A 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

email: neelon@musc.edu

and

Sara Benjamin-Neelon, PhD, MPH, RD

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

email: sara.neelon@jhu.edu

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

Biometrics 64, 1–2 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

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 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.

1. Introduction

2. Model

2.1 *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.

2.2 *Latent Class Finite Mixture*

2.3 *Multinomial Regression on Class Probabilities*

2.4 *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. Bayesian Inference

4. Simulation Studies

5. Application

6. Discussion

Put your final comments here.

Acknowledgements

Supplementary Materials

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

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

Appendix

Full Conditional Distributions