Supplementary Material to A Flexible Bayesian Approach to Monotone Missing Data in Longitudinal Studies with Nonignorable Missingness with Application to an Acute Schizophrenia Clinical Trial

July 3, 2014

A Blocked Gibbs Sampler

To implement the blocked Gibbs sampler, followingIshwaran and James (2001) we introduce a latent variable C_i with $C_i = k$ if observation i belongs to the kth class of the mixture distribution and proceed by blocked Gibbs sampling. We alter their scheme slightly by introducing additional latent variables representing $\boldsymbol{Y}_{mis,i}$ for each observation, which is blocked with C_i. This yields the following algorithm.

1. Conditional for $\boldsymbol{\theta}^{(k)}$: Simulate $\boldsymbol{\theta}^{(k)}$ from

$$p(\boldsymbol{\theta}^{(k)}|\boldsymbol{C}, \boldsymbol{Y}) \propto H(d\boldsymbol{\theta}^{(k)}) \prod_{i:C_i=k} f(\boldsymbol{Y}_i|\boldsymbol{\theta}_1^{(k)}) g(S_i|\boldsymbol{Y}, \boldsymbol{\theta}_2^{(k)}).$$

2. Conditional for $\boldsymbol{\beta}$: Let M_k denote the number of C_i equal to k, and $M_{>k}$ denote

the number of C_i strictly greater than k. First, for $k = 1, ..., K-1$, simulate

$$\beta'_k \overset{\text{ind}}{\sim} \text{Beta}(1 + M_k, \alpha + M_{>k}),$$

and set $\beta'_K \equiv 1$. Then set $\beta_k = \beta'_k \prod_{j<k}(1 - \beta'_j)$.

3. Conditional for $(C_i, \boldsymbol{Y}_{mis,i})$: First, simulate C_i according to

$$P(C_i = k) \propto \beta_k f(\boldsymbol{Y}_{obs,i}|\boldsymbol{\theta}^{(k)}) g(S_i|\boldsymbol{Y}_{obs,i}, \boldsymbol{\theta}_2^{(k)}),$$

then simulate $\boldsymbol{Y}_{mis,i}$ according to

$$p(\boldsymbol{Y}_{mis,i}|\boldsymbol{Y}_{obs,i}, C_i, \boldsymbol{\theta}_1^{(C_i)}) = f(\boldsymbol{Y}_{mis,i}|\boldsymbol{Y}_{obs,i}, \boldsymbol{\theta}_1^{(C_i)}).$$

Sampling C_i at this point involves calculating the observed data likelihood

$$L_{obs,i} = \sum_{k=1}^{K} \beta_k f(\boldsymbol{Y}_{obs,i}|\boldsymbol{\theta}_k) g(S_i|\boldsymbol{Y}_{obs,i}, \boldsymbol{\zeta}_k)$$

which may be retained if desired for model evaluation purposes.

If hyperpriors are placed on α and H, we can easily add steps corresponding to updating these parameters. The relevant likelihood of α is given by

$$L(\alpha|\beta_1, ..., \beta_k) = \alpha^{K-1} e^{\alpha \sum_k \log(1-\beta'_k)},$$

which can be given a conjugate gamma prior, and the relevant likelihood for any parameters in H is given by

$$L(H|\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K)}) = \prod_{k=1}^{K} H(d\boldsymbol{\theta}^{(k)}).$$

Any updates which cannot be done in closed form may be replaced by appropriate updates which leave these conditional distributions invariant, such as slice sampling updates Neal (2003).

B Prior Specification

B.1 Parametric Priors and the Modified Cholesky Decomposition

Parameters in parametric models in the main text are given "noninformative" priors on mean components, which we take to be a just-proper $N(0, 10^6)$ prior, and priors based on the modified Cholesky decomposition on covariance matrices (Daniels and Pourahmadi, 2002). For convenience, we review the modified Cholesky decomposition here.

