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 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 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 development and a higher percentage of

BIOMETRICS 64, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

male infants and another characterized by more steady 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.

KEY WORDS: Mixture of Experts, Pólya-Gamma, Skew-Normal, Imputation, Latent Growth, Infant Development.

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 convinient 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 yeilding 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 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.

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 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., Baley composite scores) for subject i across the J measurement occasions. For the 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 4, 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_{ik}. 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 (Azzalini and Dalla Valle, 1996).

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). As we will see 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. Let $\mathbf{Y}_k$ be an $n_k \times J$ response matrix with rows $\mathbf{y}_i^T$, $(i = 1, ..., n_k)$, where $n_k = \sum_{i=1}^{n} 1_{(z_i=k)}$ 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 $\boldsymbol{\zeta}_{ki} = \mathbf{X}_i\boldsymbol{\beta}_k$ 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 $\text{MatNorm}_{n_k \times J}(\cdot)$ denotes a $n_k \times J$ matrix normal density, $\text{vec}(\mathbf{M}_k^*) = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{t}_k \otimes \boldsymbol{\psi}_k$ is an $n_k J \times 1$ mean vector, $\mathbf{X}_k$ is an $n_k J \times Jp$ 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).

3.2 *Multinomial Regression for the Cluster Indicators*

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

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

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 this model, $z_i|\mathbf{w}_i \sim \text{Multinomial}(1, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{1K})$. During MCMC estimation, the cluster labels z_i are updated from their multinomial full conditional distribution and used in the remaining MCMC steps as cluster assignments as detailed later in section (**what section**). By allowing the cluster probabilities to vary across subjects, our model 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" (Bishop 2016).

To facilitate sampling, we adopt an 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

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

where $g_s \overset{iid}{\sim} \text{Ga}(b,1)$ for $s = 1, ..., \infty$. Polson *et al.* (2013) establish that for the $\text{PG}(b,c)$ density, and for $a, \eta \in \mathbb{R}$,

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

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

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

To make use of Pólya–Gamma data augmentation, we first write the full conditional distribution of $\boldsymbol{\delta}_k$ as the prior for $\boldsymbol{\delta}_k$ times the multinomial likelihood for all z_i.

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

where $p(\boldsymbol{\delta}_k)$ denotes the prior distribution of $\boldsymbol{\delta}_k$, $\mathbf{U}_k = (U_{k1}, ..., U_{kn})$, $U_{ik} = 1_{(z_i=k)}$ is an indicator that subject i belongs to cluster k, and π_{ki} is defined as in Section 3.4. We can rewrite π_{ki} as follows

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

where $c_{ik} = \log\sum_{h\neq k}\text{e}^{\mathbf{w}_i^T\boldsymbol{\delta}_h}$ and $\eta_{ik} = \mathbf{w}_i^T\boldsymbol{\delta}_k - c_{ki}$. We note that the sum $\sum_{h\neq k}\text{e}^{\mathbf{w}_i^T\boldsymbol{\delta}_h}$ includes the reference category, but since we fix $\boldsymbol{\delta}_K = \mathbf{0}$, we have $\text{e}^{\mathbf{w}_i^T\boldsymbol{\delta}_K} = 1$, and hence

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

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

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

which is essentially a logistic regression likelihood. We can thus apply the Pólya–Gamma sampler described in Polson et al. (2013) for logistic regression to update each $\boldsymbol{\delta}_k$ one at a

time based on the binary indicators U_{ki}. To accomplish this, 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. (2013) show that, conditional on $\boldsymbol{w} = (w_{k1}, ..., w_{kn})^T$, $\mathbf{U}_k^*$ follows a $\mathrm{N}_n(\boldsymbol{\eta}_k, \mathbf{O}_k^{-1})$ distribution with mean $\boldsymbol{\eta}_k = (\eta_{k1}, ..., \eta_{kn})^T$ and precision matrix $\mathbf{O}_k = \mathrm{Diag}(w_{k1}, ..., w_{kn})$. Thus, it follows that the full conditional distribution of $\boldsymbol{\delta}_k$ is given by

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

where $p(\boldsymbol{\delta}_k)$ is the prior distribution for $\boldsymbol{\delta}_k$, which we detail in Section 3.4 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.

