# Sparse covariance estimation in heterogeneous samples

Abel Rodriguez,* Alex Lenkoski†and Adrian Dobra‡

**Abstract**

Standard Gaussian graphical models (GGMs) implicitly assume that the conditional independence among variables is common to all observations in the sample. However, in practice, observations are usually collected form heterogeneous populations where such assumption is not satisfied, leading in turn to nonlinear relationships among variables. To tackle these problems we explore mixtures of GGMs; in particular, we consider both infinite mixture models of GGMs and infinite hidden Markov models with GGM emission distributions. Such models allow us to divide a heterogeneous population into homogenous groups, with each cluster having its own conditional independence structure. The main advantage of considering infinite mixtures is that they allow us easily to estimate the number of number of subpopulations in the sample. As an illustration, we study the trends in exchange rate fluctuations in the pre-Euro era. This example demonstrates that the models are very flexible while providing extremely interesting interesting insights into real-life applications.

Keywords: Covariance selection; Dirichlet process; Gaussian graphical model; Hidden Markov model, Nonparametric Bayes inference

## 1 Introduction

Problems with small sample sizes and large number of unknown parameters represent one of the most challenging areas of current statistical research. Graphical models deal with this type of ill-

---

*Department of Applied Mathematics and Statistics, University of California, Mailstop SOE2, Santa Cruz, California 95064, U.S.A. (email: abel@soe.ucsc.edu).

†Department of Applied Mathematics, Heidelberg University, Heidelberg, Germany (email: lenkoski@stat.washington.edu).

‡Departments of Statistics, Biobehavioral Nursing and Health Studies and the Center for Statistics and the Social Sciences, Box 354322, University of Washington, Seattle, WA 98195, U.S.A. (email: adobra@uw.edu).

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. In particular, Gaussian graphical models (GGMs), also known as covariance selection models (Dempster, 1972), have become extremely popular in applications ranging from genetics (West et al., 2001; Castelo & Roverato, 2006) to econometrics and finance (Carvalho & West, 2007; Dobra et al., 2008). GGMs 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 GGMs is that they implicitly assume a linear relationship between variables. Copulas have been used in the context graphical models to address nonlinearities. For example, Bedford & Cooke (2002) decompose the joint distribution of $X$ using pairwise copulas; however, the resulting models are computationally difficult to fit, specially when $p$ grows. An alternative to copulas is to model non-linearities through mixtures of GGMs. 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 (Müller et al., 1996; Rodriguez et al., 2009).

As a motivation for investigating mixtures of GGMs, consider the analysis of gene expression data. GGMs have been often used in the context of microarray data, where the graph encoding the conditional dependence structure provides information about expression pathways (Dobra et al., 2004; Friedman, 2004; Castelo & Roverato, 2006). 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, which 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 setting, mixtures of GGMs not only provide us with a tool to induce sparsity in heterogeneous samples, but also generate highly interpretable models.

The major challenges in implementing mixtures of GGMs are computational, and relate both

to the determination of the underlying graph associated with each component in the mixture and to the estimation of number of components. Indeed, it is well known that the number of possible partitions for the data grows exponentially, 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 Thiesson et al. (1997), who fixed the number of components in the mixture and developed a search algorithm that uses a modified Cheeseman-Stutz approximation to the marginal likelihood coupled with EM steps to estimate component-specific parameters. However, 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.

In this paper, we present the first 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 distributions (Lo, 1984; Ongaro & Cattaneo, 2004), and allows us to automatically deal with an unknown number of components/states within a simple computational framework. 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 multivariate time series models employing graphical models (Carvalho & West, 2007; Wang & West, 2009). Since the paper focuses on models that have a Pòlya urn representation, we construct marginal samplers (Neal, 2000) that explicitly integrate out the mean and variance of the individual GGMs. For problems where the the main interest is either prediction or inference the partition structure and/or the graphical structure associated with each component, this approach greatly reduces computational complexity by avoiding the explicit representation the mean and variance of the different components/states. Although the paper develops models based on GGMs, 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.

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 Section 3 and 4. We then move to discuss more general nonparametric mixture models in Section 5, including species sampling mixtures of GGMs and infinite hidden Markov models with GGM emission distributions. These models are illustrated in Section 6 with a simulated and a real-world dataset. Finally, we conclude in Section 7 with a discussion of possible extensions and future research directions.

# 2 Bayesian Framework for GGMs

Let $X = X_V$ be the vector of observed variables, where $V = \{1, 2, \ldots, p\}$. We assume that $X$ follows a multivariate Normal distribution $p(X|\mu, K) = N_p(\mu, K^{-1})$ with mean vector $\mu \in \mathbb{R}^p$ and $p \times p$ precision matrix $K = (K_{ij})$. We consider the set $\mathcal{G}_V$ of decomposable graphs associated with $V$. The Gaussian graphical model (GGM) 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 $G$ (Dempster, 1972). 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.

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}.$$

The precision matrix $K$ belongs to the cone $P_G$ of the symmetric positive definite matrices with entries equal to zero for all $(i, j) \in (V \times V) \setminus E$ (Atay-Kayis & Massam, 2005). The conditional dependence relationships implied by $G$ induce the following factorization of the joint distribution of $X$ (Dawid & Lauritzen, 1993):

$$p(X|\mu, K, G) = \frac{\prod_{C \in \mathcal{C}} p(X_C|\mu_C, K_C)}{\prod_{S \in \mathcal{S}} p(X_S|\mu_S, K_S)} \tag{1}$$

where $\mathcal{C}$ denotes the cliques of $G$ and $\mathcal{S}$ denotes separators of $G$. For an index set $V_0 \subset V$, $\mu_{V_0}$ is the subvector of $\mu$ corresponding to the entries in $V_0$, while $K_{V_0} = ((K^{-1})_{V_0})^{-1}$. We remark that 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. Similarly, the subgraph $G_S$ associated with a separator $S \in \mathcal{S}$ is also complete.

## 2.1 Prior specification

We consider the following joint prior distribution for $\mu$ and $K$:

$$p(\mu, K|G) = p(\mu|K, G)p(K|G), \tag{2}$$

where, conditional on $K$, the prior for the mean is $p(\mu|K, G) = N_p(\mu_0, (n_0 K)^{-1})$ with $\mu_0 \in \mathbb{R}^p$ and $n_0 > 0$. The prior for the precision matrix $p(K|G) = W_G(\delta_0, D_0)$ is a G-Wishart distribution with density (Roverato, 2002; Atay-Kayis & Massam, 2005; Letac & Massam, 2007)

$$\frac{1}{I_G(\delta_0, D_0)} (\det K)^{(\delta_0 - 2)/2} \exp\left\{-\frac{1}{2}\langle K, D_0\rangle\right\}, \tag{3}$$

4

with respect to the Lebesgue measure on $P_G$. Here $\langle B, C \rangle = \text{tr}(B^T C)$ denotes the trace inner product. Diaconnis & Ylvisaker (1979) 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. $G$ has only one clique $C = \{V\}$ 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

$$I_G(\delta_0, D_0) = 2^{(\delta_0+p-1)p/2} \Gamma_p \{(\delta_0 + p - 1)/2\} (\det D_0)^{-(\delta_0+p-1)/2}, \tag{4}$$

where $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma\left(a - \frac{i}{2}\right)$ for $a > (p-1)/2$ (Muirhead, 2005). If $G$ is decomposable but not necessarily complete, Dawid & Lauritzen (1993) prove 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 (Roverato, 2002):

$$I_G(\delta_0, D_0) = \frac{\prod_{C \in \mathcal{C}} I_{G_C}(\delta_0, (D_0)_C)}{\prod_{S \in \mathcal{S}} I_{G_S}(\delta_0, (D_0)_{S_j})}. \tag{5}$$

The subgraphs $G_C$ and $G_S$ associated with each clique and separator of $G$ are complete, thus $I_{G_C}(\delta_0, (D_0)_C)$ and $I_{G_S}(\delta_0, (D_0)_S)$ are explicitly calculated as in (4).