The modified Cholesky decomposition of a precision matrix $\boldsymbol{\Sigma}^{-1}$ is

$$\boldsymbol{\Sigma}^{-1} = \mathbf{L}^T \mathbf{D} \mathbf{L}$$

where $\mathbf{L}$ is a lower-triangular matrix consisting of ones on the main diagonal and the negative of the generalized autoregressive parameters off the diagonal and $\mathbf{D}$ is diagonal with elements corresponding to the inverse of the innovation variances. Noting that we

can write

$$E(Y_j|\bar{\boldsymbol{Y}}_{\boldsymbol{j-1}}) = \mu_j + \sum_{k=1}^{j-1} \phi_{jk}(Y_k - \mu_k),$$
$$\text{Var}(Y_j|\bar{\boldsymbol{Y}}_{\boldsymbol{j-1}}) = \sigma_j^2,$$

the elements of $\mathbf{L}$ correspond to $\{-\phi_{jk} : j = 1, ..., J, k = 1, ..., j-1\}$ and the elements of $\mathbf{D}$ correspond to σ_j^{-2}. In all parametric models where this decomposition is used, we set $\phi_{jk} \sim N(0, 10^6)$ and $\sigma_j \sim \text{Uniform}(0, 100)$.

B.2 Nonparametric Default Priors

We use default hierarchical priors borrowing ideas from Rasmussen (2000) and Taddy (2008). We first discuss the prior on $g(s \mid \boldsymbol{y}, \boldsymbol{\theta}_2)$. As a preprocessing step, we standardize the data so the grand observed mean (across all treatments and times) is 0 and grand observed variance is 0.5.

For the simulation in Section 4.1 we assume $g(s \mid \boldsymbol{y}, \theta_2) = g(s \mid \boldsymbol{\theta}_2) = \theta_{2s}$, i.e. $\boldsymbol{Y}$ and S are independent within cluster. We take $\boldsymbol{\theta}_2^{(k)} \sim \mathcal{D}(\boldsymbol{\zeta})$ where $\boldsymbol{\zeta}$ is chosen so that a priori $E[\theta_{2s}] = \zeta_s / \sum_j \zeta_j$ is equal to the empirical probability of $S = s$. $\sum_{j=1}^J \zeta_j$ is a smoothing parameter, analogous to α in the Dirichlet process. If $\sum_{j=1}^J \zeta_j$ is very large then the dropout distribution is essentially the same across classes, making dropout and outcome approximately independent. Conversely if $\sum_{j=1}^J \zeta_j$ is very small then only one dropout pattern will typically be represented in a given class. $\sum_{j=1}^J \zeta_j = 3$ was chosen.

In Sections 4.2 and 5 we choose $g(s \mid \boldsymbol{y}, \boldsymbol{\theta}_2)$ so that

$$\text{logit}\left\{g(S = s \mid \boldsymbol{Y}, S \geq s, \boldsymbol{\zeta}, \boldsymbol{\gamma})\right\} = \zeta_s + \gamma_{s1} Y_s + \gamma_{s2} Y_{s-1}.$$

where $\boldsymbol{\theta}_2 = (\boldsymbol{\zeta}, \boldsymbol{\gamma})$. All ζ and γ terms are given independent $N(\mu_\zeta, \sigma_\zeta^2)$ and $N(\mu_\gamma, \sigma_\gamma^2)$ distributions. μ_ζ and μ_γ are given Cauchy priors with location 0 and scales 5 and 2.5 respectively. σ_ζ^2 and σ_γ^2 are given $\Gamma^{-1}(1,1)$ priors.

