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 to identify latent clusters of infants that 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 latent class 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 Plya-Gamma data-augmentation scheme, thereby improving predictions of the cluster membership allocations. Lastly, we impute missing responses under missing at random assumption by drawing from appropriate conditional skew normal distributions. Bayesian inference is achieved through straightforward Gibbs sampling, and can be carried out 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 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: A key word; But another key word; Still another key word; Yet another key word.

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), however in these models, focus lies primary on marginal density estimation and inference on the mixture components (i.e. clusters) is 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, parameter estimation 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, namely 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.

Put lit review of Bayesian PG multinomial logistic regression here

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.

2. Nurture Study

2.1 *Baseline Demographics and Description of Variables*

2.2 *Statistical Challenges*

2.2.1 *Skewness of Bayley score residuals.*

2.2.2 *Attrition and Intermittent Missingness.*

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. 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. To this end, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$ be a $J \times 1$ vector of responses (i.e., Baley scores) for subject i ($i = 1, \ldots, n$). For the analysis of the Nurture data, we propose a finite mixture model of the form

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

where $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k ($k = 1, ..., K$) and π_{ik} 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. We also assume that class membership is fixed throughout the study period, since our focus is to cluster individuals based on their overall developmental patterns over the course of the study. In Section 6, we discuss extensions to allow for class membership to vary over time. [**We could omit these last two sentence – are they really needed? Not sure. Maybe keep for now and think about it.**]

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

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

where $MSN_J(\cdot)$ 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. The vector $\boldsymbol{\alpha}_k$ has components α_{kj}, $j = 1, ..., J$ [**let's**

use kij as our index hierarchy. Please review throughout], 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 $N_J(\boldsymbol{\zeta}_k, \boldsymbol{\Omega}_k)$, where $\boldsymbol{\Omega}_k$ 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; $t_i \sim N_{[0,\infty)}(0, 1)$ 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 $\boldsymbol{\epsilon}_{ki} \sim N_J(\mathbf{0}, \boldsymbol{\Sigma}_k)$ is a $J \times 1$ vector of error terms. Thus, conditional on t_i and $z_i = k$, $\boldsymbol{y}_i$ is distributed as $N_J(\mathbf{X}_i\boldsymbol{\beta}_k + t_i\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$. Marginally (after integrating over t_i), $\mathbf{y}_i$ is distributed $MSN_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$, where through back-transformation

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

where $\boldsymbol{\omega}_k = \text{Diag}(\sqrt{\omega_{k,11}}, ..., \sqrt{\omega_{k,JJ}})$ 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

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

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 $\mathbf{t}_k$, is jointly distributed in matrix form as

$$\mathbf{Y}_k | \mathbf{t}_k \sim MatNorm_{n_k \times J}(\mathbf{M}_k, \mathbf{I}_{n_k}, \boldsymbol{\Sigma}_k),$$

where $MatNorm_{n_k \times J}(\cdot)$ denotes a $n_k \times J$ matrix normal density, $\mathrm{vec}(\mathbf{M}_k) = \mathbf{X}_k \boldsymbol{\beta}_k + \mathbf{t}_k \otimes \boldsymbol{\psi}_k$ 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 coovariates using a multinomial logit model

$$\pi_{ik} = \Pr(z_i = k | \mathbf{w}_i) = \frac{e^{\mathbf{w}_i^T \boldsymbol{\delta}_k}}{\sum_{h=1}^K e^{\mathbf{w}_i^T \boldsymbol{\delta}_h}}, \; k = 1, \ldots, K, \tag{4}$$

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, $z_i | \mathbf{w}_i \sim Multinom(1, \boldsymbol{\pi}_i)$, 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.

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. 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, Y_i^{obs}):

$$\begin{aligned} \mathbf{y}_i^{miss}|(z_i = k, t_i, \mathbf{y}_i^{obs}) &\sim N_{J-q_i}(\boldsymbol{\mu}_{ki}^{cond}, \boldsymbol{\Sigma}_k^{cond}), \text{ where} \\ \boldsymbol{\mu}_{ki}^{cond} &= \boldsymbol{\mu}^{miss} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_i^{obs} - \boldsymbol{\mu}^{obs}) \qquad (5) \\ \boldsymbol{\Sigma}_k^{cond} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \text{ where} \end{aligned}$$

