HYPER MARKOV LAWS FOR CORRELATION MATRICES

Jeremy Gaskins

University of Louisville

Supplementary Material

S.1. Proofs of Lemmas and Theorems from Section 2

Proof of Lemma 1 and 2:

Note that the Jacobian from $\mathbf{\Sigma}$ to $(\mathbf{R}, \mathbf{D})$ is $2^p|\mathbf{D}|^p$.

$$\begin{aligned}
p_{\text{IW}}(\mathbf{D}, \mathbf{R}) &= p_{\text{IW}}(\mathbf{\Sigma}) \left| \frac{\partial \mathbf{\Sigma}}{\partial(\mathbf{D}, \mathbf{R})} \right| \propto |\mathbf{\Sigma}|^{-(\delta+2p)/2} \operatorname{etr}\left\{ -\frac{1}{2}\mathbf{\Sigma}^{-1} \right\} \times 2^p|\mathbf{D}|^p \\
&\propto |\mathbf{R}|^{-(\delta+2p)/2} |\mathbf{D}|^{-\delta-p} \operatorname{etr}\left\{ -\frac{1}{2}\mathbf{D}^{-2}\mathbf{R}^{-1} \right\} \\
&= |\mathbf{R}|^{-(\delta+2p)/2} \prod_{j=1}^{p} d_{jj}^{-\delta-p} \exp\left\{ -\frac{1}{d_{jj}^2} \frac{|\mathbf{E}_j'\mathbf{R}\mathbf{E}_j|}{2|\mathbf{R}|} \right\} \\
&\quad \big(\text{note that the } (j,j) \text{ element of } \mathbf{R}^{-1} \text{ is } |\mathbf{E}_j'\mathbf{R}\mathbf{E}_j|/|\mathbf{R}|\big) \\
p_{\text{CIW}}(\mathbf{R}) &= \int p_{\text{IW}}(\mathbf{D}, \mathbf{R}) \, \mathrm{d}\mathbf{D} \propto |\mathbf{R}|^{-(\delta+2p)/2} \prod_{j=1}^{p} \left[\frac{|\mathbf{E}_j'\mathbf{R}\mathbf{E}_j|}{|\mathbf{R}|} \right]^{-(\delta+p-1)/2} \\
&= |\mathbf{R}|^{(p+\delta)(p-1)/2-p} \prod_{j=1}^{p} |\mathbf{E}_j'\mathbf{R}\mathbf{E}_j|^{-(\delta+p-1)/2}
\end{aligned}$$

The normalizing constant and statements *(i)* and *(iv)* can be shown similarly. It is clear that *(ii)* follows from the marginalization property of IW and interchanging the order of integration, and *(iii)* is a special case of *(ii)*.

The CW distribution in Lemma 2 is derived similarly by noting $\text{etr}\{\mathbf{S}\} = \text{etr}\{\mathbf{DRD}\} = \text{etr}\{\mathbf{D}^2\}$. For $\mathbf{S} \sim \text{W}(\delta, \mathbf{V})$ with $\mathbf{V}$ diagonal, $\mathbf{D}$ and $\mathbf{R}$ are also independent with $d_{jj}^2 \sim \text{Gamma}(\frac{\delta}{2}, 2v_{jj})$ using the scale parametrization.

As discussed in the manuscript, Theorem 1 follows from Theorem 3.9 (Dawid and Lauritzen (1993)) because we construct the distribution through the Markov combination of a consistent family of distributions. Before we prove Theorem 2, we introduce the following lemma.

Lemma 3.

(i). For $\mathbf{\Sigma} \in \mathcal{Q}(G)$, $|\mathbf{\Sigma}| = \prod_{C \in \mathcal{C}} |\mathbf{\Sigma}_C| / \prod_{S \in \mathcal{S}} |\mathbf{\Sigma}_S|$.

(ii). Consider the separation strategy applied to $\mathbf{\Sigma} \in \mathcal{Q}(G)$ *producing diagonal, standard deviation matrix* $\mathbf{D}$ *and correlation matrix* $\mathbf{R} \in \mathcal{R}_p$. *Then,* $\mathbf{R} \in \mathcal{R}(G)$, *and the Jacobian of this transformation is* $|\partial \mathbf{\Sigma} / \partial(\mathbf{D}, \mathbf{R})| = 2^p \prod_{C \in \mathcal{C}} |\mathbf{D}_C|^{|C|} / \prod_{S \in \mathcal{S}} |\mathbf{D}_S|^{|S|}$.

Statement *(i)* is well known (e.g., Lauritzen (1996)). That $\mathbf{R} \in \mathcal{R}(G)$ is clear since $\mathbf{R}^{-1}$ will maintain the zero pattern of $\mathbf{\Sigma}^{-1}$. For the Jacobian, one notes that the partial derivatives are with respect to the elements that are not constrained by G: d_{jj} $(j = 1, \ldots, p)$, r_{jk} for $(j, k) \in E$, and σ_{jk} for $j = k$ or $(j, k) \in E$. For each j, we have a contribution to the Jacobian of

$2d_{jj}^{e_j+1}$ where e_j is the number of edges connected to node j. As e_j can be represented as $\sum_{C\in\mathcal{C}}|C|I(j\in C)-\sum_{S\in\mathcal{S}}|S|I(j\in S)-1$, the Jacobian is $\prod_{j\in\mathcal{V}}2d_{jj}^{e_j+1}=2^p\prod_{C\in\mathcal{C}}|\mathbf{D}_C|^{|C|}/\prod_{S\in\mathcal{S}}|\mathbf{D}_S|^{|S|}$.

