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

Biometrics 64, 1–21 DOI: 10.1111/j.1541-0420.2005.00454.x

December 2008

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 *Generic Multivariate Skew Normal Mixture*

Let $\mathbf{y}_i$ be the $J \times 1$ observation vector for subject i such that y_{ij} is the observation for subject i at timepoint j. We assume for now that $\mathbf{y}_i$ is fully observed. Later, we relax this assumption by allowing for missingness in the components of $\mathbf{y}_i$. For the analysis of the Nurture data, we propose the following MSN mixture

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

where $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k ($k = 1, ..., K$). For now we assume that K is fixed, but we discuss model selection strategies for choosing the optimal value of K in Section 4. The probability that subject i ($i = 1, ..., n$) belongs to cluster k is denoted π_{ik}. Note that π_{ik} is indexed by i, as we allow the probability of class membership to vary across subjects. However, we assume a 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.

In each cluster, we assume that the $J \times 1$ outcome vector for subject i follows an MSN distribution. Let $z_i \in \{1, ..., K\}$ denote a latent cluster membership of subject i. Conditional on $z_i = k$, we model $\mathbf{y}_i$ as

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

where $\boldsymbol{\zeta}_k$ is a $J \times 1$ vector of 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 accounts for associations between the J responses. The vector $\boldsymbol{\alpha}_k$ has components α_{jk}, $j = 1, ..., J$, 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 $N_J(\boldsymbol{\zeta}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\Sigma}_J$ is the $J \times J$ covariance matrix capturing dependence among the J outcomes.

3.2 *Multivariate Skew Normal Stochastic Representation*

We model the effect of covariates on longitudinal motor development outcomes through the use of a MSN regression model. We adopt a standard stochastic representation of a MSN random variable by conditioning on a subject-specific latent truncated normal random effect, denoted t_i (Frühwirth-Schnatter and Pyne 2010). Conditional on membership in cluster k, the stochastic representation for the j^{th} component of the multivariate skew normal distribution for subject i is

$$y_{ij} = \zeta_{ijk} + \psi_{jk} t_i + \sqrt{1 - \psi_{jk}^2}\epsilon_{ijk}. \tag{3}$$

where y_{ij} denotes the j^{th} observation for subject i, ζ_{ijk} denotes the location parameter for the j^{th} observation for subject i in cluster k, $t_i \overset{iid}{\sim} N_{[0,\infty)}(0,1)$ is a subject-specific truncated normal random effect, ψ_{jk} controls the skewness of the j^{th} outcome in cluster k, and $(\epsilon_{i1k}, ..., \epsilon_{iJk}) = \boldsymbol{\epsilon}_{ik} \sim N_J(0, \boldsymbol{\Sigma}_k)$ is a multivariate normal error term. Combining all observations for subject i into $\mathbf{y}_i$, and conditional on t_i and $z_i = k$, $\mathbf{y}_i$ follows a multivariate normal distribution with conditional mean $\mathbf{x}_i^T\boldsymbol{\beta}_k + t_i\boldsymbol{\psi}_k$ and conditional covariance $\boldsymbol{\Sigma}_k$. In the multivariate case, $\boldsymbol{\psi}_k$ is a $J \times 1$ vector containing cluster-specific skewness parameters. Marginally (after integrating over t_i), $\mathbf{y}_i$ is distributed $MSN_J(\boldsymbol{\zeta}_{ik}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$, where

$$\boldsymbol{\zeta}_{ik} = \mathbf{x}_i^T\boldsymbol{\beta}_k,$$

$$\boldsymbol{\alpha}_k = \frac{1}{\sqrt{1 - \boldsymbol{\psi}_k^T\boldsymbol{\psi}_k}}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k,$$

$$\boldsymbol{\Omega}_k = \boldsymbol{\Psi}_k\boldsymbol{\Sigma}_k\boldsymbol{\Psi}_k + \boldsymbol{\psi}_k\boldsymbol{\psi}_k^T,$$

where $\boldsymbol{\Psi}_k = \text{Diag}\left(\sqrt{1 - \psi_{k1}^2}, ..., \sqrt{1 - \psi_{kJ}^2}\right)$. To allow for time varying effects in our model, we structure $\mathbf{X}_i$ as a $J \times pJ$ vector of covariate values for subject i, and $\boldsymbol{\beta}_k$ as a $pJ \times 1$ vector of fixed effects coefficients for cluster k (think about how to explain this better).

