Statistica Sinica **29** (2019), 165-184
doi:https://doi.org/10.5705/ss.202016.0224

HYPER MARKOV LAWS FOR CORRELATION MATRICES

Jeremy Gaskins

University of Louisville

Abstract: Parsimoniously modeling dependence in multivariate data is a challenging task, particularly if the dependence parameter is a correlation matrix due to modeling assumptions or identifiability constraints. In this work, we connect the techniques of graphical models and the hyper inverse Wishart distribution to introduce hyper Markov priors for correlation matrices. The priors are formed by taking a Markov combination of non-sparse correlation matrix distributions, where these distributions come from marginalizing the diagonal elements out of an inverse Wishart or Wishart prior. These priors produce a sparse correlation matrix with zero elements in its inverse when variables are conditionally independent. An MCMC scheme for posterior inference is introduced, and the performance is considered in the context of the Gaussian copula model using a simulation study and a financial data example.

Key words and phrases: Copula model, dependence modeling, Gaussian graphical model, hyper inverse Wishart, reversible jump MCMC, sparsity.

1. Introduction

The goal of this paper is to develop theory and implementation schemes for sparse Bayesian correlation estimation. To this end we borrow from the well-studied framework of covariance estimation in graphical models. Methodology for covariance matrices with an independence structure given by a graph has grown tremendously in the last twenty years. However, the case in which one desires a sparse correlation matrix, either due to an identifiability restriction or a modeling assumption, has failed to receive much attention.

We first review some key results for graphical models; see Lauritzen (1996) for full details. Let $G = (\mathcal{V}, E)$ denote an undirected graph with vertices $\mathcal{V} = \{1, \ldots, p\}$ and edge set $E \subset \mathcal{V} \times \mathcal{V}$. The elements of $\mathcal{V}$ represent the response variables measured for each observation, and variables j and k ($j \neq k$) are neighbors if they are connected by an edge $(j, k) \in E$. As the graph is undirected, $(j, k) \in E$ implies $(k, j) \in E$. A graph or subgraph is fully connected or complete if every pair of variables are neighbors. For $S, A, B \subset \mathcal{V}$, a set S is said

to separate A and B if every path from an element $a \in A$ to an element $b \in B$ contains at least one element in S; (A, B) is a decomposition of G if $\mathcal{V} = A \cup B$, the set $S = A \cap B$ is complete, and S separates A and B. A random variable X (or its distribution) is called Markov with respect to G if X_A is conditionally independent of X_B given $X_{A\cap B}$ for all decompositions (A, B) of G, where X_A is the subset of X corresponding to the set $A \subseteq \mathcal{V}$.

Throughout we assume G is decomposable, every cycle of length greater than or equal to four possess a chord (two non-consecutive vertices sharing an edge). We denote the set of decomposable graphs on $\mathcal{V}$ by $\mathcal{G}$. A decomposable graph can be represented by a perfect ordering of cliques, where each clique $C \in \mathcal{C}$ is a maximal complete subgraph of G. The history of the graph is defined to be $H_j = \bigcup_{k=1}^{j} C_j$ $(j = 1, \ldots, |\mathcal{C}|)$, and $S_j = C_j \cap H_{j-1}$ $(j = 2, \ldots, |\mathcal{C}|)$ is the (potentially empty) separator of clique j from the history. $\mathcal{S}$ is the collection of the $|\mathcal{C}| - 1$ separators, which generally are neither distinct nor non-empty. A key benefit of using decomposable graphs is that if the random variable X is Markov with respect to decomposable G, its density factors according to its cliques and separators:

$$p(X) = \frac{\prod_{C\in\mathcal{C}} p_C(X_C)}{\prod_{S\in\mathcal{S}} p_S(X_S)}, \tag{1.1}$$

where $p_A(\cdot)$ is the marginal distribution of X_A.

Dawid and Lauritzen (1993) develop the hyper inverse Wishart distribution (HIW) for the covariance matrix $\mathbf{\Sigma}$, which is the unique hyper Markov distribution with inverse Wishart as the clique marginals. They refer to Markov distributions for model parameters as hyper Markov distributions or laws. Let $\mathcal{M}_p$ be the space of $p \times p$ positive definite matrices. For a fixed (decomposable) graph G, the space $\mathcal{Q}(G) = \{\mathbf{\Sigma} \in \mathcal{M}_p : (\mathbf{\Sigma}^{-1})_{(j,k)} = 0 \text{ if } (j, k) \notin E\}$ is the support for HIW_G. The (mean-zero) Gaussian distributions that are Markov with respect to G can be represented by the Gaussian graphical model $\mathcal{N}(G) = \{\text{N}_p(0, \mathbf{\Sigma}) : \mathbf{\Sigma} \in \mathcal{Q}(G)\}$; HIW_G is the classical conjugate prior for $\mathbf{\Sigma}$ in $\mathcal{N}(G)$. Beyond the Markov property and the corresponding conditional independence structure, improved estimation efficiency from modeling $\mathbf{\Sigma}$ on lower-dimensional support $\mathcal{Q}(G)$ has been a motivating factor in the adaptation of the HIW prior. While we refer to such a covariance/correlation matrix as sparse, it is actually the inverse that has zero elements. Similarly, the hyper Wishart (HW) distribution is the Markov law with Wishart as the clique marginals and is the sampling distribution of a covariance matrix from $\mathcal{N}(G)$ (Dawid and Lauritzen (1993)).

There are many applications where the required dependence parameter is a

correlation matrix $\mathbf{R}$, not a covariance matrix. Often, this is due to model identifiability such as in the multivariate probit model (Chib and Greenberg (1998)), Gaussian copula regression (Pitt, Chan and Kohn (2006)), or in certain latent variable models (e.g., Daniels and Normand (2006)). Other times this may be a consequence of model specification. For instance, if data is made up of multiple groups, we might assume a common, potentially sparse correlation matrix with group-specific variances (Manly and Rayner (1987); Barnard, McCulloch and Meng (2000)).

Bayesian methodology for sparse correlation matrices is extremely limited. Pitt, Chan and Kohn (2006) devise a prior with zero elements in $\mathbf{R}^{-1}$, corresponding to a (not necessarily decomposable) graphical structure G. Their sampling scheme essentially requires evaluating the volume of the space of $p \times p$ correlation matrices with zero patterns consistent with G, which becomes impractical for moderate to large p. To make the volume calculation tractable, a relatively inflexible probability model is required for G.

Gaskins, Daniels and Marcus (2014) consider a prior for $\mathbf{R}$ that imposes sparsity through the partial auto-correlations (PACs; Daniels and Pourahmadi (2009)). The PACs are depend on the ordering of $\mathcal{V}$ and are inappropriate if the responses do not have an a priori ordering.

Talhouk, Doucet and Murphy (2012) apply graphical considerations for correlation estimation in the multivariate probit model. This is similar to the approach we take, but our work differs from theirs in a number of key ways. First, they utilize a parameter expansion sampler that is not applicable outside of the probit model. We develop a more general hyper Markov laws for $\mathbf{R}$, containing their proposal as a special case. Further, we will see that their sampling scheme corresponds to an incorrect stationary distribution due to a mistaken connection to the HIW distribution. We provide details later.

The remainder of the article proceeds as follows. In the next section we propose two families of Markov laws based on the marginal distributions of $\mathbf{R}$ from the Wishart and inverse Wishart distributions. Section 3 describes the sampling algorithm for MCMC analysis under known and unknown graph structure. In Section 4, we provide some details of the Gaussian graphical model to illuminate the simulation study and data application in Sections 5 and 6. This is followed by a few concluding comments.