We now address the prior on $f(\boldsymbol{y} \mid \boldsymbol{\theta}_1)$. Here $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We use the modified Cholesky specification for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Within mixture component k we can write

$$
\begin{aligned}
Y_j &= a_j^{(k)} + \sum_{k=1}^{j-1} \phi_{\ell j}^{(k)} (Y_k - a_\ell^{(k)}) + \epsilon_j, \\
\epsilon_j &\sim N(0, \sigma_j^{(k)}).
\end{aligned}
$$

We set $a_j^{(k)} \sim N(\mu_j, \sigma_{a_j}^2)$, and $\boldsymbol{\phi}_j \sim N(\mathbf{0}, \sigma_{\phi_j}^2 \boldsymbol{I})$. We took $\mu_j \sim N(0, 0.5)$. The variance components were specified as follows.

$$
\begin{aligned}
\sigma_j^{2(k)} &\sim \Gamma^{-1}(s_1, s_1 w_j), \qquad & s_1 - 2 &\sim \Gamma^{-1}(1,1), \\
w_j &\sim \Gamma\left(1, \frac{2}{g_j}\right), & \sigma_{a_j}^2 &\sim \Gamma^{-1}(s_2, s_2 \lambda g_j), \\
s_2 - 2 &\sim \Gamma^{-1}(1,1), & \lambda_2 &\sim \Gamma(1,1), \\
\sigma_{\phi_j}^2 &\sim \Gamma^{-1}(s_3, s_3 \lambda_3), & s_3 - 2 &\sim \Gamma^{-1}(1,1), \\
\lambda_3 &\sim \Gamma(1,1),
\end{aligned}
$$

where g_j is the MLE of the conditional variance of Y_j given $\bar{\boldsymbol{Y}}_{j-1}$ under normality and MAR. The s_p's represent the shape parameters of the underlying Γ distributions, which we give $\Gamma^{-1}(1,1)$ priors. $\lambda_2 g_j$ and λ_3 represent a random scaling component for $\sigma_{a_j}^2$ and $\sigma_{\phi_j}^2$.

C Simulation Study Details

C.1 Section 4.1

In the first simulation setting in Section 4.1 data was generated according to $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (0, 0, 0)$ and $\boldsymbol{\Sigma}$ an AR-1 covariance matrix with $\text{Var}(Y_1) = 1$ and $\text{Cov}(Y_j, Y_{j+1}) = 0.7$. Missingness is MAR with discrete hazard at times $j = 1$ and $j = 2$ given by $\lambda_j(\boldsymbol{Y}) = P(S = j \mid S \geq j, \boldsymbol{Y}) = \sigma(a_j + b_j Y_{j-1})$ where $\sigma(x) = [1 + e^{-x}]^{-1}$. The values of a_1 and b_1 were chosen so that $\lambda_1(-2) = 0.5$ and $P(S = 1) = 0.2$. a_2 and b_2 were chosen so that $\lambda_2(-2) = 0.5$ and $P(S = 2 \mid S \geq 1) = .25$.

In the second simulation setting in Section 4.1, $\boldsymbol{Y}$ was drawn from a 50-50 mixture of normal distributions with means $\boldsymbol{\mu}_1 = (2, 0, -2)$, $\boldsymbol{\mu}_2 = (6, 1.5, 0)$ and covariance matrices $\boldsymbol{\Sigma}_1 = \text{diag}(2, .1, .2)$ and $\boldsymbol{\Sigma}_2$ exchangeable with variance 1 and covariance 0.8. This was chosen to make the distributions of (Y_1, Y_2) and (Y_1, Y_3) roughly shaped like an "L" rotated by 90 degrees while (Y_2, Y_3) is roughly linear. Missingness is MAR with

$$\lambda_1(\boldsymbol{Y}) = 0.4I(Y_1 \leq 2) + 0.18I(2 < Y_1 \leq 5) + 0.1I(Y_1 > 5),$$
$$\lambda_2(\boldsymbol{Y}) = 0.45I(Y_2 \leq 0) + 0.2I(0 < Y_2 \leq 2) + 0.1I(Y_2 > 2).$$

$I(Y \in A)$ here denotes the indicator function. The hazards were chosen so that $P(S = 1) \approx P(S = 2) \approx 0.2$.