3.3 *Representation as Matrix Skew Normal*

We note that the multivariate skew normal 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, this specification facilitates prior specification by admitting convenient joint prior distributions for the regression coefficients and scale matrices. In particular, we define $\mathbf{Y}_k$ as the $n_k \times J$ 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 total number of observations belonging to cluster k $(k = 1, ..., K)$.

$$\mathbf{Y}_k \sim MatSN_{n_k \times J}(\mathbf{M}, \boldsymbol{\alpha}_k, \mathbf{I}_{n_k}, \boldsymbol{\Omega}_k)$$

$$\text{vec}(\mathbf{Y}) = (\mathbf{y}_1^T, ..., \mathbf{y}_{n_k}^T)$$

$$\text{vec}(\mathbf{M}) = (\boldsymbol{\zeta}_1^T, ..., \boldsymbol{\zeta}_{n_k}^T)$$

where $\boldsymbol{\mu}_i = \mathbf{x}_i^T \boldsymbol{\beta}_k$ and $\boldsymbol{\alpha}_k = (\alpha_{k1}, ..., \alpha_{kJ})$. This implies that the i^{th} row of $\mathbf{Y}$ is $\mathbf{y}_i \sim MSN_J(\boldsymbol{\zeta}_i, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$. Using the stochastic representation presented in Section 3.2, the conditional distribution of $\mathbf{y}_i | z_i, t_i$ is $N_J(\mathbf{x}_i \boldsymbol{\beta}_k + \boldsymbol{\psi}_k t_i, \boldsymbol{\Sigma}_k)$. This implies that all responses for cluster k conditioned on $\mathbf{t}_k$ is distributed as

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

where $\text{vec}(\mathbf{M}_k)_{n_k J \times 1} = \mathbf{X}_k \boldsymbol{\beta}_k + \mathbf{t}_k \otimes \boldsymbol{\psi}_k$ and $\mathbf{X}_k$ is an $n_k \times p$ design matrix.

3.4 *Multinomial Regression on Cluster Probabilities*

A primary concern of our model is with identification of latent infant development classs. We accomplish this via multinomial logit regression model on cluster membership, which utilizes Pólya-Gamma data-augmentation, as described in Section 3.6 to allow for updating of all parameters using Gibbs sampling. The multinomial logit model is as follows for $k = 1, ..., K$.

$$P(z_i = k | \mathbf{w}_i) = \pi_{ik} = \frac{e^{\mathbf{w}_i^T \boldsymbol{\delta}_k}}{\sum_{k'=1}^{K} e^{\mathbf{w}_i^T \boldsymbol{\delta}_{k'}}}$$

where $\mathbf{w}_i$ is the vector of cluster probability covariates for subject i, $\boldsymbol{\delta}_k$ contains the multinomial regression parameters for cluster k, and K is the number of putative classs

specified *a priori*. 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.

3.5 *Conditional MVN Imputation*

We allow for missingness of outcomes in the MSN mixture model by imputing missing values from their conditional multivariate normal distributions. We note that

$$\mathbf{y}_i|z_i, \mathbf{x}_i, t_i, \boldsymbol{\beta}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim N_J(\boldsymbol{\beta}_k^T \mathbf{x}_i + t_i \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$$

where $z_i = k$. This allows us to appeal to standard conditional forms of the multivariate normal distribution to specify the distribution of missing observations. For subject i, we use q_i to denote the number of the total J possible repeated measurements are missing. We partition the full $J \times 1$ outcome vector $\mathbf{y}_i$ into the $(J - q_i) \times 1$ observed data vector $\mathbf{y}_i^{obs}$ and the $q_i \times 1$ missing data vector $\mathbf{y}_i^{miss}$. $Y_i = [Y_{i_{q\times 1}}^{miss}|Y_{i_{J-q\times 1}}^{obs}]^T$. We have

$$\mathbf{y}_i^{miss}|\mathbf{y}_i^{obs}, \mathbf{x}_i, t_i, z_i, \boldsymbol{\beta}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k \sim N(\boldsymbol{\mu}^{miss}, \boldsymbol{\Sigma}^{miss})$$

where $\boldsymbol{\mu}^{miss}$ and $\boldsymbol{\Sigma}^{miss}$ take standard forms detailed in the Appendix.

An attractive feature of this imputation approach is that each missing outcome is imputed "online", i.e. once per MCMC iteration. This provides more opportunities to explore the parameter space than multiple imputation and avoids multiplicative run-time scaling in m, the number of imputations. We demonstrate this feature using simulations detailed in Section 4.

3.6 *Bayesian Inference*

3.6.1 *Pólya–Gamma Data Augmentation.* Polson *et al.* (2013) introduced an efficient data augmentation approach to fitting several GLMs including the multinomial logit regression model specified by Holmes and Held (2006), which specifies the full conditional distributions

of the multinomial regression parameters as a function of a Bernoulli likelihood.