2. Hyper Markov Laws for Correlation Matrices

2.1. Distributions for correlation matrix on a complete graph

Our interest is the correlation matrix $\mathbf{R}$, and we consider the implied distributions of (non-sparse) $\mathbf{R}$ under Wishart and inverse Wishart. Let $\mathcal{R}_p \subset \mathcal{M}_p$ denote the space of $p \times p$ positive definite matrices with unit diagonal, the space of correlation matrices. Using the separation strategy (Barnard, McCulloch and Meng (2000)), we write $\mathbf{\Sigma} \in \mathcal{M}_p$ as $\mathbf{DRD}$ where $\mathbf{R} \in \mathcal{R}_p$ is the correlation matrix corresponding to $\mathbf{\Sigma}$ and $\mathbf{D}$ is a diagonal matrix containing the standard deviations.

Let $\mathrm{IW}_p(\delta, \mathbf{\Psi})$ denote the inverse Wishart distribution with $\delta > 0$ and scale $\mathbf{\Psi} \in \mathcal{M}_p$, and $\mathrm{W}_p(\delta, \mathbf{V})$ is the Wishart distribution with $\delta > p-1$ and $\mathbf{V} \in \mathcal{M}_p$. We describe the marginal distributions of the correlation matrix from these distributions (e.g., Barnard, McCulloch and Meng (2000); Zhang, Boscardin and Belin (2006)). We denote these by CIW and CW, respectively. Let $\Gamma_p(x) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma(x + [1-j]/2)$ be the multivariate gamma function, $\mathbf{I}_p$ denote the $p \times p$ identity matrix, and $\mathbf{E}_j$ be the matrix formed by removing column j from $\mathbf{I}_p$. We use $|\cdot|$ to denote both the determinant of a matrix and the cardinality of a set, but the relevant interpretation will be clear from context.

Lemma 1 (CIW distribution)**.** *For $\mathbf{\Sigma} = \mathbf{DRD} \sim \mathrm{IW}_p(\delta, \mathbf{I}_p)$ for $\delta > 0$, the correlation matrix $\mathbf{R}$ has distribution*

$$p_{\mathrm{CIW}}(\mathbf{R}) = \xi_{\mathrm{CIW}}(\delta, p)\, |\mathbf{R}|^{(p+\delta)(p-1)/2-p} \prod_{j=1}^{p} |\mathbf{E}_j' \mathbf{R} \mathbf{E}_j|^{-(\delta+p-1)/2}, \quad \mathbf{R} \in \mathcal{R}_p, \tag{2.1}$$

where $\xi_{\mathrm{CIW}}(\delta, p) = \Gamma\{(\delta+p-1)/2\}^p / \Gamma_p\{(\delta+p-1)/2\}$ is the normalizing constant and $|\mathbf{E}_j' \mathbf{R} \mathbf{E}_j|$ is the (j,j) minor of $\mathbf{R}$. With (2.1), we write $\mathbf{R} \sim \mathrm{CIW}_p(\delta)$.

(i). For $\mathbf{\Sigma} \sim \mathrm{IW}_p(\delta, \mathbf{\Psi})$ with diagonal $\mathbf{\Psi} \in \mathcal{M}_p$, $\mathbf{R} \sim \mathrm{CIW}_p(\delta)$.

(ii). The CIW *distribution has a marginalization property: if $\mathbf{R} \sim \mathrm{CIW}_p(\delta)$, then $\mathbf{R}_A \sim \mathrm{CIW}_{|A|}(\delta)$ for $A \subseteq \mathcal{V}$.*

(iii). For $1 \le j < k \le p$, $p(r_{jk}) \propto (1 - r_{jk}^2)^{\delta/2-1}$. If $\delta = 2$, then the marginal correlation between any two responses is uniformly distributed on $(-1, 1)$.

(iv). The conditional distribution of the variance d_j^2, given the correlation matrix $\mathbf{R}$, is $\mathrm{InvGamma}((1/2)[\delta + p - 1], |\mathbf{E}_j' \mathbf{R} \mathbf{E}_j|/2|\mathbf{R}|)$.

Lemma 2 (CW distribution)**.** *For $\mathbf{S} = \mathbf{DRD} \sim \mathrm{W}_p(\delta, \mathbf{I}_p)$ for $\delta > p-1$, the*

correlation matrix $\mathbf{R}$ *has distribution*

$$p_{\mathrm{CW}}(\mathbf{R}) = \xi_{\mathrm{CW}}(\delta, p)\,|\mathbf{R}|^{(\delta-p-1)/2}, \qquad \mathbf{R} \in \mathcal{R}_p, \tag{2.2}$$

where $\xi_{\mathrm{CW}}(\delta, p) = \Gamma(\delta/2)^p/\Gamma_p(\delta/2)$ *is the normalizing constant.* $\mathbf{R}$ *with distribution* (2.2) *is denoted by* $\mathbf{R} \sim \mathrm{CW}_p(\delta)$.

(i). For $\mathbf{S} \sim \mathrm{W}_p(\delta, \mathbf{V})$ *with diagonal* $\mathbf{V} \in \mathcal{M}_p$, $\mathbf{R} \sim \mathrm{CW}_p(\delta)$.

(ii). If $\mathbf{R} \sim \mathrm{CW}_p(\delta)$, *then* $\mathbf{R}_A \sim \mathrm{CW}_{|A|}(\delta)$ *for* $A \subseteq \mathcal{V}$.

(iii). For $1 \le j < k \le p$, $p(r_{jk}) \propto (1 - r_{jk}^2)^{(\delta+1)/2-1}$.

(iv). The variances d_j^2 *are independent of* $\mathbf{R}$, *and* $d_j^2 \sim \mathrm{Gamma}(\delta/2, 2)$.

Partial proofs can be found in the Web Appendix. From *(i)* of both lemmas, it is unnecessary to consider alternative diagonal scale parameters since CIW and CW depends only on the shape δ. If $\mathbf{\Psi}$ ($\mathbf{V}$) is non-diagonal, a closed-form representation of the distribution of $\mathbf{R}$ does not exist. For the CIW distribution, the marginal distribution of r_{jk} in *(iii)* is a shifted version of the $\mathrm{Beta}(\delta/2, \delta/2)$ distribution (Daniels and Pourahmadi (2009); Gaskins, Daniels and Marcus (2014)), which can guide the choice of the δ shape parameter. Values of $\delta < 2$ produce a U-shaped distribution with mass near the extremes -1 and 1, and $\delta > 2$ gives a unimodal distribution centered at 0. As δ increases, $\mathbf{R}$ is shrunk more strongly toward the identity matrix. For the CW distribution, r_{jk} marginally follows a shifted $\mathrm{Beta}\{(1/2)(\delta+1), (1/2)(\delta+1)\}$ distribution. Since the Wishart distribution is constrained by $\delta > p-1$, it is not possible to obtain a uniform or U-shaped distribution for r_{jk}. As the shape parameter δ increases, the marginal distribution of r_{jk} becomes tightly concentrated around 0.

