Multivariate Skew-Normal Mixture Model for Infant Development Clustering

Carter Allen[1]; Brian Neelon, PhD[1]; Sara E. Benjamin-Neelon, PhD, JD, MPH[2]

[1]Department of Public Health Sciences, Medical University of South Carolina; [2]Bloomberg School of Public Health, Johns Hopkins University

Abstract

We propose a novel Bayesian model for infant development trajectories that addresses primary research questions in this area while flexibly allowing for skewness and correlation of development outcomes. Our model is based on finite mixtures of multivariate skew normal (MSN) distributions, where covariates are allowed on both the multivariate outcomes and probability of latent class membership. We also allow for missing outcome data by drawing missing outcomes from their conditional MSN distributions. We demonstrate our method using data from the Nurture study.

Introduction

A primary goal in infant development research is to identify **latent development classes** and explain class membership in relation to covariates of interest. Additionally, it is often of interest to relate covariates to mean longitudinal growth patterns. Infant development data are inherently correlated longitudinally, often skewed, and frequently missing due to longitudinal attrition. Standard practices are ill-suited to addressing these research questions due to their ignorance of one or more of these features of development data.

Motivation

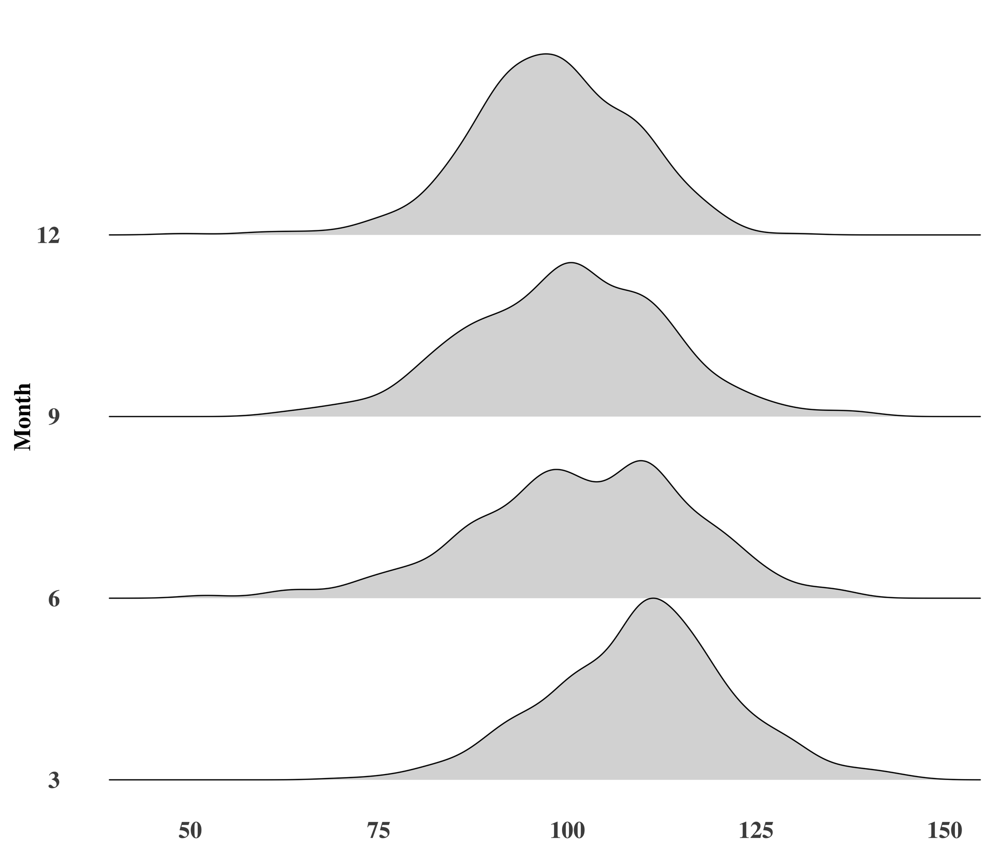

Figure 1: Density of Bayley composite scores for Nurture infants at 3, 6, 9 and 12 months of age.