Proof of Theorem 2:

First, consider the hyper Wishart case.

$$\begin{aligned}
p_{\mathrm{HW}}(\mathbf{D},\mathbf{R}) &= p_{\mathrm{HW}}(\mathbf{\Sigma})\left|\frac{\partial\mathbf{\Sigma}}{\partial(\mathbf{D},\mathbf{R})}\right| \propto \frac{\prod_{C\in\mathcal{C}}|\mathbf{\Sigma}_C|^{(\delta-|C|-1)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{\Sigma}_C\}}{\prod_{S\in\mathcal{S}}|\mathbf{\Sigma}_S|^{(\delta-|S|-1)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{\Sigma}_S\}}\left|\frac{\partial\mathbf{\Sigma}}{\partial(\mathbf{D},\mathbf{R})}\right| \\
&= \frac{\prod_{C\in\mathcal{C}}|\mathbf{R}_C|^{(\delta-|C|-1)/2}|\mathbf{D}_C|^{\delta-|C|-1}\operatorname{etr}\{-\frac{1}{2}\mathbf{D}_C\mathbf{R}_C\mathbf{D}_C\}}{\prod_{S\in\mathcal{S}}|\mathbf{R}_S|^{(\delta-|S|-1)/2}|\mathbf{D}_S|^{\delta-|S|-1}\operatorname{etr}\{-\frac{1}{2}\mathbf{D}_S\mathbf{R}_S\mathbf{D}_S\}}\frac{2^p\prod_{C\in\mathcal{C}}|\mathbf{D}_C|^{|C|}}{\prod_{S\in\mathcal{S}}|\mathbf{D}_S|^{|S|}} \\
&\propto \pi_{\mathrm{HCW}}(\mathbf{R})\prod_{j=1}^{p}d_{jj}^{\delta-1}\exp(-d_{jj}^2/2)
\end{aligned}$$

Integrating out $\mathbf{D}$ gives $p_{\mathrm{HW}}(\mathbf{R})=\pi_{\mathrm{HCW}}(\mathbf{R})$, that is, the distribution of the correlation matrix from the hyper Wishart is the same as the hyper correlation Wishart distribution. This result will also hold for any diagonal scale parameter.

Now, we consider the hyper inverse Wishart case. It follows that

$$\begin{aligned}
p_{\mathrm{HIW}}(\mathbf{D},\mathbf{R}) &= p_{\mathrm{HIW}}(\mathbf{\Sigma})\left|\frac{\partial\mathbf{\Sigma}}{\partial(\mathbf{D},\mathbf{R})}\right| \propto \frac{\prod_{C\in\mathcal{C}}|\mathbf{\Sigma}_C|^{-(\delta+2|C|)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{\Sigma}_C^{-1}\}}{\prod_{S\in\mathcal{S}}|\mathbf{\Sigma}_S|^{-(\delta+2|S|)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{\Sigma}_S^{-1}\}}\left|\frac{\partial\mathbf{\Sigma}}{\partial(\mathbf{D},\mathbf{R})}\right| \\
&\propto \frac{\prod_{C\in\mathcal{C}}|\mathbf{D}_C|^{-\delta-|C|}|\mathbf{R}_C|^{-(\delta+2|C|)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{D}_C^{-2}\mathbf{R}_C^{-1}\}}{\prod_{S\in\mathcal{S}}|\mathbf{D}_S|^{-\delta-|S|}|\mathbf{R}_S|^{-(\delta+2|S|)/2}\operatorname{etr}\{-\frac{1}{2}\mathbf{D}_S^{-2}\mathbf{R}_S^{-1}\}} \\
&= \left\{\frac{\prod_{C\in\mathcal{C}}|\mathbf{R}_C|^{-(\delta+2|C|)/2}}{\prod_{S\in\mathcal{S}}|\mathbf{R}_S|^{-(\delta+2|S|)/2}}\right\}\prod_{j=1}^{p}d_j^{-\delta-e_j-1}\exp\left\{-\frac{m_j(\mathbf{R})}{2d_j^2}\right\},
\end{aligned}$$

where e_j is again the number of edges connected to node/variable j and $m_j(\mathbf{R})=|\mathbf{E}_j'\mathbf{R}\mathbf{E}_j|/|\mathbf{R}|$ is the (j,j)-element of $\mathbf{R}^{-1}$. It is clear from the

second line that $m_j(\mathbf{R})$ can alternatively be written as

$$m_j(\mathbf{R}) = \sum_{C \in \mathcal{C}: j \in C} [\mathbf{R}_C^{-1}]_{(j,j)} - \sum_{S \in \mathcal{S}: j \in S} [\mathbf{R}_S^{-1}]_{(j,j)}, \tag{S1.1}$$

where $[\mathbf{A}]_{(j,j)}$ is the diagonal element of $\mathbf{A}$ corresponding to the variable j; see also Lauritzen (1996, Section 5.3).

$$\begin{aligned} p_{\text{HIW}}(\mathbf{R}) &= \int p_{\text{HIW}}(\mathbf{D}, \mathbf{R}) \, \mathrm{d}\mathbf{D} \\ &\propto \left\{ \frac{\prod_{C \in \mathcal{C}} |\mathbf{R}_C|^{-(\delta+2|C|)/2}}{\prod_{S \in \mathcal{S}} |\mathbf{R}_S|^{-(\delta+2|S|)/2}} \right\} \prod_{j=1}^{p} \left[\int d_j^{-\delta - e_j - 1} \exp\left\{ -\frac{m_j(\mathbf{R})}{2 d_j^2} \right\} \mathrm{d}d_j \right] \\ &= \left\{ \frac{\prod_{C \in \mathcal{C}} |\mathbf{R}_C|^{-(\delta+2|C|)/2}}{\prod_{S \in \mathcal{S}} |\mathbf{R}_S|^{-(\delta+2|S|)/2}} \right\} \prod_{j=1}^{p} \left[\frac{\Gamma\left((\delta + e_j)/2\right)}{(m_j(\mathbf{R})/2)^{(\delta+e_j)/2}} \right] \end{aligned}$$

To appropriately normalize $p_{\text{HIW}}(\mathbf{R})$, we require a factor of $\left(\prod_{C \in \mathcal{C}} k_{\text{IW}}(\delta, |C|)\right) / \left(\prod_{S \in \mathcal{S}} k_{\text{IW}}(\delta, |S|)\right)$, where $k_{\text{IW}}(\delta, p)$ is the normalizing constant of $\text{IW}_p(\delta, \mathbf{I}_p)$. Note that due to the form of $m_j(\mathbf{R})$ in (S1.1) this distribution cannot be factored as in equation (1) in terms of the cliques and the separators. Thus, $p_{\text{HIW}}(\mathbf{R})$, the distribution of the correlation matrix of $\mathbf{\Sigma} \sim \text{HIW}_G(\delta, \mathbf{I}_p)$, is not a Markov distribution. Further, it is clearly not equivalent to the $\text{HCIW}_G(\delta)$ distribution which by construction can be factored according to G.