## 2.2 Posterior distributions and the marginal likelihood of a graph

The likelihood function for $\mu$ and $K$ corresponding with an *i.i.d.* sample $x^{(1:n)} = \left(x^{(1)}, \ldots, x^{(n)}\right)$ from $X \sim N_p(\mu, K^{-1})$ is given by

$$L(\mu, K | x^{(1:n)}) = (2\pi)^{-np/2} (\det K)^{n/2} \exp\left\{-\frac{1}{2}\langle K, U + n(\bar{x} - \mu)(\bar{x} - \mu)^T \rangle\right\}, \tag{6}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$, $U = \sum_{i=1}^{n} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$. The joint prior (6) is conjugate to the likelihood (6). 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. 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. The interpretation of the resulting prior is that the components of $X$ are apriori independent and that the posterior "weight" of the prior is equivalent to one observed sample. Other possible choices for the G-Wishart prior parameters $\delta_0$ and $D_0$ are discussed in Carvalho & Scott (2009).

The marginal likelihood

$$p(x^{(1:n)}|G) = \int_{\mu \in \mathbb{R}^p} \int_{K \in P_G} L(\mu, K | x^{(1:n)}) p(\mu, K | G) \, dK \, d\mu$$

associated with a graph $G \in \mathcal{G}_V$ is

$$\frac{(2\pi)^{-\frac{(n+1)p}{2}} n_0^{p/2}}{I_G(\delta_0, D_0)} \int_{K \in P_G} J(K)(\det K)^{\frac{\delta_0+n}{2}} \exp\left\{-\frac{1}{2}\langle K, U + D_0\rangle\right\} dK, \tag{7}$$

where

$$
\begin{aligned}
J(K) &= \int_{\mu \in \mathbb{R}^p} \exp\left\{-\frac{1}{2}\langle K, n(\bar{x} - \mu)(\bar{x} - \mu)^T + n_0(\mu - \mu_0)(\mu - \mu_o)^T\rangle\right\} d\mu, \\
&= \int_{\mu \in \mathbb{R}^p} \exp\left\{-\frac{1}{2}\langle K, (n + n_0)(\mu - \bar{\mu})(\mu - \bar{\mu})^T + A\rangle\right\} d\mu.
\end{aligned}
$$

with $\bar{\mu} = \frac{n\bar{x}+n_0\mu_0}{n+n_0}$ and $A = -(n + n_0)\bar{\mu}\bar{\mu}^T + n\bar{x}\bar{x}^T + n_0\mu_0\mu_0^T$. After factorizing out $\exp\left\{-\frac{1}{2}\langle K, A\rangle\right\}$, the remaining integrand is the kernel of the posterior distribution of $\mu$:

$$p(\mu|x^{(1:n)}, K, G) = N_p\left(\bar{\mu}, [(n + n_0)K]^{-1}\right). \tag{8}$$

Therefore

$$J(K) = \frac{(2\pi)^{p/2}}{(n + n_0)^{p/2}}(\det K)^{-1/2} \exp\left\{-\frac{1}{2}\langle K, A\rangle\right\}.$$

and it follows that the marginal likelihood (7) becomes

$$\frac{(2\pi)^{-\frac{np}{2}}}{I_G(\delta_0, D_0)}\left(\frac{n_0}{n + n_0}\right)^{p/2} \int_{K \in P_G} (\det K)^{\frac{\delta_0+n-2}{2}} \exp\left\{-\frac{1}{2}\langle K, D_0 + U + A\rangle\right\} dK.$$

The integrand in the equation above is the kernel of the G-Wishart posterior distribution of $K$:

$$p(K|x^{(1:n)}, G) = W_G(\delta_0 + n, D_0 + U + A). \tag{9}$$

and therefore, the final form of the marginal likelihood of the data given $G$ is

$$p(x^{(1:n)}|G) = (2\pi)^{-\frac{np}{2}}\left(\frac{n_0}{n + n_0}\right)^{p/2} \frac{I_G(\delta_0 + n, D_0 + U + A)}{I_G(\delta_0, D_0)}. \tag{10}$$

A similar argument shows that the posterior predictive distribution of a new sample $x^{(n+1)}$ is given by

$$p(x^{(n+1)}|x^{(n)}, G) = (2\pi)^{-\frac{p}{2}}\left(\frac{n + n_0}{n + 1 + n_0}\right)^{p/2} \frac{I_G(\delta_0 + n + 1, D_0 + U + A + \widetilde{A})}{I_G(\delta_0 + n, D_0 + U + A)}, \tag{11}$$

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 $\widetilde{\mu} = \frac{x^{(n+1)}+(n+n_0)\bar{\mu}}{n+1+n_0}$. Since $G$ 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 (5), hence $p(x^{(n+1)}|G)$ and $p(x^{(n+1)}|x^{(n)}, G)$ 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.

6

# 3 Dirichlet Process Mixtures of GGMs

Consider now a mixture of GGMs

$$X|\{w_l\}, \{\mu_l^*\}, \{K_l^*\}, \{G_l^*\} \sim \sum_{l=1}^{L} w_l p(X|\mu_l^*, K_l^*, G_l^*) \tag{12}$$

where $p(X|\mu_l^*, K_l^*, G_l^*)$ is given by (1). In this model, draws from 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 $\mu_l^*$, $K_l^*$ and $G_l^*$) independently with probability $w_l$. A fully Bayesian specification of the model is completed by eliciting prior for the parameters $(\{w_l\}_{l=1}^{\infty}, \{\mu_l^*\}_{l=1}^{\infty}, \{K_l^*\}_{l=1}^{\infty}, \{G_l^*\}_{l=1}^{\infty})$. A common choice is to set $w = (w_1, \ldots, w_L) \sim \mathsf{Dir}(w^0)$ and let the component specific parameters $(\mu_l^*, K_l^*, G_l^*)$ be i.i.d. samples from some common distribution $M$.

Finite mixtures as the one described above allow for additional flexibility over regular GGMs by allowing a heterogeneous population to be divided into homogenous groups. However, estimating finite mixture models involves important practical challenges. For example, in practice 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 (RJMCMC) methods (Green, 1995), which are notoriously inefficient for high dimensional mixtures.