$\boldsymbol{\Sigma}_k$ is partitioned into four sub-matrices $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{21},$ and $\boldsymbol{\Sigma}_{22}$ such that $\boldsymbol{\Sigma}_{11}$ 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, $\boldsymbol{\Sigma}_{12}$ 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 $\boldsymbol{\Sigma}_{21},$ and $\boldsymbol{\Sigma}_{22}$ are defined in the same manner. These results follow from conventional multivariate normal theory. 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. In Section 4, we conduct simulation studies to demonstrate that imputing the missing MSN responses improves inferences over complete 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 the MSN model component, 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 $\boldsymbol{\Sigma}_k \sim \mathrm{IW}(\mathbf{V}_k, \nu_k)$ and $\boldsymbol{\beta}_k|\boldsymbol{\Sigma}_k \sim \mathrm{N}_{Jp}(\mathbf{b}_k, \mathbf{I}_p \otimes \boldsymbol{\Sigma}_k)$. For the multinomial logit model component, the regression parameters $\boldsymbol{\delta}_k$ are given a conjugate $\mathrm{N}_r(\mathbf{d}_k, \mathbf{S}_k)$ prior for $k = 1, ..., K-1$.

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 development trends can be incorporated into certain clusters while still allowing the parameters of other clusters to be almost entirely determined by the data. Additionally, prior information regarding the effect of certain covariates on development cluster membership can be incorporated in to the model by choosing informative values for $\mathbf{d}_k$ and $\mathbf{S}_k$

[**Give priors for each parameter. Be clear about the conditionally joint prior for $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$. Where appropriate, explain advantages**]

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. Additional details, including derivations can be found in the Web Appendix. [**Think about the best way to organize this section. Maybe see my Bayesian Analysis paper for guidance? We can discuss next week.**]

Pólya–Gamma Data Augmentation for z_i.. The sampler begins by updating the latent cluster indicators z_i ($i = 1, \ldots, n$) from its multinomial logit full conditional. 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. [**See my Bayesian Analysis paper for guidance on this part**].

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

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

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

where $c_{ik} = \log\sum_{h\neq k} e^{\mathbf{w}_i^T\boldsymbol{\delta}_h}$ and $\eta_{ik} = \mathbf{w}_i^T\boldsymbol{\delta}_k - c_{ik}$. 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

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

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

$$p(\boldsymbol{\delta}_k|\mathbf{Z}, \boldsymbol{\delta}_{h\neq k}) \propto p(\boldsymbol{\delta}_k)\prod_{i=1}^{n}\left(\frac{e^{\eta_{ik}}}{1+e^{\eta_{ik}}}\right)^{U_{ik}}\left(\frac{1}{1+e^{\eta_{ik}}}\right)^{1-U_{ik}} = p(\boldsymbol{\delta}_k)\prod_{i=1}^{n}\frac{(e^{\eta_{ik}})^{U_{ik}}}{1+e^{\eta_{ik}}}$$

which we note is essentially a logistic regression likelihood. We thus apply this Pólya–Gamma data augmentation scheme to update each $\boldsymbol{\delta}_k$ $(k = 1, ..., K-1)$ one at a time based on the binary indicators U_{ik}.

- Emphasize that PG data augmentation for the multinomial model results in a PG mixture of experts model, which is a computationally efficient way to model edge weights.

3.4.3 *MCMC Algorithm.*

3.4.4 *Assessment of MCMC Convergence.*

3.4.5 *Label Switching.*

Algorithm 1 Gibbs Sampler

Define n_{iter}; n_{burn}; K; $\boldsymbol{\theta}_{init}$; $\boldsymbol{\theta}_0$

$n_{sim} := n_{iter} - n_{burn}$

$\boldsymbol{\theta} := \boldsymbol{\theta}_{init}$

for $\iota = 1, ..., n_{sim}$ **do**

  I. Conditional Imputation

  for $i = 1, ..., n$ **do**

    Draw $\mathbf{y}_i^{miss}$ from $N_q(\boldsymbol{\mu}_i^{miss}, \boldsymbol{\Sigma}_i^{miss})$

    $\mathbf{y}_i := \mathbf{y}_i^{miss} \cup \mathbf{y}_i^{obs}$

  end for

  II. MSN Regression

  for $k = 1, ..., K$ **do**

    $n_k := \sum_{i=1}^{n} 1_{z_i=k}$

    for $i_k = 1, ..., n_k$ **do**

      Draw t_i from $N_{[0,\infty)}(a_i, A)$

    end for

    $\mathbf{X}^*{}_k :=$ `cbind`$(\mathbf{X}_k, \mathbf{t}_k)$

    Draw $\mathbf{B}^*{}_k$ from $\mathrm{MatNorm}(\mathbf{B}_k, \mathbf{L}_k^{-1}, \boldsymbol{\Sigma}_k)$

