

# Multivariate Skew-Normal Mixture Model for Infant Development Clustering

Carter Allen<sup>1</sup>; Brian Neelon, PhD<sup>1</sup>; Sara E. Benjamin-Neelon, PhD, JD, MPH<sup>2</sup>

<sup>1</sup>Department of Public Health Sciences, Medical University of South Carolina; <sup>2</sup>Bloomberg School of Public Health, Johns Hopkins University

## Abstract

We propose a novel Bayesian model for infant development patterns that addresses primary research questions in this area while allowing for skewness and correlation of 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 imputing missing outcomes from their conditional multivariate normal 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. It is often also 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 and 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

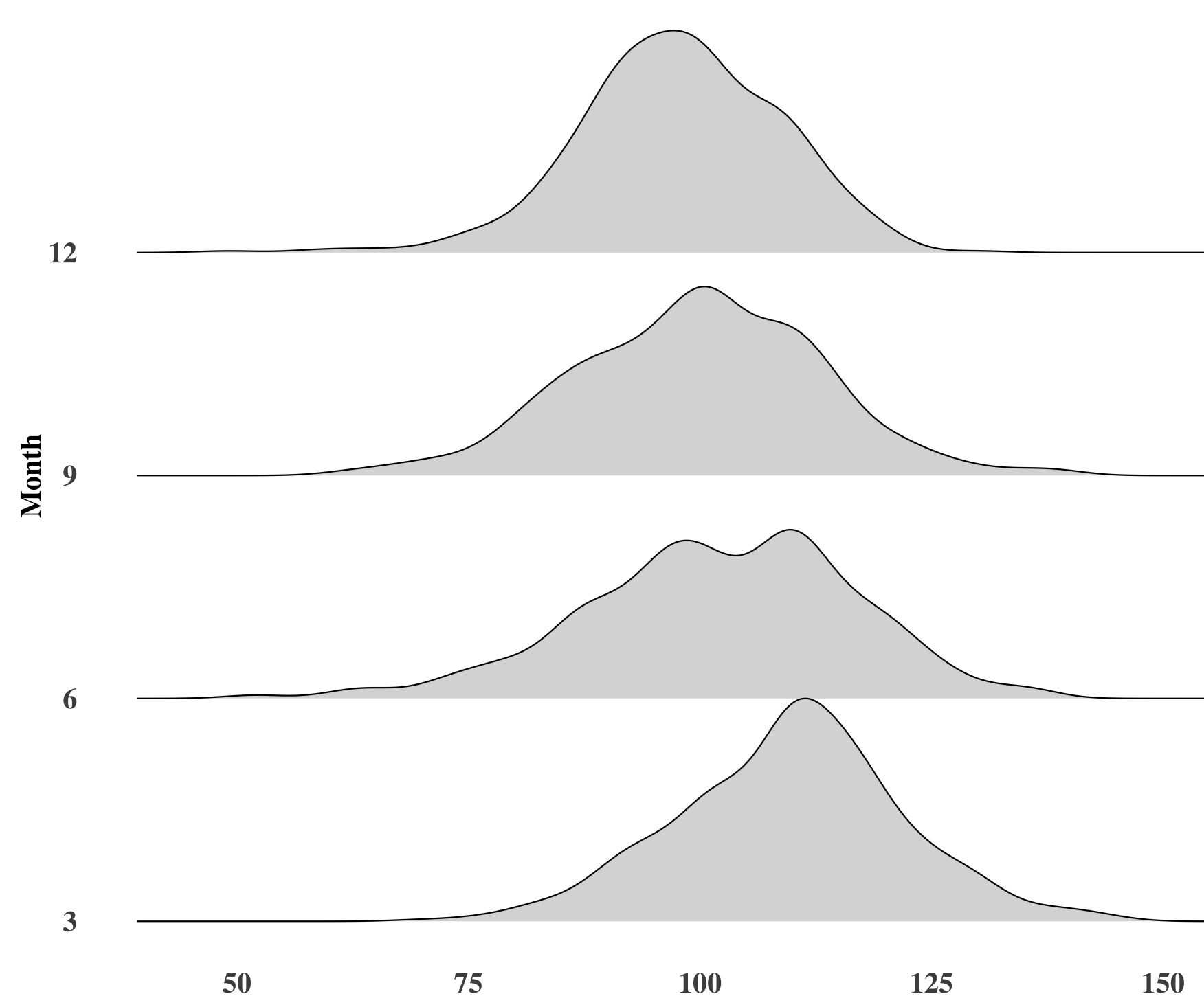


Figure 1: Density of Bayley composite scores for Nurture infants at 3, 6, 9 and 12 months of age.

