Bayesian multivariate skew-normal finite mixture model for analysis of infant development trajectories

Carter Allen

Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, U.S.A

email: allecart@musc.edu

and

Brian Neelon, PhD

Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, U.S.A

and

Sara Benjamin-Neelon, PhD, MPH, RD

Department of Health, Behavior and Society, Johns Hopkins University, Baltimore, MD, U.S.A

SUMMARY: In studies of infant motor development, a crucial research goal is the identification of latent clusters of infants who experience delayed motor development, as this is a known risk factor for adverse outcomes later in life. However, there are a number of statistical challenges in modeling infant motor development: the data are typically skewed, exhibit intermittent missingness, and are highly correlated across the repeated measurements collected during infancy. Using data from the Nurture study, a cohort of over 600 mother-infant pairs followed from pregnancy to 12 months postpartum, we develop a flexible Bayesian finite mixture model for the analysis infant motor development. Our model has a number of attractive features. First, we adopt the multivariate skew normal distribution with cluster-specific parameters that accommodate the inherent correlation and skewness in the data. Second, we model the cluster membership probabilities using a novel application of the Pólya-Gamma data-augmentation scheme, thereby improving predictions of the cluster membership allocations. Lastly, we impute missing responses under the missing at random assumption by drawing from appropriate conditional multivariate skew normal distributions. Bayesian inference is achieved through straightforward Gibbs sampling, and can be implemented in available software such as R. Through simulation studies, we show that the proposed model yields improved inferences over models that ignore skewness. In addition, our imputation method yields improvements compared to conventional missing data methods, including multiple imputation and complete or available case analysis. When applied to Nurture data, we identified two distinct development clusters: one characterized by delayed U-shaped motor development and a higher percentage

of male infants and another characterized by more steady motor development and a lower percentage of males. The clusters also differed in terms of key demographic variables, such as infant race and maternal pre-pregnancy body mass index. These findings can aid investigators in targeting interventions during this critical early-life developmental window.

CONTENTS

1. Introduction

1.1 *Infant Development Clustering*

Heterogeneity of treatment effects (HTE) (Lanza and Rhoades, 2013).

1.2 *Existing Approaches*

Mixtures of multivariate non-symmetric distributions such as the multivariate skew-normal (MSN) distribution allow for the nuances of the marginal density to be captured with a more parsimonious set of mixture components. Mixtures of MSN distributions have been dealt with previously in a Bayesian context (Frühwirth-Schnatter & Pyne, 2010 and others), however in these models, focus lies primary on marginal density estimation, and inference on the mixture components (i.e. clusters) is often not discussed. More recently, the mixtures of skew-t factor analysis (MSTFA) model has been proposed for settings in which cluster-specific inference is of primary interest (Lin *et al.* 2018). However, an important feature not included in the MSTFA is the ability to explain individual-level cluster membership as a function of covariates of interest. Additionally, the parameter estimation procedure proposed by Lin et al. for the MSTFA relies on a prohibitively complex EM algorithm and does not enjoy the inferential benefits of a Bayesian approach, including the ability to incorporate prior information into a model and make posterior probability statements. Our proposed model improves on these previous works by estimating parameters in a Bayesian framework as well as including the ability to fit a multinomial logit regression to cluster membership probabilities using a novel application of data augmentation with the Pólya-Gamma distribution.

Polson et al. (2013) introduce a data augmentation scheme using the Pólya-Gamma distribution which allows for sampling of multinomial regression parameters using straightforward Gibb's updates from Gaussian full conditional distributions. In addition to more convenient parameter estimation, the Pólya-Gamma data augmentation method for logistic regression

has the advantage of direct sampling from the posterior distributions of multinomial parameters. This approach avoids the need for approximations of the posterior distribution, thus yielding more stable sampling, especially when the number of parameters approaches the number of observations (Polson et al., 2013). Pólya-Gamma data augmentation for multinomial regression has not yet been applied to the analysis of longitudinally clustered data.

A ubiquitous feature of repeated measures studies is loss of data due to intermittent missingness and attrition. In the Bayesian setting, the standard approach to dealing with missing data is to perform multiple imputation, whereby m imputed data sets are generated from a specified imputation model. After m complete data sets are obtained, parameter estimates are combined across each data set to produce a final set of parameter estimates (Gelman *et al.* 2013). This approach is not only computationally burdensome, requiring storage and analysis of an $m \times n_{rows} \times n_{cols}$ data array in addition to multiplication of total model run time by a factor of m, but it has been shown to produce unreliable inferences (Zhou and Reiter, 2010). We instead include an "online" imputation imputation step in our Gibbs sampling procedure, whereby missing outcomes are updated at each iteration. This approach greatly increases the number of opportunities for exploration of the missing data parameter space and avoids the multiplication of total run time and number of parameters.

2. Nurture Study

2.1 *Baseline Demographics and Description of Variables*

2.2 *Statistical Challenges*

The analysis of infant motor development trends in the Nurture data presents a number of statistical challenges that motivate our proposed model. First, as depicted in Figure 1, the residuals from repeated measures models of Bayley composite scores exhibit skewness even after adjusting for covariates such as race, sex, and birthweight. This suggests that the assumption of conditional normality made by standard repeated measures models is violated, and a distinguishing feature of the data, skewness, is not being accounted for.

[Figure 1 about here.]

The Nurture data also feature intermittent missingness in Bayley composite scores throughout the study period. Of the total cohort (N = 666), 429 (64.4 %) observations were available at three months, 435 (65.3 %) observations were available at six months, 418 (62.8 %) observations were available at nine months, and 437 (65.6 %) observations were available at twelve months. As such, we require a modeling framework capable of dealing with missing data.

3. Model

3.1 *Multivariate Skew Normal Mixture Model*

A primary goal of the Nurture study is to identify clusters of infants characterized by distinct motor development trajectories throughout the first year of life. To address this aim, we propose a flexible finite mixture model that accommodates relevant features of the data, such as skewness, missing values, and dependence among the responses. For $i = 1, ..., n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$ be a $J \times 1$ vector of responses (i.e., Bayley composite scores) for subject i across the J measurement occasions. For analysis of the Nurture data, we propose a finite

mixture model of the form

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ki} f(\mathbf{y}_i | \boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k ($k = 1, ..., K$) and π_{ki} is a subject-specific mixing weight representing the probability that subject i belongs to cluster k. For now we assume that K is fixed; in Section 3.4.3, we discuss model selection strategies for choosing the optimal value of K.

