gaussian

graphical models. Countable mixture models explain nonlinearities in the conditional expectations as a consequence of hetherogeneity of the population, and can therefore be interpreted as providing adaptive local linear fits [41, 48].

As a motivation for investigating mixtures of Gaussian graphical models, consider the analysis of gene expression data. Gaussian graphical models have often been used in the context of microarray data, where the graph encoding the conditional dependence structure provides information about expression pathways [16, 22, 11]. The implicit assumptions in these models is that the expression pathways are the same for all individuals/tissues in the sample and that expression levels on different genes are linearly related. In practice, these assumptions might not be justified if the underlying population is heterogeneous. Similarly, when studying the relationship between economic variables such as exchange rates, graphical models allow us to identify groups of countries that form economic blocks and understand how these blocks interact with each other. However, as trade patterns evolve, we expect that both the block membership and the modes in which countries interact might change, making the constant-graph assumption unrealistic. In both of these settings, mixtures of Gaussian graphical models not only provide us with a tool to induce sparsity in heterogeneous samples, but also generate interpretable models.

One of the major challenges in implementing mixtures of Gaussian graphical models is computational, and relate both to the determination of the underlying graph associated with each component in the mixture and to the estimation of the number of components. It is well known that the number of possible partitions of a sample grows exponentially with the size of the dataset, a problem that is compounded when we desire to also estimate the number of components in the mixture and the graphical structure corresponding to each cluster. Work in finite mixtures of graphical models goes back at least to [55], who fixed the number of components and developed a search algorithm that used a modified Cheeseman-Stutz approximation to the marginal likelihood coupled with Expectation-Maximization steps to estimate component-specific parameters. To the best of our knowledge, the problem of determining the number of components in mixtures of graphical models has not been properly addressed before.

We present a fully Bayesian approach to inference in nonparametric mixtures and infinite hidden Markov models with Gaussian graphical models as kernel/emission distributions. Using infinite mixture models provides full support in the space of continuous distributions [39, 44], and allows us to automatically deal with an unknown number of components/states within a simple computational framework. As in [59], the hidden Markov models we discuss allow for the graph encoding the conditional independence structure of the data to change over time, an important feature that has been missing in other multivariate time series models employing graphical models [10, 60]. However, our framework allows us to deal with both decomposable and nondecomposable graphical models using collapsed Markov chain Monte Carlo algorithms that avoid explicit representation of the unknown mixing distributions, and allow us to identify structural changes in the underlying data-generation process. Al-

though the paper develops models based on Gaussian graphical models, the approaches we discuss are not restricted to multivariate continuous outcomes, but can be extended to incorporate combinations of binary, ordinal and continuous variables by introducing latent auxiliary variables.

Sparse estimation in heterogeneous samples has been a topic of recent interest in the literature. For example, [26] present a penalized likelihood approach for the joint estimation of multiple graphical models when the samples arise from known classes. In contrast, we consider the problem of estimating multiple graphical models when the classes are unknown and need to be inferred from the data. In work related to ours, [27] discussed the construction of a nonparametric prior that is Markov with respect to a given graph G. The main result in the paper is that, for absolutely continuous baseline measures, a hyper Dirichlet process with respect to G (i.e., a Dirichlet process law that is Markov with respect to a given graph G) can be generated by choosing a baseline measure to be itself Markov with respect to G. As an application, the hyper Dirichlet process is used to construct a nonparametric mixture of graphical models that uses a common graph to describe the conditional independence structure for all components. Therefore, our model can be seen as a generalization of the hyper-Dirichlet process mixture proposed by [27] where each component is Markov with respect to a (potentially) different graph.

To simplify our exposition we begin by reviewing Bayesian approaches to inference in Gaussian graphical models in Section 2 and introducing Dirichlet process mixtures of Gaussian graphical models in Sections 3 and 4. We then move to discuss infinite hidden Markov models whose emission distributions correspond to Gaussian graphical models. These models are illustrated in Section 6 using a series of simulation studies and in Section 7 by studying returns in foreign exchange markets. Finally, we conclude in Section 8 with a discussion of possible extensions and future research directions.

2. A Bayesian framework for Gaussian graphical models

Let $X = X_V$ be the vector of observed variables, where $V = \{1, 2, \ldots, p\}$, and $\mathcal{G}_V$ be the space of all graphs with vertices in the set V. We assume that X follows a multivariate Gaussian distribution $p(X \mid \mu, K) = N_p(\mu, K^{-1})$ with mean vector $\mu \in \mathbb{R}^p$ and $p \times p$ precision matrix $K = (K_{ij})$. The Gaussian graphical model associated with a graph $G = (V, E) \in \mathcal{G}_V$ is obtained by setting to zero the elements of K corresponding with missing edges in G [13]. The absence of the edge $(i, j) \in (V \times V) \setminus E$ implies $K_{ij} = K_{ji} = 0$, which in turn implies that X_i and X_j are conditionally independent given $X_{V \setminus \{i,j\}}$, i.e., the distribution of X_V is Markov with respect to the graph G. Hence, the precision matrix K belongs to the cone P_G of the symmetric positive definite matrices with entries equal to zero for all $(i, j) \in (V \times V) \setminus E$ [3].

The class of decomposable graphs $\mathcal{G}_V^D \subset \mathcal{G}_V$ is particularly appealing from a computational standpoint. Decomposable graphs are those graphs in $\mathcal{G}_V$ such that they do not contain any chordless cycles of length four or larger. Hence, if

G is decomposable, the subgraph $G_C = (C, E_C)$, $E_C = \{(i,j) \in E : i,j \in C\}$, associated with a clique $C \in \mathcal{C}$ is complete, that is, there is no edge missing from it. The dependence of the distribution of X on the graph G can be made explicit by noting that the conditional dependence relationships implied by G induce the following factorization of the joint distribution of X [12]:

$$p(X \mid \mu, K, G) = \frac{\prod_{C \in \mathcal{C}(G)} p(X_C \mid \mu_C, K_C)}{\prod_{S \in \mathcal{S}(G)} p(X_S \mid \mu_S, K_S)} \tag{2.1}$$

where $\mathcal{C}(G)$ denotes the cliques of G and $\mathcal{S}(G)$ denotes separators of G. For an index set $V_0 \subset V$, μ_{V_0} is the subvector of μ corresponding to the entries in V_0, while $K_{V_0} = ((K^{-1})_{V_0})^{-1}$.

We consider the following joint prior distribution for μ and K:

$$p(\mu, K \mid G) = p(\mu \mid K, G) p(K \mid G), \tag{2.2}$$

where, conditional on K, the prior for the mean is $p(\mu \mid K, G) = N_p(\mu_0, (n_0 K)^{-1})$ with $\mu_0 \in \mathbb{R}^p$ and $n_0 > 0$. The prior for the precision matrix $p(K \mid G) = W_G(\delta_0, D_0)$ is a G-Wishart distribution with density [50, 3, 37]

$$p(K \mid G) = \frac{1}{I_G(\delta_0, D_0)} (\det K)^{(\delta_0 - 2)/2} \exp\left\{ -\frac{1}{2} \langle K, D_0 \rangle \right\}, \tag{2.3}$$

with respect to the Lebesgue measure on P_G. Here $\langle B, C \rangle = \operatorname{tr}(B^T C)$ denotes the trace inner product. Diaconnis & Ylvisaker [14] prove that the normalizing constant $I_G(\delta_0, D_0)$ is finite if $\delta_0 > 2$ and $D_0^{-1} \in P_G$. If G is complete (i.e. G is decomposable with only one clique $\mathcal{C} = \{V\}$ and no separators), $W_G(\delta_0, D_0)$ reduces to the Wishart distribution $W_p(\delta_0, D_0)$, hence its normalizing constant is given by

$$I_G(\delta_0, D_0) = 2^{(\delta_0 + p - 1)p/2} \Gamma_p \{(\delta_0 + p - 1)/2\} (\det D_0)^{-(\delta_0 + p - 1)/2}, \tag{2.4}$$

where $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma\left(a - \frac{i}{2}\right)$ for $a > (p-1)/2$ [40]. If G is decomposable but not necessarily complete, Dawid & Lauritzen [12] showed that the G-Wishart distribution $W_G(\delta_0, D_0)$ can be factorized according to the cliques and the separators of G, hence its normalizing constant is equal to [50]:

$$I_G(\delta_0, D_0) \quad = \quad \frac{\prod_{C \in \mathcal{C}(G)} I_{G_C}(\delta_0, (D_0)_C)}{\prod_{S \in \mathcal{S}(G)} I_{G_S}(\delta_0, (D_0)_S)}. \tag{2.5}$$

Since the subgraphs G_C and G_S associated with each clique and separator of G are complete, $I_{G_C}(\delta_0, (D_0)_C)$ and $I_{G_S}(\delta_0, (D_0)_S)$ can be explicitly calculated as in (2.4). Finally, if G is nondecomposable, there is no closed-form expression for $I_G(\delta_0, D_0)$, and its value needs to be approximated either through Monte Carlo simulation or Laplace approximations [3, 36].

The joint prior (2.2) is conjugate, and the posterior distribution of $(\mu, K \mid x^{(1:n)}, G)$ is again a normal/G-Wishart distribution with

$$p(K \mid x^{(1:n)}, G) = W_G(\delta_0 + n, D_0 + U + A). \tag{2.6}$$

$$p(\mu \mid x^{(1:n)}, K, G) = N_p\left(\bar{\mu}, [(n + n_0)K]^{-1}\right), \tag{2.7}$$

with $\bar{\mu} = \frac{n\bar{x}+n_0\mu_0}{n+n_0}$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$, $U = \sum_{i=1}^{n}(x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$, and $A = -(n+n_0)\bar{\mu}\bar{\mu}^T + n\bar{x}\bar{x}^T + n_0\mu_0\mu_0^T$. On the other hand, the marginal likelihood associated with a graph $G \in \mathcal{G}_V$ is is given by

$$p(x^{(1:n)} \mid G) = (2\pi)^{-\frac{np}{2}} \left(\frac{n_0}{n+n_0}\right)^{p/2} \frac{I_G(\delta_0 + n, D_0 + U + A)}{I_G(\delta_0, D_0)}, \tag{2.8}$$

Also, the posterior predictive distribution of a new sample $x^{(n+1)}$ is given by

$$p(x^{(n+1)} \mid x^{(n)}, G) = (2\pi)^{-\frac{p}{2}} \left(\frac{n+n_0}{n+1+n_0}\right)^{p/2} \frac{I_G(\delta_0 + n + 1, D_0 + U + A + \widetilde{A})}{I_G(\delta_0 + n, D_0 + U + A)}, \tag{2.9}$$

where $\widetilde{A} = -(n+1+n_0)\widetilde{\mu}\widetilde{\mu}^T + x^{(n+1)}(x^{(n+1)})^T + (n+n_0)\bar{\mu}\bar{\mu}^T$ and $\widetilde{\mu} = \frac{x^{(n+1)}+(n+n_0)\bar{\mu}}{n+1+n_0}$. As we discussed before, if G is assumed to be decomposable, the posterior normalizing constant $I_G(\delta_0 + n, D_0 + U + A)$ can be calculated directly using a formula similar to equation (2.5), hence $p(x^{(n+1)} \mid G)$ and $p(x^{(n+1)} \mid x^{(n)}, G)$ can also be calculated directly without any numerical approximation techniques. These computations are key to a successful implementation of the sampling algorithms we describe in Section 4.

In the sequel, we assume that the data $x^{(1:n)}$ have been centered and scaled to unit variance, so that the sample mean of each X_i is zero and its sample variance is one. Hence, we complete the prior specification by taking $\mu_0 = 0$, $\delta_0 = 3$ and $D_0 = I_p$, where I_p is the p-dimensional identity matrix. With this assumption, the weight of the prior is equivalent to the weight of one observed sample. Furthermore, this choice implies that the observed variables are independent apriori.