3.3 *Conditional MSN Imputation*

To accommodate missing at random (MAR) responses, we propose a convenient imputation algorithm that can be implemented "online" 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

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

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

$\boldsymbol{\Sigma}_{k11}$ is a $(J - q_i) \times (J - q_i)$ matrix containing the rows and columns of $\boldsymbol{\Sigma}_k$ corresponding to inicies of $\mathbf{y}_i$ where missingness occurs. Similarly, $\boldsymbol{\Sigma}_{k12}$ is a $(J - q_i) \times q_i$ matrix containing the rows of $\boldsymbol{\Sigma}_k$ that correspond to missing indices of $\mathbf{y}_i$, but columns of $\boldsymbol{\Sigma}_k$ that correspond to observed indices of $\mathbf{y}_i$. The remaining partitions $\boldsymbol{\Sigma}_{k21}$, and $\boldsymbol{\Sigma}_{k22}$ are defined in the same manner. Likewise, as defined in equation 2, $\boldsymbol{\zeta}_{ki} = \mathbf{X}_i\boldsymbol{\beta}_k + t_i\boldsymbol{\psi}_k$ and is partitioned into $\boldsymbol{\zeta}_{ki}^{miss}$ and $\boldsymbol{\zeta}_{ki}^{obs}$ with respect to the missing and observed indices of $\mathbf{y}_i$, respectively. These results follow from conventional multivariate normal theory, which we detail in the Web Appendix. An attractive 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 missing values are drawn at each MCMC iteration, and often in practice $n_{sim} >> m$, where n_{sim} is the number of MCMC iterations (**find a reference**). In Section 4, we conduct simulation studies to demonstrate that imputing the missing MSN responses improves inferences over available-case analysis.

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. For $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$, we adopt a conditionally independent prior structure for $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$, where $p(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k) = p(\boldsymbol{\Sigma}_k)p(\boldsymbol{\beta}_k|\boldsymbol{\Sigma}_k)$. We choose the normal-inverse-Wishart distribution for $p(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k)$ by specifying $\boldsymbol{\Sigma}_k \sim \text{IW}(\mathbf{V}_{0k}, \nu_{0k})$ and $\boldsymbol{\beta}_k|\boldsymbol{\Sigma}_k \sim \text{N}_{Jp}(\mathbf{b}_k, \mathbf{I}_p \otimes \boldsymbol{\Sigma}_k)$. We assign the skewness coefficients $\boldsymbol{\psi}_k$ a conjugate $\text{N}_J(\mathbf{m}_k, \mathbf{P}_k)$ prior. In practice, the updates of $\boldsymbol{\beta}_k$ and $\boldsymbol{\psi}_k$ can be combined into one step by defining the $(Jp + J) \times 1$ vector $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k^T, \boldsymbol{\psi}_k^T)^T$ for which we assume a $\text{N}_{Jp+J}(\mathbf{b}_k^*, \mathbf{I}_{(p+1)} \otimes \boldsymbol{\Sigma}_k)$ prior, where $\mathbf{b}_k^*$ is formed by concatenating $\mathbf{b}_k$ and $\mathbf{m}_k$.

To make use of the Matrix Normal representation introduced previously in Section 3, we define the matrix of regression parameters $\mathbb{B}_k^*$, where $\text{vec}(\mathbb{B}_k^*) = \boldsymbol{\beta}_k^*$. We assign $\mathbb{B}_k^*|\boldsymbol{\Sigma}_k$ 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 $\text{vec}(\mathbf{B}_{0k}^*) = \mathbf{b}_k^*$, $\mathbf{I}_{p+1}$ is the $(p+1)$-dimension identity matrix, and $\boldsymbol{\Sigma}_k$ is the covariance matrix defined in equation 2. The matrix-Normal-inverse-Wishart is the conjugate joint prior for the regression parameters in the matrix Normal model given in equation 5. This induces convenient closed-form full conditional distributions that can be easily updated within our proposed Gibbs sampler.