To facilitate posterior inference, we introduce a latent cluster indicator variable z_i taking the value $k \in \{1, ..., K\}$ with probability π_{ki}. Conditional on $z_i = k$, we assume $\mathbf{y}_i$ is distributed as

$$\mathbf{y}_i | (z_i = k) \sim \text{MSN}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k), \tag{2}$$

where $\text{MSN}_J(\cdot)$ denotes the J-dimensional multivariate skew normal density, $\boldsymbol{\zeta}_{ki}$ is a $J \times 1$ vector of subject- and cluster-specific location parameters, $\boldsymbol{\alpha}_k$ is a $J \times 1$ vector of cluster-specific skewness parameters, and $\boldsymbol{\Omega}_k$ is a $J \times J$ cluster-specific scale matrix that captures dependence among the J responses for subject i. The vector $\boldsymbol{\alpha}_k$ has components α_{kj}, $j = 1, ..., J$, that control the skewness of outcome j in cluster k. When $\boldsymbol{\alpha}_k = \mathbf{0}$, the MSN distribution reduces to the multivariate normal distribution $\text{N}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k)$, where $\boldsymbol{\zeta}_{ki}$ is a $J \times 1$ mean vector and $\boldsymbol{\Omega}_k$ is a $J \times J$ covariance matrix.

We can extend model (2) to the regression setting by modeling $\boldsymbol{\zeta}_{ki}$ as a function of covariates. Here we adopt a convenient stochastic representation of the MSN density (Azzalini and Dalla Valle, 1996):

$$\mathbf{y}_i | (z_i = k, t_i) = \mathbf{X}_i \boldsymbol{\beta}_k + t_i \boldsymbol{\psi}_k + \boldsymbol{\epsilon}_{ki}, \tag{3}$$

where $\mathbf{X}_i$ is a $J \times Jp$ design matrix that includes potential time-varying covariates (e.g., indicators denoting quarterly visits); $\boldsymbol{\beta}_k = (\beta_{k11}, \ldots, \beta_{k1p}, \ldots, \beta_{kJ1}, \ldots, \beta_{kJp})^T$ is a $Jp \times 1$ vector of cluster- and outcome-specific regression coefficients; $t_i \sim \text{N}_{[0,\infty)}(0, 1)$ is a subject-specific standard normal random variable truncated below by zero; $\boldsymbol{\psi}_k = (\psi_{k1}, \ldots, \psi_{kJ})^T$ is

a $J \times 1$ vector of cluster-specific skewness parameters; and $\boldsymbol{\epsilon}_{ki} \sim \mathrm{N}_J(\mathbf{0}, \boldsymbol{\Sigma}_k)$ is a $J \times 1$ vector

of correlated error terms. Thus, conditional on t_i and $z_i = k$, $\boldsymbol{y}_i$ is distributed as $\mathrm{N}_J(\mathbf{X}_i\boldsymbol{\beta}_k +$

$t_i\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$. Marginally (after integrating over t_i), $\mathbf{y}_i|(z_i = k)$ is distributed $\mathrm{MSN}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$,

where through back-transformation

$$
\begin{aligned}
\boldsymbol{\zeta}_{ki} &= \mathbf{X}_i\boldsymbol{\beta}_k, \\
\boldsymbol{\alpha}_k &= \frac{1}{\sqrt{1 - \boldsymbol{\psi}_k^T\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k}}\boldsymbol{\omega}_k\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k, \quad \text{and} \\
\boldsymbol{\Omega}_k &= \boldsymbol{\Sigma}_k + \boldsymbol{\psi}_k\boldsymbol{\psi}_k^T,
\end{aligned}
\tag{4}
$$

where $\boldsymbol{\omega}_k = \mathrm{Diag}(\boldsymbol{\Omega}_k)^{1/2}$ is the $J \times J$ diagonal matrix containing the square root of the

diagonal entries of $\boldsymbol{\Omega}_k$. Additional details can be found in Früwirth-Schnatter and Pyne

(2010).

Of note, the MSN density can be expressed more compactly in terms of the matrix skew

normal (MatSN) density (Chen and Gupta 2005). Let $\mathbf{Y}_k$ be an $n_k \times J$ response matrix with

rows $\mathbf{y}_i^T$, $(i = 1, ..., n_k)$, where $n_k = \sum_{i=1}^n \mathbb{1}_{(z_i=k)}$ is the number of observations in cluster k.

From equation (3), it follows that $\mathbf{Y}_k$ is distributed as

$$
\begin{aligned}
\mathbf{Y}_k &\sim \mathrm{MatSN}_{n_k \times J}(\mathbf{M}_k, \boldsymbol{\alpha}_k, \mathbf{I}_{n_k}, \boldsymbol{\Omega}_k) \\
\mathrm{vec}(\mathbf{M}_k) &= (\boldsymbol{\zeta}_{k1}^T, ..., \boldsymbol{\zeta}_{kn_k}^T)^T,
\end{aligned}
\tag{5}
$$

where $\mathbf{M}_k$ is an $n_k \times J$ location matrix with rows $\boldsymbol{\zeta}_{ki} = \mathbf{X}_i\boldsymbol{\beta}_k$ as in equation (3), $\boldsymbol{\alpha}_k =$

$(\alpha_{k1}, ..., \alpha_{kJ})^T$, $\mathbf{I}_{n_k}$ is the $n_k \times n_k$ identity matrix, and $\boldsymbol{\Omega}_k$ is the $J \times J$ scale matrix defined

above in equation (2). From equation (3), it follows that $\mathbf{Y}_k$, conditional on the $n_k \times 1$ vector

of random effects $\mathbf{t}_k = (t_1, ..., t_{n_k})^T$, is jointly distributed in matrix form as

$$
\mathbf{Y}_k|\mathbf{t}_k \sim \mathrm{MatNorm}_{n_k \times J}(\mathbf{M}_k^*, \mathbf{I}_{n_k}, \boldsymbol{\Sigma}_k),
\tag{6}
$$

where $\mathrm{MatNorm}_{n_k \times J}(\cdot)$ denotes a $n_k \times J$ matrix normal density, $\mathbf{M}_k^*$ is an $n_k \times J$ matrix such

that $\mathrm{vec}(\mathbf{M}_k^*) = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{t}_k \otimes \boldsymbol{\psi}_k$ is an $n_kJ \times 1$ mean vector, $\mathbf{X}_k$ is an $n_kJ \times Jp$ design matrix,

$\boldsymbol{\beta}_k$ is the $Jp \times 1$ vector of regression coefficients defined in equation (3), and $\boldsymbol{\Sigma}_k$ is the $J \times J$

conditional covariance of $\boldsymbol{\epsilon}_{ik}$ given in equation (3). As described in Section 3.6, the matrix representation of the MSN distribution admits convenient conjugate prior distributions for the regression parameters and scale matrices, which in turn leads to efficient Gibbs sampling for posterior inference.