As an alternative, this section considers Dirichlet process mixtures of GGMs (GGM-DPM). Note that (12) can be alternative written as

$$X|H \sim \int p(X|\mu, K, G)H(\mathrm{d}\mu, \mathrm{d}K, \mathrm{d}G) \qquad H(\cdot) = \sum_{l=1}^{L} w_l \delta_{(\mu_l^*, K_l^*, G_l^*)}(\cdot) \tag{13}$$

where $\delta_a(\cdot)$ denotes the degenerate probability measure putting all its mass on $a$. Therefore, eliciting a prior on $(\{w_l\}_{l=1}^{\infty}, \{\mu_l^*\}_{l=1}^{\infty}, \{K_l^*\}_{l=1}^{\infty}, \{G_l^*\}_{l=1}^{\infty})$ is equivalent to defining a prior on the discrete probability measure $H$, one such prior is the Dirichlet process (Ferguson, 1973, 1974). A random distribution $H$ is said to follow a Dirichlet process (DP) with baseline measure $M$ and precision parameter $\alpha_0$, denoted $\mathsf{DP}(\alpha_0, M)$, if it has a representation of the form (Sethuraman, 1994)

$$H(\cdot) \sim \sum_{l=1}^{\infty} w_l \delta_{\theta_l^*}(\cdot), \tag{14}$$

where $\theta_1^*, \theta_2^*, \ldots$ are independent and identically distributed samples from the *baseline measure* $M$ and $w_l = u_l \prod_{s<l}(1 - u_s)$ where $u_1, u_2, \ldots$ is another independent and identically distributed

sample where $u_l \sim \mathsf{Beta}(1, \alpha_0)$. We refer to the joint distribution on $(w_1, w_2, \ldots)$ induced by the above construction as a stick breaking distribution with parameter $\alpha_0$, denoted $\mathsf{SB}(\alpha_0)$. The DP mixture (DPM) model is recovered from (13) when $H \sim \mathsf{DP}(\alpha_0, M)$ for appropriately chosen hyperparameters $\alpha_0$ and $M$.

Consider now an independent and identically distributed sequence $\theta_1, \ldots, \theta_n$ such that $\theta_j | H \sim H$, where $H \sim DP(\alpha_0, M)$. A useful feature of the DP 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 (Blackwell & MacQueen, 1973) where $\theta_1 \sim M$ and

$$\theta_{j+1} | \theta_j, \ldots, \theta_1, \alpha_0 \sim \sum_{i=1}^{j} \frac{1}{\alpha_0 + j} \delta_{\theta_i} + \frac{\alpha_0}{\alpha_0 + j} M, \qquad j > 1. \qquad (15)$$

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 $1 \le L \le n$ 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, (15) implies that $\theta_1^*, \theta_2^*, \ldots$ is a sequence of independent and identically distributed samples from $M$, $\xi_1 = 1$ and

$$\xi_{j+1} | \xi_j, \ldots, \xi_1, \alpha_0 \sim \sum_{l=1}^{L^j} \frac{r_l^j}{\alpha_0 + j} \delta_l + \frac{\alpha_0}{\alpha_0 + j} \delta_{L^j+1}, \qquad j > 1, \qquad (16)$$

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 (15) and (16) clearly emphasize that, for any finite sample $x^{(1:n)}$, the number of non-empty components $L^n = L$ in a DPM model is a random parameter in the model. The prior on $L$ implied by the DP (Antoniak, 1974) is given by :

$$p(L|\alpha_0, n) = S(n, L) n! \alpha_0^{L^n} \frac{\Gamma(\alpha_0)}{\alpha_0 + n} \qquad L = 1, \ldots, n, \qquad (17)$$

where $S(\cdot, \cdot)$ denotes the Stirling number of the first kind. Therefore, the mean number of non-empty components grows with $\alpha_0$, the concentration parameter.

The DP mixture model is intimately connected to the finite mixture model in (12). Consider a finite mixture with $N$ components such that

$$x^{(j)} | \{\theta_l^*\}_{l=1}^{N}, \{\xi_j\}_{j=1}^{n} \sim p(x^{(j)} | \theta_{\xi_j}^*), \qquad \xi_j | \alpha_0 \sim \mathsf{Multinom}(\alpha_0/N, \ldots, \alpha_0/N), \qquad \theta_l^* \sim M. \qquad (18)$$

As $N \to \infty$, the predictive distribution for $x^{(1:n)}$ under this model converges to the one obtained from the DP mixture (Green & Richardson, 2001; Ishwaran & Zarepour, 2002).

8

In the GGM-DPM model we explore in this paper we have $\theta = (\mu, K, G)$ and the baseline measure is defined by

$$M = p(\mu, K|G)p(G), \tag{19}$$

where $p(\mu, K|G)$ is given by (2) and $p(G) \propto 1$ is the uniform prior on $\mathcal{G}_V$. Other choices of priors on $\mathcal{G}_V$ that encourage sparsity or have desirable multiple testing properties are discussed in Jones et al. (2005); Scott & Berger (2006); Scott & Carvalho (2008).

The GGM-DPM model is a natural extension of the well-known DP mixture of multivariate normals originally presented in Müller et al. (1996), 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. The point estimates provided by the GGM-DPM model we just described 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 cluster equal to the logarithm of (17), 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. 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 (Wainwright et al., 2006) and mixture models (Fraley & Raftery, 2007).

Although the model just described induces sparsity on the structure of the component specific covariance matrices $\{K_l\}$, the fact that we are using a mixture of GGMs as the data generating model means that such sparse structure does not translate into conditional independence for the variables involved. Indeed, even if for a given pair $(i, j)$ we have $(K_l)_{ij} = 0$ for all $l$, $X_i$ and $X_j$ are not conditionally independent under the GGM-DPM model. Therefore, all conditional independence assumptions derived from the graphs $\{G_l\}$ are valid only conditional on cluster membership.

# 4 Computational implementation of GGM mixtures

As with regular GGM models, the posterior distribution arising from the DDP-GGM model is not analytically tractable because of the sheer size of the space of partitions and accompanying graphs. Therefore, we resort to MCMC algorithms to explore the features of this complicated posterior distribution. The literature on MCMC samplers for the DPM model has grown extensively in the last

15 years; the approaches can be roughly divided in three large classes: collapsed (marginal) Gibbs samplers (MacEachern, 1994; Escobar & West, 1995; Neal, 2000), which exploit the exchangeability in the data and the Pólya urn representation in (15) and (16) to construct algorithms that avoid explicitly sampling $H$, blocked samplers (Ishwaran & James, 2001; Roberts & Papaspiliopoulos, 2008; Walker, 2007), which explicitly represent the mixing distribution $H$, and Reversible Jump samplers (Jain & Neal, 2004, 2007). In this paper we focus attention on marginal samplers such as the ones described in Neal (2000) as a natural option that provides some computational advantages. Indeed, the structure of the baseline measure $M$ in (19) 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 $(L, \{\xi_j\}_{j=1}^n, \{G_l\}_{l=1}^L, \alpha_0 | x^{(1:n)})$. 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 $\xi_j$ for $j = 1, \ldots, n$ from its full conditional distribution

$$p(\xi_j | \xi^{-j} = \{\xi_{j'}\}_{j' \neq j}, x^{(1:n)}, \{G_l\}_{l=1}^{L^{-j}}) \propto \sum_{l=1}^{L^{-j}+1} q_{jl} \delta_l, \tag{20}$$

where

$$q_{jl} \propto \begin{cases} r_l^{-j} p(x^{(j)} | \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l), & l \leq L^{-j}, \\ \alpha_0 p(x^{(j)} | G_{L^{-j}+1}), & l = L^{-j} + 1. \end{cases}$$

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 $l$), $p(x^{(j)} | \{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l)$ 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 (11), and $p(x^{(j)} | G_{L^{-j}+1})$ 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 (11). 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 (19). 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 decomposable 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 (1996). We draw a candidate graph $G_l^{new}$ from the uniform distribution on $\text{nbd}_{\mathcal{G}_V}(G_l)$. We change the graph associated with cluster $l$ to $G_l^{new}$ with probability

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

otherwise the graph associated with cluster $l$ remains unchanged. Here $p(\{x^{(j)} : \xi_j = l\}|G)$ represents the marginal likelihood of the samples currently in cluster $l$ given a graph $G$ – see equation (10). 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 (typically, between 5 and 10 times seems to provide adequate mixing).

