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.S

SUMMARY: In studies of infant motor development, a crucial research goal is to identify latent classes 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 class-specific parameters that accommodate the inherent correlation and skewness in the data. Second, we model the class membership probabilities using a novel Plya-Gamma data-augmentation scheme, thereby improving predictions of the class 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 classes: 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, 1–16 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

lower percentage of males. The classes 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 *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 posterier 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.

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

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 Regression*

We model the effect of covariates on longitudinal development outcomes through the use of a MSN regression model. The MSN distribution can be represented as the superposition of a MN random variable with a latent truncated normal random effect. Let $\mathbf{Y}_{n\times k}$ be the observation matrix such that Y_{ij} is the observation for subject i at timepoint j.

$$\mathbf{Y}_{n\times k} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times k} + t_{n\times 1}\psi_{1\times k} + \boldsymbol{\epsilon}_{n\times k}$$

where X_i is the $1 \times p$ vector of covariate values for subject i, β_j is the $i \times k$ vector of fixed effects coefficients for timepoint j, $t_i \overset{iid}{\sim} N_{[0,\infty)}(0, 1)$ is a truncated normal random effect, ψ is the vector containing skewness parameters for each timepoint, and $\boldsymbol{\epsilon}_i \sim N_k(0, \boldsymbol{\Sigma}_{k\times k})$ is the correlated error term.

3.2 *Multinomial Regression on Class Probabilities*

A primary concern of our model is with identification of latent infant development clusters. We accomplish this via multinomial logit regression model on cluster membership, which utilizes Pólya-Gamma data-aumentation to allow for updating of all parameters using Gibbs sampling. The multinomial logit model is as follows for $l = 1, ..., h$.

$$P(Z_i = l|w_i) = \pi_{il} = \frac{e^{w_i^T\delta_l}}{\sum_{r=1}^{h} e^{w_i^T\delta_r}}$$

where w_i is the vector of class probability covariates for subject i, δ_l contains the multinomial regression parameters for class l, and h is the number of putative clusters specified *a priori*.

During our MCMC estimation procedure, the class labels z_i are updated from their multi-

nomial full conditional distribution and used in the remaining MCMC steps as class assignments.

3.3 *Conditional MVSN 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

$$Y_i|X_i, t_i, \boldsymbol{\beta}, \psi \sim N_k(X_i\boldsymbol{\beta} + t_i\psi, \boldsymbol{\Sigma})$$

This allows us to appeal to standard conditional forms of the multivariate normal distribution. Let $Y_i = [Y^{miss}_{i_{q\times 1}}|Y^{obs}_{i_{k-q\times 1}}]^T$. We have

$$Y_i^{miss}|Y_i^{obs}, X_i, t_i, \boldsymbol{\beta}, \psi \sim N(\mu^{miss}, \boldsymbol{\Sigma}^{miss})$$

where μ^{miss} and Σ^{miss} take standard forms. 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.

3.4 *Bayesian Inference*

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

3.4.1 *Prior Choice.*

3.4.2 *MCMC Algorithm.*

3.4.3 *Assessment of MCMC Convergence.*

3.4.4 *Label Switching.*

4. Simulation Studies

4.1 *Simulation to Compare to Multivariate Normal*

4.2 *Simulation to Compare Imputation Methods*

4.3 *Simulation to Assess Sensitivity to Mispecified 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

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_{ij} is the j^{th} covariate value for subject i.
- $\mathbf{B}$: A $m \times p$ matrix containing all multivariate skew-normal regression coefficients such that $\mathbf{B} = \left[\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p\right]$, where β_{ij} is interpreted as the effect of covariate i on outcome j for $i = 1, ..., m$ and $j = 1, ..., p$.
- $\mathbf{E}$: A $n \times p$ matrix of error terms in the multivariate skew-normal regression model component. $\mathbf{E}$ is made up of row vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{ip})$, where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_p(0, \boldsymbol{\Sigma})$ for $i = 1, ..., n$.
- $\boldsymbol{\Sigma}$: A $p \times p$ covariance matrix that defines the correlation between the p multivariate normal outcomes.
- $\boldsymbol{\Omega}$: A $p \times p$ covariance scale matrix that defines the correlation between the p multivariate skew-normal outcomes.
- $\boldsymbol{\psi}$: A $p \times 1$ vector containing the skewness parameter for each outcome.
- $\boldsymbol{\alpha}$: A $p \times 1$ 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 $n \times (m + 1)$ matrix constructed by column binding $\mathbf{t}$ to $\mathbf{X}$
- $\mathbf{B}^*$: A $(m + 1) \times p$ 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 $n \times p$. For convenience, we define $\mathbf{X}^*$ as a $n \times (m+1)$ matrix constructed by column binding $\mathbf{t}$ to $\mathbf{X}$, and $\mathbf{B}^*$ as a $(m+1) \times p$ 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 $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{ip})$ for $i = 1, ..., n$, where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N_p(0, \boldsymbol{\Sigma})$.

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 dervive the joint posterior distribution of the parameters $\mathbf{B}^*$ and $\mathbf{\Sigma}$.

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

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 practive, the data and parameters in this section would be replaced by cluster specific estimates in the case of clustering.

For a given observation vector $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, we allow for missingness in at most $p - 1$ 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 $(p - q) \times 1$ 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 blockwise 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.

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.

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.

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.

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

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.