3.2 *Multinomial Regression on Cluster Indicators*

To accommodate heterogeneity in the cluster-membership probabilities, we model π_{ki} as a function of covariates using a multinomial logit model

$$\pi_{ki} = \Pr(z_i = k | \mathbf{w}_i) = \frac{e^{\mathbf{w}_i^T \boldsymbol{\delta}_k}}{\sum_{h=1}^{K} e^{\mathbf{w}_i^T \boldsymbol{\delta}_h}}, \ \ k = 1, \ldots, K, \tag{7}$$

where $\mathbf{w}_i$ is an $r \times 1$ vector of subject-level covariates, $\boldsymbol{\delta}_k$ is a $r \times 1$ vector of regression parameters associated with membership in cluster k. For identifiability purposes, we fix the reference category $k = K$ and set $\boldsymbol{\delta}_K = \mathbf{0}$. Under model (7), $z_i | \boldsymbol{\pi}_i \sim \text{Multinomial}(1, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{1i}, ..., \pi_{Ki})$. As detailed below in Section 3.4.2, each cluster label z_i is updated during MCMC estimation from its multinomial full conditional distribution and used in the remaining MCMC steps as the working cluster assignment for subject i. By allowing the cluster probabilities to vary across subjects, model (1) can be viewed as a *mixture of experts* model, in which π_{ki} acts as a *gating function* controlling the prior probability of membership in cluster k, and $f(\mathbf{y}_i | \boldsymbol{\theta}_k)$ in equation (1) is the "expert" providing information on the within-cluster distribution of $\mathbf{y}_i$ (Bishop 2006).

To facilitate sampling, we adopt the efficient data-augmentation approach introduced by Polson et al. (2013), which expresses the inverse-logit function as a mixture Pólya–Gamma densities. By using Pólya–Gamma data augmentation for the multinomial model, we obtain a *Pólya–Gamma mixture of experts model* – a computationally efficient way to obtain inferences for the mixing weights in the Bayesian setting. A random variable w is said to follow a Pólya–

Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$ if

$$w \sim \text{PG}(b, c) \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{s=1}^{\infty} \frac{g_s}{(s - 1/2)^2 + c^2/(4\pi^2)}, \tag{8}$$

where $g_s \stackrel{iid}{\sim} \text{Ga}(b, 1)$ for $s = 1, ..., \infty$. Polson et al. establish that for $a, \eta \in \mathbb{R}$,

$$\frac{(e^\eta)^a}{(1 + e^\eta)^b} = 2^{-b} e^{\kappa \eta} \int_0^\infty e^{-w\eta^2/2} p(w|b, c = 0) dw, \tag{9}$$

where $\kappa = a - b/2$ and $p(w|b, c = 0)$ denotes a $\text{PG}(b, 0)$ density. Polson et al. further show that the conditional distribution $p(w|b, c)$ results from an "exponential tilting" of the $\text{PG}(b, 0)$ density, thus

$$p(w|b, c) = \frac{e^{-c^2 w/2} p(w|b, 0)}{E_w[e^{-c^2 w/2}]} = \frac{e^{-c^2 w/2} p(w|b, 0)}{\int_0^\infty e^{-c^2 w/2} p(w|b, 0) dw}. \tag{10}$$

Polson et al. use these results to show that, for the logistic model, the Bernoulli likelihood can be written as a scale-mixture of normal densities with Polya-Gamma precision terms, w, resulting in a normal full conditional distribution for the logistic regression parameters. To extend the Pólya–Gamma data augmentation approach to the multinomial setting, we first introduce the binary indicators U_{ki}, such that $U_{ki} = \mathbb{1}_{(z_i = k)}$ – i.e., U_{ki} is an indicator variable taking the value 1 if subject i belongs to cluster k, and 0 otherwise. The full conditional distribution of $\boldsymbol{\delta}_k$ is then given as

$$p(\boldsymbol{\delta}_k|\mathbf{U}_k, \boldsymbol{\delta}_{h \neq k}) \propto p(\boldsymbol{\delta}_k) \prod_{i=1}^n \pi_{ki}^{U_{ki}} (1 - \pi_{ki})^{1 - U_{ki}},$$

where $p(\boldsymbol{\delta}_k)$ denotes the prior distribution of $\boldsymbol{\delta}_k$, and π_{ki} is defined as in equation (7). We can rewrite π_{ki} in terms of U_{ki} as

$$\pi_{ki} = P(U_{ki} = 1) = \frac{e^{\mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}}}{1 + e^{\mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}}} = \frac{e^{\eta_{ki}}}{1 + e^{\eta_{ki}}},$$

where $c_{ki} = \log \sum_{h \neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$ and $\eta_{ki} = \mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}$. We note that the sum $\sum_{h \neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$ includes

the reference category, but since we fix $\boldsymbol{\delta}_K = \mathbf{0}$, we have $e^{\mathbf{w}_i^T \boldsymbol{\delta}_K} = 1$, and hence

$$c_{ki} = \log \sum_{h \neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_h} = \log \left(1 + \sum_{h \notin \{k,K\}} e^{\mathbf{w}_i^T \boldsymbol{\delta}_h} \right)^T.$$

We can use these quantities to re-express the full conditionals for $\boldsymbol{\delta}_k$ as

$$\begin{aligned} p(\boldsymbol{\delta}_k | \mathbf{U}_k, \boldsymbol{\delta}_{h \neq k}) & \propto & p(\boldsymbol{\delta}_k) \prod_{i=1}^{n} \left(\frac{e^{\eta_{ki}}}{1 + e^{\eta_{ki}}} \right)^{U_{ki}} \left(\frac{1}{1 + e^{\eta_{ki}}} \right)^{1 - U_{ki}} \\ & = & p(\boldsymbol{\delta}_k) \prod_{i=1}^{n} \frac{(e^{\eta_{ki}})^{U_{ki}}}{1 + e^{\eta_{ki}}}, \end{aligned} \tag{11}$$

which is a logistic likelihood. We can thus apply the Pólya–Gamma sampler described by Polson et al. for logistic regression to update each $\boldsymbol{\delta}_k$ one at a time based on the binary indicators U_{ki}. To do so, we first define for $k = 1, ..., K$, the $n \times 1$ vector $\mathbf{U}_k^* = \left(\frac{U_{k1} - 1/2}{w_{k1}}, ..., \frac{U_{kn} - 1/2}{w_{kn}} \right)^T$. Polson et al. show that, conditional on $\boldsymbol{w} = (w_{k1}, ..., w_{kn})^T$, $\mathbf{U}_k^*$ follows a $N_n(\boldsymbol{\eta}_k, \mathbf{O}_k^{-1})$ distribution with mean $\boldsymbol{\eta}_k = (\eta_{k1}, ..., \eta_{kn})^T$ and precision matrix $\mathbf{O}_k = \text{Diag}(w_{k1}, ..., w_{kn})$. Thus, it follows that the full conditional distribution of $\boldsymbol{\delta}_k$ is given by

$$p(\boldsymbol{\delta}_k | \mathbf{z}, \mathbf{W}, \mathbf{O}_k) \propto p(\boldsymbol{\delta}_k) \exp \left[-\frac{1}{2} (\mathbf{U}_k^* - \mathbf{W} \boldsymbol{\delta}_k)^T \mathbf{O}_k (\mathbf{U}_k^* - \mathbf{W} \boldsymbol{\delta}_k) \right]. \tag{12}$$

As detailed in Section 3.4, if we assume $p(\boldsymbol{\delta}_k)$ to be multivariate normal, then $\boldsymbol{\delta}_k$ has a closed-form multivariate normal full conditional distribution that can be easily embedded within our proposed Gibbs sampling routine. For more details on the Pólya-Gamma Gibbs sampler for logistic and multinomial models, see Polson et al. (2013).