These two sequences of steps produce a sample from the posterior distribution of interest, $(L, \{\xi_j\}_{j=1}^n, \{G_l\}_{l=1}^L, \alpha_0|x^{(1:n)})$ 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 conditional on $\{\xi_j\}_{j=1}^n$ by noting that for every $l = 1, \ldots, L$ (see equations (8) and (9))

$$K_l|G_l, \{x^{(j)} : \xi_j = l\} \sim W_{G_l}(\delta_0 + r_l, D_0 + U_l + A_l),$$

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

independently of other components. 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 $l$).

Also, additional flexibility can be obtained by sampling some of the hyperparameters associated with the DP prior. For example, the concentration parameter $\alpha_0$ controls the expected number of components, and therefore has an important effect on the inferences generated by the model. Since eliciting values for $\alpha_0$ can be extremely difficult in practice, it is recommendable to try to infer it from the data. For example, we can assume a vague $\mathsf{G}(a_0, b_0)$ prior for the precision parameter $\alpha_0$, in which case the full conditional distribution can be easily sampled using an auxiliary-variable Gibbs sampling step (Escobar & West, 1995) (see Appendix).

# 5 More general nonparametric mixtures of graphical models

The ideas just described for Dirichlet process mixtures of Gaussian graphical models can be directly extended to other nonparametric mixture models where Pòlya urn representations similar to (15) and (16) can be exploited to analytically integrate out the random distributions out of the model. Some example include species sampling models (McCloskey, 1965; Pitman, 1996; Lijoi et al., 2007; Lee et al., 2009), hierarchical Dirichlet processes (Teh et al., 2006), nested Dirichlet processes (Rodriguez et al., 2008) and linear combinations of Dirichlet process (Müller et al., 2004; Dunson et al., 2007). In this section we consider in detail two such extensions.

## 5.1 Species sampling models

As a first example, consider replacing the Dirichlet process mixture of GGMs with a more general species sampling mixture of GGMs. An *exchangeable* sequence $\theta_1, \ldots, \theta_n$ is said to follow a species sampling model (McCloskey, 1965; Pitman, 1996; Lijoi et al., 2007; Lee et al., 2009) with baseline measure $M$ if $\theta_j = \theta^*_{\xi_j}$ where $\theta^*_1, \theta^*_2, \ldots$ is a sequence of independent and identically distributed samples from $M$ and the indicators $\xi_1, \ldots, \xi_n$ are sampled according to the predictive formula

$$\xi_{j+1} | \xi_j, \ldots, \xi_1 \sim \sum_{l=1}^{L^j} v_{jl}(r^j_1, \ldots, r^j_{L^j}) \delta_l + v_{j,L^j+1} \delta_{L^j}, \tag{21}$$

where the weights $v_{j1}, \ldots v_{j,L^j+1}$ satisfy $\sum_{i=1}^{L^j+1} v_{jl} = 1$ for all $i$ and $r^j_l$ and $L^j$ are defined as in Section 3. The Dirichlet process is the best known member of the class of species sampling models, which also includes the two parameter Poisson-Dirichlet process (Pitman, 1996; Ishwaran & James, 2001) and the normalized inverse-gamma priors(Lijoi et al., 2005), among others.

Moving beyond DPM models is of interest because the prior on the partition structure induced by the Dirichlet process can be somewhat restrictive. For example, (17) implies that, a priori and for a given precision parameter $\alpha_0$, the expected number of occupied clusters grows with the logarithm of $n$, which might be inappropriate for certain application such as computer vision (Sudderth & Jordan, 2009). Also, the Dirichlet process favors partitions that consist of a small number of clusters with a large number of observations along with a larger number of small clusters.

Since samples from a species sampling model are exchangeable, the predictive distribution (21) also provides the full conditional distribution required to implement the Gibbs sampling algorithm discussed in Section 3. For example, for the Poisson Dirichlet process with baseline measure $M$,

discount $0 \leq \alpha < 1$ and strength $\alpha_0 \geq -\alpha$ we have

$$\xi_{j+1}|\xi_j, \ldots, \xi_1 \sim \sum_{l=1}^{L^j} \frac{r_l^j - \alpha}{\alpha_0 + j}\delta_l + \frac{\alpha_0 - \alpha L^j}{\alpha_0 + j}\delta_{L^j+1}$$

Note that if $\alpha = 0$, the Poisson-Dirichlet process reduces to the standard Dirichlet process. Modifying the algorithm in Section 3 to fit a Poisson-Dirichlet mixture of GGMs is straightforward. In particular, we only need to slightly modify the posterior weights in (20) to reflect the new prior distribution,

$$q_{jl} = \begin{cases} (r_l^{-j} - \alpha)p(x^{(j)}|\{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l), & l \leq L^{-j}, \\ (\alpha_0 - \alpha L^{-j})p(x^{(j)}|G_{L^{-j}+1}), & l = L^{-j} + 1. \end{cases}$$

## 5.2 Infinite Hidden Markov Gaussian Graphical models

Recently, multivariate time series models that use graphical models to improve estimation of the crosssectional covariance structure have been developed (Carvalho & West, 2007; Wang & West, 2009). These approaches rely on extensions of the dynamic linear model (DLM) (West & Harrison, 1997) and assume that the graph underlying the model is constant in time which, as our first illustration in Section 6.2 demonstrates, might not be an appropriate assumption in practical application. As an alternative we focus on a nonparametric version of the popular hidden Markov model where the emission distribution corresponds to a GGM.

Hidden Markov models (HMMs) (Cappé et al., 2005), are hierarchical mixture models where

$$x^{(j)}|\{\theta_l^*\}_{l=1}^L, \{\xi_j\}_{j=n}^L \sim p(x^{(j)}|\theta_{\xi_j}^*), \quad \xi_j|\xi_{j-1}, \{\pi^l\}_{l-1}^L \sim \mathsf{Multinom}(\pi^{\xi_{j-1}}), \quad \xi_0 \sim \mathsf{Multinom}(\pi^0) \quad \theta_l^* \sim M.$$

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 $\mathsf{Pr}(\xi_j = l|\xi_{j-1} = l') = \pi_l^{l'}$. 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 $p(\cdot|\theta_{\xi_1}^*), \ldots, p(\cdot|\theta_{\xi_n}^*)$.