2.2. Markov priors for the correlation matrix

We seek sparse correlation matrices in the same way HIW_G and HW_G provide distributions on the sparse $\mathcal{Q}(G)$. Let $\mathcal{R}(G) = \mathcal{Q}(G) \cap \mathcal{R}_p$ denote the space of correlation matrices with zero pattern consistent with the graph G, i.e., $\{\mathbf{R} \in \mathcal{R}_p : (\mathbf{R}^{-1})_{(j,k)} = 0 \text{ if } (j,k) \notin E\}$. We call this the correlation selection problem. Clearly, if $\mathbf{R} \in \mathcal{R}(G)$ and $X \sim \mathrm{N}_p(0, \mathbf{R})$, X is Markov with respect to G. It is therefore natural to ask that the prior for $\mathbf{R}$ be hyper Markov to gain results from Dawid and Lauritzen (1993) that imply the posterior $\pi(\mathbf{R}|X)$ is a Markov law. Requiring the prior to be Markov is a stronger condition than merely having a zero pattern consistent with G, as the zero pattern only implies independence relationships in X. The Markov assumption additionally implies independence

relationships in the elements of $\mathbf{R}$, both in the prior and posterior. In particular, for a fixed decomposable graph G, the prior of Pitt, Chan and Kohn (2006) is not hyper Markov, even though $\mathbf{R}^{-1}$ has a zero pattern given by G.

We construct a Markov distribution through (1.1) by specifying distributions on the clique marginals that satisfy a consistency requirement (Dawid and Lauritzen (1993, Thm. 2.6)). A pair of distributions $p_A(\cdot)$ on X_A and $q_B(\cdot)$ on X_B are consistent if $p_{A\cap B}(x_{A\cap B}) = q_{A\cap B}(x_{A\cap B})$ for all x; clearly, the CIW and CW distributions are consistent families from the marginalization property *(ii)* of Lemmas 1 and 2. Thus, we can define a hyper Markov law on $\mathcal{Q}(G)$ through the Markov combination

$$\pi(\mathbf{R}|G) = \frac{\prod_{C\in\mathcal{C}} p_C(\mathbf{R}_C)}{\prod_{S\in\mathcal{S}} p_S(\mathbf{R}_S)}, \tag{2.3}$$

where $p_A(\mathbf{R}_A)$ is the density evaluated at the submatrix of $\mathbf{R}$ corresponding to the clique/separator A. When $p_A(\cdot)$ is $\mathrm{CIW}_{|A|}(\delta)$ for each A, we refer to distribution (2.3) as the hyper-CIW distribution, denoted by $\mathrm{HCIW}_G(\delta)$. Similarly, $\mathrm{HCW}_G(\delta)$ is the hyper law where $p_A(\mathbf{R}_A)$ is the density $\mathrm{CW}_{|A|}(\delta)$. To maintain the consistency requirement in both HCIW and HCW, δ is a common parameter across all clique/separator distributions and is referred to as the shape parameter for the full distribution. For $\mathrm{HCW}_G(\delta)$ we require $\delta > \max_{C\in\mathcal{C}}\{|C|-1\}$, so each marginal is well defined. By definition, $p_S(\cdot) = 1$ if the separator S is empty or a singleton as there are are no correlation parameters. Finally, the normalizing constant of $\pi(\mathbf{R}|G)$ is known in closed-form since the normalizing constants for the clique-marginals are known (Lemmas 1 and 2).

As $\pi(\mathbf{R}|G)$ is a function of only r_{jk} with $(j,k)\in E$, there is no contribution from $r_{j'k'}$ for $(j',k')\notin E$. However, the independence relationships from G require that $(\mathbf{R}^{-1})_{jk} = 0$ for $(j,k)\notin E$. Because G is decomposable there is a unique $\hat{\mathbf{R}}\in\mathcal{R}(G)$ such that $r_{jk} = \hat{r}_{jk}$ if $(j,k)\in E$ and $\{(\hat{\mathbf{R}})^{-1}\}_{(j,k)} = 0$ for $(j,k)\notin E$ (Letac and Massam (2007, Prop. 2.1)). Without confusion $\mathbf{R}$ from (2.3) is taken to be this unique completion matrix $\mathbf{R}\in\mathcal{R}(G)$. We summarize these results in a theorem that follows from Theorem 3.9 of Dawid and Lauritzen (1993).

Theorem 1 (Hyper Markov laws for correlation matrices)**.** *The* $\mathrm{HCIW}_G(\delta)$ *law, in* (2.3) *with* $p_A(\cdot)$ *the density* $\mathrm{CIW}_{|A|}(\delta)$*, is the unique hyper Markov law on* $\mathcal{R}(G)$ *with respect to* G *that has* $\mathrm{CIW}(\delta)$ *as the clique marginals. Likewise,* $\mathrm{HCW}_G(\delta)$*, given by* (2.3) *with* $p_A(\cdot)$ *the density* $\mathrm{CW}_{|A|}(\delta)$*, is the unique hyper Markov law on* $\mathcal{R}(G)$ *with respect to* G *that has* $\mathrm{CW}(\delta)$ *as the clique marginals.*

The use of the CIW and CW distributions in (2.3) is not arbitrary. Other common distribution choices for $p(\mathbf{R})$ cannot be used because they do not represent consistent families. The uniform prior $p(\mathbf{R}) \propto 1$ is not consistent as the marginal distributions for r_{jk} are not uniform, but highly concentrated near zero (Barnard, McCulloch and Meng (2000)). The Jeffreys' prior $p(\mathbf{R}) \propto |\mathbf{R}|^{-(p+1)/2}$ is improper, so the marginal distribution over a separating set is not a well-defined concept.

It is natural to consider the connection between our Markov laws on $\mathcal{R}(G)$ and the distribution of $\mathbf{R}$ from the standard Markov distributions on $\mathcal{Q}(G)$. Of main interest is whether HCIW is the marginalization of the hyper inverse Wishart distribution, and likewise for the HCW and hyper Wishart distributions.

Theorem 2 (Marginalization of hyper Markov laws on $\mathcal{Q}(G)$)**.**

(i) If $\mathbf{S} = \mathbf{DRD} \sim \mathrm{HW}_G(\delta, \mathbf{I}_p)$*, the marginal distribution of* $\mathbf{R}$ *is* $\mathrm{HCW}_G(\delta)$*.*

(ii) If $\mathbf{\Sigma} = \mathbf{DRD} \sim \mathrm{HIW}_G(\delta, \mathbf{I}_p)$*, the marginal distribution of* $\mathbf{R}$*,* $p_{\mathrm{HIW}}(\mathbf{R}) = \int p_{\mathrm{HIW}}(\mathbf{\Sigma})\, \mathrm{d}\mathbf{D}$*, is not the* HCIW *distribution. Furthermore,* $p_{\mathrm{HIW}}(\mathbf{R})$ *is not a hyper Markov law.*

Proof of this theorem and the form of $p_{\mathrm{HIW}}(\mathbf{R})$ are in the Web Appendix. Constructing the hyper law from Wishart marginals is equivalent to taking the marginal from the hyper law of Wisharts. This correspondence is unsurprising since $\mathbf{R}$ and $\mathbf{D}$ are independent for Wishart with diagonal $\mathbf{V}$. This does not carry over to the inverse Wishart case. Intuitively, parameters in different cliques are independent given the covariance parameters in their separating set but, after marginalizing out the standard deviations in the separator, independence is lost. (See Web Appendix for a detailed example.) While $p_{\mathrm{HIW}}(\mathbf{R})$ does have the desired support (the zero pattern in $\mathbf{R}^{-1}$ corresponds to G), it does not admit the Markov factorization (1.1), and the posterior distribution is not Markov.