3.3 Conditional MSN Imputation

To accommodate intermittent missing at random (MAR) responses, we propose a convenient imputation algorithm that can be implemented "online" – that is, as part of the Gibbs sampler. In Section 6, we discuss extensions to allow for non-ignorable missingness (i.e., observations missing not at random). Suppose $\mathbf{y}_i$ has $q_i \in (1, \ldots, J)$ observed values, denoted $\mathbf{y}_i^{obs}$, and $J - q_i$ intermittent missing values, denoted $\mathbf{y}_i^{miss}$. We can make use of the stochastic representation given in equation (3) to impute $\mathbf{y}_i^{miss}$ from its conditional multivariate normal

distribution given $(z_i, t_i, \mathbf{y}_i^{obs})$:

$$
\begin{aligned}
\mathbf{y}_i^{miss}|(z_i = k, t_i, \mathbf{y}_i^{obs}) \quad &\sim \quad \mathrm{N}_{J-q_i}(\boldsymbol{\mu}_{ki}^{cond}, \boldsymbol{\Sigma}_k^{cond}), \text{ where} \\
\boldsymbol{\mu}_{ki}^{cond} \quad &= \quad \boldsymbol{\mu}_{ki}^{miss} + \boldsymbol{\Sigma}_{k12}\boldsymbol{\Sigma}_{k22}^{-1}(\mathbf{y}_i^{obs} - \boldsymbol{\mu}_{ki}^{obs}) \\
\boldsymbol{\Sigma}_k^{cond} \quad &= \quad \boldsymbol{\Sigma}_{k11} - \boldsymbol{\Sigma}_{k12}\boldsymbol{\Sigma}_{k22}^{-1}\boldsymbol{\Sigma}_{k21}, \qquad\qquad (13) \\
\boldsymbol{\mu}_{ki} \quad &= \quad \begin{pmatrix} \boldsymbol{\mu}_{ki}^{miss} \\[2mm] \boldsymbol{\mu}_{ki}^{obs} \end{pmatrix}, \text{ and} \\
\boldsymbol{\Sigma}_k \quad &= \quad \begin{pmatrix} \boldsymbol{\Sigma}_{k11} & \boldsymbol{\Sigma}_{k12} \\[2mm] \boldsymbol{\Sigma}_{k21} & \boldsymbol{\Sigma}_{k22} \end{pmatrix}, \text{ where}
\end{aligned}
$$

The location vector $\boldsymbol{\mu}_{ki}$ is defined as $\boldsymbol{\mu}_{ki} = \mathbf{X}_i\boldsymbol{\beta}_k + t_i\boldsymbol{\psi}_k$, and is partitioned into $\boldsymbol{\mu}_{ki}^{miss}$ and $\boldsymbol{\mu}_{ki}^{obs}$ with respect to the missing and observed indices of $\mathbf{y}_i$, respectively. The partition $\boldsymbol{\Sigma}_{k11}$ is a $(J-q_i) \times (J-q_i)$ matrix containing the rows and columns of $\boldsymbol{\Sigma}_k$ corresponding to $\mathbf{y}_i^{miss}$. Similarly, $\boldsymbol{\Sigma}_{k12}$ is a $(J-q_i) \times q_i$ matrix containing the rows of $\boldsymbol{\Sigma}_k$ that correspond to $\mathbf{y}_i^{miss}$, but columns of $\boldsymbol{\Sigma}_k$ that correspond to $\mathbf{y}_i^{obs}$. The remaining partitions $\boldsymbol{\Sigma}_{k21}$, and $\boldsymbol{\Sigma}_{k22}$ are defined in the same manner. These results follow from conventional multivariate normal theory. We note that, while conditional on t_i, $\mathbf{y}_i$ follows a MVN distribution, after marginalizing over t_i the vector $\mathbf{y}_i$ follows a MSN distribution. Thus, this proposed online conditional imputation method provides a convenient way of imputing MSN responses using samples from more standard densities.

An attractive practical feature of this imputation algorithm is that it avoids multiplicative run-time scaling in m, the number of imputations (Gelman et al. 2013; Zhou and Reiter, 2010). Our approach also provides more opportunities to explore the missing data parameter space than does multiple imputation, since each missing component is drawn once per MCMC iteration, and often in practice $n_{sim} >> m$, where n_{sim} is the total number of MCMC iterations (**find a reference**). In Section 4, we conduct simulation studies to demonstrate that imputing the missing MSN responses with our online conditional imputation method

improves inferences over standard Bayesian multiple imputation as outlined by Gelman et al. (2013).

3.4 *Bayesian Inference*

3.4.1 *Prior Specification.* We adopt a fully Bayesian inferential approach and assign prior distributions to all model parameters. Conveniently, all parameters admit conditionally conjugate priors, which greatly improves posterior computation via a data-augmented Gibbs sampler. To make use of the matrix normal representation introduced previously in Section 3, we define the matrix of regression parameters $\mathbb{B}_k^*$, where $\text{vec}(\mathbf{B}_k^*) = \boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k^T, \boldsymbol{\psi}_k^T)^T$. We assign $\mathbf{B}_k^*|\boldsymbol{\Sigma}_k$ a $\text{MatNorm}(\mathbf{B}_{0k}^*, \mathbf{I}_{p+1}, \boldsymbol{\Sigma}_k)$ prior, where $\mathbf{B}_{0k}^*$ is a matrix of location parameters such that $\text{vec}(\mathbf{B}_{0k}^*)$ is a vector of prior location parameters for the components of $\boldsymbol{\beta}_k^*$, $\mathbf{I}_{p+1}$ is the $(p+1)$-dimensional identity matrix, and $\boldsymbol{\Sigma}_k$ is the covariance matrix defined in equation (2), for which we specify an $\text{IW}(\mathbf{V}_{0k}, \nu_{0k})$ prior. This leads to a matrix-normal-inverse-Wishart joint prior for $\mathbb{B}_k^*$ and $\boldsymbol{\Sigma}_k$, which is the conjugate joint prior for the regression parameters in the matrix normal model given in equation (5). This conjugate prior specification induces convenient closed-form full conditional distributions that can be easily updated within our proposed Gibbs sampler.