3. Dirichlet process mixtures of Gaussian graphical models

Consider now a finite mixture

$$X \mid \{w_l\}, \{\mu_l^*\}, \{K_l^*\}, \{G_l^*\} \sim \sum_{l=1}^{L} w_l p(X \mid \mu_l^*, K_l^*, G_l^*), \tag{3.1}$$

where $p(X \mid \mu_l^*, K_l^*, G_l^*)$ is given in (2.1). In this model, draws from X come from one of L potentially different graphical models; a realization $x^{(i)}$ comes from the l-th graphical model (which is defined by the parameters μ_l^*, K_l^* and G_l^*) independently with probability w_l. A fully Bayesian specification of the model is completed by eliciting a prior for the parameters $\{w_l\}_{l=1}^L$, $\{\mu_l^*\}_{l=1}^L$, $\{K_l^*\}_{l=1}^L$, and $\{G_l^*\}_{l=1}^L$. A common choice is to set $w = (w_1, \ldots, w_L) \sim \mathsf{Dir}(w^0)$ and let the component specific parameters (μ_l^*, K_l^*, G_l^*) be independent and identically distributed samples from some common distribution M.

Finite mixtures, as the one described above, allow for additional flexibility over regular Gaussian graphical models by allowing a heterogeneous population to be divided into homogenous groups. However, estimating finite mixture

models involves important practical challenges. For example, we generally do not know how many components are present in the population. We could allow L to be random and assign a prior distribution to it, but fitting the resulting model involves the use of reversible-jump Markov chain Monte Carlo methods [25], which are notoriously inefficient for high dimensional mixtures.

As an alternative, we consider Dirichlet process (DP) mixtures of Gaussian graphical models. Note that (3.1) can be alternatively written as

$$X \mid H \sim \int p(X \mid \mu, K, G) H(\mathrm{d}\mu, \mathrm{d}K, \mathrm{d}G) \quad H(\cdot) = \sum_{l=1}^{L} w_l \delta_{(\mu_l^*, K_l^*, G_l^*)}(\cdot) \tag{3.2}$$

where $\delta_a(\cdot)$ denotes the degenerate probability measure putting all of its mass on a. Therefore, eliciting a prior on $(\{w_l\}_{l=1}^L, \{\mu_l^*\}_{l=1}^L, \{K_l^*\}_{l=1}^L, \{G_l^*\}_{l=1}^L)$ is equivalent to defining a prior on the discrete probability measure H, one such prior is the Dirichlet process [19, 20]. A random distribution H is said to follow a Dirichlet process with baseline measure M and precision parameter α_0, denoted $\mathsf{DP}(\alpha_0, M)$, if it has a representation of the form [52]

$$H(\cdot) = \sum_{l=1}^{\infty} w_l \delta_{\theta_l^*}(\cdot), \tag{3.3}$$

where $\theta_1^*, \theta_2^*, \ldots$ are independent and identically distributed samples from the *baseline measure* M and $w_l = v_l \prod_{s<l}(1 - v_s)$ where $v_1, v_2, \ldots$ is another independent and identically distributed sample for which $v_l \sim \mathsf{Beta}(1, \alpha_0)$. We refer to the joint distribution on $(w_1, w_2, \ldots)$ induced by the above construction as a stick breaking distribution with parameter α_0, denoted $\mathsf{SB}(\alpha_0)$. The Dirichlet process mixture model is recovered from (3.2) when $H \sim \mathsf{DP}(\alpha_0, M)$ for appropriately chosen hyperparameters α_0 and M.

Consider now an independent and identically distributed sequence $\theta_1, \ldots, \theta_n$ such that $\theta_j \mid H \sim H$, where $H \sim \mathsf{DP}(\alpha_0, M)$. A useful feature of the Dirichlet process prior is that the joint distribution for $(\theta_1, \ldots, \theta_n)$ obtained after integrating out the random H is given by a sequence of predictive distributions [7] where $\theta_1 \sim M$ and

$$\theta_{j+1} \mid \theta_j, \ldots, \theta_1, \alpha_0 \sim \sum_{i=1}^{j} \frac{1}{\alpha_0 + j} \delta_{\theta_i} + \frac{\alpha_0}{\alpha_0 + j} M, \qquad j > 1. \tag{3.4}$$

The presence of ties in the sequence $\theta_1, \ldots, \theta_n$ sometimes makes it convenient to use an alternative representation where $\theta_1^*, \ldots, \theta_L^*$ denotes the set of $1 \leq L \leq n$ unique values among $\theta_1, \ldots, \theta_n$ and $\xi_1, \ldots, \xi_n$ is a sequence of indicator variables such that $\theta_j = \theta_{\xi_j}^*$. Under this representation, $\theta_1^*, \theta_2^*, \ldots$ is a sequence of independent and identically distributed samples from M, $\xi_1 = 1$ and

$$\xi_{j+1} \mid \xi_j, \ldots, \xi_1, \alpha_0 \sim \sum_{l=1}^{L^j} \frac{r_l^j}{\alpha_0 + j} \delta_l + \frac{\alpha_0}{\alpha_0 + j} \delta_{L^j + 1}, \qquad j > 1, \tag{3.5}$$

where $L^j = \max_{i \le j}\{\xi_i\}$ is the number of distinct values among $\theta_1, \ldots, \theta_j$, and $r_l^j = \sum_{i=1}^{j} \mathbf{1}_{(\xi_i = l)}$ is the number of samples among the first l with $\xi_j = l$. Expressions (3.4) and (3.5) clearly emphasize that, for any finite sample $x^{(1:n)}$, the number of non-empty components $L^n = L$ in a Dirichlet process mixture model is a random parameter in the model. The prior on L implied by the Dirichlet process is given by:

$$p(L \mid \alpha_0, n) = S(n, L) n! \alpha_0^L \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \qquad L = 1, \ldots, n, \tag{3.6}$$

where $S(\cdot, \cdot)$ denotes the unsigned Stirling number of the first kind [1]. Therefore, the mean number of non-empty components grows with α_0, the concentration parameter.

The Dirichlet process mixture model is intimately connected to the finite mixture model in (3.1). Consider a finite mixture with N components such that

$$\begin{aligned} x^{(j)} \mid \{\theta_l^*\}_{l=1}^N, \{\xi_j\}_{j=1}^n &\sim p(x^{(j)} \mid \theta_{\xi_j}^*), \\ \xi_j \mid \alpha_0 \sim \mathsf{Multno}\left(\frac{\alpha_0}{N}, \ldots, \frac{\alpha_0}{N}\right), &\quad \theta_l^* \sim M. \end{aligned} \tag{3.7}$$

As $N \to \infty$, the predictive distribution for $x^{(1:n)}$ under this model converges to the one obtained from the DP mixture [24, 29].

In the Dirichlet process mixture of Gaussian graphical models we explore in this paper, $\theta = (\mu, K, G)$ and the baseline measure is defined by

$$M = p(\mu, K \mid G) p(G), \tag{3.8}$$

where $p(\mu, K \mid G)$ is given by (2.2) and $p(G)$ is the prior on the graph space.

Our framework is capable of accommodating any prior $p(G)$ on the set of graphs. A usual choice is the uniform prior $p(G) \propto 1$, but this prior is biased toward middle size graphs and gives small probabilities to sparse graphs and to graphs that are almost complete. Here the size of a graph G is defined as the number of edges in G and denoted by $size(G) \in \{0, 1, \ldots, p(p-1)/2\}$. [16, 32] assume that the probability of inclusion of any edge in G is constant and equal to $\psi \in (0, 1)$, which leads to the prior

$$p(G) \quad \propto \quad \psi^{size(G)} (1 - \psi)^{p(p-1)/2 - size(G)}. \tag{3.9}$$

Sparser graphs can be favored with prior (3.9) by choosing a small value for ψ. Alternatively, [2] suggested a hierarchical prior on the graph space that gives equal probability to the size of a graph and equal probability to graphs of each size, i.e.

$$p(G) = p(G \mid size(G) = k) p(size(G) = k) \tag{3.10}$$

where $p(size(G) = k) = 1/\{1 + p(p-1)/2\}$ and $p(G \mid size(G) = k) \propto 1$. We note that, for the class of general graphs, the expected size of a graph under size based prior is $m/2$, which is also the expected size of a graph under the uniform prior on $\mathcal{G}_p$. In what follows we will consider both the uniform prior and the prior given by equation (3.10).

3.1. Model properties and interpretation

The model discussed in the previous section is a natural extension of the well-known Dirichlet process mixture of multivariate normals originally presented in [41], but the introduction of the component-specific graphical structure allows us to induce sparsity in the estimation of the precision matrix associated with the mixture components. Hence, the point estimates provided by the Dirichlet process mixture of Gaussian graphical models can be interpreted as providing doubly-regularized estimates of the cluster-specific covariance matrices; one level of regularization arises because of the introduction of the prior distribution on the number of components, which introduces a penalty structure on the number of clusters equal to the logarithm of (3.6), while the second level of regularization arises because of the introduction of the prior $p(G)$ on the graph encoding the cluster-specific conditional independence structure. This is important because it is well known that, for high dimensional problems, estimation of the covariance matrices $\{K_l^{-1}\}_{l=1}^L$ can be extremely unstable and that regularized estimators produce improved results. Similar approaches to regularization have recently proved effective in both graphical models [56] and mixture models [21].

In the classical framework for Gaussian graphical models, the same set of graphs characterize the associations in all the samples. However, in a Dirichlet process mixture of Gaussian graphical models, for each sample there is a (potentially different) ordering of all the graphs with respect to their posterior probabilities. Hence, posterior inferences about the conditional independence among outcomes has to be made with regard to each specific observation rather than with respect to the whole population. This is clearer when we rewrite the model by introducing cluster indicators $\xi_1, \ldots, \xi_n$, so that the Dirichlet process mixture can be written in terms of a random partition [46], where

$$p(x^{(1:n)} \mid \xi_1, \ldots, \xi_n) = \prod_{l=1}^{L} \left\{ \int \left[\prod_{\{i:\xi_i=l\}} p(x_i \mid \mu_l^*, K_l^*, G_l^*) \right] M(\mathrm{d}\mu_l^*, \mathrm{d}K_l^*, \mathrm{d}G_l^*) \right\}$$

and $p(\xi_1, \ldots, \xi_n)$, which defines the prior probability on the partition, is given by (3.5). This representation highlights that the model groups observations into homogeneous classes, with samples on each class being generated from a standard Bayesian Gaussian graphical model, so that when $L = 1$, the mixture reduces to a standard Gaussian graphical model. Therefore, although the joint distribution is not Markov with respect to any single graph G, the distribution of the l-th class is Markov with respect to an (unknown) graph G_l^* (potentially distinct for each class), which is assigned the prior $p(G)$. The conditional independence graph associated with sample i is then obtained by setting $G_i = G_{\xi_i}^*$; hence, uncertainty about G_i arises not only from uncertainty about the structure of the cluster-specific G_l^*, but also from the uncertainty about the cluster indicator ξ_i.