For the multinomial logit model component, the regression parameters $\boldsymbol{\delta}_k = (\delta_{k1}, ..., \delta_{kr})$ are assigned a $\text{N}_r(\mathbf{d}_{0k}, \mathbf{S}_{0k})$ prior for $k = 1, ..., K-1$, which is conditionally conjugate as described in Section (**what section**). We allow the normal-inverse-Wishart and multinomial hyperparameters 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 that can be efficiently updated as part of a Gibbs sampler outlined below. A programatic sketch of our MCMC algorithm is given in Table (**??**). Additional details, including derivations of full conditionals and MCMC diagnostics can be found in the Web Appendix.

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, 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 for all $i = 1, ..., n$. 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 updating t_i, the truncated Normal random effect used in the stochastic representation of $\mathbf{y}_i$ given in equation 3. For cluster k, 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 $i = 1, ..., n_k$, compute $a_i = A_k\boldsymbol{\psi}_k^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\zeta}_{ik})$, where $\boldsymbol{\zeta}_{ik} = X_i\boldsymbol{\beta}_k$. Finally, 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

$$\mathbf{V}_k = (\mathbf{Y} - \mathbf{X}_k^*\mathbf{B}_k^*)^T(\mathbf{Y}_k - \mathbf{X}_k^*\mathbf{B}_k^*)(\mathbf{B}_k^* - \mathbf{B}_{0k}^*)^T\mathbf{I}_{p+1}(\mathbf{B}_k^* - \mathbf{B}_{0k}^*) + \mathbf{V}_{0k}.$$

The matrix $\mathbf{X}_k^*$ is formed by adding We then make use of the Matrix Normal representation introduced in Section 3.2 to draw $\mathbb{B}_k^*$ from a $\mathrm{MatNorm}_{p+1,J}(\mathbf{B}_k^*, \mathbf{L}_k^*, \boldsymbol{\Sigma}_k)$ density, where $\mathrm{vec}(\mathbb{B}_k^*) = \boldsymbol{\beta}_k^* = (\beta_{k11}, ..., \beta_{kJp}, \psi_{k1}, ..., \psi_{kJ})^T$ and

$$\begin{aligned}\mathbf{B}_k^* &= \mathbf{L}_k^*(\mathbf{X}_k^{*T}\mathbf{Y}_k + \mathbf{I}_{p+1}\mathbf{B}_{0k}^*)\\ \mathbf{L}_k^* &= (\mathbf{X}_k^{*T}\mathbf{X}_k^* + \mathbf{I}_{p+1})^{-1}.\end{aligned}$$

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 $U_{ki} = 1_{z_i=k}$ for $i = 1, ..., n$ and $k = 1, ..., K-1$. Next, we update w_{ki} from a $\text{PG}(1, \eta_{ki})$ density, where $\eta_{ki} = \mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}$, and $c_{ki} = \log \sum_{h \neq k} \text{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$. 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 $\text{N}_r(\mathbf{d}_k, \mathbf{S}_k)$ density, where $\mathbf{S}_k = (\mathbf{S}_{k0} + \mathbf{W}^T \mathbf{O}_k \mathbf{W})^{-1}$, $\mathbf{O}_k = \text{Diag}(w_{k1}, ..., w_{kn})$, $\mathbf{d}_k = \mathbf{S}_k(\mathbf{S}_{k0}\mathbf{d}_{k0} + \mathbf{W}^T \mathbf{O}_k \boldsymbol{\gamma}_k)$, and $\mathbf{W}$ is the $n \times r$ matrix of multinomial logit regression covariates such that the i^{th} 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

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

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 $\boldsymbol{\upsilon}_i = (\upsilon_{1i}, ..., \upsilon_{Ki})$, where

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

The cluster labels z_i are then updated from a $\text{Multinomial}(1, \boldsymbol{\upsilon}_i)$ density for $i = 1, ...n$. A schematic outline of the Gibbs sampler is given in the Web Appendix.