Infinite hidden Markov models (iHMMs) (Beal et al., 2001; Teh et al., 2006; van Gael et al., 2008) generalize HMMs to models with an infinite number of states, in a similar way as how Dirichlet process models generalize finite mixture models, allowing us to estimate the number of

states $L$. In particular, we can build a GGM-iHMM where

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

$$\xi_j|\xi_{j-1}, \{\pi^l\}_{l=1}^{\infty} \sim \mathsf{Multinom}(\pi^{\xi_{j-1}})$$

$$\pi^l|\alpha, \gamma \sim \mathsf{DP}(\alpha, \gamma)$$

$$\gamma|\alpha_0 \sim \mathsf{SB}(\alpha_0),$$

and $\theta_l^* = (\mu_l, K_l, G_l) \sim M$, where $M$ is defined as in (19). This GGM-iHMM has some distinct advantages over the dynamic linear models with graphical structure discussed in Carvalho & West (2007) and Wang & West (2009). 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.

Again, a marginal Gibbs sampler similar to the one described in Section 3 can be devised for the GGM-iHMM. We denote by $r_{ll'}^{j_1:j_2}$ the number of transitions from state $l$ to state $l'$ in the sub-trajectory $\{\xi_j\}_{j=j_1}^{j_2}$ and by $r_{l.}^{j_1:j_2}$ the number of transitions out of state $l$ in the same sub-trajectory. Given the base DP parameters $\gamma = \{\gamma_1, \ldots, \gamma_{L+1}\}$ and the precision parameters $\alpha_0$ and $\alpha$, the states $\xi_1, \ldots, \xi_n$ are sequentially updated using the full conditional distributions:

$$q_{jl} = \begin{cases} \left(r_{\xi_{j-1},l}^{1:(j-1)} + r_{\xi_{j-1},l}^{(j+1):n} + \alpha\gamma_l\right) \frac{r_{l,\xi_{j+1}}^{1:(j-1)} + r_{l,\xi_{j+1}}^{(j+1):n} + \alpha\gamma_{\xi_{j+1}}}{r_{l.}^{1:(j-1)} + r_{l.}^{(j+1):n} + \alpha} p(x^{(j)}|\{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l), & l \leq L^{-j}, l \neq \xi_{t-1}, \\[3mm] \left(r_{\xi_{j-1},l}^{1:(j-1)} + r_{\xi_{j-1},l}^{(j+1):n} + \alpha\gamma_l\right) \frac{r_{l,\xi_{j+1}}^{1:(j-1)} + r_{l,\xi_{j+1}}^{(j+1):n} + \alpha\gamma_{\xi_{j+1}} + 1}{r_{l.}^{1:(j-1)} + r_{l.}^{(j+1):n} + \alpha + 1} p(x^{(j)}|\{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l), & l = \xi_{j-1} = \xi_{j+1}, \\[3mm] \left(r_{\xi_{j-1},l}^{1:(j-1)} + r_{\xi_{j-1},l}^{(j+1):n} + \alpha\gamma_l\right) \frac{r_{l,\xi_{j+1}}^{1:(j-1)} + r_{l,\xi_{j+1}}^{(j+1):n} + \alpha\gamma_{\xi_{j+1}}}{r_{l.}^{1:(j-1)} + r_{l.}^{(j+1):n} + \alpha + 1} p(x^{(j)}|\{x^{(j')} : j' \neq j, \xi_{j'} = l\}, G_l), & l = \xi_{j-1} \neq \xi_{j+1}, \\[3mm] \alpha\gamma_l\gamma_{\xi_{j+1}} p(x^{(j)}|G_{L+1}), & l = L^{-j} + 1. \end{cases}$$

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 $\gamma$ by setting $\gamma_{L+1}^{new} = v\gamma_{L+1}$, $\gamma_{L+2}^{new} = (1-v)\gamma_{L+1}$ were $v \sim \mathsf{Beta}(\alpha_0, 1)$.

Given a trajectory $\{\xi_j\}_{j=1}^{n}$, $\alpha_0$ and $\alpha$, we sample $\gamma$ by introducing the independent auxiliary variables $\{m_{ll'}\}$ for $l, l' \in \{1, \ldots, L\}$ such that

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

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

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

where $m_{.l'} = \sum_{l=1}^{L} m_{ll'}$. We use vague gamma priors for the precision parameters $\alpha_0$ and $\alpha$ and update them as described in the Appendix.

# 6 Examples

## 6.1 Simulated data

We consider first a small simulation that involves data arising from a 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 samples are from a star graphical model $N_{10}(0.5, K_1^{-1})$ with every variable $X_j$, $j > 2$, connected to $X_1$, while the second cluster contains 100 samples from a cycle model $N_{10}(-0.5, K_2^{-1})$. The non-zero elements of the two precision matrices are

$$
\begin{aligned}
(K_1)_{j,j} = (K_2)_{j,j} = 1, && j = 1, \ldots, 10, \\
(K_1)_{1,j} = (K_1)_{j,1} = 0.3, && j = 2, \ldots, 10, \\
(K_2)_{j-1,j} = (K_2)_{j,j-1} = 0.3, && j = 2, \ldots, 10, \\
(K_2)_{1,10} = (K_2)_{10,1} = 0.3.
\end{aligned}
$$

We are interested in recovering the two clusters and their corresponding conditional independence graphs. Since the first 100 observations belong to the first cluster and the remainder to the second cluster, we are also able to employ the iHMM discussed above and compare its performance to the DPM model. After sampling a dataset, we ran both DPM and iHMM samplers for 20,000 iterations after 5,000 iterations of burn-in. To determine the utility of including GGMs in the Dirichlet Process framework we also ran both the DPM and iHMM models with attention restricted to the full graph.

Figure 1 shows the clustering for the DPM and iHMM both using the full space of decomposable GGMs and with attention restricted to the full graph. The upper lefthand panel, which corresponds to the results from the DPM model using the full graph space, shows two well defined groups with a few observations incorrectly classfied. The upper righthand panel shows that restricting attention to the full graph decreases the ability to cluster observations properly. The bottom row of Figure 1 shows that using the iHMM model in this case dramatically improves clustering. The iHMM model using the unrestricted graph space clusters the observations perfectly, while the iHMM model using the full graph does nearly as well, though it misclassifies one observation.

Figure 2 shows the edge probabilities in each of the two clusters for the DPM and iHMM models, along with the true edges from the underlying model. We see that for the most part, the edges with the highest probability in both models correspond to the true edges. However, the improved

15

clustering capabilities of the iHMM model translates into sharper edge probabilities, especially in the second cluster.

## 6.2   Real data: understanding trends in exchange rate fluctuations

First, we consider a dataset that follows the exchange rate of 11 currencies relative to the US dollar between November 1993 and August 1996. This dataset consists of 1000 daily observations and includes three Asian currencies (NZD,AUD, JPY), five European currencies that eventually became part of the Euro (DEM, FRF, BEF, NLG, ESP) and three additional European currencies (GBP, SEK, CHF). These data have previously been used in a variety of contexts related to graphical models (see e.g. Carvalho et al., 2007 and Carvalho & West, 2007). In Carvalho et al. (2007) the authors present the graph shown in Figure 3, which is determined using stochastic search methods first discussed in Jones et al. (2005) 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 single clique, while the British Pound (GBP), Swedish Krona (SEK) and Swiss Franc (CHZ) connect with only some of these counties, most notably the currencies of the largest Euro-area countries, the Deutsch Mark (DEM) and French Franc (FRF) (the Swiss Franc is also connected to the Netherlands Gilder (NLG), being more integrated with mainland European economies). Carvalho & West (2007) 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 $\Sigma$, evidence of the effectiveness of the graphical models approach.