C.2 Section 4.2

Parameters under M1 are

$$\boldsymbol{\mu} = (95.5, 94.1, 91.6, 89.0, 86.2, 81.3)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 114.0 & & & & & \\ 98.5 & 143.2 & & & & \\ 101.5 & 149.0 & 222.7 & & & \\ 115.3 & 156.6 & 225.0 & 335.1 & & \\ 119.8 & 145.6 & 220.6 & 355.8 & 444.3 & \\ 118.3 & 142.1 & 210.9 & 337.0 & 420.2 & 441.6 \end{pmatrix}$$

$$\boldsymbol{\zeta} = (-16.4, 0.7, -11.5, -9.9, -27.6)^T,$$

$$\boldsymbol{\gamma}_1 = (-0.1, -0.0, 0.2, 0.4, 0.4)^T,$$

$$\boldsymbol{\gamma}_2 = (\text{Not Applicable}, -0.0, -0.1, -0.4, -0.1).$$

These parameters come from fitting the selection model to the data and taking the posterior mean of each parameter. M2 is a 5 component mixture that was obtained by fitting a Dirichlet mixture of lag-1 selection models and taking the parameters corresponding to the 5 components of highest posterior probability (we do not take posterior means

because the likelihood is invariant under permutations of component labels).

$$\boldsymbol{\beta} = (0.119, 0.578, 0.001, 0.115, 0.186)^T,$$

$$\boldsymbol{\zeta} = \begin{pmatrix} -9.58 & -10.65 & -9.25 & -9.86 & -9.42 \\ -9.45 & -10.31 & -9.55 & -9.08 & -9.72 \\ -9.61 & -9.46 & -9.77 & -8.79 & -9.80 \\ -9.51 & -9.08 & -10.45 & -9.33 & -9.44 \\ -9.03 & -8.91 & -10.19 & -9.21 & -9.58 \end{pmatrix},$$

$$\boldsymbol{\gamma} = \begin{pmatrix} -0.18 & -0.56 & -0.09 & -0.54 & 0.10 \\ -0.45 & -0.77 & -0.63 & -0.25 & -0.34 \\ -1.02 & -0.16 & -0.41 & -0.28 & -0.57 \\ -0.90 & 0.11 & -0.11 & -0.39 & 0.11 \\ -0.67 & -0.16 & 0.11 & 0.13 & 0.09 \end{pmatrix},$$

$$\boldsymbol{\mu} = \begin{pmatrix} 98.11 & 95.94 & 91.85 & 91.68 & 100.94 & 75.71 \\ 95.58 & 93.09 & 89.34 & 82.97 & 78.82 & 76.86 \\ 75.95 & 64.27 & 63.69 & 59.69 & 57.83 & 34.60 \\ 85.11 & 83.13 & 72.67 & 67.32 & 64.35 & 61.12 \\ 97.49 & 99.48 & 101.83 & 107.23 & 94.34 & 104.05 \end{pmatrix}.$$

Each row of a given matrix corresponds to a mixture component. The covariance matrices for each class are given by

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 245 & 222 & 207 & 199 & 195 & 183 \\ 222 & 268 & 250 & 240 & 235 & 220 \\ 207 & 250 & 270 & 259 & 254 & 237 \\ 199 & 240 & 259 & 295 & 288 & 269 \\ 195 & 235 & 254 & 288 & 374 & 349 \\ 183 & 220 & 237 & 269 & 349 & 354 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 72 & 68 & 64 & 61 & 56 & 55 \\ 68 & 117 & 110 & 105 & 98 & 94 \\ 64 & 110 & 174 & 166 & 155 & 147 \\ 61 & 105 & 166 & 194 & 182 & 172 \\ 56 & 98 & 155 & 182 & 199 & 187 \\ 55 & 94 & 147 & 172 & 187 & 212 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 120 & 111 & 104 & 98 & 97 & 89 \\ 111 & 156 & 146 & 138 & 136 & 125 \\ 104 & 146 & 194 & 184 & 179 & 165 \\ 98 & 138 & 184 & 239 & 231 & 211 \\ 97 & 136 & 179 & 231 & 274 & 250 \\ 89 & 125 & 165 & 211 & 250 & 251 \end{pmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{pmatrix} 73 & 69 & 65 & 63 & 59 & 52 \\ 69 & 122 & 115 & 112 & 105 & 94 \\ 65 & 115 & 188 & 182 & 171 & 153 \\ 63 & 112 & 182 & 227 & 214 & 191 \\ 59 & 105 & 171 & 214 & 220 & 197 \\ 52 & 94 & 153 & 191 & 197 & 221 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_5 = \begin{pmatrix} 106 & 100 & 96 & 91 & 83 & 84 \\ 100 & 125 & 119 & 113 & 104 & 105 \\ 96 & 119 & 167 & 159 & 147 & 148 \\ 91 & 113 & 159 & 360 & 335 & 336 \\ 83 & 104 & 147 & 335 & 337 & 339 \\ 84 & 105 & 148 & 336 & 339 & 367 \end{pmatrix}.$$