3.4.3 *Assessment of MCMC Convergence, Label Switching, and Model Selection.* We monitor conver convergence of our MCMC algorithm through the use of standard approaches such Geweke's (1992) Z-diagnostic. In simulation studies under realistic parameter settings, we observed relatively fast convergence of all MCMC chains (i.e. within 1,000 iterations).

A common challenge for Bayesian mixture models is the so-called "label switching" problem, in which draws of cluster-specific parameters may be assigned to the wrong cluster at some point during the MCMC simulation, rendering summaries of class-specific parameters incoherent. After conducting simulation studies under a wide variety of realistic parameter

settings, as detailed in Sections 4 and 5, we found little evidence of label switching in posterior draws of $\mathbf{z}$. When label switching was observed, we implemented *post hoc* relabling algorithms as described in Papastamoulis 2016 and implemented in the `label.switching` package in `R` (Papastamoulis 2016).

To specify the number of cluster K, we follow Neelon and Chung (2016), who use WAIC to determine the optimal number of latent factors in a Bayesian latent factor zero-inflated Poisson model for the analysis of correlated and zero-inflated count data. In Section 6, we discuss possible extentions to other methods for choosing K.

4. Simulation Studies

4.1 *Simulation to Compare to Multivariate Normal*

Our first simulation study will be concerned with comarison of the MSN mixture model to a standard multivariate normal mixture model. Here, the primary goal is to validate our parameter estimation scheme in a setting that resembles our application to the Nurture data. Our secondary goal is to investigate to what degree allowing for skewness in outcome components reduces the number of mixture components necessary to estimate the marginal density. For the purposes of illustration, we consider the following generative model

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

where $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k, for $k \in \{1, 2, 3\}$, and $\mathbf{y}_i|\boldsymbol{\theta}_k \sim \text{MSN}_4(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$, where $\boldsymbol{\zeta}_{ki} = (\zeta_{ki1}, \zeta_{ki2}, \zeta_{ki3}, \zeta_{ki4}) = \mathbf{x}_i\boldsymbol{\beta}_k$. We note from the above model specification that the number of clusters $K = 3$ and the number of measurement occasions $J = 4$. For the MSN regression model, we fit a main effect for time (i.e. measurement occasion) in addition to a baseline covariate, whose value does not vary with time, but whose effect does vary with time. Under this setting with $p = 2$ covariates, the design matrix $\mathbf{x}_i$ for subject i and the matrix of regression coefficients for cluster k, $\mathbf{B}_k$ are structured as

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

where $\text{vec}(\mathbf{B}_k) = \boldsymbol{\beta}_k$. Thus, for subject i in cluster k, the $J = 4$ measurements have location parameters given by

$$\begin{aligned}
\zeta_{ki1} &= \beta_{k11} + \beta_{k12} x_i \\
\zeta_{ki2} &= \beta_{k21} + \beta_{k22} x_i \\
\zeta_{ki3} &= \beta_{k31} + \beta_{k32} x_i \\
\zeta_{ki4} &= \beta_{k41} + \beta_{k42} x_i.
\end{aligned} \tag{14}$$

For cluster k, we can collect 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 model the class labels z_i as a function of an intercept and one baseline covariate, thus $r = 2$. We leave the problem of missing data to simulation study #2 and do not incorporate missing data into this simulation.

For our second goal in this simulation, to compare model performance when skewness is accounted for skewness to model performance when skewness is ignored, we fit both the proposed MSN finite mixture model as well as a classic multivariate Normal finite mixture model to data generated from an MSN distribution. We evaluate model fit in terms of absolute difference of parameter estimates as well as WAIC.

4.2 *Simulation to Compare Imputation Methods*

In our second simulation study, we compare our online conditional multivariate normal imputation method to standard Bayesian multiple imputation.

4.3 *Simulation to Assess Sensitivity to Misspecified K*

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.

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, 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.

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.

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

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.

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.

Acknowledgements

Received October 2007. *Revised February* 2008. *Accepted March* 2008.