We used the exchange 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 GGM-iHMM model discussed in Section 5.2. We ran the model for 100000 iterations after a burn-in period of 20000 iterations, and ran five separate instances of the algorithm from separate starting points. After completion we assessed the results from each chain and verified they returned the same estimates, Figure 4 shows the convergence in $\alpha$ and $\alpha_0$ across chains. When run using the iHMM 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 first 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
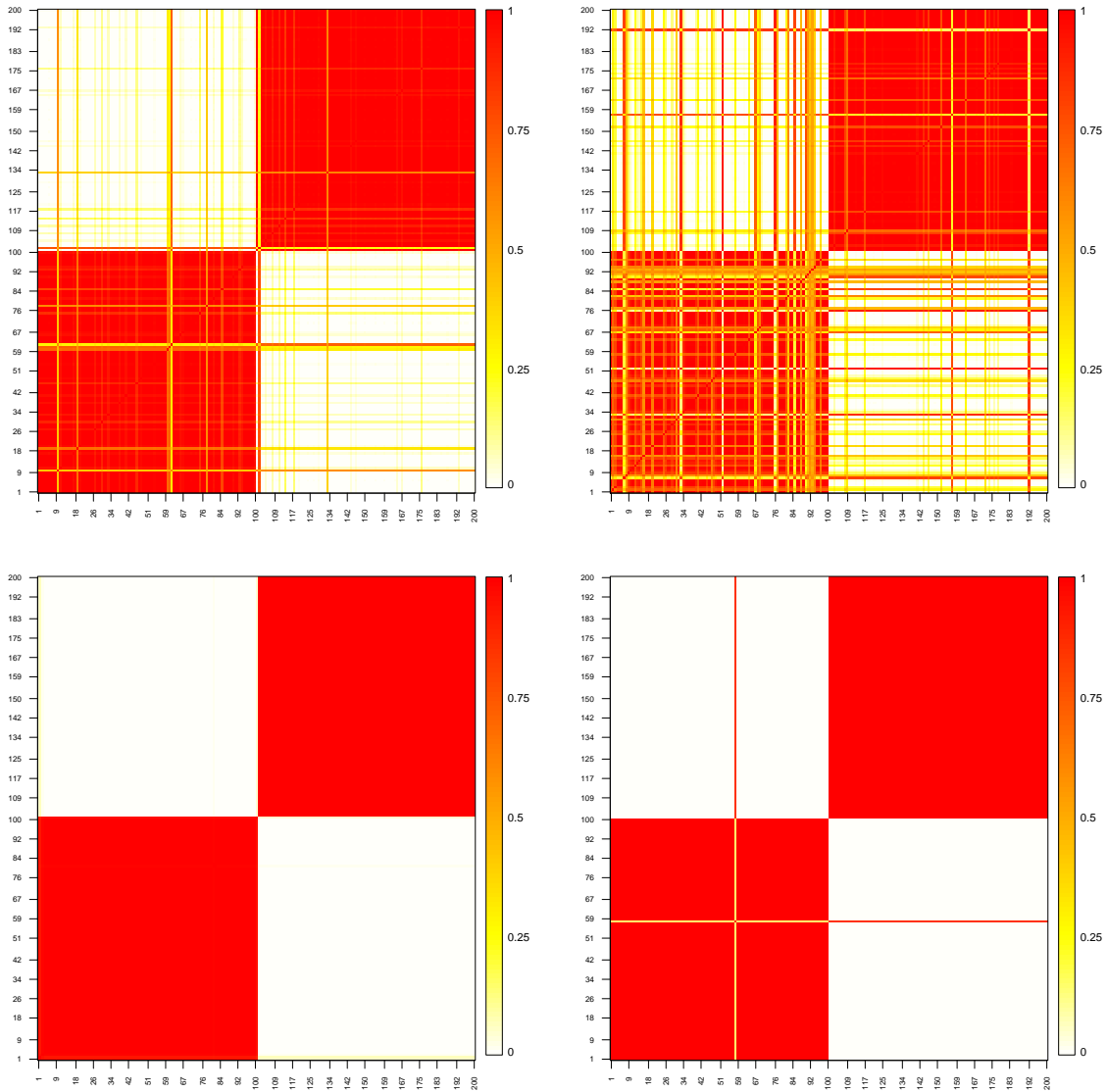
Figure 1: Probability that two observations belong to the same cluster in the simulated example. The panel shows, in a clockwise order, the results using the DPM model in the unrestricted case, the DPM model restricting to the full graph, the iHMM model restricting to the full model and the iHMM model in the unrestricted case. Note that the true underlying clustering is shown by the iHMM model in the unrestricted case.
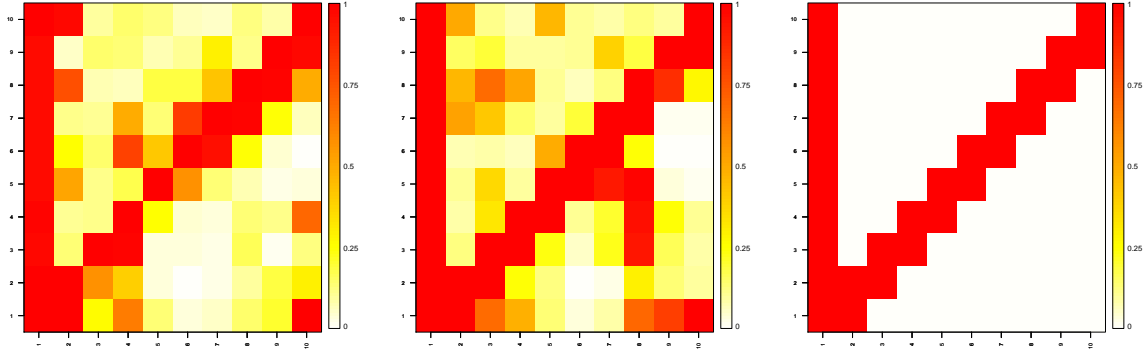
Figure 2: Edge probabilities in the simulated example. In each panel, the upper triangle shows the average edge probability of observations in the first cluster, while the lower triangle shows the average edge probability of observations in the second cluster. The lefthand panel corresponds to the results from the DPM search, the middle panel corresponds to the iHMM model and the righthand panel displays the true edges in the underlying model.

a brief period roughly comprising timepoints 623 to 700 (7/29/95 to 10/14/95). These two regimes show a good deal of similarity in the associated graphical models, but also present some key differences. Figure 6 displays the graphical model associated with timepoint 40 (belonging to the first regime) and timepoint 540 (belonging to the second). The graphs shown are assembled from those edges that had a greater than 80% chance of inclusion for the respective observation, which is similar in spirit to displaying the top graphical model from a stochastic search, as performed by Carvalho et al. (2007).