For the multinomial logit model, the regression parameters $\boldsymbol{\delta}_k = (\delta_{k1}, ..., \delta_{kr})^T$ are assigned a $\text{N}_r(\mathbf{d}_{0k}, \mathbf{S}_{0k})$ prior for $k = 1, ..., K - 1$, which is conditionally conjugate under the Pólya-Gamma sampling scheme described in Section 3.2. We allow the normal-inverse-Wishart and multinomial hyperparameters (e.g., $\mathbf{B}_{0k}^*$ and $\mathbf{V}_{0k}$) to vary across clusters, though they may be shared across clusters in practice. An advantage of allowing for cluster-specific prior parameters is that *a priori* knowledge of motor development trends can be incorporated into certain clusters while allowing the priors for other clusters to be less informative. Additionally, prior information regarding the effect of certain covariates on cluster membership can be incorporated in to the model by choosing informative values for $\mathbf{d}_{0k}$ and $\mathbf{S}_{0k}$

3.4.2 *Posterior Inference.*   The above prior specification induces closed-form full conditionals for all model parameters, which can be efficiently updated as part of the Gibbs sampler outlined below. Additional details on derivations of full conditionals can be found in the Web Appendix A. We report MCMC diagnostics in Sections 4 and 5.

Step 1*: Conditional MSN Imputation.*   The sampler begins by imputing missing values $\mathbf{y}_i^{miss}$ conditional on current values of $z_i = k$ and t_i as well as the associated $\mathbf{y}_i^{obs}$ observed data vector. Specifically, for $i = 1, ..., n$, we draw $\mathbf{y}_i^{miss}$ from $\mathrm{N}_{J-q_i}(\boldsymbol{\mu}_{ki}^{cond}, \boldsymbol{\Sigma}_k^{cond})$ as described in equation (13). We conclude by constructing a complete outcome vector $\mathbf{y}_i$ that merges $\mathbf{y}_i^{miss}$ with $\mathbf{y}_i^{obs}$.

Step 2: *Update of MSN Regression Parameters.*   We begin the update of MSN regression parameters by first updating t_i, the truncated normal random effect used in the stochastic representation given in equation (3). For cluster k, we compute $A_k = (1 + \boldsymbol{\psi}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\psi}_k)^{-1}$ using current values of $\boldsymbol{\psi}_k$ and $\boldsymbol{\Sigma}_k$. Next, for $i = 1, ..., n_k$, we compute $a_i = A_k \boldsymbol{\psi}_k^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\zeta}_{ki})$, where $\boldsymbol{\zeta}_{ki} = \mathbf{X}_i \boldsymbol{\beta}_k$. Finally, we update t_i from $\mathrm{N}_{[0,\infty)}(a_i, \mathbf{A}_k)$. Repeat these updates for $k = 1, ..., K$.

The remaining MSN regression parameters are updated from their full conditionals as follows. First, we form the $n_k \times (p+1)$ matrix $\mathbf{X}_k^*$ by column-binding $\mathbf{X}_k$ and $\mathbf{t}_k$. For each cluster k ($k = 1, ..., K$), we update $\boldsymbol{\Sigma}_k$ from an $\mathrm{IW}(\nu_k, \mathbf{V}_k)$ density, where $\nu_k = \nu_{0k} + n_k$ and

$$\mathbf{V}_k \;\; = \;\; (\mathbf{Y}_k - \mathbf{X}_k^* \mathbf{B}_k^*)^T (\mathbf{Y}_k - \mathbf{X}_k^* \mathbf{B}_k^*) + (\mathbf{B}_k^* - \mathbf{B}_{0k}^*)^T (\mathbf{B}_k^* - \mathbf{B}_{0k}^*) + \mathbf{V}_{0k}.$$

We then make use of the matrix normal representation introduced in Section 3.2 to draw $\mathbf{B}_k^*$ from a $\mathrm{MatNorm}_{p+1,J}(\mathbb{B}_k^*, \mathbf{L}_k^*, \boldsymbol{\Sigma}_k)$ density, where

$$\mathbb{B}_k^* \;\; = \;\; \mathbf{L}_k^* (\mathbf{X}_k^{*T} \mathbf{Y}_k + \mathbf{I}_{p+1} \mathbf{B}_{0k}^*)$$

$$\mathbf{L}_k^* \;\; = \;\; (\mathbf{X}_k^{*T} \mathbf{X}_k^* + \mathbf{I}_{p+1})^{-1}.$$

For efficient sampling of the matrix normal distribution, we use the R package `matrixsampling` (Laurent 2018).

Step 3: *Pólya–Gamma Data Augmentation for z_i.* The sampler concludes with updates of the multinomial regression parameters $\boldsymbol{\delta}_k$, for $k = 1, ..., K-1$, followed by updates of each latent cluster indicator z_i $(i = 1, \ldots, n)$ from its multinomial logit full conditional. First, we define $U_{ki} = \mathbb{1}_{(z_i = k)}$ for $i = 1, ..., n$ and $k = 1, ..., K-1$. Next, we update w_{ki} from a $\mathrm{PG}(1, \eta_{ki})$ density, where $\eta_{ki} = \mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}$, and $c_{ki} = \log \sum_{h \neq k} \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$, using an efficient Pólya-Gamma sampling algorithm implemented in the `R` package `pgdraw` (Makalic and Schmidt, 2016). Next, define $U_{ki}^* = \frac{U_{ki} - 1/2}{w_{ki}}$ and let $\mathbf{U}_k^* = (U_{k1}^*, ..., U_{kn}^*)^T$. Finally, for $k = 1, ..., K-1$, update $\boldsymbol{\delta}_k$ from a $\mathrm{N}_r(\mathbf{d}_k, \mathbf{S}_k)$ density, where $\mathbf{S}_k = (\mathbf{S}_{k0} + \mathbf{W}^T \mathbf{O}_k \mathbf{W})^{-1}$, $\mathbf{O}_k = \mathrm{Diag}(w_{k1}, ..., w_{kn})$, $\mathbf{d}_k = \mathbf{S}_k(\mathbf{S}_{k0}\mathbf{d}_{k0} + \mathbf{W}^T \mathbf{O}_k \mathbf{U}_k^*)$, and $\mathbf{W}$ is the $n \times r$ matrix of multinomial logit regression covariates such that the ith row of $\mathbf{W}$ is $\mathbf{w}_i$.