For additional clarity, consider the following simple example. Let G be the graph with variables $\mathcal{V} = \{1, 2, 3\}$ and edges $E = \{(1, 2), (2, 3)\}$. Then $C_1 = \{1, 2\}$, $C_2 = \{2, 3\}$, and $S_2 = \{2\}$ provide a perfect ordering of the cliques. For notational convenience, let $x = r_{12}$ and $y = r_{23}$. It is easy to show that $|\mathbf{R}_{C_1}| = 1 - x^2$, $|\mathbf{R}_{C_2}| = 1 - y^2$, $|\mathbf{R}_{S_2}| = 1$, $e_1 = e_3 = 1$, $e_2 = 2$,

$m_1 = (1-x^2)^{-1}$, $m_3 = (1-y^2)^{-1}$, and $m_2 = (1-x^2)^{-1}(1-y^2)^{-1}(1-x^2y^2)$. Then,

$$\begin{aligned}
\pi_{\mathrm{HCIW}}(\mathbf{R}) &= \frac{p_{C_1}(\mathbf{R}_{C_1})\, p_{C_2}(\mathbf{R}_{C_2})}{p_{S_2}(\mathbf{R}_{S_2})} \propto |\mathbf{R}_{C_1}|^{\delta/2-1}\, |\mathbf{R}_{C_2}|^{\delta/2-1} = (1-x^2)^{\delta/2-1}\,(1-y^2)^{\delta/2-1} \\
p_{\mathrm{HIW}}(\mathbf{R}) &\propto \frac{|\mathbf{R}_{C_1}|^{-\delta/2-2}|\mathbf{R}_{C_2}|^{-\delta/2-2}}{|\mathbf{R}_{S_2}|^{-\delta/2-1}} \prod_{j=1}^{p} m_j(\mathbf{R})^{-\delta/2-e_j/2} \\
&= (1-x^2)^{(\delta-1)/2}(1-y^2)^{(\delta-1)/2}(1-x^2y^2)^{-(\delta+2)/2} \neq \pi_{\mathrm{HCIW}}(\mathbf{R}).
\end{aligned}$$

Again, it is clear that $\pi_{\mathrm{HCIW}}(\mathbf{R})$ and $p_{\mathrm{HIW}}(\mathbf{R})$ are not the same. Further, p_{HIW} is not Markov. The separating set S_2 has no correlation parameters, so for $p_{\mathrm{HIW}}(\mathbf{R})$ to be Markov, the parameters from C_1 ($x = r_{12}$) and C_2 ($y = r_{23}$) would have to be independent. Note that the Markov property of HIW implies that $(\sigma_{11}, \sigma_{12}) \perp\!\!\!\perp (\sigma_{23}, \sigma_{33}) \,|\, \sigma_{22}$, which implies $r_{12} \perp\!\!\!\perp r_{23} \,|\, d_2$. However, marginalizing out the standard deviation destroys the independence between the correlations.

S.2. Additional computational details from simulation study

In Section 5 of the manuscript, we introduced a simulation study to evaluate the performance of our proposal. Here we provide additional details.

We choose $\mu_j \sim \mathrm{N}(0, 50^2)$, $\sigma_j^2 \sim \mathrm{InvGamma}(0.1, 0.1)$, and $\nu_j \sim \mathrm{Unif}(2, 30)$ as relatively uninformative prior distributions. The degrees of freedom is bounded by 2, so that Y_j has at least two moments. Table S.1 contains the

Table S.1: Tuning parameter specification and average run time for simulation study of Section 5. Burn-in is the number of burn-in iterations. A thinning value of k means that every kth iteration is retained for inference; the total number of iterations run is burn-in plus $2000k$. ϵ is the tuning parameter for the block sampler for $\mathbf{R}$, and graph steps indicates the number of edge proposal made per iteration (see final paragraph of Section 3). The standard deviation for the dimension-matching parameter U in the reversible jump proposal is σ (Section 3.2). ζ_β, ζ_σ, and ζ_ν represent the standard deviation for the (univariate) random walk Metropolis-Hasting steps required to sample the marginal distributions of the Gaussian copula. Time denotes the average time per data set that the sampler runs.

Model		Correlation A		Correlation B		Correlation C	
		$N = 100$	$N = 500$	$N = 100$	$N = 500$	$N = 100$	$N = 500$
$\text{HCIW}_G(\delta)\pi(G)$	burn-in	250	200	300	120	200	400
	thinning	10	10	15	20	40	80
	graph steps	100	150	50	80	25	20
	ϵ	50	50	50	150	150	1100
	σ	0.1	0.1	0.1	0.1	0.1	0.1
	ζ_μ	0.3	0.1	0.2	0.1	0.2	0.1
	ζ_σ	0.4	0.2	0.4	0.2	0.4	0.3
	ζ_ν	5	3	5	3	5	3
	time($\delta = 2$)	10.1 h	15.4 h	12.1 h	22.2 h	17.3 h	38.3 h
	time($\delta = 1$)	10.1 h	14.8 h	11.7 h	17.5 h	21.4 h	36.6 h
$\text{HCW}_G(\delta)\pi(G)$	burn-in	200	200	300	1000	1200	500
	thinning	5	5	15	20	20	50
	graph steps	100	100	50	75	50	20
	ϵ	50	50	50	250	200	1200
	σ	0.1	0.1	0.1	0.1	0.1	0.1
	ζ_μ	0.3	0.1	0.2	0.1	0.2	0.1
	ζ_σ	0.4	0.2	0.4	0.2	0.4	0.2
	ζ_ν	5	3	5	3	5	3
	time($\delta = 25$)	5.0 h	5.6 h	12.4 h	22.1 h	16.7 h	23.3 h
	time($\delta = 10$)	5.0 h	5.6 h	12.1 h	20.2 h	14.7 h	22.5 h
$\pi(\mathbf{R}) \propto I(\mathbf{R} \in \mathcal{R})$	burn-in	600	25,000	30,000	1000	30,000	1,200
	thinning	300	250	200	200	200	150
	ϵ	6000	30,000	7000	30,000	7000	30,000
	ζ_μ	0.2	0.1	0.2	0.1	0.2	0.1
	ζ_σ	0.4	0.2	0.4	0.2	0.4	0.2
	ζ_ν	5	3	5	3	5	3
	time	16.1 h	19.6 h	11.2 h	15.4 h	11.1 h	11.7 h
$\mathbf{R} = \mathbf{I}_p$	burn-in	100	1200	100	120	100	100
	thinning	2	3	3	3	2	2
	ζ_μ	0.3	0.2	0.4	0.2	0.3	0.2
	ζ_σ	0.4	0.3	0.5	0.4	0.4	0.3
	ζ_ν	5	4	5	3	5	3
	time	0.05 h	0.2 h	0.04 h	0.2 h	0.05 h	0.1 h