The special case of $\mathrm{HCIW}_G(\delta)$ with $\delta = 2$ was previously considered in Talhouk, Doucet and Murphy (2012) in the context of a multivariate probit model. However, the authors incorrectly claim $\pi_{\mathrm{HCIW}}(\mathbf{R})$ as the marginal from $\mathrm{HIW}_G(2, \mathbf{I}_p)$ in contradiction to Theorem 2. By sampling with parameter expansion to HIW, their MCMC algorithm has as its stationary distribution the non-Markovian prior $p_{\mathrm{HIW}}(\mathbf{R})$, not the desired $\pi_{\mathrm{HCIW}}(\mathbf{R})$. In the next section we introduce an MCMC algorithm corresponding to the correct $\pi_{\mathrm{HCIW}}(\mathbf{R})$ distribution that can be used generally for any model with dependence defined by a correlation matrix.

Table 1. MCMC Algorithm to sample $\mathbf{R}$ from HCIW prior for fixed G.

For each clique $k = 1, \ldots, |\mathcal{C}|$:

1. For variable $j \in C_k$:
   Sample $d_j^2 | \mathbf{R}_{C_k} \sim \text{InvGamma}\{(1/2)[\delta + |C_k| - 1], |\mathbf{E}_j' \mathbf{R}_{C_k} \mathbf{E}_j|/2 | \mathbf{R}_{C_k}|\}$.
2. Sample $\mathbf{\Sigma}^\star_{C_k} = \mathbf{D}^\star \mathbf{R}^\star_{C_k} \mathbf{D}^\star \sim \text{IW}_{|C_k|}(\epsilon, (\epsilon - 2)\mathbf{D}\mathbf{R}_{C_k}\mathbf{D})$
   where $\mathbf{D} = \text{diag}\{d_1, \ldots, d_{|C_k|}\}$.
3. For all other cliques $l = 1, \ldots, |\mathcal{C}|$ $(l \neq k)$:
   Form the candidate $\mathbf{R}^\star_{C_l}$.
   If edge $(i_1, i_2) \in C_l \cap C_k$, let $[\mathbf{R}^\star_{C_l}]_{(i_1,i_2)} = [\mathbf{R}^\star_{C_k}]_{(i_1,i_2)}$.
   If edge $(i_1, i_2) \in C_l$ and $(i_1, i_2) \notin C_k$, let $[\mathbf{R}^\star_{C_l}]_{(i_1,i_2)} = [\mathbf{R}_{C_l}]_{(i_1,i_2)}$.
   Check if $\mathbf{R}^\star_{C_l}$ is positive definite. If not, immediately reject move.
4. Form the candidate $\mathbf{R}^\star$ by combining $\mathbf{R}^\star_1, \ldots, \mathbf{R}^\star_{|\mathcal{C}|}$ and obtaining the completion.
5. Accept the move from $\mathbf{R}$ to $\mathbf{R}^\star$ with probability given by equation (3.1).

3. MCMC Sampling

3.1. Sample R with fixed graph G

Most Bayesian methodology for correlation matrices, particularly those that allow sparse $\mathbf{R}$, is hindered by MCMC algorithms that require sampling each of the relevant parameters one at a time: either the marginal correlations (Barnard, McCulloch and Meng (2000)), the partial correlation (Pitt, Chan and Kohn (2006)), or the partial auto-correlation (Daniels and Pourahmadi (2009); Gaskins, Daniels and Marcus (2014)). When the correlation matrix is non-sparse, parameter expansion techniques provide a partial solution by introducing unidentifiable variance parameters to expand the parameter space to correspond to conjugate or well-known distributions (Liu (2001); Liu and Daniels (2006); Zhang, Boscardin and Belin (2006)). The only block sampler for sparse correlation matrices to our knowledge is the work of Talhouk, Doucet and Murphy (2012), but their parameter expansion algorithm for multivariate probit data has stationary distribution proportional to $p_{\text{HIW}}(\mathbf{R})$ not $\text{HCIW}_G(2)$ as desired. Most parameter expansion methods are specific to the multivariate probit model, but our goal is to propose a general sampling scheme for $\mathbf{R}$ that can be applied across a variety of modeling situations.

To sample $\mathbf{R}$ with a fixed graph G under the HCIW_G prior, we follow the algorithm summarized in Table 1. This block algorithm seeks to update the correlation matrix $\mathbf{R}_C$ associated with a clique $C \in \mathcal{C}$. Such a block sampler is generally more computationally efficient than one-at-a-time algorithms, as we perform a step for each of the $|\mathcal{C}|$ cliques which is generally much smaller than the $|E|$ steps needed for sampling r_{jk} individually.

To update the correlation matrix corresponding to clique k, we first sample variance parameters given the current $\mathbf{R}_{C_k}$ (Step 1) such that $\mathbf{\Sigma}_{C_k} = \mathbf{D}\mathbf{R}_{C_k}\mathbf{D} \sim \mathrm{IW}_{|C_k|}(\delta, \mathbf{I})$ in the prior (Lemma 1). Using an approach similar to that used in Zhang, Boscardin and Belin (2006), we draw the candidate $\mathbf{\Sigma}^\star_{C_k}$ from an inverse Wishart with mean $\mathbf{\Sigma}_{C_k}$. In this way, our Step 2 mimics a random walk, where ϵ is a tuning parameter with large values corresponding to small steps. The corresponding correlation matrix is the proposal $\mathbf{R}^\star_{C_k}$ for clique k.

We now need a $p \times p$ candidate correlation matrix $\mathbf{R}^\star$ based on the proposed $\mathbf{R}^\star_{C_k}$, that is, candidate correlation matrices for all cliques C_l $(l = 1, \ldots, |\mathcal{C}|, l \neq k)$ that are consistent with the proposed $\mathbf{R}^\star_{C_k}$. If a correlation r_{i_1,i_2} corresponds to an edge in both C_l and C_k, the candidate value is the one proposed in $\mathbf{R}^\star_{C_k}$. If edge (i_1, i_2) is not in C_k, we keep the current value of the correlation from $\mathbf{R}_{C_l}$. It is necessary to check that this updated $\mathbf{R}^\star_{C_l}$ is positive definite but, in our experience, this is almost always the case. If not, we reject the Metropolis-Hastings (MH) step. The full candidate correlation $\mathbf{R}^\star$ is found by combining all the $\mathbf{R}_C$'s and obtaining the completion using the algorithm of Carvalho, Massam and West (2007).

The move from $\mathbf{R}$ to $\mathbf{R}^\star$ is accepted according to the MH probability

$$\min\left\{1, \frac{\pi_{\mathrm{CIW}}(\mathbf{R}^\star)\, p(\mathbf{y}|\mathbf{R}^\star)}{\pi_{\mathrm{CIW}}(\mathbf{R})\, p(\mathbf{y}|\mathbf{R})} \frac{p(\mathbf{D}^\star|\mathbf{R}^\star_{C_k})}{p(\mathbf{D}|\mathbf{R}_{C_k})} \frac{p_{\mathrm{IW}}(\mathbf{\Sigma}_{C_k}|\mathbf{\Sigma}^\star_{C_k})\,|\mathbf{D}|^{|C_k|}}{p_{\mathrm{IW}}(\mathbf{\Sigma}^\star_{C_k}|\mathbf{\Sigma}_{C_k})\,|\mathbf{D}^\star|^{|C_k|}}\right\}. \tag{3.1}$$

We repeat this step for each clique in $\mathcal{C}$ in random order. Sampling for the HCW model is performed similarly by replacing inverse Gamma with $\mathrm{Gamma}(\delta/2, 2)$ in Step 1, replacing IW with Wishart in Step 2, and making the appropriate corrections in the probability (3.1).

