1 Model

1.1 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.

1.2 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).

1.3 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 $n_k = \sum_{i=1}^n 1_{(z_i=k)}$ 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 $\boldsymbol{\zeta}_i = \mathbf{x}_i^T\boldsymbol{\beta}_k$ and $\boldsymbol{\alpha}_k = (\alpha_{k1},...,\alpha_{kJ})$. This implies that the i^{th} row of $\mathbf{Y}_k$, a $n_k \times J$ matrix of responses for cluster k, 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

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

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

1.4 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.

1.5 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.

1.6 Bayesian Inference

1.6.1 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 $\boldsymbol{\delta}_k$, $U_{ik} = 1_{z_i=k}$ 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.

1.6.2 Prior Choice

1.6.3 MCMC Algorithm

1.6.4 Assessment of MCMC Convergence

1.6.5 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**
    $n_k := \sum_{i=1}^{n} 1_{z_i = k}$
    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