To generate from model M3 we first generate data under M1 and apply the appropriate normal distribution function to each component to get data which is marginally uniform. Next we apply the skew-t quantile function to each component to get data

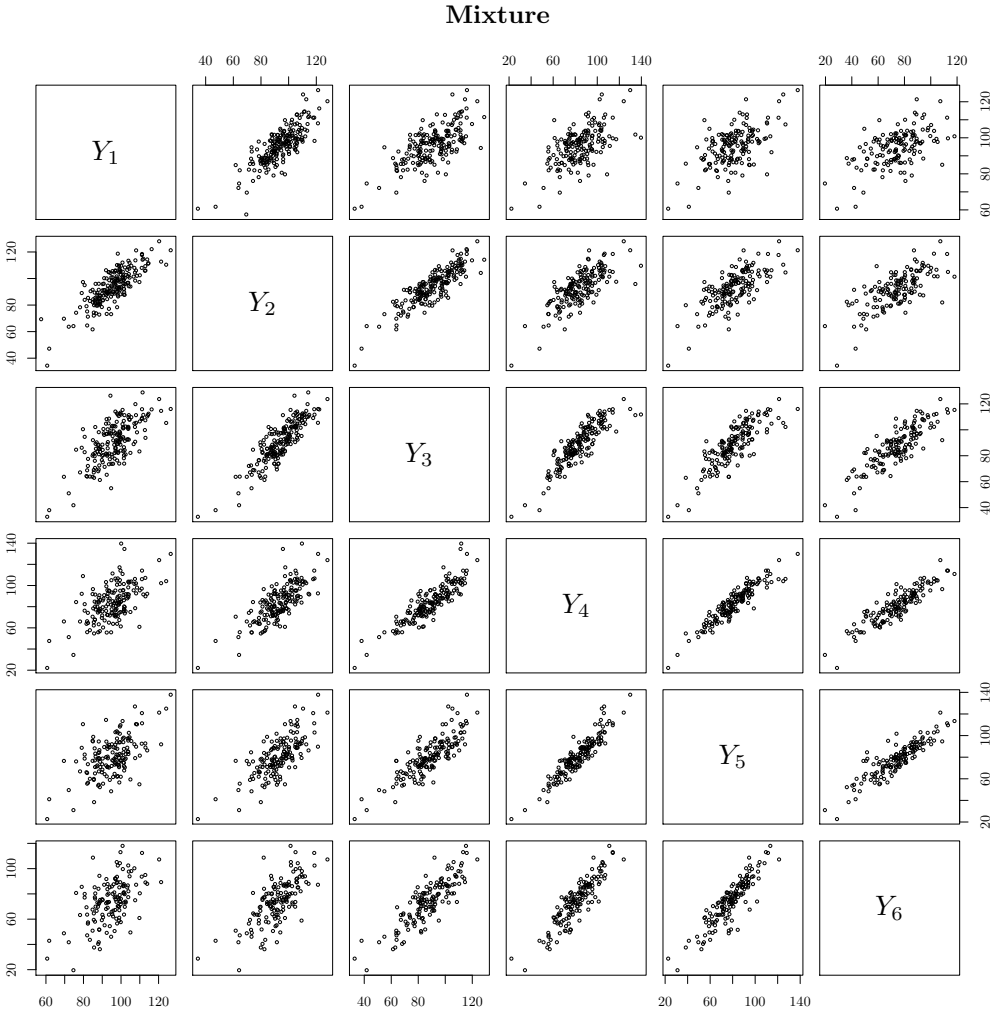

Figure 1: Dataset generated under M2.