3.2. Sampling with unknown graph G

Thus far, we have assumed the graph structure to be fixed and known, which is rarely the case in practice. Typically, we jointly model the graph G and the correlation matrix $\mathbf{R} \in \mathcal{R}(G)$ hierarchically through a prior $\pi(G)$ on $\mathcal{G}$, the space of decomposable graphs, and a prior for $\mathbf{R}$ conditional on G. In our case the prior $\pi(\mathbf{R}|G)$ is either $\mathrm{HCIW}_G(\delta)$ or $\mathrm{HCW}_G(\delta)$.

One common choice for $\pi(G)$ is the uniform distribution over $\mathcal{G}$. However, this prior places most of its weight on graphs with an intermediate number of edges. As a remedy the prior $\pi(G|\beta) \propto \beta^{|E|}(1-\beta)^{J-|E|} I(G \in \mathcal{G})$ has been proposed to encourage sparse graphs, where $J = p(p-1)/2$ is the number of edges in the complete graph. The choice $\beta = 2/(p-1)$ was suggested by Dobra et al.

Table 2. RJMCMC Algorithm to sample G.

1. Sample the pair (j,k) uniformly over all possible edges.
2. Create candidate graph $G^\star$: $e^\star_{jk} = 1 - e_{jk}$, $e^\star_{j'k'} = e_{j'k'}$ for all other (j',k').
3. Check that $G^\star$ is decomposable. If not, immediately reject move.
4. Create candidate $\mathbf{R}^\star$:
4a. If $e_{jk} = 0$ and $e^\star_{jk} = 1$:
Sample $r^\star_{jk} \sim \mathrm{N}(r_{jk}, \sigma^2)$ and set $u = r^\star_{jk}$.
Form the $\mathbf{R}^\star$ by taking $r^\star_{j'k'} = r_{j'k'}$ for $(j',k') \in E$, obtain completion, and check positive definiteness. If not positive definite, immediately reject move.
4b. If $e_{jk} = 1$ and $e^\star_{jk} = 0$:
Set $u = r_{jk}$.
Form the $\mathbf{R}^\star$ by taking $r^\star_{j'k'} = r_{j'k'}$ for $(j',k') \in E^\star$, obtain completion, and check positive definiteness. If not positive definite, immediately reject move.
5. Accept the move from $\mathbf{R}$ to $\mathbf{R}^\star$ with probability given by equation (3.2).

(2004), so that graphs with p edges are a priori most likely. Carvalho and Scott (2009) consider a hierarchical structure by letting $\beta \sim \mathrm{Beta}(a,b)$. Marginally, $\pi(G) \propto B(|E|+a, J-|E|+b)I(G \in \mathcal{G})$ with $B(\cdot,\cdot)$ the beta function. In our empirical work, we use this prior with $a = b = 1$.

When the dependence parameter is a covariance matrix with $\mathrm{HIW}_G(\delta, \mathbf{\Psi})$ prior, estimation of the graph is simplified by the fact that the covariance matrix can be integrated out. The probability of graph G given data $\mathbf{y}$ (marginally over the covariance matrix) can be written as a ratio of HIW normalizing constants, and a Metropolis-Hastings step is used to traverse the G-space (Carvalho and Scott (2009)). As we noted in Section 2, we are unable to marginalize $\mathbf{D}$ out of HIW with a non-diagonal scale parameter, so a conjugate step of this sort is not available to sample the posterior of the correlation selection problem.

When sampling requires traversing models with differing numbers of parameters and marginalization is unavailable, reversible jump MCMC (RJMCMC; Green (1995)) often provides an accessible remedy. We propose a RJMCMC algorithm to update the graph G in Table 2.

We propose the candidate graph $G^\star$ by uniformly choosing (j,k) over all possible edges (Step 1). For notational convenience we introduce the variables e_{jk} $(j < k)$ where $e_{jk} = 1$ if $(j,k) \in E$ and $e_{jk} = 0$ otherwise. If $e_{jk} = 0$, then we add the edge to form $G^\star$, and if the edge is in the current graph, we remove it. G and $G^\star$ differ only by a single edge, and it is necessary to check that the proposed $G^\star$ is decomposable (Steps 2 and 3).

To accept the new graph $G^\star$, we must simultaneously propose a new correla-

tion matrix $\mathbf{R}^\star$ since the support $\mathcal{R}(G^\star)$ has changed. As G and $G^\star$ differ only in the edge e_{jk}, $\mathbf{R}$ and $\mathbf{R}^\star$ differ by only one unconstrained parameter r_{jk}. When we propose to add the edge (j,k) (Step 4a), we draw the candidate $r^\star_{jk} \sim \mathrm{N}(r_{jk}, \sigma^2)$ depending on the current constrained value r_{jk} and a tuning variance σ^2. The remaining $r^\star_{j'k'}$ are set to their values in $\mathbf{R}$. Using these $r^\star_{j'k'}$s, we form the candidate $\mathbf{R}^\star$ from the completion. If we propose to remove the edge (Step 4b), $r^\star_{jk}$ becomes a constrained parameter, and its value is determined when taking the completion $\mathbf{R}^\star$. RJMCMC requires a dimension-matching parameter to maintain the detailed balance condition; the parameter u, the correlation between (j,k) in the larger model, plays this role.

The move from $(G, \mathbf{R})$ to $(G^\star, \mathbf{R}^\star)$ is accepted with probability

$$\min\left\{1, \frac{\pi(G^\star)\,\pi_{\mathrm{HCIW}}(\mathbf{R}^\star|G^\star)\,p(\mathbf{y}|\mathbf{R}^\star)}{\pi(G)\,\pi_{\mathrm{HCIW}}(\mathbf{R}|G)\,p(\mathbf{y}|\mathbf{R})}\,\frac{p(u|r^\star_{jk})^{1-e^\star_{jk}}}{p(u|r_{jk})^{1-e_{jk}}}\right\}, \tag{3.2}$$

where $p(u|r_{jk})$ is the $\mathrm{N}(r_{jk}, \sigma^2)$ density evaluated at u.

Our general sampling strategy is to alternate between updating the graph G (Table 2), updating $\mathbf{R}$ through each clique C (Table 1), and updating all other model parameters. The graph update step is often accepted infrequently relative to the correlation step, so we typically perform multiple graph steps per iteration. We treat the number of edge proposals per iteration as an MCMC tuning parameter to be optimized along with ϵ and σ^2. Under the HCW prior, sampling is the same with the obvious adjustment to (3.2).

4. Gaussian Copula Model

Our focus is multivariate modeling when the appropriate dependence structure is constrained to be a correlation matrix. In some scenarios such as the probit model, parameter expansion is possible, and simpler, specialized algorithms may be available. However, our goal is methodology that can be used across a variety of modeling schemes, particularly those that do not lead to computational simplification. For illustration we consider the case of the Gaussian copula model.