Lastly, we update $z_1, ..., z_n$ by first computing $\boldsymbol{\pi}_i = (\pi_{1i}, ..., \pi_{Ki})$ as

$$\pi_{ki} = \frac{\mathrm{e}^{\mathbf{w}_i \boldsymbol{\delta}_k}}{1 + \sum_{h=1}^{K-1} \mathrm{e}^{\mathbf{w}_i \boldsymbol{\delta}_h}},$$

for $i = 1, ...n$ and $k = 1, ..., K$. We also compute, according to the multivariate normal density of $\mathbf{y}_i | t_i$, the probability $P(\mathbf{y}_i | \boldsymbol{\zeta}_{ki}^*, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\zeta}_{ki}^* = \mathbf{X}_i^* \boldsymbol{\beta}_k^*$. We use these quantities to compute $\boldsymbol{v}_i = (v_{1i}, ..., v_{Ki})^T$, where

$$v_{ki} = P(z_i = k | \mathbf{y}_i, \boldsymbol{\zeta}_{ki}^*, \boldsymbol{\Sigma}_k) = \frac{\pi_{ki} P(\mathbf{y}_i | \boldsymbol{\zeta}_{ki}^*, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^K \pi_{hi} P(\mathbf{y}_i | \boldsymbol{\zeta}_{hi}^*, \boldsymbol{\Sigma}_h)}.$$

The cluster labels z_i are then updated from a $\mathrm{Multinomial}(1, \boldsymbol{v}_i)$ density for $i = 1, ...n$. A schematic outline of the Gibbs sampler is given in the Web Appendix B. An `R` package for implementing the proposed model is currently in development. We provide `R` scripts for implementing the simulations and applications described below at `carter-allen.github.io/MVSN-FMM/`.

3.4.3 *Assessment of MCMC Convergence, Label Switching, and Model Selection.* We monitor convergence of the MCMC algorithm through the use of standard approaches such as trace plots and Geweke's (1992) Z-diagnostic, implemented in the `R` package `coda` (Plummer et al. 2006). In simulation studies under realistic parameter settings, we observed relatively fast convergence of all MCMC chains (i.e., within 500 iterations).

A common challenge for Bayesian mixture models is "label switching," in which draws of cluster-specific parameters may be associated with different cluster labels at various points during the MCMC simulation, rendering summaries of class-specific parameters incoherent. To address label switching, we implemented the *post hoc* ECR relabeling algorithm included in the `label.switching` package in `R` (Papastamoulis 2016). In simulation studies and application to the Nurture data, we observed fast convergence of the ECR relabeling algorithm, which is indicative of properly labeled parameter estimates throughout MCMC estimation.

Because our primary objective is to identify a small number of clinically meaningful motor development clusters, we use Bayesian model selection criteria to choose the optimal K from among a small number of possible values (e.g., $K = 1, ...4$). To this end, we propose the use of the "widely applicable information criterion" (WAIC) introduced by Watanabe (2010) for model selection. WAIC has the desirable property of penalizing complexity in models – a feature congruent with our goal of explaining infant motor development heterogeneity with a parsimonious number of clusters. See Gelman et al. (2014) for a detailed discussion of WAIC and comparison to other popular model fit criteria, such as DIC. In Section 4, we show through a simulation study that this approach recovered the true value of K under realistic parameter settings. In Section 6, we discuss alternative methods for choosing the optimal value of K.

4. Simulation Studies

4.1 *Simulation to Compare the MSN Model to the MVN Model*

Our first simulation study compared the MSN mixture model to a MVN mixture model, with the primary goal being to validate our parameter estimation scheme in a setting that resembles the Nurture data. Our secondary goal was to investigate to what degree ignoring skewness in outcome components leads to poor posterior inferences. To emulate the Nurture study, we simulated $n = 1,000$ subjects from the following model

$$f(\mathbf{y}_i) = \sum_{k=1}^{3} \pi_{ki} f(\mathbf{y}_i | \boldsymbol{\theta}_k), \tag{14}$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{i4})^T$ to conform to the $J = 4$ measurement occasions in the Nurture study; $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k ($k = 1, 2, 3$), and $f(\mathbf{y}_i | \boldsymbol{\theta}_k) \overset{\mathrm{d}}{=}$ $\mathrm{MSN}_4(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$; $\boldsymbol{\zeta}_{ki} = (\zeta_{ki1}, ..., \zeta_{ki4})^T = \mathbf{X}_i \boldsymbol{\beta}_k$, $\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{B}_k)$, and $\mathbf{X}_i$ and $\mathbf{B}_k$ are given by.

$$\mathbf{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & x_i & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & x_i & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & x_i & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & x_i \end{bmatrix}, \text{ and } \mathbf{B}_k = \begin{bmatrix} \beta_{k11} & \beta_{k12} \\ \beta_{k21} & \beta_{k22} \\ \beta_{k31} & \beta_{k32} \\ \beta_{k41} & \beta_{k42} \end{bmatrix}.$$

We fit a dummy indicator for each measurement occasion and a time invariant baseline covariate x_i (e.g., birthweight for gestational age z-score), whose effect may vary across the J measurement occasions. Thus, for subject i in cluster k, the $J = 4$ measurements have location parameters given by

$$
\begin{aligned}
\zeta_{ki1} &= \beta_{k11} + \beta_{k12} x_i \\
\zeta_{ki2} &= \beta_{k21} + \beta_{k22} x_i \\
\zeta_{ki3} &= \beta_{k31} + \beta_{k32} x_i \\
\zeta_{ki4} &= \beta_{k41} + \beta_{k42} x_i.
\end{aligned}
$$

To emulate z-score-transformed covariates present in the Nurture data such as birthweight for gestational age z-score, we sampled x_i from a $N(0, 1)$ density for $i = 1, ..., n$.

For cluster k, we collected all observations $\mathbf{y}_i$ ($i = 1, ..., n_k$) into the $n_k \times 4$ matrix $\mathbf{Y}_k$ and all covariates $\mathbf{X}_i$ into the $4n_k \times 8$ matrix $\mathbf{X}_k$ to utilize the matrix normal model specification detailed in Section 3, which allows for simultaneous updates of all MSN regression parameters in cluster k. For the multinomial regression model component of this simulation, we modeled the class labels z_i as a function of an intercept and one baseline covariate, w_{i1}, implying that $r = 2$. We did introduce missing data into this simulation, as missing data will be dealt with in the second simulation study. As a result, the final sample size was $N = n \times J = 4,000$.

We chose the MSN hyperparameters to be homogeneous across the three clusters by setting, for $k = 1, 2, 3$, $\mathbf{B}_{0k}^* = \mathbf{0}_{4\times 3}$, $\mathbf{V}_{0k} = \mathbb{1}_{4\times 4}$, and $\nu_{0k} = J + 2 = 6$, which gives $\mathrm{E}(\mathbf{\Sigma}_k) = \mathbb{1}_{4\times 4}$. Similarly, we set $\mathbf{d}_{01} = \mathbf{d}_{02} = (0, 0)^T$ and $\mathbf{S}_{01} = \mathbf{S}_{02} = \mathbb{1}_{2\times 2}$, noting that $k = 3$ is the reference cluster. To investigate the effect of ignoring skewness, we fit both the MVN and MSN mixture models to data generated from model (14). The values of WAIC for the MSN mixture model and MVN mixture model were 37,221.7 and 324,705.3, respectively, indicating model fit suffered when skewness was ignored. Furthermore, we report below in Table 1 a selection of parameter estimates from the MSN and MVN models that show incorrect inference (i.e., true parameter values not contained in 95% posterior credible intervals) resulted from ignoring skewness. A full table showing all model parameters estimates and 95% posterior credible intervals for the MSN and MVN models in addition to ground truth values is given in Web Appenxix B.