In the first regime, an association structure broadly consistent with the graph used in Carvalho and West Carvalho & West (2007) is present. We see a tight grouping of the Euro adopters (not quite a clique, but missing only an ESP NLG connection), but with increased connection between the Swiss Franc and the Euro countries. A clique amongst the Asian currencies is connnected to only three of the Euro adopters. Furthermore the Pound is only connected to the Euro through DEM, the currency of the economic leader of this area. 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 (ERM), 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 US (as suggested by Carvalho et al Carvalho et al., 2007), leads to a separation of
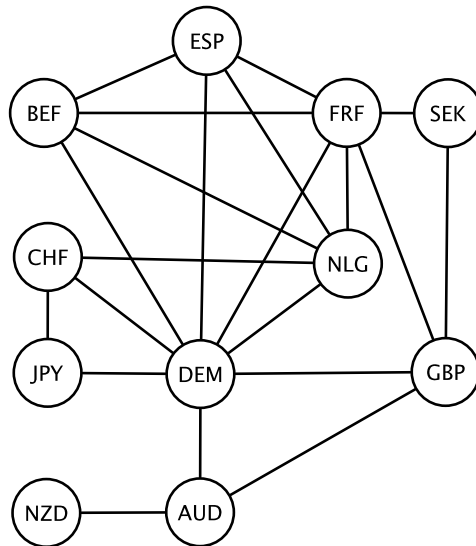
18

Figure 3: Graphical model presented by Carvalho et al Carvalho et al. (2007), 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.
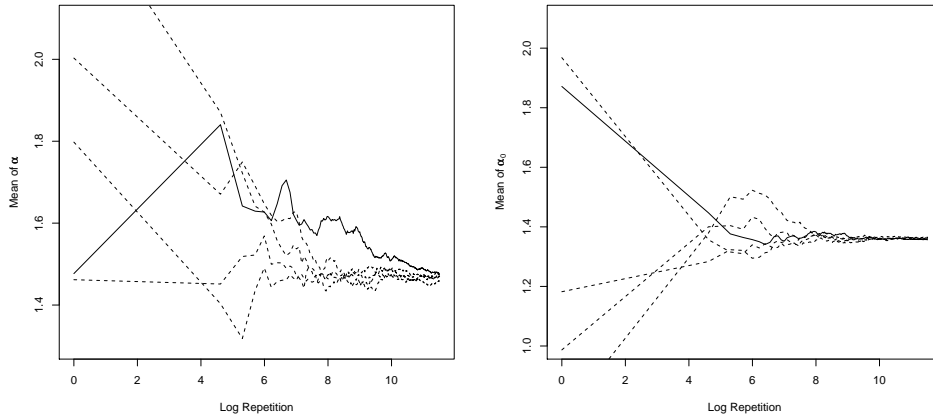
Figure 4: Convergence plot for $\alpha$ and $\alpha_0$ across chains by log iteration. This plot shows the running average of these two parameters across five separate instances of the GGM-iHMM algorithm. Their mutual agreement implies the settings used are sufficient to assure convergence.
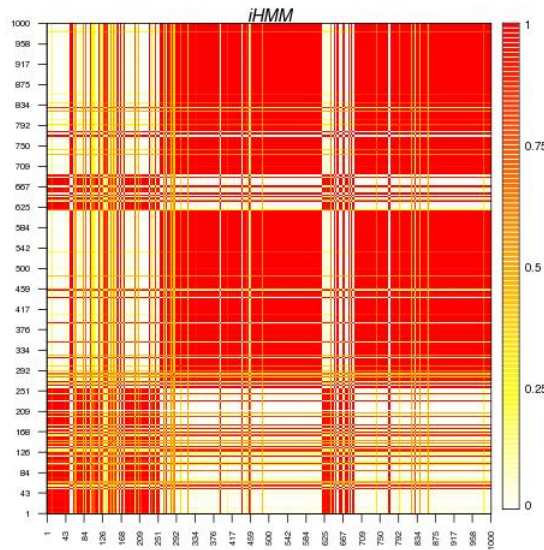


Figure 5: Heatmap displaying probability that two observations belong to the same cluster for the exchange example. The figure shows two dominant regimes, one that runs for the most part from timepoints 1 to 251 and again roughly between timepoints 623 and 700 and the other which is present most of the remaining time periods.
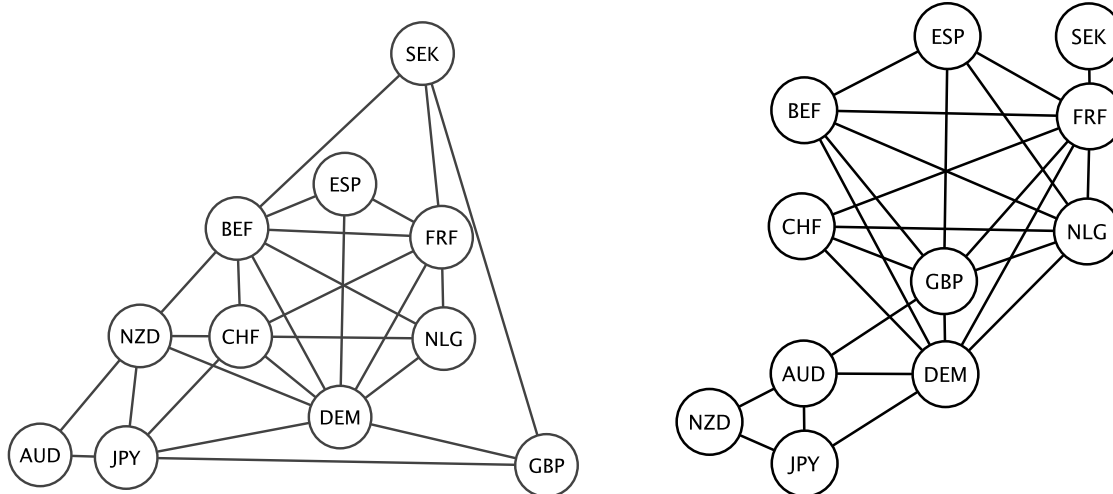
Figure 6: Graphical models associated with timepoint 40 (left) and timepoint 540 (right) in the exchange rate example. These graphs were constructed by adding any edge that had greater than 80% posterior inclusion probability for the respective timepoint. Such a high threshold was chosen in order to make a suitable comparison to the graph shown in Figure 3, which is the highest probability graph determined using stochastic search methods.

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 still is 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 CHF) and subsequently connections between ESP and DEM and BEF and CHF are no longer present. Furthermore, the Asian countries lose two neighbors, CHF and BEF, and the AUD is now the sole currency connecting the Asian countries to both the DEM and GBP, though the JPY maintains its association with DEM.

The greater connectedness of the GBP 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 ERM 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 ERM

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 the Pound would join 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.

Carvalho & West (2007) use the exchange rate data to show that minimum variance portfolios will yield better return when GGMs are employed to estimate covariances. We considered a similar analysis and used the expectation over the sampled values of $\mu_{T+1}$ and $K_{T+1}$ for each $T$ between 20 and 300 as the first two moments of the predictive distribution and calculated portfolio weights $w_T$ assuming a target return of $m = 0.1\%$ per day (see Carvalho & West (2007) for details regarding the construction of these portfolios). Figure 7 shows that the predictive distributions from the GGM-iHMM model give portfolio weights that have higher yields than both the iHMM only model and the GGM only model. 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 the GGMs into the model formulation, we are able to induce sparser estimates of covariation, which likewise improve predictive performance.

# 7 Discussion

Although this paper has focused on two relatively simple models (nonparametric mixtures of GGMs and infinite hidden Markov models with GGM 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 Rodriguez & Vuppala (2009) to include GGM kernels as a way to improve classification rates in high-dimensional problems. Also, in the spirit of Müller et al. (1996), Müller et al. (2004) and Rodriguez et al. (2009), sparse nonlinear regression models can be generated by using GGM 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 Dobra et al. (2008) on sparse regression to allow for adaptive local linear fits.