values chosen for the tuning parameters for this simulation study. These are chosen by trail and error from running the sampler for a small number of iterations, evaluating the mixing, and adjusting the values. The value of ϵ determines how similar the block proposal $\mathbf{R}^{\star}_{C_k}$ is to the current value $\mathbf{R}_{C_k}$, and following the advice of Zhang et al. (2006), we choose ϵ for an acceptance rate near 15–20%. As the sample size gets larger, the posterior is more concentrated, and smaller moves (larger ϵ) are required. Additionally, larger clique sizes $|C|$ typically require larger ϵ to achieve acceptance. One could allow ϵ to vary by clique size (Table 1, Step 2), but due to the additional tuning needed, we did not pursue this. When using the flat prior on the full graph, we employ the sampling scheme of Section 3.1 using the single clique $C = \mathcal{V}$ (similar to Zhang et al. (2006)). Updating the full $\mathbf{R}$ requires very large ϵ (greater than 6000 for $N = 100$ and 30,000 for $N = 500$) to obtain acceptance rates between 15–20% as this represents a large clique. We select the number of times to repeat the graph-update step (Section 3.2), so that we average between 0.5 and 1 accepted edge changes per iteration. Burn-in values are chosen by trace plots and the Geweke tests on the data likelihood from preliminary runs; chain length and thinning values are taken so that 2000 iterations, having an effective sample size of at least 400, are retained for inference. Parameters of the marginal

distributions are updated through independent Metropolis-Hastings steps.

Table S.2 contains the estimated risk of the four quantities of interest: the location parameters, scale parameters, correlation matrix, and graph structure. For the location and scale parameters, we use sum of squared error loss: $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \sum_j (\mu_j - \hat{\mu}_j)^2$ and $L(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) = \sum_j (\sigma_j - \hat{\sigma}_j)^2$. For the correlation matrix, we employ the log-likelihood loss function $L(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}\{\hat{\mathbf{R}}\mathbf{R}^{-1}\} - \log|\hat{\mathbf{R}}\mathbf{R}^{-1}| - p$. The Bayes estimators for the locations and scales are the posterior means, and for $\mathbf{R}$, the Bayes estimator is $\hat{\mathbf{R}} = [\text{E}(\mathbf{R}^{-1}|\mathbf{y}) - \boldsymbol{\Lambda}]^{-1}$ where $\boldsymbol{\Lambda}$ is a diagonal matrix of Lagrange multipliers constraining $\hat{\mathbf{R}}$ to have unit diagonal (Pitt et al. (2006)). To evaluate the accuracy of graph recovery, we consider 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. Box plots containing the location, scale, correlation, and graph losses from the simulation study in Section 5 of the manuscript are contained in Figure S.1 for $N = 100$ and in Figure S.2 for $N = 500$.

In addition to this simulation study, we also consider another small simulation study to explore the use of the sampling algorithm in Section 3.1 when the graph structure is fixed. Using the data drawn under true cor-

Table S.2: Estimated risk from the simulation study. Monte Carlo standard errors are in parentheses. Loss functions are sum of squared errors for location and scale parameters, $L(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}\{\hat{\mathbf{R}}\mathbf{R}^{-1}\} - \log|\hat{\mathbf{R}}\mathbf{R}^{-1}| - p$ for correlation matrix, and average number of error for the graph.