[Table 1 about here.]

An attractive feature of the MSN model is that it includes the MVN model as a special case. As such, the MSN model can be used in place of the MVN model even when data are not clearly skewed. We demonstrate this by fitting both the MSN and MVN model to data

generated from model (14), where all skewness parameters are set to zero. As shown Table 2 of Web Appendix B, parameter estimates are similar between the MSN and MVN models, and all 95% credible intervals for the α parameters of the MSN models contain 0, indicating the MSN model performs well even when data are not skewed.

4.2 *Simulation to Compare Imputation Methods*

In the second simulation study, we compare our online conditional imputation approach to Bayesian multiple imputation, in which M (large) imputed data sets are generated, MCMC is performed on each imputed data set, and the data are pooled for final posterior inference. Details of this approach are nicely summarized in (Zhou and Reiter, 2010). Our goal here is to demonstrate that our proposed online conditional imputation method performs comparable to Bayesian multiple imputation with respect to accuracy of parameter estimates. If the performance of the two methods are qualitatively alike, we might argue that the online conditional imputation method is preferable due to its computational efficiency. Additionally, we compare both methods to the default analysis based on available cases, despite extensive literature cautioning against this approach (Gelman and Hill, 2006)

To demonstrate the advantages of our proposed imputation method, we first generated $n = 1,000$ observations from a simple three-cluster ($K = 3$) MSN mixture model with $J = 4$ repeated measurements, one main effect for time ($p = 1$), and two multinomial regression predictors ($r = 2$). We then removed observations intermittently across the 4 measurement occasions according to a MAR mechanism, whereby the occurence of missing data depends only on the values of the observed data (Gelman and Hill 2006). The `ampute` function from the `mice` package in `R` was used to generate missing observations. We specified that each of the 4 repeated measures for a given individual were equally likely to be missing. To further approximate the Nurture data, we set the frequency of missing data at each measurement occasion to be 30%. We then fit our proposed MSN finite mixture model to the

amputed data using both online conditional imputation and Bayesian multiple imputation as defined by (Zhou and Reiter, 2010). As is described by Zhou and Reiter, Bayesian multiple imputation requires the number of imputations M to be large. Accordingly, we generated $M = 1,000$ imputed data sets using conditional MVN imputation, where the multivariate mean and covariance parameters were estimated from the available case sample sub-sample. The resultant multiply imputed data array occupied over 31 megabytes of disk space, illustrating the computational burden of Bayesian multiple imputation compared to online imputation.

4.3 *Simulation to Assess Sensitivity to Misspecified K*

Our final simulation is concerned with validating the use of WAIC for determining the number of clusters K. To this end, we simulate data from the same MSN mixture as in simulation study #2, that is – we simulate data from a $K = 3$ cluster model and fit MSN mixtures of $K \in \{2, 3, 4\}$, comparing the WAIC under each setting of K.

Need to start this simulation before writing more.

5. Application

- Include both time varying and non-time varying covariates for the within cluster covariate set.

6. Discussion

- Discuss how we handle non-ignorable missingness

- Discuss other label switching approaches

- Discuss skew-t?

References

Arellano-Valle RB, Azzalini A. On the unification of families of skewnormal distributions. *Scandinavian Journal of Statistics.* 2006 Sep;33(3):561-74.

Azzalini A. A class of distributions which includes the normal ones. *Scandinavian journal of statistics.* 1985 Jan 1:171-8.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew normal distribution. Biometrika 83, 715726.

Bishop, CM. Pattern recognition and machine learning. Springer; 2006.

Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software.* 2010:1-68.

Chen JT, Gupta AK. Matrix variate skew normal distributions. *Statistics.* 2005 Jun 1;39(3):247-53.

Neelon SE, Østbye T, Bennett GG, Kravitz RM, Clancy SM, Stroo M, Iversen E, Hoyo C. Cohort profile for the Nurture Observational Study examining associations of multiple caregivers on infant growth in the Southeastern USA. *BMJ Open.* 2017 Feb 1;7(2):e013939.

Franczak BC, Tortora C, Browne RP, McNicholas PD. Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters.* 2015 Jun 1;58:69-76.

Frühwirth-Schnatter S, Pyne S. Bayesian inference for finite mixtures of univariate and

multivariate skew-normal and skew-t distributions. *Biostatistics*. 2010 Jan 27;11(2):317-36.

Ganjali M, Baghfalaki T. A Bayesian shared parameter model for analysing longitudinal skewed responses with nonignorable dropout. *International Journal of Statistics in Medical Research*. 2014 Apr 1;3(2):103.

Geweke, J. Evaluating the accuracy of sampling-based ap- proaches to calculating posterior moments. *Bayesian Statistics* Vol. 4, J. M. Bernardo, J. Berger, A. P. Dawid, and A.F.M. Smith (eds), 169193. 1992. Cambridge, U.K.: Oxford University Press.

Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. *Cambridge university press*; 2006 Dec 18.

Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. *Chapman and Hall/CRC*; 2013 Nov 27.

Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. 2014 Nov 1;24(6):997-1016.

Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*. 2006;1(1):145-68.

Lagona F, Picone M. Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics*. 2012 May 1;39(5):927-45.

Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*. 2013 Apr 1;14(2):157-68.

Laurent, S. `matrixsampling`: Simulations of Matrix Variate Distributions. R package version 1.1.0. https://CRAN.R-project.org/package=matrixsampling.

Lee SX, McLachlan GJ. Model-based clustering and classification with non-normal mixture distributions. Statistical Methods & Applications. 2013 Nov 1;22(4):427-54.

Lee SX, Mclachlan GJ. On mixtures of skew normal and skew-*t* distributions. *Advances in*

Data Analysis and Classification. 2013 Sep 1;7(3):241-66.

Lin TI, Wang WL, McLachlan GJ, Lee SX. Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution. *Statistical Modelling.* 2018 Feb;18(1):50-72.

Luo S, Lawson AB, He B, Elm JJ, Tilley BC. Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. *Statistical Methods in Medical Research.* 2016 Apr;25(2):821-37.

Makalic E, Schmidt DF. High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649.* 2016 Nov 21.

Melnykov V, Maitra R. Finite mixture models and model-based clustering. *Statistics Surveys.* 2010;4:80-116.

Neelon B, Chung D. The LZIP: A Bayesian latent factor model for correlated zeroinflated counts. Biometrics. 2017 Mar;73(1):185-96.