The implementation of GGM 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. This has allowed us to avoid the explicit repre-
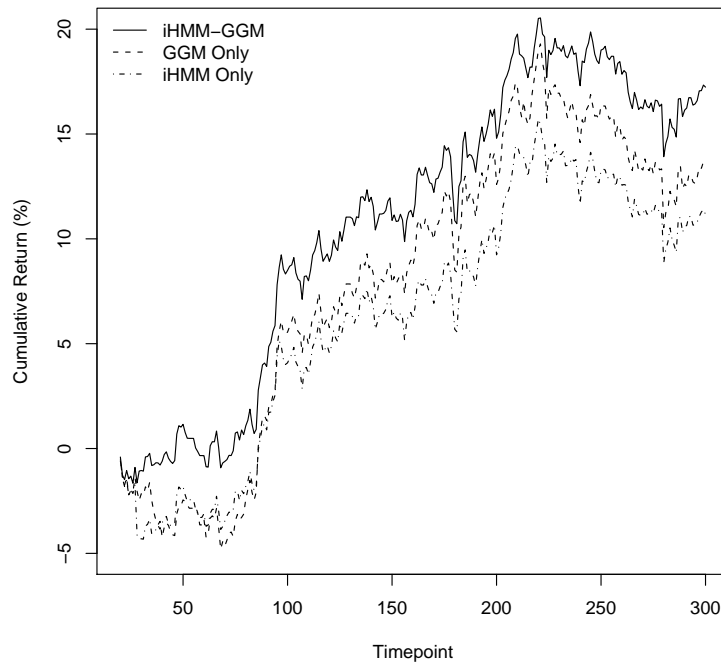
Figure 7: Cumulative returns from forming the optimal portfolio weights based on running the full GGM-iHMM model as well as the iHMM only and GGM only models, run over time periods 20 to 300. The final cumulative return was 17.2% for the GGM-iHMM, 13.5% for the GGM only and 11.2% for the iHMM only models.

sentation (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 Jain & Neal (2004) and Jain & Neal (2007).

The data analyses in this paper suggest that implementing mixtures of GGMs using Markov chain Monte Carlo algorithms is feasible for a moderate numbers of variables. However, many interesting applications of GGM mixtures (e.g., gene-expression data) involve outcomes in much higher dimensions, where previous experience suggest that random-walk MCMC algorithms will be inefficient. We are currently exploring deterministic search algorithms based on heuristics, such as the feature-inclusion (Berger & Molina, 2005; Scott & Carvalho, 2008), that might allow us to identify high-probability partitions and their associated graphs.

Our framework extends to include nondecomposable graphs in a straightforward manner. The only difference comes in the implementation of our sampling algorithms: for nondecomposable graphs the marginal likelihood (10) and the predictive distribution (11) must be numerically approximated instead of being calculated directly through formulas. We are currently working on this extension based on the recent developments from Lenkoski & Dobra (2010).

# Acknowledgments

# Appendix

We give a brief description of an auxiliary variable scheme for sampling from the posterior distributions of the single concentration parameter $\alpha_0$ in Section 4 and the two concentration parameters $\alpha$ and $\alpha_0$ from Section 5.2 – see Escobar & West (1995) and Teh et al. (2006) for full details. We assume that the priors for $\alpha$ and $\alpha_0$ are $\mathsf{G}(a, b)$ and $\mathsf{G}(a_0, b_0)$, respectively.

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

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

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 $\alpha$, $\varsigma_l | \alpha \sim \mathsf{Beta}(\alpha + 1, r_{l.})$ and $u_l | \alpha \sim \mathsf{Bernoulli}(r_{l.}/(\alpha + r_{l.}))$, where $r_{l.} = \sum_{l'=1}^{L} r_{ll'}$. Then, $\alpha$ is sampled from its full conditional distribution,

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

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

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

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

# References

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

ATAY-KAYIS, A. & MASSAM, H. (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92**, 317–35.

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*.

BEDFORD, T. & COOKE, R. M. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics* **30**, 1031–1068.

BERGER, J. O. & MOLINA, G. (2005). Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**, 3–15.

BLACKWELL, D. & MACQUEEN, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

CAPPÉ, O., MOULINES, E. & RYDEN, T. (2005). *Inference in Hidden Markov Models*. Springer.

CARVALHO, C. M., MASSAM, H. & WEST, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.

CARVALHO, C. M. & SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* To appear.

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

CASTELO, R. & ROVERATO, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.* **7**, 2621–2650.

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**, 157–75.

DIACONNIS, P. & YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–81.

DOBRA, A., EICHER, T. S. & LENKOSKI, A. (2008). Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models. Technical Report 55, Center for Statistics and the Social Sciences, University of Washington.

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.

DUNSON, D. B., PILLAI, N. & PARK, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B.* **69**, 163–183.

ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* **90**, 577–588.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2** , 615–629.

FRALEY, C. & RAFTERY, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* **24**, 155–181.

FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **6**, 799–805.

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)*.

GREEN, P. & RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**.

ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

ISHWARAN, H. & ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

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.

JAIN, S. & NEAL, R. M. (2007). Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis* **2**, 445–472.

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 University Press.

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.

LENKOSKI, A. & DOBRA, A. (2010). Computational aspects related to inference in Gaussian graphical models with the G-wishart prior. *Journal of Computational and Graphical Statistics* To appear.

LETAC, G. & MASSAM, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35**, 1278–323.

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *Journal of American Statistical Association* **100**, 1278–1291.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* **12**, 351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commnunications in Statistics, Part B - Simulation and Computation* **23**, 727–741.

McCloskey, J. W. (1965). *A Model for the Distribution of Individuals by Species in an Environment*. Ph.D. thesis, Michigan State University.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons.

Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P., Quintana, F. & Rosner, G. (2004). Hierarchical meta-analysis over related nonparametric Bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.

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

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

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Eds. T. S. Ferguson, L. S. Shapeley & J. B. MacQueen, pp. 245–268. Hayward, CA:IMS.

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

Rodriguez, A., Dunson, D. B. & Gelfand, A. E. (2008). The nested Dirichlet process, with discussion. *Journal of American Statistical Association* **103**, 1131–1144.

RODRIGUEZ, A., DUNSON, D. B. & GELFAND, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika* **96**, 149–162.

RODRIGUEZ, A. & VUPPALA, R. (2009). Probabilistic classification using bayesian nonparametric mixture models. Technical report, University of California, Santa Cruz.

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

SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plan. Infer.* **136**, 2144–2162.

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

SETHURAMAN, J. (1994). A constructive definition of Dirichelt priors. *Statistica Sinica* **4**, 639–650.

SUDDERTH, E. B. & JORDAN, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems 21*, Eds. D. Koller, D. Schuurmans, Y. Bengio & L. Bottou.

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.

THIESSON, B., MEEK, C., CHICKERING, D. M. & HECKERMAN, D. (1997). Learning mixtures of DAG models. In *Proc. of the Conf. on Uncertainty in AI*, pp. 504–513. Morgan Kaufmann, Inc.

WAINWRIGHT, M. J., RAVIKUMAR, P. & LAFFERTY, J. D. (2006). High-dimensional graphical model selection using 1-regularized logistic regression. In *In Neural Information Processing Systems*. MIT Press.

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

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

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.

WEST, M. & HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer - Verlag, New York, second edition edition.