## Clustering

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-augmentation to allow for updating of all parameters using Gibbs sampling. The multinomial logit model is as follows for  $l = 1, \dots, h$ .

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

where  $w_i$  is the vector of class probability covariates for subject  $i$ ,  $\delta_l$  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 multinomial full conditional distribution and used in the remaining MCMC steps as class assignments.

## Conditional 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

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

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

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

where  $\mu^{miss}$  and  $\Sigma^{miss}$  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.

## 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 three distinct development classes characterized by differences in development trajectories and demographics.

## MSN 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  $j$ .

$$\mathbf{Y}_{n \times k} = \mathbf{X}_{n \times p} \beta_{p \times k} + \mathbf{t}_{n \times 1} \psi_{1 \times k} + \epsilon_{n \times k}$$

where  $X_i$  is the  $1 \times p$  vector of covariate values for subject  $i$ ,  $\beta_j$  is the  $i \times k$  vector of fixed effects coefficients for timepoint  $j$ ,  $t_i \stackrel{iid}{\sim} N_{[0, \infty)}(0, 1)$  is a truncated normal random effect,  $\psi$  is the vector containing skewness parameters for each timepoint, and  $\epsilon_i \sim N_k(0, \Sigma_{k \times k})$  is the correlated error term.

## Simulation Study

Table 1: Parameter estimates for  $n = 1000$ ,  $k = 4$ ,  $p = 2$ ,  $h = 3$ ,  $v = 3$  simulation.

| Component    | Parm.           | Class 1 |       |                | Class 2 |       |                | Class 3 |       |                |
|--------------|-----------------|---------|-------|----------------|---------|-------|----------------|---------|-------|----------------|
|              |                 | True    | Est.  | (95% CrI)      | True    | Est.  | (95% CrI)      | True    | Est.  | (95% CrI)      |
| MSN Reg.     | $\beta_{11}$    | -0.54   | -0.51 | (-0.61, -0.41) | -0.19   | -0.22 | (-0.35, -0.09) | 1.84    | 1.81  | (1.65, 1.99)   |
|              | $\beta_{12}$    | -1.06   | -1.07 | (-1.15, -0.98) | -0.16   | -0.13 | (-0.25, -0.02) | 2.03    | 2.09  | (1.93, 2.23)   |
|              | $\beta_{13}$    | -1.28   | -1.22 | (-1.32, -1.11) | 0.59    | 0.58  | (0.46, 0.69)   | 1.8     | 1.74  | (1.58, 1.9)    |
|              | $\beta_{14}$    | -1.91   | -1.83 | (-1.92, -1.74) | -0.2    | -0.12 | (-0.25, -0.01) | 1.43    | 1.37  | (1.22, 1.52)   |
|              | $\sigma_{11}^2$ | 1       | 0.9   | (0.78, 1.03)   | 1       | 1.04  | (0.88, 1.26)   | 1       | 1.39  | (1.08, 1.76)   |
|              | $\sigma_{12}^2$ | -0.1    | -0.12 | (-0.22, -0.02) | -0.32   | -0.31 | (-0.44, -0.18) | 0.08    | 0.27  | (0.04, 0.52)   |
|              | $\sigma_{13}^2$ | -0.18   | -0.17 | (-0.28, -0.08) | -0.52   | -0.5  | (-0.66, -0.38) | 0.78    | 1.19  | (0.87, 1.54)   |
|              | $\sigma_{14}^2$ | 0.39    | 0.31  | (0.22, 0.43)   | 0.04    | 0.05  | (-0.07, 0.22)  | 0.41    | 0.72  | (0.47, 0.99)   |
|              | $\sigma_{22}^2$ | 1       | 0.97  | (0.84, 1.12)   | 1       | 0.94  | (0.77, 1.14)   | 1       | 1.18  | (0.91, 1.5)    |
|              | $\psi_1$        | -0.33   | -0.28 | (-0.4, -0.18)  | 0.67    | 0.64  | (0.5, 0.78)    | -1      | -0.91 | (-1.12, -0.68) |
| Multi. Logit | $\psi_2$        | -0.33   | -0.21 | (-0.34, -0.09) | 0.67    | 0.7   | (0.55, 0.84)   | -1      | -0.96 | (-1.16, -0.76) |
|              | $\psi_3$        | -0.33   | -0.27 | (-0.39, -0.14) | 0.67    | 0.72  | (0.58, 0.86)   | -1      | -0.91 | (-1.14, -0.69) |
|              | $\psi_4$        | -0.33   | -0.18 | (-0.3, -0.07)  | 0.67    | 0.73  | (0.6, 0.87)    | -1      | -0.93 | (-1.13, -0.73) |
|              | $\delta_{11}$   | -0.53   | -0.5  | (-0.74, -0.24) | -0.53   | -0.5  | (-0.74, -0.24) | -0.53   | -0.5  | (-0.74, -0.24) |
| Clust.       | $\delta_{12}$   | 0.44    | 0.28  | (0.02, 0.56)   | 0.44    | 0.28  | (0.02, 0.56)   | 0.44    | 0.28  | (0.02, 0.56)   |
|              | $\delta_{21}$   | 0.34    | 0.38  | (0.09, 0.63)   | 0.34    | 0.38  | (0.09, 0.63)   | 0.34    | 0.38  | (0.09, 0.63)   |
|              | $\pi_1$         | 0.41    | 0.41  | (0.4, 0.42)    | 0.36    | 0.36  | (0.35, 0.37)   | 0.23    | 0.23  | (0.22, 0.24)   |