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

where $p(\boldsymbol{\delta}_k)$ denotes the prior distribution of $\boldsymbol{\delta}_k$, $U_{ik} = \mathbb{1}_{z_i=k}$ 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 $c_{ik} = \log \sum_{k'\neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_{k'}}$ and $\eta_{ik} = \mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ik}$. We note that the sum $\sum_{k'\neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_{k'}}$ 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

$$c_{ik} = \log \sum_{k'\neq k} e^{\mathbf{w}_i^T \boldsymbol{\delta}_{k'}} = \log \left(1 + \sum_{k' \notin \{k, K\}} e^{\mathbf{w}_i^T \boldsymbol{\delta}_{k'}} \right)$$

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

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

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.

3.6.2 *Prior Choice.*

3.6.3 *MCMC Algorithm.*

3.6.4 *Assessment of MCMC Convergence.*

3.6.5 *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**

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

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

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

end for

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

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

Draw $\boldsymbol{\Sigma}_k$ from 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 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.

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

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.

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

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 $\mathbf{\Sigma}$.

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

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 $\mathbf{y} \sim N_J(\boldsymbol{\mu}, \mathbf{\Sigma})$, 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 $\mathbf{\Sigma}$ as follows.

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

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

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

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

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

		Cluster 1		Cluster 2		Cluster 3	
Model Component	Parameter	True	Est. (95% CrI)	True	Est. (95% CrI)	True	Est. (95% CrI)
MVSN Regression	β_0	-1.35	-1.59 (-2.3, -0.94)	-0.18	-0.24 (-0.53, 0.38)	0.72	0.79 (0.52, 1.02)
	β_1	-1.19	-1.35 (-1.99, -0.84)	-0.09	-0.33 (-0.62, 0.78)	1.64	1.65 (1.33, 1.91)
	β_2	-1.65	-1.81 (-2.66, -1.41)	-0.47	-0.62 (-0.89, 0.3)	1.44	1.32 (0.97, 1.6)
	β_3	-1.75	-1.89 (-2.52, -1.37)	-0.22	-0.32 (-0.63, 0.55)	2.28	2.26 (1.96, 2.5)
	σ_{11}	1	1.02 (0.78, 1.19)	1	0.98 (0.77, 1.23)	1	1.06 (0.85, 1.29)
	σ_{12}	-0.32	-0.19 (-0.33, -0.02)	0.16	0.14 (-0.01, 0.4)	0.72	0.82 (0.62, 1.05)
	σ_{13}	-0.65	-0.55 (-0.68, -0.35)	0.72	0.7 (0.51, 0.94)	0.14	0.27 (0.1, 0.48)
	σ_{14}	-0.44	-0.33 (-0.46, -0.13)	0.5	0.48 (0.31, 0.72)	-0.01	-0.02 (-0.16, 0.16)
	σ_{22}	1	0.92 (0.72, 1.06)	1	0.94 (0.72, 1.22)	1	1.11 (0.87, 1.38)
	σ_{23}	0.56	0.49 (0.33, 0.6)	0.53	0.46 (0.29, 0.73)	-0.1	0.08 (-0.1, 0.28)
	σ_{24}	0.98	0.9 (0.7, 1.04)	0.24	0.14 (-0.03, 0.41)	0.19	0.17 (0.01, 0.37)
	σ_{33}	1	0.9 (0.66, 1.04)	1	0.92 (0.72, 1.19)	1	1.26 (1.03, 1.52)
	σ_{34}	0.56	0.51 (0.35, 0.62)	0.86	0.79 (0.58, 1.05)	-0.65	-0.63 (-0.78, -0.45)
	σ_{44}	1	0.93 (0.72, 1.07)	1	0.93 (0.7, 1.2)	1	1.11 (0.87, 1.36)
	ψ_1	-0.33	-0.02 (-0.84, 0.87)	0.67	0.69 (-0.09, 1.01)	-1	-0.98 (-1.25, -0.67)
	ψ_2	-0.33	-0.16 (-0.8, 0.62)	0.67	0.81 (-0.6, 1.14)	-1	-0.98 (-1.28, -0.6)
	ψ_3	-0.33	-0.15 (-0.65, 0.89)	0.67	0.72 (-0.35, 1.02)	-1	-0.8 (-1.15, -0.36)
	ψ_4	-0.33	-0.18 (-0.84, 0.59)	0.67	0.71 (-0.32, 1.06)	-1	-1 (-1.32, -0.62)
Multinom.	δ_{11}	-0.84	-0.78 (-0.96, -0.59)	-0.84	-0.78 (-0.96, -0.59)	-0.84	-0.78 (-0.96, -0.59)
	δ_{12}	-0.24	-0.26 (-0.42, -0.1)	-0.24	-0.26 (-0.42, -0.1)	-0.24	-0.26 (-0.42, -0.1)
Clustering	π_l	0.39	0.39 (0.38, 0.4)	0.26	0.26 (0.25, 0.27)	0.34	0.35 (0.33, 0.36)