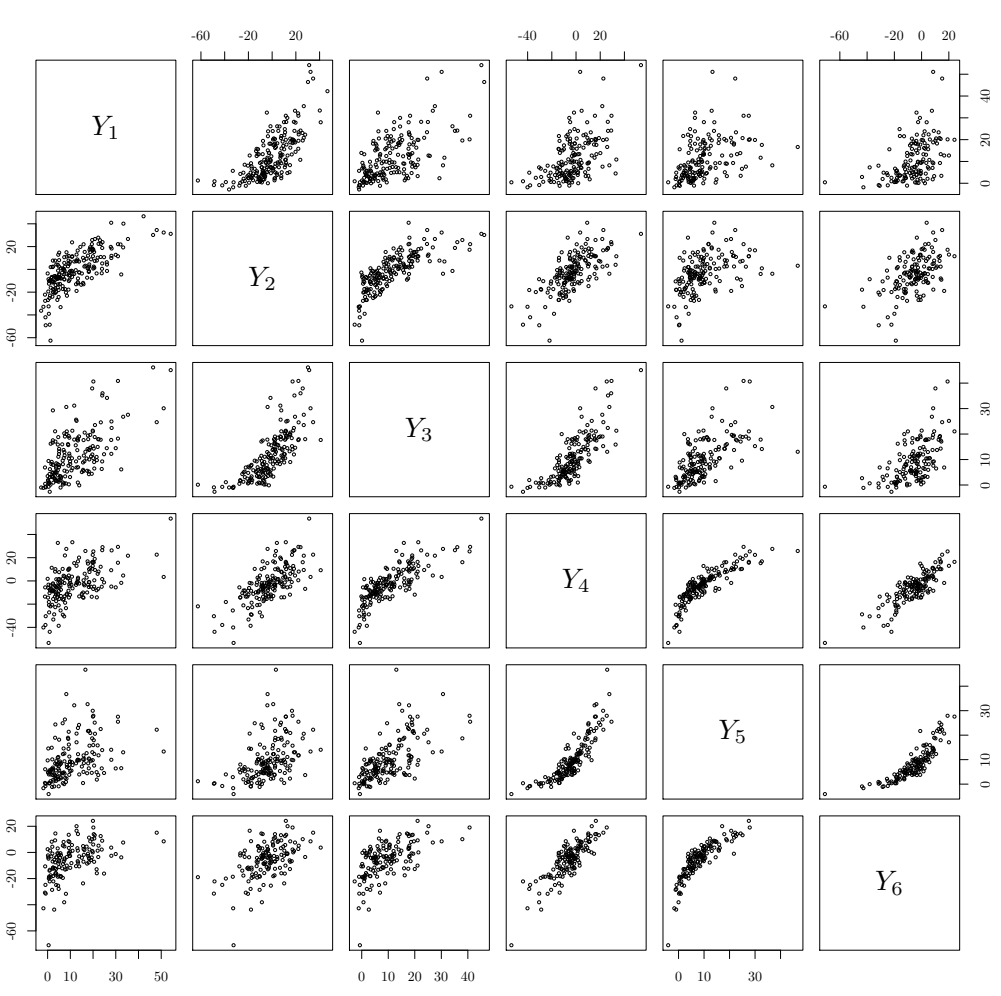

Figure 2: Dataset generated under M3.