Correlation Matrix	N	Loss Fcn	Prior Choice					
			HCIW$_G$(2)	HCIW$_G$(1)	HCW$_G$(10)	HCW$_G$(25)	Flat	Indep
A	100	Location	$0.294_{(0.012)}$	$0.294_{(0.012)}$	$0.293_{(0.012)}$	$0.295_{(0.012)}$	$0.307_{(0.013)}$	$0.327_{(0.013)}$
A	100	Scale	$0.269_{(0.008)}$	$0.271_{(0.009)}$	$0.263_{(0.008)}$	$0.297_{(0.009)}$	$0.385_{(0.012)}$	$0.362_{(0.014)}$
A	100	Corr	$0.402_{(0.013)}$	$0.399_{(0.013)}$	$0.572_{(0.018)}$	$1.006_{(0.026)}$	$4.114_{(0.035)}$	$29.957_{(-)}$
A	100	Graph	$1.02_{(0.04)}$	$0.89_{(0.04)}$	$1.53_{(0.05)}$	$2.45_{(0.06)}$	$276_{(-)}$	$24_{(-)}$
A	500	Location	$0.055_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.066_{(0.003)}$
A	500	Scale	$0.063_{(0.002)}$	$0.063_{(0.002)}$	$0.065_{(0.002)}$	$0.071_{(0.002)}$	$0.065_{(0.002)}$	$0.090_{(0.003)}$
A	500	Corr	$0.076_{(0.002)}$	$0.075_{(0.002)}$	$0.091_{(0.003)}$	$0.133_{(0.004)}$	$0.663_{(0.006)}$	$29.957_{(-)}$
A	500	Graph	$0.44_{(0.02)}$	$0.38_{(0.02)}$	$0.68_{(0.02)}$	$1.16_{(0.03)}$	$276_{(-)}$	$24_{(-)}$
B	100	Location	$0.272_{(0.011)}$	$0.272_{(0.011)}$	$0.276_{(0.011)}$	$0.281_{(0.012)}$	$0.296_{(0.013)}$	$0.338_{(0.017)}$
B	100	Scale	$0.284_{(0.009)}$	$0.282_{(0.009)}$	$0.320_{(0.010)}$	$0.405_{(0.012)}$	$0.526_{(0.015)}$	$0.359_{(0.014)}$
B	100	Corr	$1.069_{(0.027)}$	$1.078_{(0.027)}$	$1.387_{(0.035)}$	$2.016_{(0.044)}$	$5.168_{(0.046)}$	$26.726_{(-)}$
B	100	Graph	$10.73_{(0.22)}$	$10.46_{(0.22)}$	$13.17_{(0.26)}$	$17.87_{(0.29)}$	$268_{(-)}$	$32_{(-)}$
B	500	Location	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.003)}$	$0.057_{(0.003)}$	$0.069_{(0.004)}$
B	500	Scale	$0.075_{(0.002)}$	$0.075_{(0.002)}$	$0.085_{(0.003)}$	$0.111_{(0.003)}$	$0.101_{(0.003)}$	$0.089_{(0.003)}$
B	500	Corr	$0.108_{(0.004)}$	$0.106_{(0.004)}$	$0.157_{(0.006)}$	$0.279_{(0.009)}$	$0.819_{(0.009)}$	$26.726_{(-)}$
B	500	Graph	$1.43_{(0.05)}$	$1.17_{(0.04)}$	$2.42_{(0.07)}$	$4.71_{(0.11)}$	$268_{(-)}$	$32_{(-)}$
C	100	Location	$0.232_{(0.012)}$	$0.231_{(0.012)}$	$0.236_{(0.012)}$	$0.240_{(0.013)}$	$0.256_{(0.015)}$	$0.328_{(0.023)}$
C	100	Scale	$0.259_{(0.008)}$	$0.261_{(0.009)}$	$0.402_{(0.015)}$	$0.618_{(0.019)}$	$0.400_{(0.013)}$	$0.367_{(0.018)}$
C	100	Corr	$1.951_{(0.040)}$	$2.062_{(0.044)}$	$2.972_{(0.063)}$	$3.824_{(0.072)}$	$5.548_{(0.052)}$	$30.591_{(-)}$
C	100	Graph	$40.4_{(0.5)}$	$41.0_{(0.5)}$	$42.1_{(0.5)}$	$44.3_{(0.6)}$	$210_{(-)}$	$90_{(-)}$
C	500	Location	$0.043_{(0.002)}$	$0.043_{(0.002)}$	$0.044_{(0.002)}$	$0.044_{(0.002)}$	$0.045_{(0.002)}$	$0.069_{(0.006)}$
C	500	Scale	$0.062_{(0.002)}$	$0.063_{(0.002)}$	$0.111_{(0.004)}$	$0.182_{(0.006)}$	$0.183_{(0.006)}$	$0.091_{(0.003)}$
C	500	Corr	$0.295_{(0.005)}$	$0.303_{(0.006)}$	$0.496_{(0.012)}$	$0.718_{(0.015)}$	$1.133_{(0.015)}$	$30.591_{(-)}$
C	500	Graph	$9.22_{(0.09)}$	$9.21_{(0.09)}$	$10.05_{(0.08)}$	$11.63_{(0.11)}$	$210_{(-)}$	$90_{(-)}$

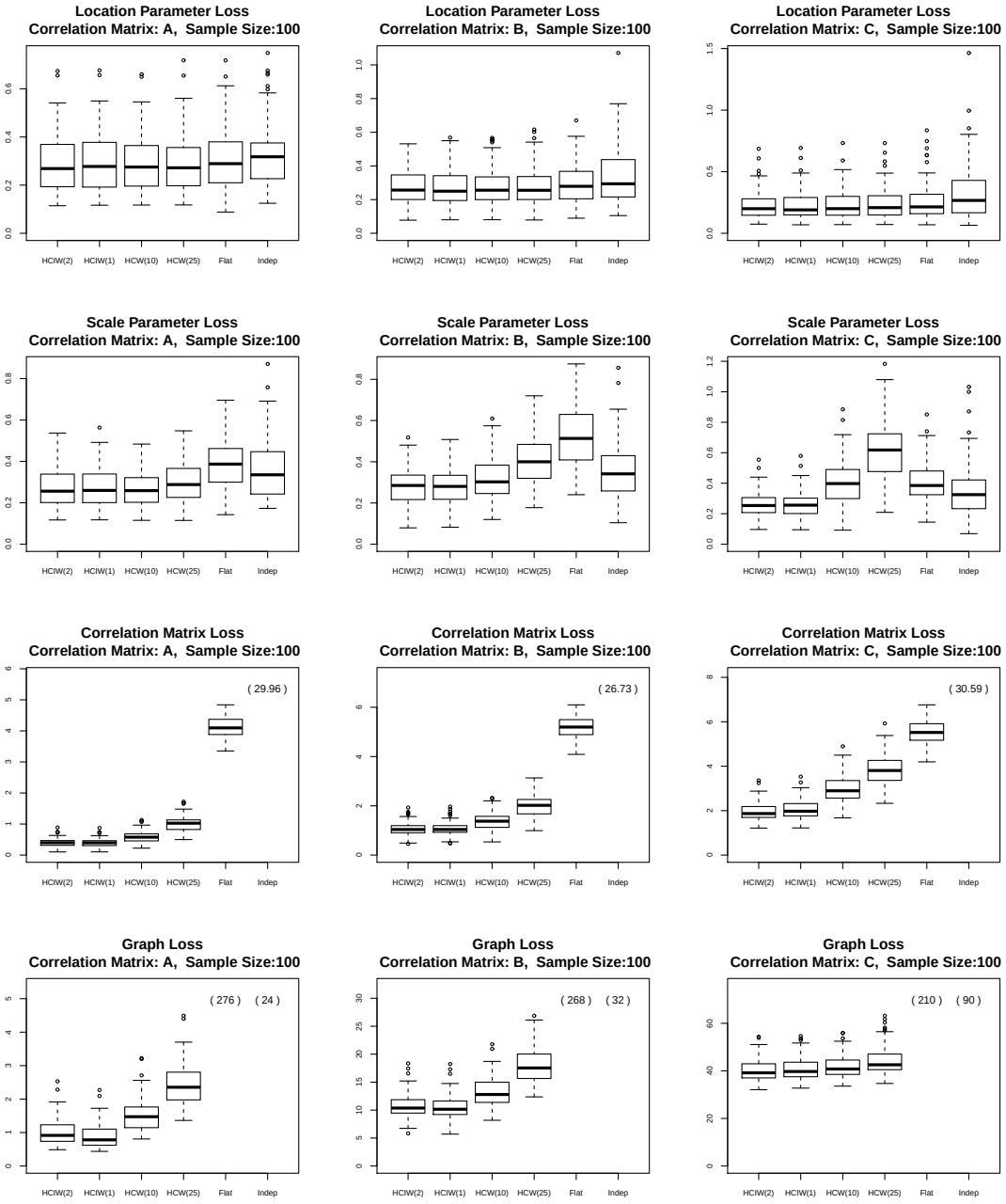

Figure S.1: Box plot of losses for the 100 data sets under each prior choice from the simulation study of Section 5 for $N = 100$. The rows gives plots for the location, scale, correlation, and graph loss. Columns designate the different correlation/graph choices.