Moreover, note that since observations are exchangeable under a Dirichlet Process model, the Pólya urn representation in (3.4) immediatly shows that the

prior marginal distribution $p(X_i) = \mathsf{E}_H\left\{\int \mathsf{N}(X_i \mid \mu, K, G)H(\mathrm{d}\mu, \mathrm{d}K, \mathrm{d}G)\right\}$, also follows a standard Gaussian graphical model with respect to a randomly chosen subject-specific graph G_i, which implies that our procedure generates a prior that is Markov with respect to G_i. In addition, since the posterior distribution of the mixing distribution for H is a mixture of Dirichlet process [1],

$$H \mid x^{(1:n)} \sim \int \mathsf{DP}\left(\alpha_0 + n, \frac{\alpha_0}{\alpha_0 + n}G_0 + \frac{1}{\alpha_0 + n}\sum_{i=1}^{n}\delta_{(\mu_i, K_i, G_i)}\right) p(\mu_i, K_i, G_i \mid x^{(1:n)})\mathrm{d}\mu_i\mathrm{d}K_i\mathrm{d}G_i$$

the same argument implies the posterior predictive distribution for a new observation $x^{(n+1)}$,

$$p(x^{(n+1)} \mid x^{(1:n)}) = \mathsf{E}_{H|x^{(1:n)}}\left\{\int \mathsf{N}(X_{n+1} \mid \mu, K, G)H(\mathrm{d}\mu, \mathrm{d}K, \mathrm{d}G)\right\},$$

is also Markov with respect to a newly sampled graph G_{n+1}. This new graph is equal to one of the previously observed graphs G_l^* with probability $r_l/(n + \alpha_0)$, or corresponds to a newly sampled graph from the baseline measure with probability $\alpha_0/(n + \alpha_0)$.

4. Posterior inference for mixtures of Gaussian graphical models

As with regular Gaussian graphical models, the posterior distribution arising from the nonparametric mixture of Gaussian graphical models is not analytically tractable because of the sheer size of the space of partitions and accompanying graphs. Therefore, we resort to Markov chain Monte Carlo algorithms to explore the features of this complicated posterior distribution. The literature on sampling algorithms for the Dirichlet process mixture model has grown extensively in the last 15 years; in this paper we focus attention on marginal samplers such as the ones described in [43], and the slice sampler introduced in [57].

4.1. Collapsed samplers for mixtures of decomposable Gaussian graphical models

The structure of the baseline measure M in (3.8) is such that we can easily integrate the means $\{\mu_l\}$ and precision matrices $\{K_l\}$ out of the model and create a sampler that acts on the space of partitions and graphs directly, which can dramatically reduce the computational burden. Given an initial state where the data $x^{(1:n)}$ has been divided into L clusters through indicator variables $\xi_1, \ldots, \xi_n$, and where graphs $G_1^*, \ldots, G_L^*$ are associated with each of the components, the algorithm proceeds to sample from the joint distribution of $(L, \{\xi_j\}_{j=1}^n, \{G_l^*\}_{l=1}^L, \alpha_0 \mid x^{(1:n)})$. As a first stage we update the sequence of indicators $\{\xi_j\}_{j=1}^n$ (and, implicitly, the number of components L) by sequentially

sampling each ξ_j for $j = 1, \ldots, n$ from its full conditional distribution

$$\mathsf{Pr}(\xi_j = l \mid \xi^{-j} = \{\xi_{j'}\}_{j' \neq j}, x^{(1:n)}, \{G_l^*\}_{l=1}^{L^{-j}}) \propto \begin{cases} r_l^{-j} p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l^*), & l \leq L^{-j} \\ \alpha_0 p(x^{(j)} \mid G_{L^{-j}+1}^*) \delta_{L^{-j}+1} & l = L^{-j} + 1 \end{cases} \tag{4.1}$$

In the previous expression, L^{-j} is the number of clusters in the sample (excluding observation $x^{(j)}$), $r_l^{-j} = \sum_{j' \neq j} \mathbf{1}_{\{\xi_j' = l\}}$ is the number of observations included in cluster l (excluding observation j if this sample currently belongs to cluster l), $p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l^*)$ is the posterior predictive distribution of sample $x^{(j)}$ given the samples that are currently in the l-th cluster (excluding $x^{(j)}$ if it happens to belong to this cluster) and the graph G_l^* associated with this cluster – see equation (2.9), and $p(x^{(j)} \mid G_{L^{-j}+1}^*)$ is the posterior predictive distribution of sample $x^{(j)}$ given an empty cluster, which is calculated by setting $n = 0$, $\bar{\mu} = 0_{(p \times 1)}$ and $U = 0_{(p \times p)}$ in equation (2.9). The graph $G_{L^{-j}+1}^*$ is to be randomly sampled from our baseline measure on $\mathcal{G}_V$, which we labeled $p(G)$ in (3.8). If the last observations has been moved out of a cluster, that cluster is deleted and L is decreased by 1. Similarly, if an observation is moved to a new cluster that is currently empty, L is increased by 1.

Once the cluster assignment has been updated, the graph G_l^* associated with each cluster $l = 1, \ldots, L$ is also updated as follows. We let the neighborhood of G_l^*, denoted by $\mathrm{nbd}_{\mathcal{G}_V}(G_l^*)$, be the set of graphs that can be obtained from G_l^* by adding or deleting one edge. These neighborhood sets connect any two graphs in $\mathcal{G}_V$ through a sequence of graphs that differ by exactly one edge – see, for example, Lauritzen [34]. We draw a candidate graph G_l^{*new} from the uniform distribution on $\mathrm{nbd}_{\mathcal{G}_V}(G_l^*)$. We change the graph associated with cluster l to G_l^{*new} with probability

$$\min\left\{1, \frac{p(\{x^{(j)} : \xi_j = l\} \mid G_l^{*new}) / |\mathrm{nbd}_{\mathcal{G}_V}(G_l^{*new})|}{p(\{x^{(j)} : \xi_j = l\} \mid G_l^*) / |\mathrm{nbd}_{\mathcal{G}_V}(G_l^*)|}\right\}, \tag{4.2}$$

otherwise the graph associated with cluster l remains unchanged. Here $p(\{x^{(j)} : \xi_j = l\} \mid G)$ represents the marginal likelihood of the samples currently in cluster l given a graph G – see equation (2.8). We denote by $|B|$ the number of elements of a set B. To improve mixing, we update the graphs associated with each cluster multiple times before another cluster assignment update is carried out (in our experience, between 5 and 10 updates seem to provide adequate mixing).

These two sequences of steps produce a sample from the posterior distribution of interest, $(L, \{\xi_j\}_{j=1}^n, \{G_l^*\}_{l=1}^L, \alpha_0 \mid x^{(1:n)})$ without any need to sample the means $\{\mu^*\}_{l=1}^L$ or precisions $\{K^*\}_{l=1}^L$. Therefore, if we are only interested in inferences about the clustering structure or the graphical structure associated with the clusters, or on predictive inference, the previous algorithm is sufficient

and can dramatically reduce the computational burden of the algorithm. However, if needed, the mean and variances of each mixture component can be easily sampled conditionally on $\{\xi_j\}_{j=1}^n$ using equation (2.6) and (2.7),

$$K_l^* \mid G_l^*, \{x^{(j)} : \xi_j = l\} \sim \mathsf{Wis}_{G_l^*}(\delta_0 + r_l, D_0 + U_l + A_l), \tag{4.3}$$

$$\mu_l^* \mid K_l^*, G_l^*, \{x^{(j)} : \xi_j = l\} \sim \mathsf{N}_p(\bar{\mu}_l, [(r_l + n_0)K_l^*]^{-1}), \tag{4.4}$$

independently of other components. As before, the subscript l denotes the corresponding values computed using only the observations assigned to component l (for example, r_l is the number of observations assigned to component l).

Since a graph proposed from the prior distribution $p(G)$ is highly unlikely to provide a good description of the conditional independence structure in a new cluster, the algorithm described above could potentially mix very slowly, especially in high dimensions. However, since conditional independence graphs are seriously under-determined when the sample size is small, and our algorithm evaluates any proposed graph on the basis of a single observation, in practice the probability of creating a new cluster is actually relatively insensitive to the sampled graph. Similarly, when expanding a cluster with a small number of observations assigned to it, the quality of the graph plays a very minor role in determining the probability of acceptance. Once a new cluster has been created, our use of multiple Metropolis-Hastings updates for the graph on each component tends to dramatically improve the quality of the graph, allowing the newly created component to persist. The empirical evidence from our simulations seems to support this intuitive argument.

In any case, we can improve the mixing of the algorithm by using multiple samples from the baseline measure to improve the probability that a "good" graph is generated; this approach is reminiscent of the multiple try methods described in [38]. The resulting sampling scheme involves just a slight modification of the algorithm described above, where the new component is represented by $T \geq 1$ graphs randomly sampled from the baseline measure, which we arbitrarily label $G^*_{L^{-j}+1}, \ldots G^*_{L^{-1}+T}$. For $l \geq L^{-j}+1$, the full conditional probability of creating a new component that has associated with it the graph G_l^* is then proportional to $\alpha_0 p(x^{(j)} \mid G_l^*)/T$. When $T = 1$ we recover our original algorithm. In the simulation study contained in the Section 6 and in the data analysis in Section 7 we worked with $T = 1$ and obtained Markov chains with excellent mixing times.

An additional avenue for improvement that we do not explored in this paper is to implement split/merge reversible jump Markov chain Monte Carlo algorithms in conjunction with the collapsed sampler we have focused on. In particular, since we cannot explicitly integrate G_i^*, we require an algorithm for non-conjugate models such as the one described in [31]. This algorithm would use restricted Gibbs sampling split-merge proposals that first sample from the baseline measure and then perform a series of updates on the graph to improve the quality of the proposal.

Additional flexibility can be obtained by sampling some of the hyperparameters associated with the DP prior. For example, the concentration parameter α_0

controls the expected number of components, and therefore has an important effect on the inferences generated by the model. Since eliciting values for α_0 can be difficult in practice, it is recommendable to try to infer it from the data. For example, we can assume a vague $\mathsf{Gam}(a_0, b_0)$ prior for the precision parameter α_0, in which case the full conditional distribution can be easily sampled using an auxiliary-variable Gibbs sampling step [18] (see Appendix). Finally, note that, although the model and computational algorithms has been described in terms of Dirichlet process mixtures, they can be easily extended to include any other species sampling priors on H (for examples, see 45 and 35).

4.2. Collapsed samplers for mixtures of arbitrary Gaussian graphical models

The sampling schemes discussed in Section 4.1 are particularly attractive for inference with decomposable graphs. As we have discussed in Section 2, the normalizing constant $I_G(\delta, D)$ of the G-Wishart $\mathsf{Wis}_G(\delta, D)$ distribution associated with a decomposable graph G can be calculated using formula (2.5), which implies that marginal likelihood (2.8) and posterior predictive distribution (2.9) are also calculated using formulas. Unfortunately, for a nondecomposable graph G, the normalizing constant $I_G(\delta, D)$ is no longer readily available and needs to be numerically approximated. This problem has been studied in Lenkoski & Dobra [36], who show that the Monte Carlo method of [3] is fast and accurate for calculating $I_G(\delta, D)$ for small values of δ and D set to the identity matrix, but it can be slow to converge otherwise. They also discuss the Laplace approximation for $I_G(\delta, D)$, but point out that it is accurate only for larger values of δ. As such, the sampling methods from Section 4 are difficult to implement for arbitrary graphs in $\mathcal{G}_V$ due to the numerical difficulties related to the calculation of the normalizing constants of G-Wishart distributions.

To this end, collapsed samplers that keep track of both the precision matrix K_l^* and graph G_l^* associated with each cluster $l = 1, \ldots, L$ can be developed instead. Given the observations $\{x^{(j)} : \xi_j = l\}$ that currently belong to cluster l, we update K_l^* based on the G-Wishart distribution (4.3) by employing the Metropolis-Hastings algorithm of Dobra et al. [17]. Given the updated K_l^*, we can update the graph G_l^* given K_l^* is performed using the reversible jump Markov chain method of Dobra et al. [17] instead of using equation (4.2). Once an edge is changed in G_l^*, the corresponding element of K_l^* must also be updated as it either becomes constrained to zero (if the edge is deleted) or becomes free (if the edge is added). Thus a candidate state that comprises the new graph G_l^{**} and a precision matrix K_l^{**} in the cone defined by G_l^{**} must be generated. The Metropolis-Hastings acceptance probability involves a change in the dimensionality of the parameter space, hence the reversible jump approach of [25] is required to decide whether the Markov chain transitions to the candidate state (G_l^{**}, K_l^{**}) or stays at the current state (G_l^*, K_l^*). The calculation of this acceptance probability involves the calculation of a ratio of two normalizing constants associated with G-Wishart prior distributions $\mathsf{Wis}_{G_l^*}(\delta_0, D_0)$ and $\mathsf{Wis}_{G_l^{**}}(\delta_0, D_0)$ which are efficiently approximated with the Monte Carlo method of [3].