Definitions

Let ϕ and Φ be the standard normal pdf and cdf, respectively. Azzalini (1985) defined the density of a skew-normal random variable Z follows.

$$f(z;\lambda) = 2\phi(z)\Phi(\lambda z)$$

Similar to the construction of the familiar student's t random variable, Azzalini (2014) defines a skew-t random variable as the ratio of a skew normal and the square root of a χ^2 divided by its degrees of freedom. The resultant density is

$$t(x;\lambda,\nu) = 2t_0(x;\nu)T_0(\lambda x\sqrt{\frac{\nu+1}{\nu+x^2}};\nu+1)$$

where t_0 and T_0 are the density and mass functions of the student's t distribution, respectively. A linear regression model with skew error terms is a modification of classical regression with the modification of either assuming $\mathcal{SN}$ or $\mathcal{ST}$ random errors.

Full Conditionals

Let $Y_i = \mathbf{x}_i\beta + \psi z_i + \sigma\epsilon$ with $Z \sim N_+(0,1)$, and $\epsilon \sim N(0,1)$. Define $\mathbf{X}^* = [\mathbf{X}|\mathbf{z}]$, and $\beta^* = [\beta_0, \beta_1, ..., \beta_p, \psi]$. Then,

$$\beta^*|Y,\mathbf{X}^*,\tau \sim N_{p+2}\left(\frac{(T_0\beta_0 + \tau\mathbf{X}^{*T}Y)}{\tau\mathbf{X}^{*T}\mathbf{X}^* + T_0}, \tau\mathbf{X}^{*T}\mathbf{X}^* + T_0\right)$$

$$\tau|\beta^*,\mathbf{X}^*,Y \sim \Gamma(n/2+\alpha, \frac{1}{2}(Y-\mathbf{X}^*\beta^*)^T(Y-\mathbf{X}^*\beta^*)+b)$$

$$z_i|y_i,\mathbf{x}_i,\beta,\tau \sim N_+\left(\frac{\psi\tau(y_i-\mathbf{x}_i\beta)}{\tau\psi^2+1}, \frac{1}{\tau\psi^2+1}\right)$$

Important Results

We developed a novel Bayesian MSN mixture model, and showed superior performance compared to standard approaches. We applied the MSN mixture model to data from the Nurture study and discovered two distinct development classes characterized by differences in development trajectories and demographics.

Modeling Approaches

1. *Maximum Likelihood Estimation*: For a random sample $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} \mathcal{SN}(\xi, \omega^2, \alpha)$, where $\xi = \mathbf{x}^T\beta$ for a collection of predictors $x_1, x_2, ..., x_n$ and a vector of unknown regression coefficients $\beta \in \mathbb{R}^p$. Azzalini (2014) describes procedures for obtaining MLE estimates of β and α, our primary parameters of interest. Azzalini's `R` package `sn` contains a function `selm` for fitting regression models with $\mathcal{SN}$ or $\mathcal{ST}$ random errors.
2. *Bayesian Gibbs Sampler*: We introduce the following stochastic representation of the skew normal distribution

$$Y_i = \mathbf{x}_i\beta + \psi z_i + \sigma\epsilon$$

where $z_i \sim N_+(0,1)$ and $\epsilon \sim N(0,1)$. The

Gibbs Sampler Simulation

sim-post.pdf

Figure 2: Posterior distributions of α, β_0, β_1, σ^2 in simulation.

Param.	True	MLE	Gibbs	SLR
α	4.00	0.3928	4.016	–

sim-trace.pdf

Figure 3: Trace plots of parameter estimates in simulation

Modeling Results

nurt_posteriors.pdf

Figure 4: Posterior distributions of model parameters

Param.	Est.	95% CI
α	0.239	(-0.169,0.624)
β_{low}	0.009	(-0.225,0.245)
$\beta_{v.low}$	0.230	(0.020,0.579)

References

(1) Azzalini, S. (1985). A class of distributions which includes the normal ones. SJS; **(2) Fruhwirth-Schnatter, S and Pyne, S. (2010)**. Bayesian inference for finite mixtures of univariate ... Biostatistics; **(3) Benjamin-Neelon SE, Ostbye T, Bennett GG, et al.** Cohort profile for the Nurture Observational Study ... BMJ Open 2017; **(4) Neelon, B. (2015)** Bayesian two-part spatial models

Funding: *This work is supported by a grant from the NIH (**R01DK094841**).*

Further Resources

https://carter-allen.github.io/MVSN-FMM

frame.png