    Draw $\boldsymbol{\Sigma}_k$ from $\mathrm{InvWish}(\nu_k, \mathbf{V}_k)$

  end for

  III. Multinomial Logit

  for $i = 1, ..., n$ **do**

    for $k = 1, ..., K$ **do**

      $\pi_{ik} := P(z_i = k | \mathbf{w}_i, \boldsymbol{\delta}_k)$

      $p_{ik} := P(\mathbf{y}_i | \boldsymbol{\beta}_k^{*T} \mathbf{x}_i^*, \boldsymbol{\Sigma}_k)$

    end for

    $\mathbf{p}_{z_i} := \frac{\mathbf{p}_i \circ \boldsymbol{\pi}_i}{\mathbf{p}_i \cdot \boldsymbol{\pi}_i}$

    Draw z_i from $\mathrm{Categorical}(\mathbf{p}_{z_i})$

    for $k = 1, ..., K-1$ **do**

      Draw $\boldsymbol{\delta}_k$ from $N(\mathbf{M}, \mathbf{S})$

    end for

  end for

  $\boldsymbol{\theta} := \{\mathbf{B}^*, \boldsymbol{\Sigma}, \mathbf{Z}, \boldsymbol{\delta}\}$

  Store $\boldsymbol{\theta}$

end for

4. Simulation Studies

4.1 *Simulation to Compare to Multivariate Normal*

[Table 1 about here.]

4.2 *Simulation to Compare Imputation Methods*

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 non-ignorable missingness here

- Discuss how we handle non-ignorable missingness

7. Appendix

Put your final comments here.

Acknowledgements

Supplementary Materials

7.1 *Glossary of Notation*

- $\mathbf{Y}$: A $n \times J$ matrix containing all multivariate skew-normal outcomes such that y_{ij} is the j^{th} outcome observed for subject i, where $i = 1, ..., n$ and $j = 1, ...J$.
- $\mathbf{X}$: A $n \times P$ matrix containing all multivariate skew-normal regression covariates such that x_{ip} is the p^{th} covariate value for subject i, where $i = 1, ..., n$ and $p = 1, ...P$.
- $\mathbf{B}$: A $P \times J$ matrix containing all multivariate skew-normal regression coefficients such that $\mathbf{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J]$, where β_{pj} is interpreted as the effect of covariate p on outcome j for $p = 1, ..., P$ and $j = 1, ..., J$.
- $\mathbf{E}$: A $n \times J$ matrix of error terms in the multivariate skew-normal regression model component. $\mathbf{E}$ is made up of row vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{iJ})$, where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_J(0, \boldsymbol{\Sigma})$ for $i = 1, ..., n$.
- $\boldsymbol{\Sigma}$: A $J \times J$ covariance matrix that defines the correlation between the p multivariate normal outcomes.
- $\boldsymbol{\Omega}$: A $J \times J$ covariance scale matrix that defines the correlation between the p multivariate skew-normal outcomes.
- $\boldsymbol{\psi}$: A $J \times 1$ vector containing the skewness parameter for each outcome.
- $\boldsymbol{\alpha}$: A $J \times 1$ vector containing the skewness parameter for each outcome.
- $\mathbf{t}$: An $n \times 1$ vector of truncated normal random effects used in the stochastic representation of the multivariate skew-normal distribution. For $i = 1, ..., n$, $t_i \overset{iid}{\sim} T_{[0,\infty)}(0, 1)$
- $\mathbf{X}^*$: A $n \times (P + 1)$ matrix constructed by column binding $\mathbf{t}$ to $\mathbf{X}$
- $\mathbf{B}^*$: A $(P + 1) \times J$ matrix constructed by row binding $\boldsymbol{\psi}^T$ to $\mathbf{B}$.

7.2 *Derivation of Full Conditional Distributions*

7.2.1 *Multivariate Skew-Normal Regression.* Without loss of generality, we derive the full conditional distributions for the multivariate skew-normal regression model component under the assumption that all observations belong to a single cluster. To make the extension to the case where more than one cluster is specified, simply apply these distributional forms to cluster specific parameters and data. Finally, we assume for the moment that we have complete data for all outcomes for each subject. We extend consider the case of missing data in section (INSERT SECTION).

The multivariate skew-normal regression model can be written as follows in matrix form.