When the creation of a new component is proposed in equation (4.1), we employ the direct sampling algorithm from the G-Wishart distribution of Wang & Carvalho [58] to sample a new precision matrix associated with a graph randomly sampled from the baseline measure on $\mathcal{G}_V$. Finally, the predictive distributions (2.9) in the sampling scheme from Section 4 are replaced by multivariate normal distributions $p(x^{(j)} \mid K_l^*)$.

4.3. Slice samplers

A number of alternatives to the collapsed Gibbs sampler have been presented in the literature. Particularly interesting are algorithms that explicitly sample the mixing distribution H; examples include the blocked Gibbs sampler [28], the retrospective sampler [47], and the slice sampler [57]. In this section we focus on the slice sampler, as it maintains the simplicity of the blocked Gibbs sampler but allows us to adaptively select the number of mixture components that are explicitly represented during execution.

To construct a slice sampler for the Dirichlet process mixture of Gaussian graphical models we introduce two sets of auxiliary variables, the indicator variables $\xi_1, \ldots, \xi_n$ already described in the previous sections, and uniformly distributed slice variables $u_1, \ldots, u_n$ so that

$$p(x^{(i)}, u_i, \xi_i \mid \{\mu_l^*\}_{l=1}^{\infty}, \{K_l^*\}_{l=1}^{\infty}, \{G_l^*\}_{l=1}^{\infty}) = p(x^{(i)} \mid \mu_{\xi_i}^*, K_{\xi_i}^*, G_{\xi_i}^*)\mathbf{1}(u_i \leq w_{\xi_i})$$

where $p(x^{(i)} \mid \mu_{\xi_i}^*, K_{\xi_i}^*, G_{\xi_i}^*)$ is given in equation (2.1) and $\mathbf{1}(A)$ is the indicator function on the set A. Note that integrating over u_i and ξ_i leads to our original mixture representation.

The algorithm is initialized by selecting the initial number of explicitly represented mixture component L^* (this number will adaptively change as the sampler progresses) and initializing the value of all model parameters. Given the rest of the parameters, the component-specific parameters for the explicitly represented components $\{\mu_l^*\}_{l=1}^{L^*}$, $\{K_l^*\}_{l=1}^{L^*}$, and $\{G_l^*\}_{l=1}^{L^*}$ can be updated using exactly the same procedures discussed in the previous subsections. On the other hand, conditionally on the auxiliary variables, we can easily update the weights by first sampling the stick-breaking ratios

$$v_l \mid \{\xi_i\}_{i=1}^n \sim \mathsf{Beta}\left(1 + r_l, \alpha_0 + \sum_{k>l} r_k\right)$$

and letting $w_l = v_l \prod_{k<l}(1 - v_k)$. In particular, note that if $N^* = \max\{\xi_i\}$, then $v_l \mid \{\xi_i\}_{i=1}^n \sim \mathsf{Beta}(1, \alpha_0)$ for $l > N^*$.

Finally, the slice variables can be updated by sampling them from $u_i \mid \{w_l\}_{l=1}^{\infty}, \xi_i \sim \mathsf{Uni}(0, w_{\xi_i})$, and the indicator variable ξ_i can be sampled by noting that

$$\begin{aligned}\mathsf{Pr}(\xi_i = l \mid u_i, \{w_l\}_{l=1}^{\infty}, \{\mu_l^*\}_{l=1}^{\infty}, \{K_l^*\}_{l=1}^{\infty}, \{G_l^*\}_{l=1}^{\infty}) \\ \propto p(x^{(i)} \mid \mu_l^*, K_l^*, G_l^*)\mathbf{1}(u_i \leq w_l)\end{aligned}$$

Although in principle the normalization constant for this posterior probability involves an infinite sum, note that in practice just a finite number of w_ls can satisfy $u_i \leq w_l$. Hence, in order to sample every ξ_i we only need to explicitly represent L^* components, where L^* satisfies

$$\sum_{l=1}^{L^*} w_l > \max_{1 \leq i \leq n} \{1 - u_i\} \tag{4.5}$$

If, at a given iteration, additional mixture components need to be explicitly represented to ensure that (4.5) is satisfied, the corresponding weights and component-specific parameters can be simply sampled from the baseline measure. Similarly, if too many components are currently being represented, the excess ones can be discarded.

4.4. Sampling from the baseline measure over graphs

All the algorithms described above require that we sample graphs from the baseline measure. In the case of general graphs this is straightforward under any of the prior in (3.9) or (3.10). However, if we restrict attention to non-decomposable graphs, some care needs to be exercised.

To illustrate this, consider sampling graphs uniformly on $\mathcal{G}_V^D$ (which corresponds to using prior (3.9) with $\psi = 1/2$). For graphs with a small number of vertices the following accept-reject algorithm works very well:

1. Sample an arbitrary graph from the uniform distribution (this is done by independently sampling the occurrence of each edge with probability 0.5).
2. Accept the graph if it is decomposable; otherwise repeat 1 and 2.

For example, for $p = 8$ vertices, the ratio between the number of decomposable graphs and the total number of graphs is 0.12. Hence the probability of acceptance of a graph sampled with the accept-reject algorithm is 0.12. More generally, even though the acceptance rate declines as $p \to \infty$, the ratio converges to a non-zero constant.

For graphs with a large number of vertices, the accept-reject algorithm could be less efficient. To this end, we devised the following Metropolis-Hastings algorithm that works directly on the set of decomposable graph with p vertices. Given a current decomposable graph G, identify the neighbors of G (denoted $\text{nbd}(G)$) which comprise all the decomposable graphs that are obtained by adding or deleting an edge in G. Uniformly sample a decomposable graph G' from $\text{nbd}(G)$. The chain moves to G' with probability:

$$\min\left\{1, \frac{|\text{nbd}(G)|}{|\text{nbd}(G')|}\right\}$$

The Metropolis-Hastings sampler gives approximate samples from the uniform distribution on decomposable graphs. We run such a chain as a separate program and saved a large number of graphs in a separate file. To reduce the

dependence between two consecutive draws we discard 1000 sampled graphs before saving the next graph. Our main code reads this output file to retrieve sampled decomposable graphs as needed. In a more efficient parallel implementation, the Metropolis-Hastings sampler can be run as a separate process that returns a graph as needed.

4.5. Discussion

It is important to consider the trade-offs associated with the choice of computational algorithms for posterior inference on Dirichlet process mixtures of graphical models. Generally speaking, algorithms that explicitly represent the mixing distribution H tend to generate samples with higher autocorrelations than collapsed samplers. In addition, they have higher memory requirements (because of the need to explicitly represent the mean vectors and variance matrices associated not only with the occupied component, but also with a potentially large number of unoccupied ones). However, in the case when inference is restricted to decomposable graphs, the slice sampler avoids the need to compute the normalizing constants associated with the graphs, which can potentially lead to speedups. Indeed, even though the normalizing constant for a decomposable graph breaks down to the evaluation of many gamma functions and determinants of positive definite matrices, our experience indicates these operations can represent up to 40% of the computational effort on some datasets. These observations suggest that for relatively small sample sizes, large number of mixture components, and sparse component-specific graphs, collapsed samplers should be preferred. In this type of situation, the computation of the normalizing constants involved in the predictive distribution is extremely efficient, and collapsed algorithms are faster (in CPU time), and produce samples with smaller autocorrelations. On the other hand, for problems where the data supports a relatively small number of mixture components with dense graphs, slice samplers would seem to be more efficient. These conclusions seem to be supported by the empirical comparison carried out in the context of the simulation study included in Section 6.

Another important issue for posterior inference in mixture models is label-switching, i.e., the invariance of the posterior distribution to different combinations of values used to label the mixture components [53]. A simple solution to this problem is to present posterior summaries that are invariant to label switching. For example, the posterior distribution over data partitions can be summarized through the pairwise incidence matrix Υ, where $\Upsilon_{ij} = \mathsf{Pr}(\xi_i = \xi_j \mid x^{(1:n)})$. This matrix can then be used to generate point estimates of cluster membership (for example, see 33). Similarly, we avoid presenting inferences about the component-specific graph G_k^* and focus instead on the *observation specific* graph $G_i = G_{\xi_i}^*$. This can potentially be done for any observation $i = 1, \ldots, n$, but in practice you would typically do it for just a few "representative" observations, which can be selected using the pairwise clustering probability matrix Υ described above. This approach is illustrated in Section 7.

5. Infinite hidden Markov Gaussian graphical models

Multivariate time series models that use graphical models to improve estimation of the crosssectional covariance structure have been recently developed. Two recent pertinent examples are [10] and [60]. These approaches rely on extensions of the dynamic linear model [62] and assume that the graph underlying the model is constant in time which, as the example in Section 7 illustrates, might not be an appropriate assumption in practical applications. Also, [59] consider time series models where the underlying conditional independence graph evolves smoothly time. As an alternative, we develop a nonparametric version of the popular hidden Markov model where the emission distribution corresponds to a Gaussian graphical model. This class of models are an extension of the Dirichlet process mixtures of Gaussian graphical models from Section 3 that explicitly accounts for the temporal dependence among the observations.

Hidden Markov models [8] are hierarchical mixture models where

$$x^{(j)} \mid \{\theta_l^*\}_{l=1}^L, \{\xi_j\}_{j=n}^L \sim p(x^{(j)} \mid \theta_{\xi_j}^*),$$
$$\xi_j \mid \xi_{j-1}, \{\pi^l\}_{l-1}^L \sim \mathsf{Multno}(\pi^{\xi_{j-1}}), \quad \theta_l^* \sim M.$$

and $\xi_0 \sim \mathsf{Multno}(\pi^0)$. In this context the latent indicator $\xi_j \in \{1, \ldots, L\}$ is called a hidden state, while the entire set of indicators $\{\xi_j\}_{j=1}^n$ is called a trajectory. The ordering of the states is implicitly defined by the ordering of their indices; trajectories evolve according to a Markov process with transition probabilities $\mathsf{Pr}(\xi_j = l \mid \xi_{j-1} = l') = \pi_l^{l'}$. The initial state probabilities are $\mathsf{Pr}(\xi_0 = l) = \pi_l^0$. Conditionally on a set of states $\{\xi_j\}_{j=1}^n$, the observations $x^{(1)}, \ldots, x^{(n)}$, are independently distributed from state dependent distributions $p(\cdot \mid \theta_{\xi_1}^*), \ldots, p(\cdot \mid \theta_{\xi_n}^*)$.

Infinite hidden Markov models [4, 54, 23] generalize hidden Markov models to allow for an infinite number of states, in a similar way as how Dirichlet process mixture models generalize finite mixture. Hence, infinite hidden Markov models allow us to treat the number of states L as a random variable that is to be estimated from the data. More specifcally, we consider a model where

$$x^{(j)} \mid \{\mu_l^*\}_{l=1}^\infty, \{K_l^*\}_{l=1}^\infty, \{G_l^*\}_{l=1}^\infty, \{\xi_j\}_{j=1}^n \sim \mathsf{N}_p(x^{(j)} \mid \mu_{\xi_j}^*, (K_{\xi_j}^*)^{-1}),$$
$$\xi_j \mid \xi_{j-1}, \{\pi^l\}_{l=1}^\infty \sim \mathsf{Multno}(\pi^{\xi_{j-1}}),$$

$\pi^l \mid \alpha, \gamma \sim \mathsf{DP}(\alpha, \gamma)$, $\gamma \mid \alpha_0 \sim \mathsf{SB}(\alpha_0)$, and $\theta_l^* = (\mu_l^*, K_l^*, G_l^*) \sim M$, where M is defined as in (3.8). As before, π^l corresponds to the vector of transition probabilities leaving state l and γ is the vector of average transition probabilities. A model of this form has some distinct advantages over the dynamic linear models with graphical structure discussed in [10] and [60]. In particular, it allows for the graph controlling the conditional independence structure of the data to evolve in time while still taking into account the sequential nature of the problem.