Papastamoulis, P (2016). label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software*, 69(1), 1-24. doi:10.18637/jss.v069.c01

Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News* 2006, vol 6, 7-11.

Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya - Gamma latent variables. Journal of the American statistical Association. 2013 Dec 1;108(504):1339-49.

R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing.* 2019. https://www.R-project.org/.

Tiao GC, Zellner A. On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society*: Series B (Methodological). 1964 Jul;26(2):277-85.

Viroli C. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*. 2011 Oct 1;21(4):511-22.

Vrbik I, Mcnicholas PD. Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics & Data Analysis*. 2014 Mar 1;71:196-210.

Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*. 2010;11(Dec):3571-94.

Zeller CB, Cabral CR, Lachos VH. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *Test*. 2016 Jun 1;25(2):375-96.

Zhou X, Reiter JP. A note on Bayesian inference after multiple imputation. *The American Statistician*. 2010 May 1;64(2):159-63.
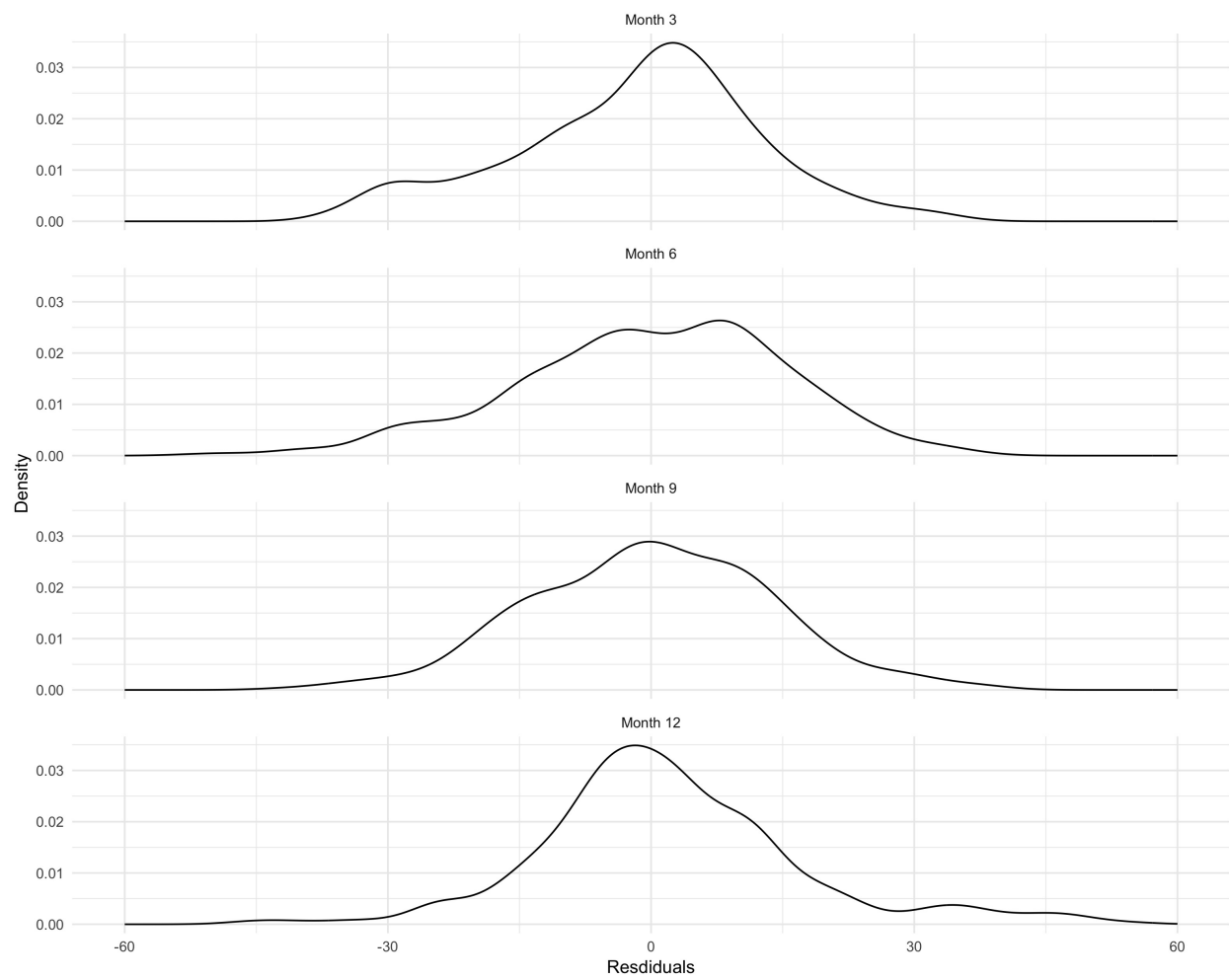
7. Appendix

Put your final comments here.

Table 1

Model results for simulated data with n = 1,000, J = 4, p = 2, K = 3, r = 2. 1,000 iterations were run with a burn in of 100. Missingness mechanism was MAR and P(miss) = 0. Model results for the multivariate skew normal (MSN) and multivariate normal (MN) mixtures are presented.

			Class 1	
Component	Param.	True	MSN Est. (95% CrI)	MN Est. (95% CrI)
MVSN	β_{11}	11	11.07 (10.74, 11.39)	9.42 (8.91, 9.77)
Regression	β_{21}	12	12.02 (11.87, 12.17)	11.98 (11.77, 12.18)
	β_{31}	13	13.06 (12.75, 13.36)	11.39 (10.7, 11.78)
	β_{41}	14	14.06 (13.91, 14.22)	14.02 (13.78, 14.22)
	β_{12}	2	2.11 (1.82, 2.35)	0.42 (0.03, 0.83)
	β_{22}	2	2.03 (1.88, 2.17)	2.02 (1.86, 2.22)
	β_{32}	2	2.13 (1.8, 2.43)	0.49 (0.14, 0.86)
	β_{42}	2	2.08 (1.93, 2.23)	2.08 (1.92, 2.28)
	α_1	-0.99	-0.81 (-2.12, 0.05)	/
	α_2	-0.5	-0.22 (-1.3, 0.75)	/
	α_3	-0.5	-0.96 (-2.14, 0.01)	/
	α_4	-0.99	-1.18 (-2.44, -0.06)	/
Multinom.	δ_{11}	-0.08	-0.07 (-0.27, 0.12)	-0.54 (-0.77, -0.32)
	δ_{12}	0.51	0.25 (-0.04, 0.53)	-0.26 (-0.6, 0.05)
	δ_{21}	-0.97	-0.91 (-1.15, -0.68)	-0.07 (-0.28, 0.14)
	δ_{22}	0.84	0.39 (0.09, 0.71)	0.24 (-0.04, 0.5)
Clustering	π_l	0.38	0.38 (0.38, 0.38)	0.38 (0.13, 0.41)