Normal Lag-2 Selection Model

	95% CI Width		Coverage Probability		Root Mean Squared Error	
ξ	Normal	Dirichlet	Normal	Dirichlet	Normal	Dirichlet
0	6.80(0.46)	6.54(0.41)	93.7(1.4)	92.0(1.6)	1.73(0.07)	1.71(0.07)
0.5	7.55(0.53)	7.24(0.48)	94.3(1.3)	93.0(1.3)	1.93(0.08)	1.92(0.07)
1	8.54(0.61)	8.18(0.58)	94.3(1.3)	93.3(1.5)	2.19(0.09)	2.16(0.08)
1.5	9.70(0.69)	9.34(0.66)	93.7(1.4)	93.7(1.4)	2.50(0.10)	2.55(0.09)
2	10.99(0.78)	10.62(0.73)	94.0(1.4)	93.7(1.4)	2.83(0.12)	2.86(0.11)

Mixture of Lag-1 Selection Models

	95% CI Width		Coverage Probability		Root Mean Squared Error	
ξ	Normal	Dirichlet	Normal	Dirichlet	Normal	Dirichlet
0	5.8(0.4)	5.9(0.4)	91.3(1.6)	95.3(1.2)	1.65(0.07)	1.40(0.064)
0.5	6.2(0.4)	6.3(0.4)	92.0(1.6)	95.3(1.2)	1.76(0.07)	1.51(0.068)
1	6.8(0.5)	6.8(0.5)	90.7(1.7)	95.3(1.2)	2.00(0.08)	1.67(0.076)
1.5	7.6(0.6)	7.6(0.7)	90.3(1.7)	95.3(1.2)	2.28(0.09)	1.86(0.085)
2	8.6(0.7)	8.5(0.8)	91.7(1.6)	95.3(1.2)	2.51(0.10)	2.10(0.095)

Skew-T Copula Lag-2 Selection Model

	95% CI Width		Coverage Probability		Root Mean Squared Error	
ξ	Normal	Dirichlet	Normal	Dirichlet	Normal	Dirichlet
0	5.4(0.4)	5.5(0.9)	89.9(1.7)	95.6(1.2)	1.62(0.07)	1.35(0.06)
0.5	5.9(0.5)	6.0(1.1)	91.2(1.6)	96.6(1.0)	1.69(0.07)	1.51(0.06)
1	6.5(0.6)	6.7(1.3)	88.9(1.8)	95.3(1.2)	2.03(0.09)	1.65(0.07)
1.5	7.3(0.7)	7.6(1.6)	88.6(1.8)	94.9(1.3)	2.34(0.10)	1.86(0.08)
2	8.4(0.8)	8.9(1.9)	86.9(2.0)	96.0(1.1)	2.35(0.12)	2.10(0.06)

Table 1: Results from the simulation study in Section 4.2. Normal refers to M1, Mixture to M2, and Skew-T to M3.

which is marginally skew-t. Recall the density of the skew-t distribution (Azzalini, 2013) with location 0, scale 1, degrees of freedom ν, and shape ω is

$$f(z \mid \nu, \omega) = 2t_\nu(z)T_{\nu+1}\left(\omega z\sqrt{\frac{\nu+1}{z^2+\nu}}\right),$$

where t_ν is the students-t density with ν degrees of freedom and $T_{\nu+1}$ is the students-t distribution function with $\nu+1$ degrees of freedom. We set $\nu = 15$ for each component and $\omega = (10, 0, 10, 0, 10, 0)$ to induce a nonlinear relationship between components. The data were then returned approximately to their original scale by multiplying by 15. Sample datasets of data generated under M2 and M3 are given in Figures 1 and 2. Detailed simulation results are given in Table 1.

References

Azzalini, A. (2013). *The Skew-Normal and Related Families*, volume 3. Cambridge University Press.

Daniels, M. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

Neal, R. (2003). Slice sampling. *The Annals of Statistics*, 31:705–767.

Rasmussen, C. (2000). The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12.

Taddy, M. (2008). *Bayesian Nonparametric Analysis of Conditional Distributions and*

Inference for Poisson Point Processes. PhD thesis, University of California, Santa Cruz.