Briefly, the Gaussian copula model (e.g., Song (2000)) provides a joint distribution for $\mathbf{Y}$ allowing separate specification of the marginal distributions. The model can be defined by the following construction. First, correlated normal scores $\boldsymbol{\epsilon}$ are drawn from $\mathrm{N}_p(\mathbf{0}, \mathbf{R})$, where for identifiability $\mathbf{R}$ is constrained to be a correlation matrix. The jth component Y_j is a function of ϵ_j given by $Y_j = F_j^{-1}(\Phi(\epsilon_j))$, where $\Phi(\cdot)$ is the cumulative distribution for a standard nor-

mal and $F_j^{-1}(\cdot)$ is the inverse of the distribution of the jth margin. The marginal distribution of Y_j is $F_j(\cdot)$ depending on parameters θ_j, and the Y_js are correlated because they are functions of the correlated normal scores $\boldsymbol{\epsilon}$. The joint density of $\mathbf{Y}$ (assuming each margin is absolutely continuous) is given by

$$f(\mathbf{y}|\mathbf{R},\boldsymbol{\theta}) = |\mathbf{R}|^{-1/2} \exp\left\{\frac{1}{2}\boldsymbol{\epsilon}(\mathbf{y},\boldsymbol{\theta})'(\mathbf{I}-\mathbf{R}^{-1})\boldsymbol{\epsilon}(\mathbf{y},\boldsymbol{\theta})\right\} \prod_{j=1}^{p} f_j(y_j|\theta_j), \qquad (4.1)$$

where $f_j(\cdot|\theta_j) = F_j'(\cdot)$ is the jth marginal density and $\epsilon_j(\mathbf{y}) = \Phi^{-1}\{F_j(y_j)\}$.

In the next two sections we apply this copula model to data analysis using the MCMC scheme proposed in Section 3. Unlike the multivariate probit case, there are no general parameter expansion algorithms to sample (sparse or non-sparse) $\mathbf{R}$ in this model. There are a two exceptions in the literature: when $\mathbf{R}$ is the only parameter of interest and the marginal distributions are a nuisance parameter (Hoff (2007)) or when all Y_js are discrete (Dobra and Lenkoski (2011)). In both cases, the key is that the relationship between the normal scores ϵ and the data Y is many-to-one, but this approach is not available when the marginal distributions are of interest and/or continuous. However, we can apply the algorithms introduced in Tables 1 and 2 using the density (4.1) for $p(\mathbf{y}|\mathbf{R})$.

5. Simulation Study

To evaluate the empirical performance of our correlation hyper laws, we performed a simulation study using the Gaussian copula model with dimension $p = 25$. Due to space constraints, we briefly comment on the results. Full details are available in the Web Appendix.

The distribution for Y_{ij}, the jth component of $\mathbf{Y}_i$ $(i = 1,\ldots,N)$, is a t-distribution with location $\mu_j = j$, scale $\sigma_j = 1$, and degrees of freedom $\nu_j = 5$ $(j = 1,\ldots,p)$. Figure 1 shows the three choices of the true graph structure $\tilde{G}$. The graphs have decreasing sparsity (increasing complexity) with $|E| = 24, 32, 90$ edges out of a possible $J = p(p-1)/2 = 300$. For $\mathbf{R}_A$, the correlation matrix corresponding to graph $\tilde{G}_A$, connected variables have marginal correlation 0.7, producing an autoregressive structure. For $\mathbf{R}_B$ the correlations corresponding to the six edges connected to the central node have value 0.8, and all others are set to 0.5. In $\mathbf{R}_C$ the correlations corresponding to all edges are set to 0.7. The remaining elements of each $\mathbf{R}$ are determined by the completions.

For each $\mathbf{R}$ we generated 100 data sets containing $N = 100$ observations and 100 data sets with $N = 500$. We chose relatively uninformative priors for μ_j, σ_j, ν_j. For the prior of the correlation matrix conditional on the ran-

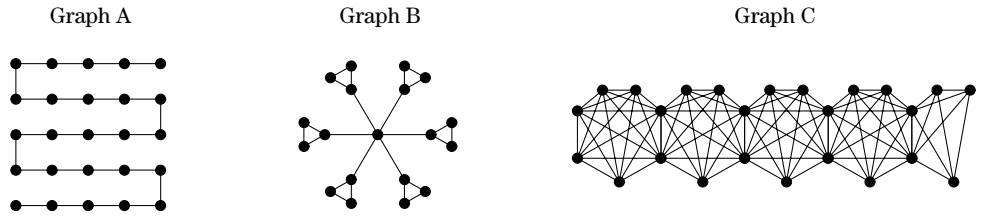

Figure 1. The true graph structure for each of our three choices $\tilde{G}_A$, $\tilde{G}_B$, and $\tilde{G}_C$.

dom graph, we used $\text{HCIW}_G(2)$, $\text{HCIW}_G(1)$, $\text{HCW}_G(25)$, and $\text{HCW}_G(10)$. For HCIW, $\delta = 2$ produces a uniform distribution for r_{jk} with $(j,k) \in E$, and $\delta = 1$ favors values toward -1 and 1. The HCW prior requires the shape parameter to be larger than one less the maximum clique size, and $\delta = p = 25$ provides a default choice when the graph is unknown; we included $\delta = 10$ to yield less shrinkage toward $\mathbf{I}_p$. For the $\text{HCW}_G(10)$ choice, we modified the prior on $\pi(G)$ to place no probability on graphs containing cliques with $|C| > 10$. As a comparison we also considered an analysis with dense correlation matrix by using the flat prior $\pi(\mathbf{R}) \propto I(\mathbf{R} \in \mathcal{R})$, as well as assuming independence by fixing $\mathbf{R} = \mathbf{I}_p$. Details regarding the choice of tuning parameters, MCMC specifications, and computational time can be found in the appendix.

To compare methods we estimated the risk (average loss) for four quantities of interest: the location parameters, scale parameters, correlation matrix, and graph structure. We used sum of squared error for the locations μ and the scales σ. For the correlation matrix, we employed the log-likelihood loss function $L(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}\{\hat{\mathbf{R}}\mathbf{R}^{-1}\} - \log|\hat{\mathbf{R}}\mathbf{R}^{-1}| - p$. To evaluate the accuracy of graph recovery, we considered the total number of errors, the sum of false positives (edges included in G not in the true graph $\tilde{G}$) and false negatives (edges excluded from G that are in $\tilde{G}$), averaged across iterations in the posterior sample.

Full details are available in the Web Appendix (see Table S.1 and Figures S.1 and S.2), but we provide a few general comments here. There is very little difference in the estimation of the location parameters across the six different choices for $\mathbf{R}$; for the larger sample size the $\mathbf{R} = \mathbf{I}$ choice is slightly worse. For the scale parameters, HCIW(2) and HCIW(1) produce lower risk, and the flat prior and HCW(25) are the worst performers. When estimating $\mathbf{R}$, failing to allow sparsity is a significant disadvantage. The flat prior leads to the highest risk, followed by HCW(25) and then HCW(10).

Looking to graph estimation, the HCIW priors are able recover the graph well, and the $\text{HCW}_G(\delta)$ performs somewhat less favorably. This is especially true

for HCW(25), which tends to have many more false positives than its competitors. This appears to be related to the large amount of shrinkage of r_{jk} associated with large δ, as (correctly) excluding an edge may look similar to (incorrectly) including it with r_{jk} near zero. Using $\delta = 10$ partly corrects this over-shrinkage but requires restricting the support of G to graphs with maximal clique size less than 10. We find estimating the more complex $\tilde{G}_C$ to be the most challenging of the three graphs, especially for the smaller sample size. When $N = 100$, we consistently underestimate the graph (around 6 false positives to 35 false negatives) across all methods, and the average number of edges in the estimated graph is 61 compared to 90 in the true graph. But in cases such as graph C where we have a dense graph relative to a small sample, we prefer to err on the side of overly sparse model (and the prior on G also encourages this); with $N = 500$ observations, the graph is well estimated. Overall, we determine that $\text{HCIW}_G(\delta)$ is the ideal hyper law to use in situations requiring sparse correlation estimation with $\delta = 1$ and $\delta = 2$ performing similarly. In the copula context, using a sparse correlation matrix also leads to more efficient estimation the parameters of the marginal distributions.