A marginal Gibbs sampler similar to the one described in Section 3 can be devised for the infinite hidden Markov model. To do so, note that if $r_l =$

$(r_{l1}, r_{l2}, \ldots)$ with $r_{ll'}$ denoting the number of transitions between state l and state l', the posterior distribution for the vector $\pi^l = (\pi^l_1, \pi^l_2, \ldots)$ is given by

$$\pi^l \mid r_l \sim \mathsf{DP}\left(\alpha + r_{l\cdot}, \frac{r_l + \alpha\gamma}{r_{l\cdot} + \alpha}\right).$$

Hence, we can explicitly integrate the unknown transition probabilities $\{\pi^l\}_{l=1}^{\infty}$ yielding

$$\mathsf{Pr}(\xi_t = l' \mid \xi_{t-1} = l, \alpha, \gamma, r_l) = \mathsf{E}\left\{\pi^l_{l'} \mid r_l\right\} = \frac{\alpha\gamma_l + r_{ll'}}{\alpha + r_{l\cdot}},$$

an expression that is reminiscent of the Pólya urn in (3.4). Due to the Markovian structure of the model, this implies that the full conditional distribution for ξ_t is given by

$$p(\xi_t \mid \xi^{-t}, \alpha, \gamma, \{r_l\}_{l=1}^{\infty}) \propto p(\xi_{t+1} \mid \xi_t, \alpha, \gamma, \{r_l\}_{l=1}^{\infty}) p(\xi_t \mid \xi_{t-1}, \alpha, \gamma, \{r_l\}_{l=1}^{\infty}).$$

In the case of decomposable GGMs, combining the full conditional prior with the likelihood, and integrating over the mean and variance of each state, the full conditional distribution for ξ_t reduces to

$$\mathsf{Pr}(\xi_t = l \mid \xi^{-t}, \alpha, \gamma, x^{(1:n)}) \propto$$
$$\begin{cases} \dfrac{r^{1:(j-1)}_{\xi_{j-1},l} + r^{(j+1):n}_{\xi_{j-1},l} + \alpha\gamma_l}{r^{1:(j-1)}_{\xi_{j-1}\cdot} + r^{(j+1):(n)}_{\xi_{j-1}\cdot} + \alpha} \dfrac{r^{1:(j-1)}_{l,\xi_{j+1}} + r^{(j+1):n}_{l,\xi_{j+1}} + \alpha\gamma_{\xi_{j+1}}}{r^{1:(j-1)}_{l\cdot} + r^{(j+1):n}_{l\cdot} + \alpha} \times & \\ \qquad p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G^*_l) & l \leq L^{-j}, l \neq \xi_{t-1}, \\ \dfrac{r^{1:(j-1)}_{\xi_{j-1},l} + r^{(j+1):n}_{\xi_{j-1},l} + \alpha\gamma_l}{r^{1:(j-1)}_{\xi_{j-1}\cdot} + r^{(j+1):(n)}_{\xi_{j-1}\cdot} + \alpha} \dfrac{r^{1:(j-1)}_{l,\xi_{j+1}} + r^{(j+1):n}_{l,\xi_{j+1}} + \alpha\gamma_{\xi_{j+1}} + 1}{r^{1:(j-1)}_{l\cdot} + r^{(j+1):n}_{l\cdot} + \alpha + 1} \times & \\ \qquad p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G^*_l) & l = \xi_{j-1} = \xi_{j+1}, \\ \dfrac{r^{1:(j-1)}_{\xi_{j-1},l} + r^{(j+1):n}_{\xi_{j-1},l} + \alpha\gamma_l}{r^{1:(j-1)}_{\xi_{j-1}\cdot} + r^{(j+1):(n)}_{\xi_{j-1}\cdot} + \alpha} \dfrac{r^{1:(j-1)}_{l,\xi_{j+1}} + r^{(j+1):n}_{l,\xi_{j+1}} + \alpha\gamma_{\xi_{j+1}}}{r^{1:(j-1)}_{l\cdot} + r^{(j+1):n}_{l\cdot} + \alpha + 1} \times & \\ \qquad p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G^*_l) & l = \xi_{j-1} \neq \xi_{j+1}, \\ \dfrac{\alpha\gamma_l}{r^{1:(j-1)}_{\xi_{j-1}\cdot} + r^{(j+1):(n)}_{\xi_{j-1}\cdot} + \alpha} \gamma_{\xi_{j+1}} p(x^{(j)} \mid G^*_{L+1}) & l = L^{-j} + 1. \end{cases} \tag{5.1}$$

where $r^{j_1:j_2}_{ll'}$ denotes the number of transitions from state l to state l' in the sub-trajectory $\{\xi_j\}_{j=j_1}^{j_2}$ and by $r^{j_1:j_2}_{l\cdot}$ the number of transitions out of state l in the same sub-trajectory, and G^*_{L+1} is a graph randomly sampled from the baseline distribution. If a new empty cluster needs to be created, we update the number of clusters L by setting $L^{new} = L+1$ and the vector γ by setting $\gamma^{new}_{L+1} = u\gamma_{L+1}$, $\gamma^{new}_{L+2} = (1-u)\gamma_{L+1}$ were $u \sim \mathsf{Beta}(1, \alpha_0)$. To justify the update to γ, remember that $\gamma \sim \mathsf{SB}(\alpha_0)$ and note that γ_{L+1} reflects the combined prior probability that an observation is assigned to one of the (countably many) empty states. When one of the empty components becomes active, we must split the probability of the newly occupied state from the current estimate of the combined probability

according to a randomly selected random variate $u \sim \mathsf{Beta}(1, \alpha_0)$ (see the stick-breaking construction in equation (3.3)). As we discussed in Section 4.2, in the case of arbitrary graphs the sampler can be modified by explicitly representing the precision matrices $\{K_l\}$ associated with the different states, replacing $p(x^{(j)} \mid G^*_{L+1})$ and $p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G^*_l)$ by $p(x^{(j)} \mid K^*_{L+1}, G^*_{L+1})$ and $p(x^{(j)} \mid \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, K^*_l, G^*_l)$ in (5.1), and jointly sampling (K^*_l, G^*_l) for each l using the algorithm described in [17].

In any case, given a trajectory $\{\xi_j\}_{j=1}^n$, α_0 and α, we sample γ by introducing the independent auxiliary variables $\{m_{ll'}\}$ for $l, l' \in \{1, \ldots, L\}$ such that

$$\mathsf{Pr}(m_{ll'} = m) \propto S(r^{(1:n)}_{ll'}, m)(\alpha\gamma_{l'})^m, \qquad m \in \{1, \ldots, r^{(1:n)}_{ll'}\},$$

where $S(\cdot, \cdot)$ denotes the Stirling number of the first kind. Conditional on these auxiliary variables we can update γ by sampling

$$(\gamma_1, \ldots, \gamma_{L+1}) \sim \mathsf{Dir}(m_{\cdot 1}, \ldots, m_{\cdot L}, \alpha_0),$$

where $m_{\cdot l'} = \sum_{l=1}^{L} m_{ll'}$. We use vague gamma priors for the precision parameters α_0 and α and update them as described in the Appendix. An alternative slice sampler for this model can also be developed along the lines described in [23].

6. Simulation studies

We consider first a small simulation that involves data arising from two Gaussian clusters. For brevity, the results we present in this Section correspond to a single run of the simulation, but these are representative of those obtained over multiple runs. In the first cluster n samples are from a star graphical model $N_{11}(\mu, K_1^{-1})$ with every variable X_j, $j > 2$, connected to X_1, while the second cluster contains n samples from an AR(2) model $N_{11}(-\mu, K_2^{-1})$. The non-zero elements of the two precision matrices are

$$\begin{aligned}
(K_1)_{j,j} &= (K_2)_{j,j} = 1, && j = 1, \ldots, 11,\\
(K_1)_{1,j} &= (K_1)_{j,1} = 0.3, && j = 2, \ldots, 11,\\
(K_2)_{j-1,j} &= (K_2)_{j,j-1} = 0.3, && j = 2, \ldots, 11,\\
(K_2)_{j-2,j} &= (K_2)_{j,j-2} = 0.3, && j = 3, \ldots, 11.
\end{aligned}$$

We consider four settings of $n = (50, 100, 200, 500)$ and four settings for $\mu = (0, .1, .2, .5)$ and thus run 16 simulations in total. Results are shown using the uniform graph prior. For each data set, we ran the Dirichlet Process Mixture procedure for 5000 iterations after a burn-in of 2000 iterations. Figure 1 shows the resulting clustering under each combination of n and μ.

Figure 1 shows that both the number of samples from each cluster as well as the difference in mean values affects the ability of our method to separate the two groups. When $n = 50$, and $\mu = 0$ (the upper left plot) the figure shows

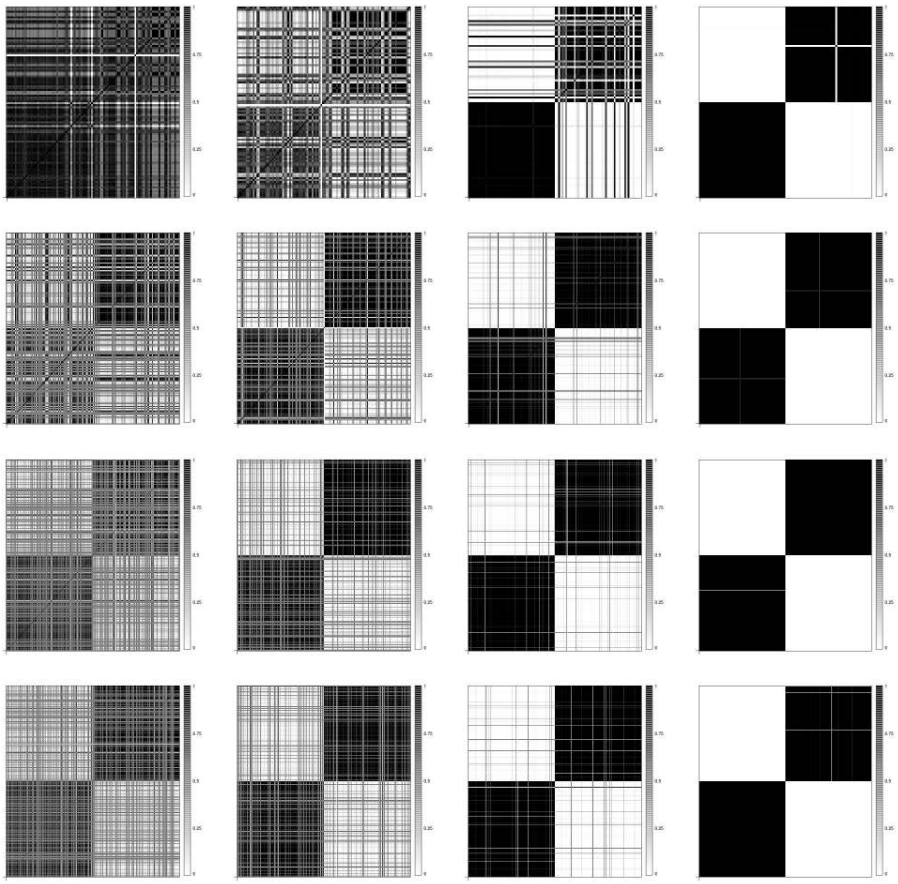

FIG 1. *Clustering Results from Simulation Study. Each row of results corresponds to setting* $n = 50, 100, 200, 500$ *and each column corresponds to setting* $\mu = 0, .1, .2, .5$

that there is only a slight differentiation between the two groups. However, even when $n = 50$ but $\mu = 0.5$, the two groups are almost perfectly clustered. By contrast, when $n = 500$ the method performs considerably better (though far from perfectly) at discerning the two groups when $\mu = 0$ and clustering is essentially correct when $\mu = 0.5$.