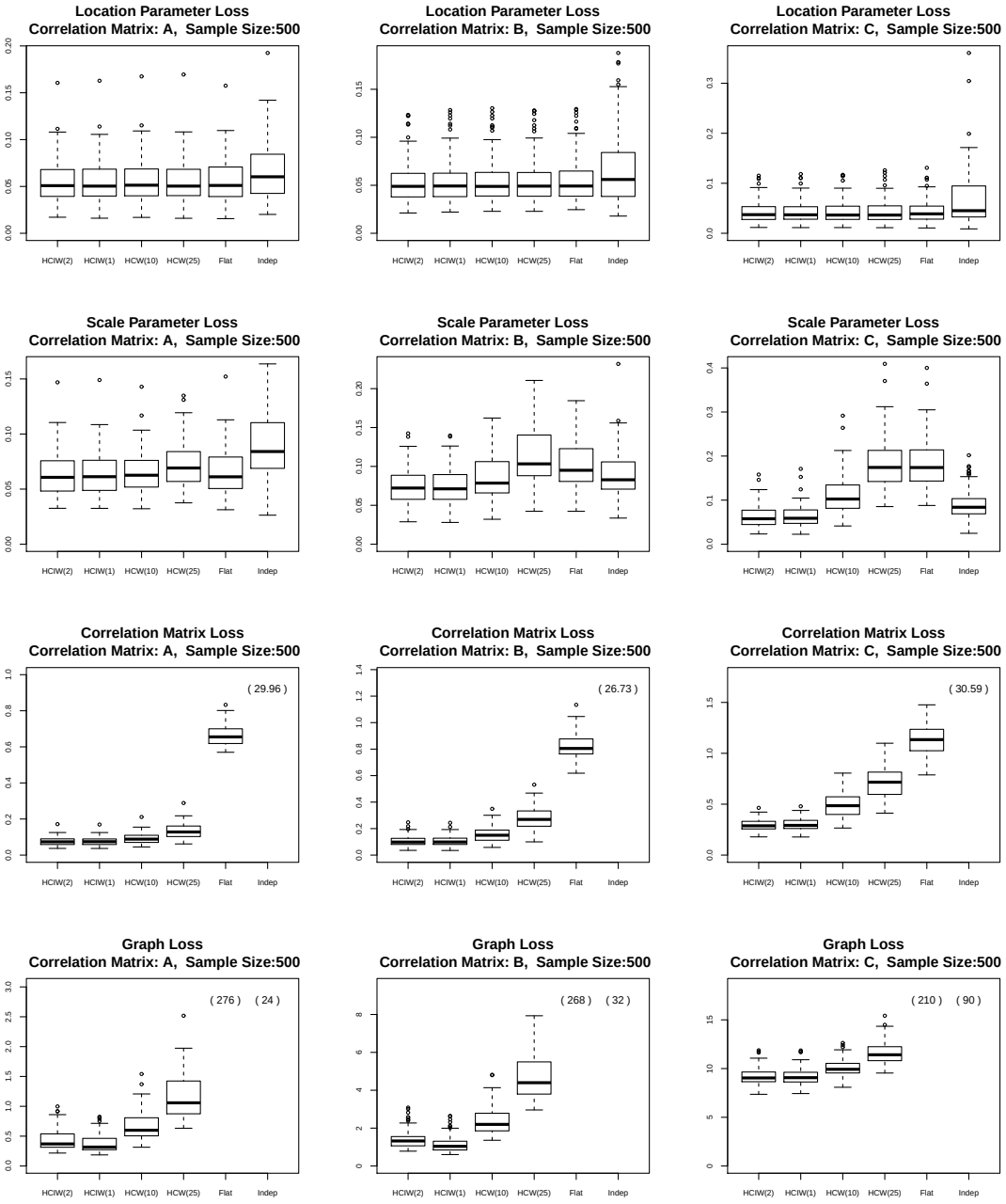

Figure S.2: Box plot of the losses for the 100 data sets under each prior choice from the simulation study of Section 5 for $N = 500$. The rows gives plots for the location, scale, correlation, and graph loss. Columns designate the different correlation/graph choices.

Table S.3: Estimated risk from simulation study with fixed true graph A (Monte Carlo standard errors in parentheses). Loss functions are sum of squared errors for location and scale parameters and $L(\hat{\mathbf{R}}, \mathbf{R}) = \text{tr}\{\hat{\mathbf{R}}\mathbf{R}^{-1}\} - \log|\hat{\mathbf{R}}\mathbf{R}^{-1}| - p$ for correlation matrix. Compare with Table 1 from the manuscript for values when graph is unknown.