6. Capital Asset Pricing Model Application

6.1. Data and model specification

The Capital Asset Pricing Model (CAPM) is widely used in finance to model the expected excess return for a particular asset from the excess return of the full market. For an introduction to the CAPM, see Fama and French (2004). Gibbons (1982) provided a multivariate extension allowing multiple assets to be jointly modeled. An additional consideration is that the normality assumption underlying much of the CAPM theory is known to perform poorly in practice due to the heavy tails exhibited in many financial data. Pitt, Chan and Kohn (2006) consider a Gaussian copula with unique t-distributions as the marginal distributions to jointly model the returns of a number of financial sectors. We employ a similar model to demonstrate the use of our HCIW priors.

We applied the CAPM in a copula framework to data obtained from Kenneth French's data library website. The data we used considers monthly percentage excess returns from January 1950 through December 1999 across 30 industry profiles. The response y_{ij} $(i = 1, \ldots, 600; j = 1, \ldots, 30)$ is the excess return for industry j at time i, the difference between the return and the risk-free market return (U.S. Treasury bills are used as a proxy for risk-free return). Marginally,

y_{ij} is assumed to follow a t-distribution with ν_j degrees of freedom, location/mean parameter of $\beta_j z_i$, and scale parameter σ_j. z_i is the excess market return at time i, the difference between the market return and the risk-free return, and the parameter β_j represents how sensitive the returns for industry j are to variability in the market. As in the simulation study, we applied a Gaussian copula to introduce dependence across the marginal t-distributions.

This model has a couple of important advantages relative to competitors. By using a copula with t marginals, we accommodate the heavy-tailed nature of the data without sacrificing the ability to jointly model the industries. Also, by using a copula versus a multivariate t-distribution, we have separate degrees of freedom for each industry, allowing some industries to be more likely than others to exhibit extreme departures from the mean. By incorporating a sparse structure for $\mathbf{R}^{-1}$, $e_{jk} = 0$ implies condition independence of ϵ_{ij} and ϵ_{ik}, and consequently, the transformations Y_{ij} and Y_{ik}.

We compared a number of models for the data, considering combinations of two choices for the joint distribution and three for the dependence model. We fit the data using the copula model, as well as a multivariate normal model (MVN) with $\mathrm{E}(Y_{ij}) = z_i\beta_j$ and $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{\Sigma}$. For each, we used an assumption of independence ($\mathbf{R}$ is diagonal unit matrix for the copula model; $\mathbf{\Sigma}$ is diagonal for MVN), an assumption of a complete graph ($\pi(\mathbf{R}) \propto I(\mathbf{R} \in \mathcal{R})$ for the copula; $\mathbf{\Sigma} \sim \mathrm{IW}_p(2, \mathbf{I}_p)$ for MVN), and an assumption of a sparse dependence parameter. For the sparse-MVN model, we used $\mathbf{\Sigma} \sim \pi(G)\mathrm{HIW}_G(2, \mathbf{I}_p)$, and for the sparse-copula model, we considered three prior choices for $\mathbf{R}$: $\pi(G)\mathrm{HCIW}_G(2)$, $\pi(G)\mathrm{HCIW}_G(1)$, and $\pi^\star(G)\mathrm{HCW}_G(10)$, where $\pi^\star(G)$ is the restriction of the prior to decomposable graphs with a maximum clique size of 10. We used $\beta_j \sim \mathrm{N}(0, 10^2)$, $\sigma_j^2 \sim \mathrm{InvGamma}(0.1, 0.1)$, and $\nu_j \sim \mathrm{Unif}(2{,}30)$ as (relatively uninformative) prior distributions.

With the increase in p from 25 to 30, the number of potential edges in the graph increases from 300 to $J = 435$. This exacerbates the difficulty in graph selection, and the algorithm from Section 3.2 that uniformly selects candidate edges struggles to mix well. To that end, we replace Step 1 in Table 2 with an adaptive sampler that proposes edges relative to the uncertainty of its inclusion in G. We provide details about this adaptive algorithm in the Web Appendix. This adaptive strategy is used for both the copula and MVN sparse models.

6.2. Modeling results

To compare the model choices we used the deviance information criteria

Table 3. Model comparison statistics for finance data, ordered by increasing prediction error. Prediction accuracy is the sum of log-scores (log predictive density) over the $N_{\text{test}} = 72$ prediction months.

Model Specification		Prediction Accuracy	Model Fit Statistics		
Joint Dist.	Dependence	Log-Score	*Dev*	p_D	DIC
Copula	Sparse - $\text{HCIW}_G(2)$	−6,738	88,300	354	89,008
Copula	Sparse - $\text{HCIW}_G(1)$	−6,744	88,305	350	89,005
Copula	Complete Graph	−6,767	87,973	523	89,019
Copula	Sparse - $\text{HCW}_G(10)$	−6,794	88,526	322	89,169
Copula	Independence	−6,968	92,753	78	92,910
MVN	Complete Graph	−7,153	88,829	480	89,790
MVN	Sparse	−7,248	89,952	183	90,317
MVN	Independence	−7,369	93,637	60	93,758

(DIC; Spiegelhalter et al. (2002)) applied to the fitted data and a measure of out of sample predictive accuracy using data from the next five years (January 2000 to December 2005; $N_{\text{test}} = 72$). At each i in the test set, we measured the log-score of $\tilde{\mathbf{y}}_i$ by the log predictive density $\log f(\tilde{\mathbf{y}}_i)$ averaged over the posterior sample (Gneiting and Raftery (2007); Zhou et al. (2015)). We let the predictive accuracy be the sum over the log-scores for the 72 months in the test data, with larger (less negative) values indicating better prediction. The DIC is the sum of the model deviance *Dev* at the posterior means of the parameters and twice p_D which measures model complexity; smaller DIC are favored. Table 3 contains DIC statistics and the predictive accuracy for each model.

Of the models considered the copula model performs best in terms of both fit to the modeled data and out-of-sample prediction. Applying the copula model with $\text{HCIW}_G(2)$ has the best predictive performance, although $\text{HCIW}_G(1)$ has a slightly lower DIC. Ultimately, there is little practical difference between using $\delta = 1$ or $\delta = 2$. Assuming the independence across industries is clearly invalid, whereas using a full correlation matrix is somewhat competitive. None of the multivariate normal models perform well.