Figure 2 shows the estimated edge probabilities for each combination of n and μ. In each plot, the lower triangle corresponds to the first n observations, while the upper triangle corresponds to the second n observations. Figure 2 shows several interesting features. First, even when μ is near zero and n is small, a situation in which the model does not cluster the observations perfectly, the structure of K_1 is quickly discerned by the model. As clustering improves, the structure of K_1 continues to be recovered well. As both n and μ grow, the method is also able to recover the structure of K_2.

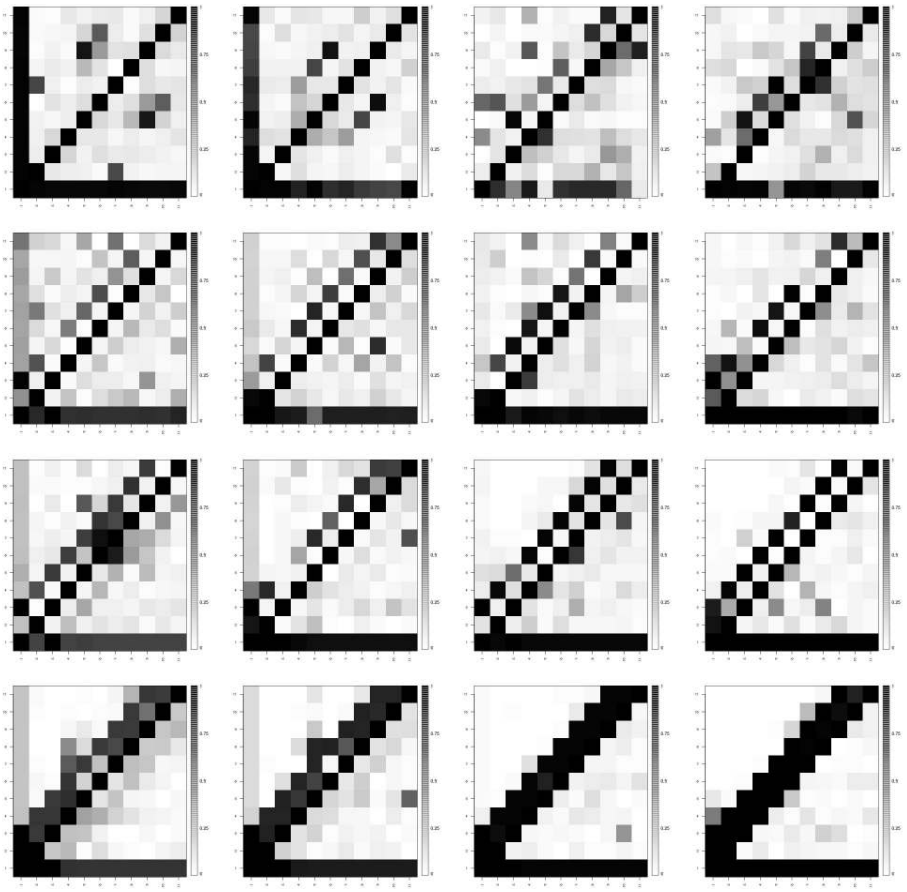

FIG 2. *Estimated edge probabilities from the simulation study. Each row of results corresponds to setting* $n = 50, 100, 200, 500$ *and each column corresponds to setting* $\mu = 0, .1, .2, .5$. *The lower triangle of each plot is the edge probability associated with the first* n *observations, while the upper triangle is the edge probability associated with the second* n *observations*

6.1. Timing study

We now compare the reduced and slice samplers discussed in Section 3 of the main report. Using the simulation study described above, we set $\mu = 0.1$ and considered datasets with n set to values between 50 and 500 observations at increasing increments of 50 observations. For each setting of n we sampled 20 datasets and ran the algorithm, as above, for 5000 repetitions after a burn-in of 2000 iterations under both the reduced sampler and the slice sampler.

Table 1 shows the results of this comparison. This table shows the average time that each algorithm took to complete the run for each value of n. In addition to the timing comparison, we also compute the effective sample size (ESS)

TABLE 1
Average time to completion (seconds) and average Effective Sample Size (ESS) with standard deviations across 20 replications for the reduced and slice sampler versions of the DPM model.

	Time		ESS	
n	Reduced	Slice	Reduced	Slice
50	450.31	395.61	3851.94	3725.7
	(50.47)	(68.96)	(1451.13)	(993.71)
100	866.29	415.13	4249.98	3923.88
	(13.06)	(59.75)	(888.71)	(1222.84)
150	1296.35	422.5	4495.99	3757.88
	(26.36)	(77.73)	(497.4)	(1106.85)
200	1722.39	442.9	4510.52	4067.71
	(19.09)	(61.85)	(447.16)	(1198.38)
250	2144.79	454.1	4922.04	3998.93
	(12.75)	(59.64)	(438.91)	(735.74)
300	2583.13	484.36	4658.79	4004.52
	(61.87)	(58.11)	(903.07)	(1538.46)
350	2993.49	507.94	4648.73	4237.84
	(23.29)	(46.72)	(478.66)	(1094.65)
400	3417.02	542.32	4834.01	4025.85
	(16.18)	(42.75)	(796.26)	(918.99)
450	3848.35	574	4630.43	4327.83
	(43.8)	(49.87)	(944.87)	(1091.51)
50	4300.59	614.27	4692.48	3731.38
	(130.47)	(40.9)	(351)	(1405.63)

of the parameter α that is returned from the 5000 final repetitions (thus ignoring burn-in) under each algorithm. When $n = 50$ we see that the two methods have essentially the same timing. As n grows, the time taken by the reduced sampler appears to increase nearly linearly, while the slice sampler shows considerably less time increase. However, the parameter α exhibits somewhat less autocorrelation, using the reduced sampler, as show in the ESS calculation.

7. Illustration: Modeling trends in exchange rate fluctuations

We consider a dataset that follows the returns on exchange rates of 11 currencies relative to the United States dollar between November 1993 and August 1996. This dataset consists of 1000 daily observations and includes three Asian currencies – the New Zealand Dollar (NZD), the Australian Dollar (AUD), and the Japanese Yen (JPY) – five European currencies that eventually became part of the Euro – the Deutsch Mark (DEM), French Franc (FRF), Belgian Franc (BEF), Netherlands Gilder (NLG) and Spanish Peso (ESP) – as well as three additional European currencies – the British Pound(GBP), Swedish Krona (SEK), and Swiss Franc (CHF). These data have previously been used in a variety of contexts related to graphical models (see e.g. 9 and 10). In [9] the authors present the graph shown in Figure 3, which is determined using stochastic search methods first discussed in [32] over the final 100 timepoints. The authors note that this graph is sensible from the standpoint of known trading relations: the mainland European countries that join the Euro are closely linked in a sin-

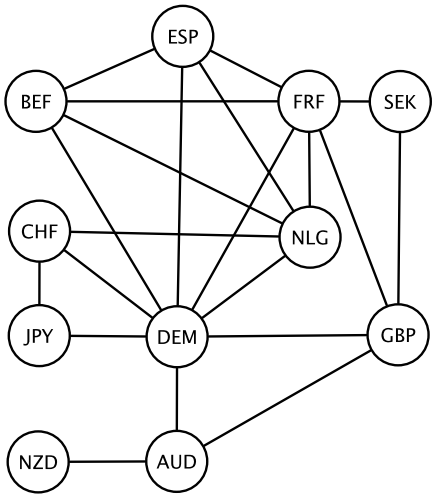

FIG 3. *Graphical model presented by [9], which represents the highest probability graph found using stochastic search methods over the last 100 timepoints of the exchange dataset, using the author's prior specifications. This graph has been used to show that investment strategies based on graphical models often have lower variability and higher yield than methods based on the full covariance matrix.*

gle clique, while the British Pound, Swedish Krona and Swiss Franc connect with only some of these counties, most notably the currencies of the largest Euro-area countries, the Deutsch Mark and French Franc (the Swiss Franc is also connected to the Netherlands Gilder, being more integrated with mainland European economies). [10] then show that portfolio weights based on estimates from this graphical model give an investment strategy with increased return and reduced variability when compared to using an approach that does not impose graphical structure on the estimates of Σ, evidence of the effectiveness of the graphical models approach.

We used this dataset to explore the possibility of alternating regimes with separate patterns of interaction during these 1000 days. Given that the data form a natural timecourse, we employed the infinite hidden Markov model with Gaussian graphical model emissions distributions discussed in Section 5. To make comparisons against [10] fair, we concentrate of iHMMs of decomposable GGMs. We ran the sampler for 100000 iterations after a burn-in period of 20000 iterations, and ran five separate instances of the algorithm from separate starting points. On a quad-core 2.8 gHz computer with 4 GB of RAM running Linux, each instance of the algorithm took approximately 8 hours to run for the full dataset. We ran this example using both the uniform prior on the graph space as well as the size-based prior given in equation (3.10). In what follows we refer to these as iHMM-U and iHMM-S, respectively.

After completion we assessed the results from each chain and verified they returned the same estimates. For example, Figure 4 shows the convergence in

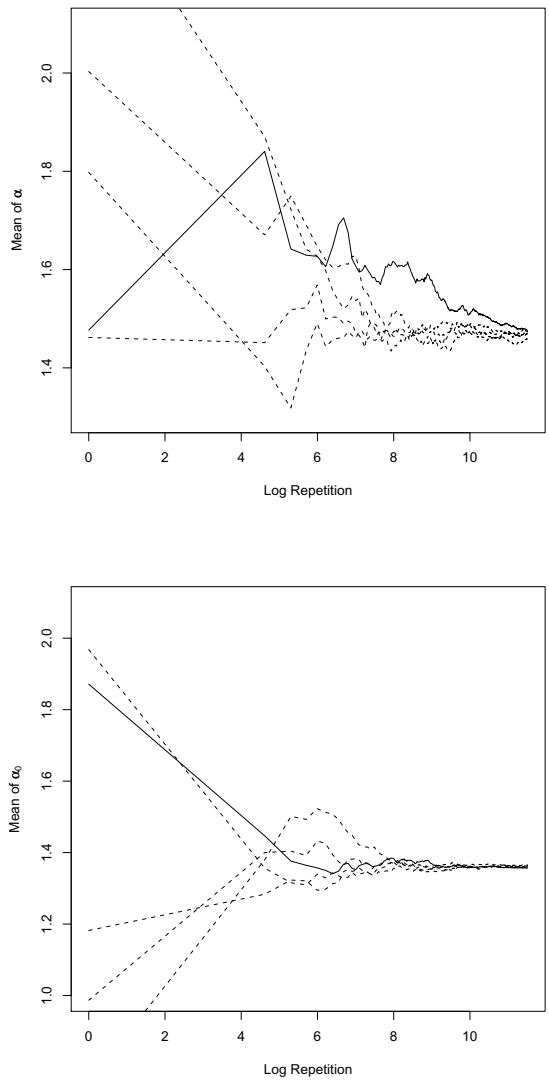

FIG 4. *Convergence plot for α and α_0 across chains by log iteration. This plot shows the running average of these two parameters across five separate instances of the algorithm. Their mutual agreement implies the settings used are sufficient to assure convergence.*

α and α_0 across chains for iHMM-U. The convergence plots for iHMM-S are nearly identical.