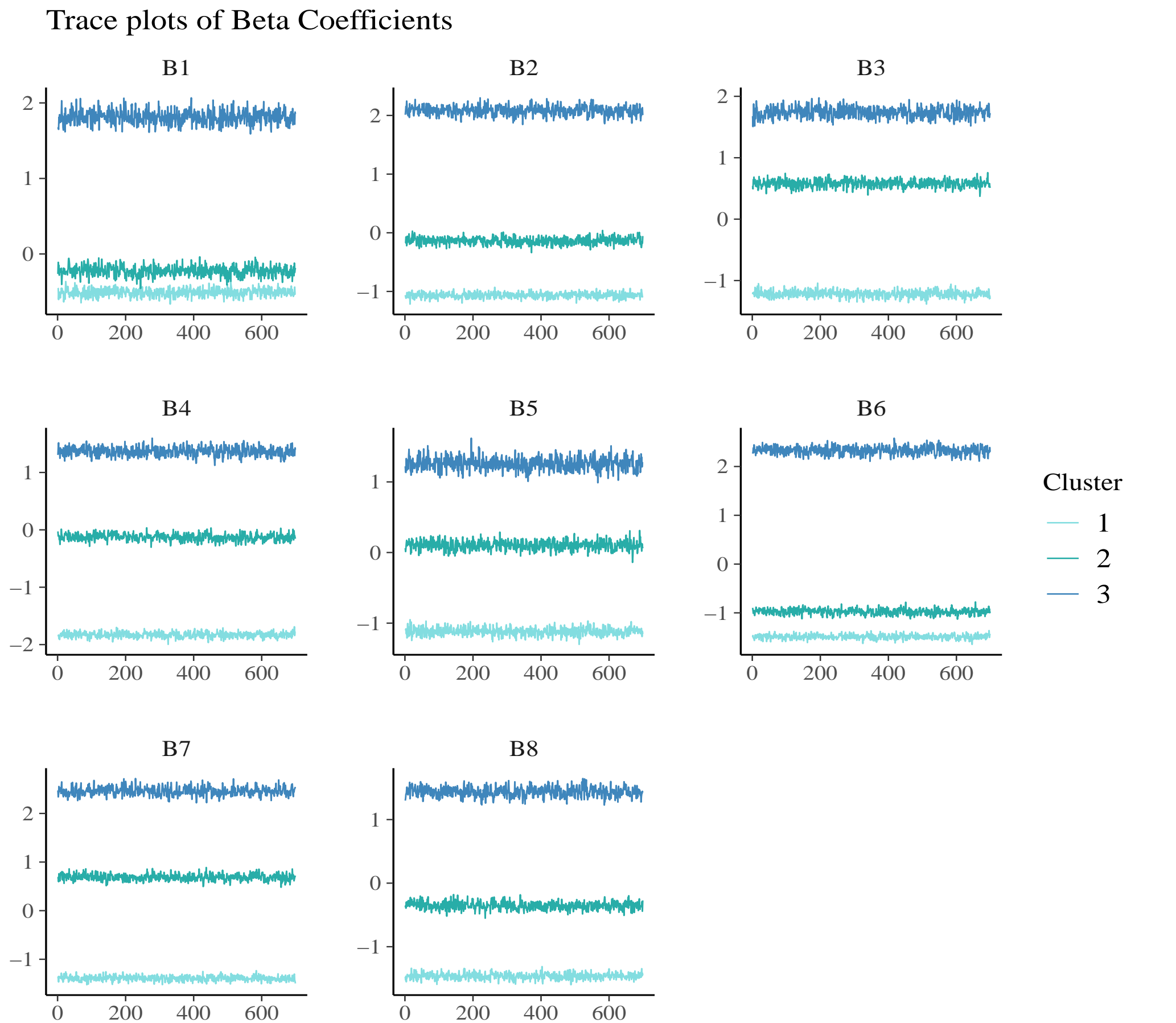


Figure 2: Trace plots of regression parameter estimates in simulation

## Application Results

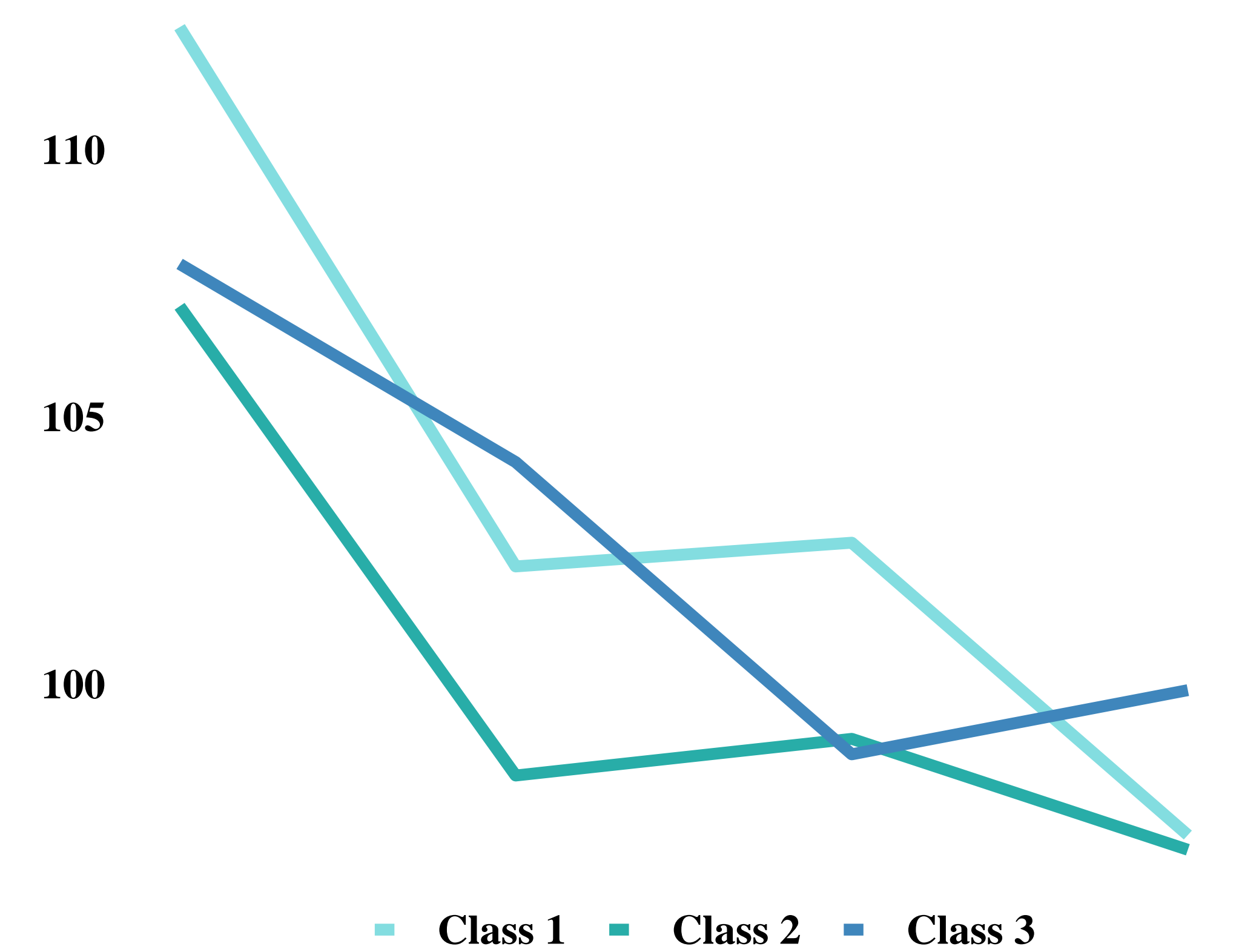


Figure 3: Mean development patterns in each estimated class

## Further Resources

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

(1) Fruhwirth-Schnatter, S and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate ... Biostatistics; (2) Benjamin-Neelon SE, Ostbye T, Bennett GG, et al. Cohort profile for the Nurture Observational Study ... BMJ Open 2017.

Funding: This work is supported by a grant from the NIH (R01DK094841) and a grant from the NLM (1R21LM012866-01).