N	Loss Fcn	Prior Choice					
		$\text{HCIW}_{\tilde{G}}(2)$	$\text{HCIW}_{\tilde{G}}(1)$	$\text{HCW}_{\tilde{G}}(10)$	$\text{HCW}_{\tilde{G}}(25)$	$\text{HCIW}_{G}(2)\pi(G)$	Flat
100	Location	$0.294_{(0.012)}$	$0.293_{(0.012)}$	$0.294_{(0.012)}$	$0.295_{(0.012)}$	$0.294_{(0.012)}$	$0.307_{(0.013)}$
100	Scale	$0.269_{(0.009)}$	$0.270_{(0.009)}$	$0.263_{(0.008)}$	$0.293_{(0.009)}$	$0.269_{(0.008)}$	$0.385_{(0.012)}$
100	Corr	$0.393_{(0.013)}$	$0.389_{(0.012)}$	$0.561_{(0.018)}$	$0.982_{(0.026)}$	$0.402_{(0.013)}$	$4.114_{(0.035)}$
500	Location	$0.055_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.056_{(0.002)}$	$0.055_{(0.002)}$	$0.056_{(0.002)}$
500	Scale	$0.063_{(0.002)}$	$0.063_{(0.002)}$	$0.065_{(0.002)}$	$0.071_{(0.002)}$	$0.063_{(0.002)}$	$0.065_{(0.002)}$
500	Corr	$0.075_{(0.002)}$	$0.075_{(0.002)}$	$0.090_{(0.003)}$	$0.132_{(0.004)}$	$0.076_{(0.002)}$	$0.663_{(0.006)}$

relation $\mathbf{R}_A$, we repeat the simulation where the graph G is fixed to be the true graph $\tilde{G}_A$. Table S.3 contains the location, scale, and correlation losses using the four hyper laws: $\text{HCIW}_{\tilde{G}}(2), \text{HCIW}_{\tilde{G}}(1), \text{HCW}_{\tilde{G}}(10), \text{HCW}_{\tilde{G}}(25)$. For comparison, we repost the results for $\text{HCIW}_G(2)\pi(G)$ (unknown graph) and the flat prior (full correlation matrix) from Table 3.

Overall, the results are quite similar to the case when the graph is unknown. We do find the losses to be slightly smaller when the graph is known compared to that seen in Table 3 when the graph was estimated. Slightly improved risk is to be expected when one knows the true graph. The small magnitude of the difference can be explained by the graph being so well estimated for scenario A. Using $\mathbf{R}_B$ and $\mathbf{R}_C$, one would expect qualitatively similar results, although using the fixed true graph $\tilde{G}$ may be more strongly favored relative to estimating the graph as in Table 3. Overall, this further confirms the ability of the sampling algorithm to correctly estimate

the correlation dependence parameter from the Gaussian copula model.

S.3. Additional computational details from data example

The financial data application in Section 6 considers a Gaussian copula with $p = 30$. Due to the decreased mixing over $\mathcal{G}$ as p increases, it is necessary to adjust the sampler by incorporating an adaptive Metropolis proposal. To that end, we incorporate an adaptive sampler that proposes edges relative to the uncertainty of its inclusion in G. Adaptive MCMC has been found to be effective in a wide array of problems, particularly in tuning the variance terms of random walk Metropolis-Hastings (Andrieu and Thoms (2008); Roberts and Rosenthal (2009)) and variable selection (Nott and Kohn (2005); Lamnisos et al. (2013)).

In our adaptive sampler which replace Step 1 in Table 2, we associate a weight γ_{jk} to each edge e_{jk}, and the probability of an attempt to swap that edge in the graph step is proportional to this weight. During each batch of 1000 iterations, we retain how many times we propose changing e_{jk}, how many times the resulting graph $G^\star$ is decomposable, and how many times we accept the move to $G^\star$. After running a batch of 1000 iterations, we increase γ_{jk} by $t^{-1/2}$ if more than 20% of proposals where changing e_{jk} yielded a decomposable $G^\star$ were accepted, or we decrease γ_{jk} by $t^{-1/2}$ if fewer than 20% of decomposable proposals were accepted (t denotes the

iteration number). We also bound $\gamma_{jk} \in [0.1, 1.0]$, so that all edges have some chance of being proposed and that no edge dominates.

As the adaptive algorithm changes the proposal distribution, our general theory about MH sampling no longer applies. Let $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{p-1,p})$ and $P_{\boldsymbol{\gamma}}(\mathbf{R}, \cdot)$ be the transition function with starting point $\mathbf{R}$ (and its corresponding graph G) and proposal weights $\boldsymbol{\gamma}$. Roberts and Rosenthal (2007) provides sufficient conditions for the ergodicity of adaptive samplers. The adaptive algorithm is ergodic if we have simultaneous uniform ergodicity and diminishing adaptation (Roberts and Rosenthal (2007, Theorem 1)). The diminishing adaptation is easy to establish since $|\gamma_{jk}^{(t+1)} - \gamma_{jk}^{(t)}| \to 0$ in t, so $\sup_{\mathbf{R}} ||P_{\boldsymbol{\gamma}^{(t+1)}}(\mathbf{R}, \cdot) - P_{\boldsymbol{\gamma}^{(t)}}(\mathbf{R}, \cdot)|| \to 0$. While simultaneous uniform ergodicity is not easy to show, Roberts and Rosenthal (2007, Corollary 3 and Lemma 1) show ergodicity if the support of $\mathbf{R}$ and the support of $\boldsymbol{\gamma}$ are compact with a continuity condition. Clearly $\boldsymbol{\gamma} \in [0.1, 1]^J$ has compact support, but $\mathcal{R}$ does not without an extension to semi-definite correlation matrices. Alternatively, we could create a compact support by restricting $\mathcal{R}$ to contain only those correlation matrices whose smallest eigenvalue or determinant is at or above some very small threshold; in essence, any computer algorithm works in such a way.

Further work exploring ergodicity more formally with these adaptive

samplers would be of great benefit. However, the mathematical analysis is beyond the scope of the current article and is far from trivial. We seek only to provide a flavor of the issues involved. We additionally note that one can also stop adjusting $\boldsymbol{\gamma}$ after burn-in and apply standard theory, or one could use the original sampler of Section 5 with all γ_{jk} equal. However, preliminary analysis indicates the original sampler will require three to four times the computational time to produce a similar effective sample size.

As a final note, we also make use of the adaptive sampler for graph selection with the MVN-HIW model. Because the HIW sampler updates G marginally over $\boldsymbol{\Sigma}$, it is only necessary to update G and γ. Applying Corollary 3 and Lemma 1 of Roberts and Rosenthal (2007) to the finite space $\mathcal{G}$ and the compact support of γ immediately shows that this algorithm is ergodic. A similar result for adaptive algorithms in the context of covariate selection for the linear regression model was shown by Lamnisos et al. (2013).

Table S.4 shows tuning parameters and MCMC specifications used in the algorithms that fit the financial data. As in the simulation study, we retain 2000 iterations after burn-in using a thinning value such that we have an effective sample size of at least 400.