When run using this model, the observations for the most part clearly fall into one of two regimes. Figure 5 shows the posterior probabilities that two observations belong to the same state. The upper triangle corresponds to clustering probabilities from iHMM-U and the lower triangle to iHMM-S. The first

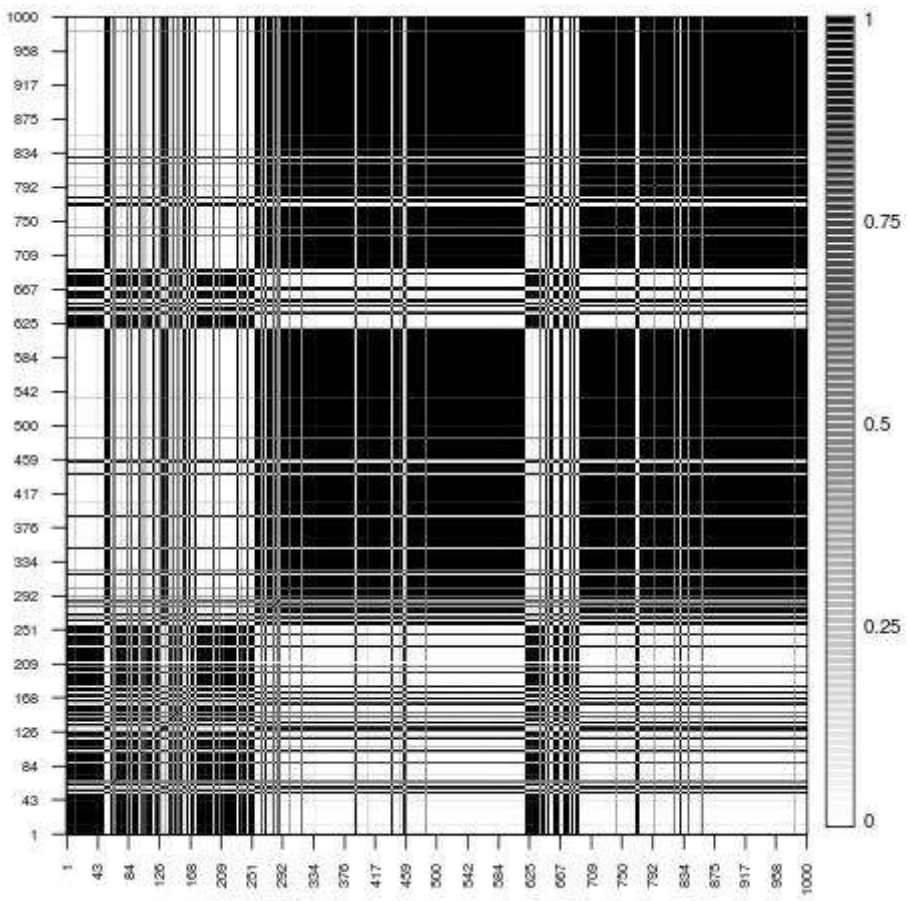

FIG 5. *Heatmap displaying the probability that two observations belong to the same cluster for the exchange example. The Figure shows two dominant regimes, one that runs for the most part from timepoints 1 to 251 and again roughly between timepoints 623 and 700 and the other which is present most of the remaining time periods.*

state is comprised roughly of the time points one through 251 (11/14/1993 to 7/21/1994) at which point a second regime takes over. Interestingly, the first interaction regime arises again for a brief period roughly comprising timepoints 623 to 700 (7/29/95 to 10/14/95). As shown in the figure the graph prior has little effect on clustering probabilities. The maximum difference between the estimated probability of two points belonging to the same group was 3% under the alternative graph priors.

Figure 6 shows a heatmap of the edge probabilities for timepoint 40 (belonging to the first regime) and timepoint 540 (belonging to the second regime) as well as point estimate graphical models which are assembled from those edges that had a greater than 50% probability of inclusion for the respective observa-

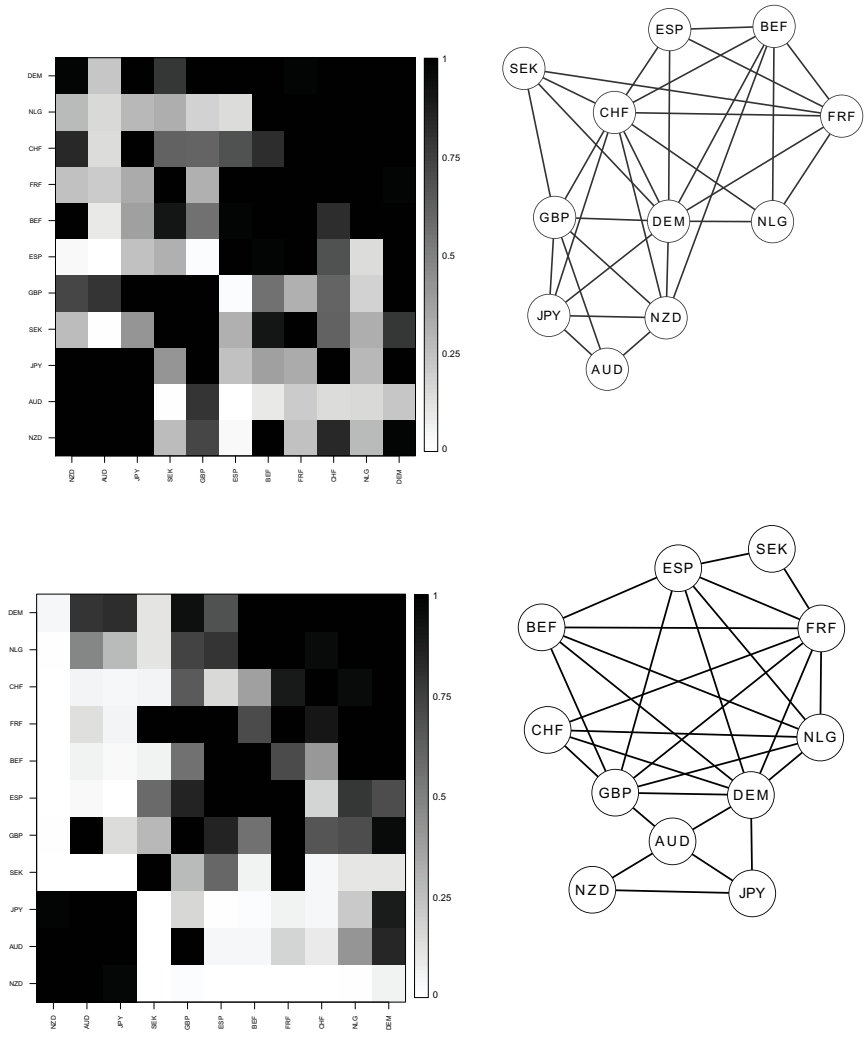

FIG 6. *Edge probability heatmaps and graphical models associated with timepoint 40 (upper row) and timepoint 540 (lower row) in the exchange rate example. In each heatmap, the upper triangle corresponds to iHMM-U and the lower triangle to iHMM-S, though the differences are negligible. The point estimate graphs in this figure were constructed by adding any edge that had greater than 50% posterior inclusion probability for the respective timepoint.*

tion. As with Figure 5, upper triangles correspond to estimates from iHMM-U, and lower triangles to iHMM-S. Again, there is negligible difference between estimates under the two graphs prior (the largest difference was an estimate of .88 and .81 under iHMM-U and iHMM-S, respectively, for the edge between JPY and DEM in timepoint 540). This suggests that the dataset is large enough to negate the effect of prior graph assumptions.

The edge probabilities associated with these two timepoints show a large degree of similarity. In particular, the Asian currencies show edge probabilities of 1 in both regimes as do edges between many of the European currencies. However, there are some differences, in particular the second period places much higher inclusion probabilities on edges between the British Pound and Euro area currencies. These differences are easiest to see in the point estimate graphical models.

In the first regime, an association structure broadly consistent with the graph used in [10] is present. We see a tight grouping of the Euro adopters, but with increased connection between the Swiss Franc and the Euro countries. A clique amongst the Asian currencies is connected to only three of the Euro adopters. Furthermore the British Pound is only connected to the Euro area through the Deutsch Mark, the currency of the economic leader of this group and the Swiss Franc. The interpretation of this graph is similar to that reported earlier: the fluctuations in the exchange rate of mainland European currencies to the Dollar roughly track one another. However, at this point the British Pound was no longer part of the European Exchange Rate Mechanism, following the Pound's crash on "Black Wednesday", September 16th of 1992. This reason, along with the greater integration of trade between Britain and the United States (as suggested by 9), leads to a separation of the Pound from the mainland European currencies.

The second regime has a similar structure but a markedly different interpretation of the interactions between the Pound and the smaller Euro adopters. At this point, the graph is still comprised of a large group consisting of the Euro countries, however the Pound has joined–and become a central part of–this grouping. It connects with each member of the Euro countries (as well as the Swiss Franc). Furthermore, the Asian countries lose several neighbors.

The greater connectedness of the Pound to the Euro area may have been a result of the uncertainty regarding the specifics of the Euro's implementation in the mid-nineties. In particular, the crash of the Pound in 1992 and Britain's subsequent withdrawal from the European Exchange Rate Mechanism left a looming uncertainty regarding if, and when, Britain would again agree to join the common currency. The initial switch from a "UK-excluding" to a "UK-inclusive" regime in the exchange rate data occurs on July 22, 1994. What is curious about this date is that Tony Blair was elected to lead the Labour party on July 21, 1994. Blair would eventually run a campaign based, in part, on rejoining the Exchange Rate Mechanism and adopting the Euro, a stance he held until the events of 2001. The graphical models displayed in Figure 6 suggest that currency markets began integrating the possibility that Britain would adopt the Euro by exhibiting greater covariation with mainland European countries. This new regime was, itself, somewhat unstable, as evidenced by the return of a "UK-exclusive" regime during the summer of 1995.