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{t}\boldsymbol{\psi}^T + \mathbf{E} = \mathbf{X}^*\mathbf{B}^* + \mathbf{E}$$

The matrix $\mathbf{Y}$ is of dimension $n \times J$. For convenience, we define $\mathbf{X}^*$ as a $n \times (P+1)$ matrix constructed by column binding $\mathbf{t}$ to $\mathbf{X}$, and $\mathbf{B}^*$ as a $(P+1) \times J$ matrix constructed by row binding $\boldsymbol{\psi}^T$ to $\mathbf{B}$. We assume that $t_i \overset{iid}{\sim} T_{[0,\infty)}(0,1)$ and that $\mathbf{E}$ is made of row vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{iJ})$ for $i = 1, ..., n$, where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_J(0, \boldsymbol{\Sigma})$.

The conditional likelihood for this model is given below.

$$p(\mathbf{Y}|\mathbf{X}^*, \mathbf{B}^*, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{Y} - \mathbf{X}^*\mathbf{B}^*)^T(\mathbf{Y} - \mathbf{X}^*\mathbf{B}^*)\boldsymbol{\Sigma}^{-1}\right\}$$

We choose conjugate priors for $\mathbf{B}^*$ and $\boldsymbol{\Sigma}$ as follows.

$$\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\mathbf{V}_0, \nu_0)$$

$$\mathbf{B}^*|\boldsymbol{\Sigma} \sim MatNorm_{(m+1),p}(\mathbf{B}_0^*, \mathbf{L}_0^{-1}, \boldsymbol{\Sigma})$$

We now derive the joint posterior distribution of the parameters $\mathbf{B}^*$ and $\boldsymbol{\Sigma}$.

$$\begin{aligned} p(\mathbf{B}^*, \boldsymbol{\Sigma}|\mathbf{X}^*, \mathbf{Y}) &\propto p(\mathbf{Y}|\mathbf{X}^*, \mathbf{B}^*, \boldsymbol{\Sigma})p(\mathbf{B}^*|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}) \\ &\propto |\boldsymbol{\Sigma}|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{Y}-\mathbf{X}^*\mathbf{B}^*)^T(\mathbf{Y}-\mathbf{X}^*\mathbf{B}^*)\boldsymbol{\Sigma}^{-1}\right]\right\} \\ &\times |\boldsymbol{\Sigma}|^{-(P+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[(\mathbf{B}^*-\mathbf{B}_0^*)^T\mathbf{L}_0(\mathbf{B}^*-\mathbf{B}_0^*)\boldsymbol{\Sigma}^{-1}\right]\right\} \\ &\times |\boldsymbol{\Sigma}|^{(\nu_0+J+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{V}_0\boldsymbol{\Sigma}^{-1})\right\} \end{aligned}$$

7.2.2 *Multinomial Logit Regression.*

7.2.3 *Multivariate Normal Conditional Imputation.* The multivariate normal conditional imputation derivations are given for a single cluster without loss of generality. In practice, the data and parameters in this section would be replaced by cluster specific estimates in the case of clustering.

For a given observation vector $\mathbf{y} \sim N_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we allow for missingness in at most $J-1$ of the multivariate outcomes through the use of a conditional imputation step embedded within our Gibbs sampler. Suppose $\mathbf{y}$ contains q missing observations and can be partitioned into two vectors $\mathbf{y_1}$ and $\mathbf{y_2}$ such that $\mathbf{y_1}$ is a $q \times 1$ vector of missing observations and $\mathbf{y_2}$ is a $(J-q) \times 1$ vector of complete observations. Similarly, partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows.

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

We will use these quantities to derive the conditional distribution $f(\mathbf{y_1}|\mathbf{y_2}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$\begin{aligned}
f(\mathbf{y_1}|\mathbf{y_2}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\propto f(\mathbf{y_1}, \mathbf{y_2}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}^T \boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Sigma}^*_{11} & \boldsymbol{\Sigma}^*_{12}\\ \boldsymbol{\Sigma}^*_{21} & \boldsymbol{\Sigma}^*_{22}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1-\boldsymbol{\mu}_1\\ \mathbf{y}_2-\boldsymbol{\mu}_2\end{bmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\mathbf{y}_1-\boldsymbol{\mu}_{cond})^T\boldsymbol{\Sigma}_{cond}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_{cond})\right]\right\} \\
&\Rightarrow \mathbf{y_1}|\mathbf{y_2}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_q(\boldsymbol{\mu}_{cond}, \boldsymbol{\Sigma}_{cond})
\end{aligned}$$

$$\boldsymbol{\mu}_{cond} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2), \qquad \boldsymbol{\Sigma}_{cond} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

The block-wise inversion formula was used to invert $\boldsymbol{\Sigma}$ according to the following reparameterizations.

$$\begin{aligned}
\boldsymbol{\Sigma}^*_{11} &= \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \\
\boldsymbol{\Sigma}^*_{12} &= -\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\
\boldsymbol{\Sigma}^*_{21} &= -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \\
\boldsymbol{\Sigma}^*_{22} &= (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}
\end{aligned}$$

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.