It is initially surprising that the sparse MVN model is outperformed by the MVN with complete graph. Here, the MVN model tends to favor highly sparse covariances matrices; the posterior mean of $|E|$ is 55.0 with a 95% credible interval of (52, 58) out of a potential $J = 435$. Conversely, the $\text{HCIW}_G(2)$ copula model favors graphs with an average of 167 edges (150, 187). While it is perhaps counter-intuitive that the posterior graphs would differ so greatly, the copula model looks for correlations in $\epsilon_{ij} = \Phi^{-1}(F_{ij}(Y_{ij}))$ which do not necessarily

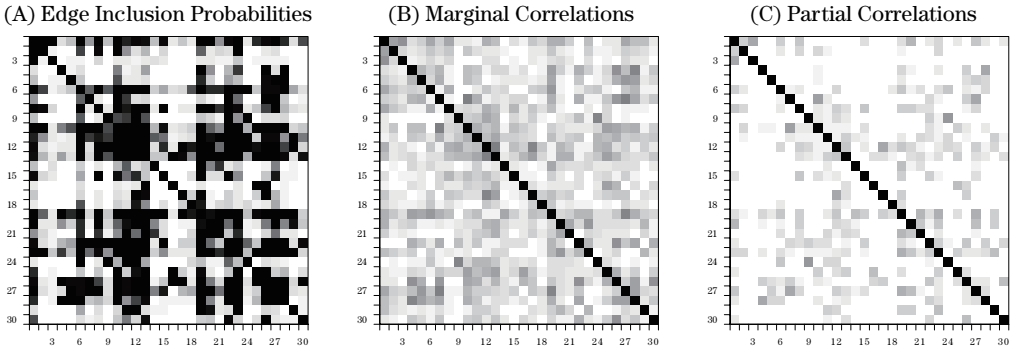

Figure 2. From the copula-HCIW(2) model, heat maps of (a) the posterior edge inclusion probability $P\{(j,k) \in E \,|\, \mathbf{y}\}$, (b) the absolute value of the posterior marginal correlation r_{jk}, and (c) the absolute value of the posterior partial correlations, $\mathbf{R}^{-1}$ rescaled to unit diagonal.

match the correlation in the untransformed Y_{ij}s.

We now explore the estimated dependence and correlation matrix under the best-predicting model: the Gaussian copula with $\text{HCIW}_G(2)$ correlation prior. See the Web Appendix for conclusions regarding the marginal distribution parameters. Figure 2 contains heat maps of the posterior edge inclusion probabilities, the marginal correlations, and the partial correlations. It is clear from Figure 2(a) that there are a number of industries that exhibit large connectivity to other industries. For example, textiles (10), steel (12), fabricated products (13), oil (19), business equipment (23), and wholesale (26) each average more than 20 edges. Sectors with fewer than 5 edges on average (tobacco products (3), electrical equipment (14), aircraft, ships, and railroad equipment (16), coal (18)) are less dependent on other industries. Figure 2(c) contains the partial correlations which respect the zero pattern from the graph G showing that we are able to use relatively few parameters to describe the dependence across industries.

7. Discussion

As one reviewer noted, a sparse prior for $\boldsymbol{\Sigma}$ such as HIW will automatically induce a sparse distribution on the correlation matrix, so we wish to further clarify why our approach of directly specifying a distribution for $\mathbf{R}$ is preferable. First, using an induced distribution may obscure the properties of $\mathbf{R}$, whereas we can easily interpret both the clique-marginal distributions and the marginals for the non-zero r_{ij}. As we show in Theorem 2, properties of the induced marginal distributions do not necessarily follow from the covariance prior.

Most importantly, we seek to develop an "off the shelf" method that can be applied across many situations constrained by $\mathbf{R}$. Specialized parameter expansion algorithms with specific choices for $\pi(\mathbf{R})$ have been considered by Talhouk, Doucet and Murphy (2012) for probit regression, and by Hoff (2007) and Dobra and Lenkoski (2011) for copula models with only discrete outcomes, but these approaches cannot be applied in other contexts. Developing a parameter expansion procedure specific for each new problem is not generally easy, and the usual intuition is not always a trusty guide (as shown by the discrepancy in the sampler of Talhouk, Doucet and Murphy (2012)). While we demonstrate our methodology in the Gaussian copula model, our approach and algorithms are general and do not require a particular choice for the data likelihood $p(\mathbf{y}|\mathbf{R})$; see equations (3.1) and (3.2).

One of the most important difficulties that remains is computational. Due to the inability to marginalize over $\mathbf{R}$, the MCMC algorithm we implement is of the reversible jump type. A number of authors have questioned the ability of RJ-MCMC to deal with problems of moderate dimension (e.g., Scott and Carvalho (2008)). While diagnostic checks indicate adequate computational performance in our examples, our methodology may struggle in terms of computational time and mixing as p continues to grow. Further work improving the proposed algorithm and/or developing new prior distributions that yield faster algorithms is needed. The introduction of an adaptive step for proposing graph edges was effective in improving mixing in the data example, and there may be potential for additional improvements using other techniques such as tempering, parallel chains, and/or junction trees in the edge proposal distribution (Green and Thomas (2013)). Despite these challenges our HCIW priors are shown to have good performance in the situation of small-to-moderate p whereas previous Bayesian attempts have tended to consider cases with smaller dimension ($p \leq 12$; Barnard, McCulloch and Meng (2000); Pitt, Chan and Kohn (2006); Talhouk, Doucet and Murphy (2012); Gaskins, Daniels and Marcus (2014)).

By using the graphical model framework, our HCIW priors are only defined for decomposable models. However, as most quantities of interest involve averaging over G in addition to $\mathbf{R}$, our methodology is a form of Bayesian model averaging. Giudici and Green (1999) showed the average over decomposable graphs can reliably estimate a covariance matrix whose true graph is non-decomposable, and Fitch, Jones and Massam (2014) demonstrated that the top decomposable models are comparable, and sometimes superior, to the top non-decomposable models. Hence, our restriction to decomposable G is not a significant drawback.

Supplementary Materials

An online Web Appendix contains proofs for the results from Section 2 and additional computational details from the simulation study and data application.

Acknowledgment

The author thanks Dr. Mike Daniels (University of Florida) and Dr. Matt Souther (University of Missouri) for their helpful comments and discussion.

References

Barnard, J., McCulloch, R. and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

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

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

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

Daniels, M. J. and Normand, S.-L. (2006). Longitudinal profiling of health care units based on mixed multivariate patient outcomes. *Biostatistics* **7**, 1–15.

Daniels, M. J. and Pourahmadi, M. (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis* **100**, 2352–2363.

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

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

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics* **5**, 969–993.

Fama, E. F. and French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives* **18**, 25–46.

Fitch, A. M., Jones, M. B. and Massam, H. (2014) The performance of covariance selection methods that consider decomposable models only *Bayesian Analysis*, **9**, 659–684.

Gaskins, J. T., Daniels, M. J. and Marcus, B. H. (2014). Sparsity inducing prior distributions for correlation matrices of longitudinal data. *Journal of Computational and Graphical Statistics* **23**, 966–984.

Gibbons, M. R. (1982). Multivariate tests of financial models: A new approach. *Journal of Financial Economics* **10**, 3–27.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika* **100**, 91–110.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* **1**, 265–283.

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

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *The Annals of Statistics* **35**, 1278–1323.

Liu, C. (2001). Comment on "The art of data augmentation" by D. A. van Dyk and X.-L. Meng. *Journal of Computational and Graphical Statistics* **10**, 75–81.

Liu, X. and Daniels, M. J. (2006). A new algorithm for simulating a correlation matrix based on parameter expansion and re-parametrization. *Journal of Computational and Graphical Statistics* **15**, 897–914.

Manly, B. F. J. and Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika* **74**, 841–847.

Pitt, M., Chan, D. and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554.

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

Song, P. X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27**, 305–320.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **64**, 583–639.

Talhouk, A., Doucet, A. and Murphy, K. (2012). Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *Journal of Computational and Graphical Statistics* **21**, 739–757.

Zhang, X., Boscardin, W. J. and Belin, T. R. (2006). Sampling correlation matrices in Bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics* **15**, 880–896.

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G. and Micheas, A. C. (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association* **110**, 9–15.

Department of Bioinformatics and Biostatistics, University of Louisville, 485 E. Gray Street, Louisville, KY 40202, USA.

E-mail: jeremy.gaskins@louisville.edu

(Received May 2016; accepted June 2017)