In [10], the exchange rate data is used to show that minimum variance portfolios will yield better return when Gaussian graphical models are employed to estimate covariances. We considered a similar analysis and used the out-of-sample expectation over the sampled values of μ_{T+1} and K_{T+1} for each T

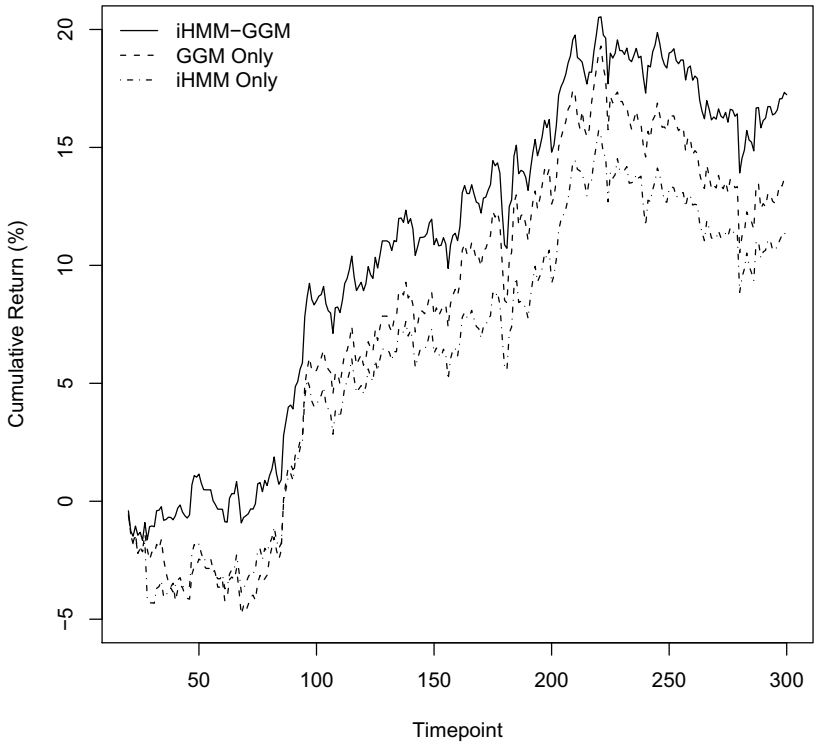

FIG 7. *Cumulative out-of-sample returns from forming the optimal portfolio weights based on running the full infinite hidden Markov model with Gaussian graphical model emission distributions (iHMM-GGM) as well as the infinite hidden Markov model with full covariance emission distribution only (iHMM only) and a Gaussian graphical model only (GGM only), run over time periods* 20 *to* 300. *The final cumulative return was 17.2% for the iHMM-GGM, 13.5% for the GGM only and 11.2% for the iHMM only models.*

between 19 and 299 as the first two moments of the predictive distribution and calculated portfolio weights w_{T+1} assuming a target return of $m = 0.1\%$ per day (see [10] for details regarding the construction of these portfolios). The portfolio allocation exercise reported was run out-of-sample. More specifically, for every $T = 19, \ldots, 299$, a separate Markov chain Monte Carlo run was conducted; the corresponding posterior predictive distribution was used to design the portfolio at time $T + 1$, and the returns computed from the observed returns at that time point (which were not used to fit the model). Figure 7 shows that the predictive distributions from the infinite hidden Markov model with Gaussian graphical model emission distributions gives portfolio weights that have higher yields than both the infinite hidden Markov model with the full covariance matrix and the Gaussian graphical model only approach. This shows the ulitility of

our framework: by considering a mixture model we are able to adapt to changing conditions, thereby leading to better specified predictive distributions. Furthermore, by incorporating Gaussian graphical models into the model formulation, we are able to induce spars estimates of covariation, which likewise improve predictive performance. Results are shown using the iHMM-U model, but they are essentially unchanged under the iHHM-S model.

8. Discussion

Although this paper has focused on two relatively simple models (nonparametric mixtures of Gaussian graphical models and infinite hidden Markov models with Gaussian graphical model emission distributions), the basic structure can be employed to generalize many other nonparametric models. For example, we plan to extend the nonparametric mixture classifier developed in [49] to include Gaussian graphical model kernels as a way to improve classification rates in high-dimensional problems. Also, in the spirit of [41, 42] and [48], sparse nonlinear regression models can be generated by using Gaussian graphical model mixtures as the joint model for outcomes and predictors, from which the regression function can be derived by computing the conditional expectation of the outcome given the predictors. This generalizes the work of [15] on sparse regression to allow for adaptive local linear fits.

The implementation of Gaussian graphical model mixtures we have discussed in this paper exploits the Pólya urn representation available for many nonparametric models to construct a Gibbs sampler that updates the grouping structure one observation at a time. In the case of decomposable models, this allowed us to avoid the explicit representation (and the sampling) of means and covariance parameters, which can be computationally intensive. However, Pólya urn samplers can suffer from slow mixing and we plan to explore in the near future alternative computational algorithms, in particular those employing split-merge moves such as those developed in [30] and [31].

The data analyses in this paper suggest that implementing mixtures of Gaussian graphical models using Markov chain Monte Carlo algorithms is feasible for a moderate numbers of variables. However, many interesting applications of Gaussian graphical model mixtures (e.g., gene-expression data) involve outcomes in much higher dimensions, where previous experience suggest that random-walk Markov chain Monte Carlo algorithms will be inefficient. We are currently exploring other search algorithms based on heuristics, such as the feature-inclusion [6, 51], that might allow us to identify high-probability partitions and their associated graphs.

Appendix

We give a brief description of an auxiliary variable scheme for sampling from the posterior distributions of the single concentration parameter α_0 in Section 4 and the two concentration parameters α and α_0 from Section 5 – see [18] and

[54] for full details. We assume that the priors for α and α_0 are $\mathsf{Gam}(a, b)$ and $\mathsf{Gam}(a_0, b_0)$, respectively.

In the case of the GGM-DPM, we sample α_0 by introducing an auxiliary variable η. Conditional on α_0, we have $\eta \mid \alpha_0 \sim \mathsf{Beta}(\alpha_0 + 1, n)$. Conditional on η, α_0 follows a mixture distribution

$$\alpha_0 \mid \eta \sim d_\eta \mathsf{Gam}(a_0 + L, b_0 - \log \eta) + (1 - d_\eta)\mathsf{Gam}(a_0 + L - 1, b_0 - \log \eta),$$

where $d_\eta/(1 - d_\eta) = (a_0 + L - 1)/[n(b_0 - \log(\eta))]$.

In the case of the GGM-iHMM we additionally introduce auxiliary variables $\varsigma_1, \ldots, \varsigma_L$ and $u_1, \ldots, u_L$. Conditionally of α, $\varsigma_l \mid \alpha \sim \mathsf{Beta}(\alpha + 1, r_{l\cdot})$ and $u_l \mid \alpha \sim \mathsf{Bernoulli}(r_{l\cdot}/(\alpha + r_{l\cdot}))$, where $r_{l\cdot} = \sum_{l'=1}^{L} r_{ll'}$. Then, α is sampled from its full conditional distribution,

$$\alpha \mid \{u_l\}, \{\varsigma_l\} \sim \mathsf{Gam}\left(a + m_{\cdot\cdot} - \sum_{l=1}^{L} u_l, b - \sum_{l=1}^{L} \log \varsigma_l\right)$$

where $m_{\cdot\cdot} = \sum_{l=1}^{L} \sum_{l'=1}^{L} m_{ll'}$. To sample α_0, we follow a procedure that is very similar to the one we used for the GGM-DPM. Again, we introduce an auxiliary variable η. Conditional on α_0, we have $\eta \mid \alpha_0 \sim \mathsf{Beta}(\alpha_0 + 1, m_{\cdot\cdot})$. Conditional on η, α_0 follows a mixture distribution

$$\alpha_0 \mid \eta \sim d_\eta \mathsf{Gam}(a_0 + L, b_0 - \log \eta) + (1 - d_\eta)\mathsf{Gam}(a_0 + L - 1, b_0 - \log \eta),$$

where $d_\eta/(1 - d_\eta) = (a_0 + L - 1)/[m_{\cdot\cdot}(b_0 - \log(\eta))]$.

References

[1] Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174. MR0365969

[2] Armstrong, H., Carter, C. K., Wong, K. F. & Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing* **19**, 303–316. MR2516221

[3] Atay-Kayis, A. & Massam, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92**, 317–35. MR2201362

[4] Beal, M. J., Ghahramani, Z. & Rasmussen, C. E. (2001). The infinite hidden markov model. In *Proceedings of Fourteenth Annual Conference on Neural Information Processing Systems*.

[5] Bedford, T. & Cooke, R. M. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics* **30**, 1031–1068. MR1926167

[6] Berger, J. O. & Molina, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**, 3–15. MR2137378

[7] Blackwell, D. & MacQueen, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355. MR0362614

[8] Cappé, O., Moulines, E. & Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer. MR2159833

[9] Carvalho, C. M., Massam, H. & West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659. MR2410014

[10] Carvalho, C. M. & West, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis* **2**, 69–98. MR2289924

[11] Castelo, R. & Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *Journal of Machine Learning Research* **7**, 2621–2650. MR2274453

[12] Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* **21**, 1272–1317. MR1241267

[13] Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–75.

[14] Diaconnis, P. & Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics* **7**, 269–81. MR0520238

[15] Dobra, A., Eicher, T. & Lenkoski, A. (2010). Modeling uncertainty in macroeconomic growth determinants using gaussian graphical models. *Statistical Methodology* **7**, 292–306.

[16] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212. MR2064941

[17] Dobra, A., Lenkoski, A. & Rodríguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association* To appear.

[18] Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588. MR1340510

[19] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230. MR0350949

[20] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2** , 615–629. MR0438568

[21] Fraley, C. & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* **24**, 155–181. MR2415725

[22] Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **6**, 799–805.

[23] van Gael, J., Saatci, Y., Teh, Y.-W. & Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*.

[24] Green, P. & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375. MR1842255

[25] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**. MR1380810
[26] Guo, J., Levina, E., Michailidis, G. & Zhu (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.
[27] Heinz, D. (2009). Building hyper Dirichlet processes for gaphical models. *Electronic Journal of Statistics* **3**, 290–315. MR2495840
[28] Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173. MR1952729
[29] Ishwaran, H. & Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963. MR1929973
[30] Jain, S. & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Graphical and Computational Statistics* **13**, 158–182. MR2044876
[31] Jain, S. & Neal, R. M. (2007). Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis* **2**, 445–472. MR2342168
[32] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. & West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statististical Science* **20**, 388–400. MR2210226
[33] Lau, J. W. & Green, P. (2007). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics* **16**, 526–558. MR2351079
[34] Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991
[35] Lee, J., Müller, P., Trippa, L. & Quintana, F. A. (2009). Defining predictive probability functions for species sampling models. Technical report, Pontificia Universidad Católica de Chile.
[36] Lenkoski, A. & Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-wishart prior. *Journal of Computational and Graphical Statistics* **20**, 140–157.
[37] Letac, G. & Massam, H. (2007). Wishart distributions for decomposable graphs. *Annals of Statistics* **35**, 1278–323. MR2341706
[38] Liu, J. S., Liang, F. & Wong, W. H. (2000). The use of multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association* **95**, 121–134. MR1803145
[39] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* **12**, 351–357. MR0733519
[40] Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons. MR0652932
[41] Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79. MR1399156
[42] Müller, P., Quintana, F. & Rosner, G. (2004). Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of the Royal Statistical Society, Series B* **66**, 735–749. MR2088779

[43] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265. MR1823804

[44] Ongaro, A. & Cattaneo, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters* **67**, 33–45. MR2039931

[45] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. MR1337249

[46] Quintana, F. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B.* **65**, 557–574. MR1983764

[47] Roberts, G. & Papaspiliopoulos, O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186. MR2409721

[48] Rodríguez, A., Dunson, D. B. & Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96**, 149–162. MR2482141

[49] Rodríguez, A. & Vuppala, R. (2009). Probabilistic classification using Bayesian nonparametric mixture models. Technical report, University of California, Santa Cruz.

[50] Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics* **29**, 391–411. MR1925566

[51] Scott, J. G. & Carvalho, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics* **17**, 790–808. MR2649067

[52] Sethuraman, J. (1994). A constructive definition of Dirichelt priors. *Statistica Sinica* **4**, 639–650. MR1309433

[53] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B.* **62**, 795–809. MR1796293

[54] Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581. MR2279480

[55] Thiesson, B., Meek, C., Chickering, D. M. & Heckerman, D. (1997). Learning mixtures of DAG models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pp. 504–513. Morgan Kaufmann, Inc.

[56] Wainwright, M. J., Ravikumar, P. & Lafferty, J. D. (2006). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *In Neural Information Processing Systems*. MIT Press.

[57] Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation* **36**, 45–54. MR2370888

[58] Wang, H. & Carvalho, C. M. (2010). Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics* **4**, 1470–1475. MR2741209

[59] Wang, H., Reeson, C. & Carvalho, C. M. (2011). Dynamic Financial Index Models: Modeling Conditional Dependences via Graphs. *Bayesian Analysis* To appear.

[60] Wang, H. & West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96**, 821–834. MR2564493

[61] West, M., Blanchette, H., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R. & Nevings, J. R. (2001). Predicting the clinical status of human breast cancer by using gne expression profiles. *Proceedings of the National Academi of Sciences* **98**, 11462–11467.

[62] West, M. & Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer - Verlag, New York, 2nd edition. MR1482232