Finally, we discuss some of the conclusions about the marginal distri-

Table S.4: Tuning parameter specification and average run time for the financial data application of Section 6. See Table S.1 for definitions. Table entries denoted by "—" indicates that this value is not relevant to the sampler of the given model.

	Copula with t-marginals					Multivariate Normal		
	Complete	Sparse HCIW(2)	Sparse HCIW(1)	Sparse HCW(10)	Indep.	Complete	Sparse	Indep.
burn-in	15000	30,000	15,000	21,000	100	100	18,000	100
thinning	350	60	30	75	5	1	15	1
graph steps	—	35	75	75	—	—	250	—
ϵ	45,000	2400	2400	1200	—	—	—	—
σ	—	0.1	0.1	0.1	—	—	—	—
ζ_β	0.08	0.08	0.08	0.8	0.08	—	—	—
ζ_σ	0.15	0.2	0.2	0.2	0.2	—	—	—
ζ_ν	4	4	4	4	4	—	—	—
time	2.2 d	4.0 d	1.7 d	5.5 d	39 m	12 s	2.0 d	14 s

bution parameters under the best-predicting model (the Gaussian copula with $\text{HCIW}_G(2)$ correlation prior). Figure S.3 displays the parameters of the marginal density for each industry. From the values of β, we find utilities (industry number 20) and communications (21) to be the least sensitive industries to market fluctuations, while recreation (4) is the most affected. From an economic perspective this is quite reasonable as utilities and communications are believed to have relatively inelastic demand curves, whereas recreation spending increases (decreases) during economic booms (downturns). Figure S.3(c) clearly indicates the need of the copula model as the degrees of freedom varies considerably across industries. Some industries, including electrical equipment (14), business equipment (23), and retail (27), have wide credible intervals with larger values of ν_j. While these

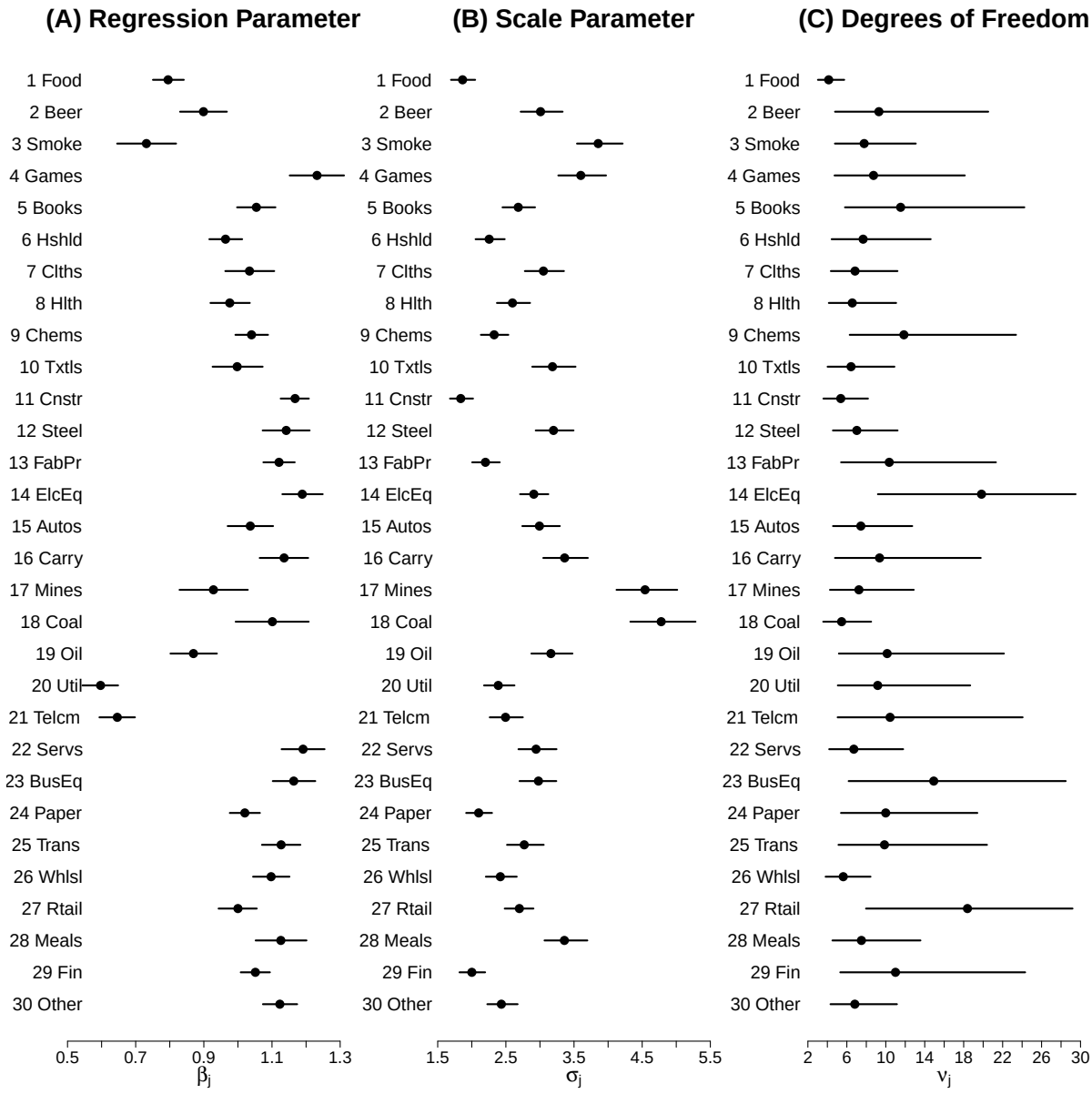

Figure S.3: Posterior means and 95% credible intervals for the marginal distribution parameters as fit by the copula model with sparse correlation matrix.

industries may be reasonably modeled with a normal distribution, other industries (food products (1), construction (11), coal (18), wholesale (26)) are highly concentrated around small values of ν_j due to their heavy tailed behavior.

References

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

Lamnisos, D., Griffin, J. E., and Steel, M. F. J. (2013). Adaptive MC^3 and Gibbs algorithms for Bayesian model averaging in linear regression models. arXiv:1306.6028.

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

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3):537–554.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Compu-*

tational and Graphical Statistics, 18(2):349–367.

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(4):880–896.