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.

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.

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.

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.

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

Table 1

Model results for simulated data with $n = 1500$, $k = 4$, $p = 1$, $h = 3$, $v = 1$. 5000 iterations were run with a burn in of 1000. Missingness mechanism was MAR and P(miss) = 0

		Class 1		Class 2		Class 3	
Model Component	Parameter	True	Est. (95% CrI)	True	Est. (95% CrI)	True	Est. (95% CrI)
MVSN Regression	β_0	-3.21	-3.34 (-3.8, -2.99)	-0.32	-0.33 (-0.48, -0.14)	3.35	3.33 (3.22, 3.44)
	β_1	-3.08	-3.3 (-3.73, -2.87)	-0.75	-0.72 (-0.87, -0.52)	2.6	2.5 (2.39, 2.6)
	β_2	-2.97	-3.18 (-3.58, -2.76)	-0.45	-0.44 (-0.58, -0.26)	3.43	3.42 (3.31, 3.53)
	β_3	-2.91	-3.08 (-3.49, -2.68)	-0.66	-0.68 (-0.83, -0.48)	3.04	2.98 (2.87, 3.09)
	σ_{11}	1	0.95 (0.84, 1.02)	1	1 (0.89, 1.11)	1	1.06 (0.97, 1.16)
	σ_{12}	0.74	0.7 (0.59, 0.76)	0.68	0.68 (0.59, 0.78)	-0.45	-0.41 (-0.47, -0.36)
	σ_{13}	0.74	0.69 (0.58, 0.75)	-0.16	-0.13 (-0.2, -0.06)	0.82	0.88 (0.79, 0.97)
	σ_{14}	0.98	0.93 (0.81, 0.99)	0.64	0.65 (0.56, 0.75)	0.7	0.75 (0.67, 0.83)
	σ_{22}	1	0.94 (0.82, 1.01)	1	1.03 (0.93, 1.13)	1	1.07 (0.99, 1.16)
	σ_{23}	0.83	0.79 (0.67, 0.85)	-0.43	-0.4 (-0.46, -0.34)	-0.66	-0.62 (-0.68, -0.57)
	σ_{24}	0.81	0.77 (0.66, 0.83)	0.63	0.67 (0.58, 0.77)	0.01	0.07 (0.01, 0.13)
	σ_{33}	1	0.96 (0.84, 1.03)	1	1 (0.91, 1.09)	1	1.05 (0.96, 1.15)
	σ_{34}	0.85	0.81 (0.69, 0.87)	0.15	0.15 (0.08, 0.23)	0.59	0.64 (0.56, 0.72)
	σ_{44}	1	0.95 (0.83, 1.01)	1	1.02 (0.92, 1.13)	1	1.06 (0.97, 1.15)
	ψ_1	-0.33	-0.33 (-0.62, 0.69)	0.67	0.7 (0.46, 0.89)	-1	-1.01 (-1.13, -0.87)
	ψ_2	-0.33	-0.32 (-0.61, 0.64)	0.67	0.63 (0.38, 0.82)	-1	-0.88 (-1.01, -0.75)
	ψ_3	-0.33	-0.33 (-0.61, 0.69)	0.67	0.64 (0.43, 0.82)	-1	-1.01 (-1.14, -0.88)
	ψ_4	-0.33	-0.31 (-0.63, 0.67)	0.67	0.7 (0.45, 0.89)	-1	-0.94 (-1.07, -0.81)
Multinom.	δ_{11}	0.9	0.88 (0.81, 0.95)	0.9	0.88 (0.81, 0.95)	0.9	0.88 (0.81, 0.95)
	δ_{12}	0.23	0.22 (0.14, 0.3)	0.23	0.22 (0.14, 0.3)	0.23	0.22 (0.14, 0.3)
Clustering	π_l	0.28	0.28 (0.27, 0.28)	0.42	0.43 (0.42, 0.43)	0.3	0.3 (0.3, 0.3)