Biometrics 64, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

A Bayesian multivariate mixture model for skewed longitudinal data with intermittent missing observations: An application to infant motor development

Carter Allen

Department of Biomedical Informatics, The Ohio State University, Columbus, OH, U.S.A.

Sara E. Benjamin-Neelon, PhD, JD

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

Brian Neelon, PhD*

Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, U.S.A.

**email:* neelon@musc.edu

Summary: In studies of infant growth, an important research goal is to identify latent clusters of infants with delayed motor development — a risk factor for adverse outcomes later in life. However, there are numerous statistical challenges in modeling motor development: the data are typically skewed, exhibit intermittent missingness, and are correlated across repeated measurements over time. Using data from the Nurture study, a cohort of approximately 600 mother-infant pairs, we develop a flexible Bayesian mixture model for the analysis of infant motor development. First, we model developmental trajectories using matrix skew normal distributions with cluster-specific parameters to accommodate dependence and skewness in the data. Second, we model the cluster membership probabilities using a Pólya-Gamma data-augmentation scheme, which improves predictions of the cluster membership allocations. Lastly, we impute missing responses from conditional multivariate skew normal distributions. Bayesian inference is achieved through straightforward Gibbs sampling. Through simulation studies, we show that the proposed model yields improved inferences over models that ignore skewness or adopt conventional imputation methods. We applied the model to the Nurture data and identified two distinct developmental clusters, as well as detrimental effects of food insecurity on motor development. These findings can aid investigators in targeting interventions during this critical early-life developmental window.

Key words: Conditional ignorability; Intermittent missing; Matrix skew normal; Mixture of experts; Motor development; Pólya-Gamma distribution.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Infant motor development is an important predictor of health later in life. Early motor development is associated with improved physical activity, cognitive function, and educational attainment (Aaltonen et al., 2015; Taanila et al., 2005), while delayed motor development is associated with increased sedentary time (Sánchez et al., 2017) and has been linked to adult cognitive disorders such as schizophrenia (Filatova et al., 2017). Thus, there is growing interest in identifying developmental patterns that may place infants at risk for long-term adverse health outcomes. One approach to tackling this problem is to identify underlying subgroups of infants with delayed motor development, and to isolate important predictors of subgroup membership. Our goal, therefore, is to introduce a flexible latent growth mixture model to detect high-risk developmental patterns and associated risk factors.

Our work is motivated by the Nurture study, a birth cohort of predominately black women and their infants residing in the southeast United States (Benjamin-Neelon et al., 2017). The overall aim of the study was to examine how infant feeding, physical activity, motor development, sleep, and stress contribute to infant weight gain. A second aim was to identify infant subpopulations that exhibit unique motor development trajectories, and to examine cluster-specific associations between household food security and motor development.

The Nurture data pose several statistical challenges. First, the repeated outcomes are correlated across measurement occasions. In the Nurture study, the pairwise correlations vary across time points, suggesting the need for a flexible error covariance structure. Second, the Nurture data are skewed, with the direction of skewness varying over time. The Nurture data also feature intermittent missingness throughout the study period. As such, we require a framework capable of addressing potentially non-ignorable missing data. Finally, we seek to develop a model that incorporates covariate information into both the multivariate regression model of infant development trajectories, as well as the clustering model.

To address these challenges, we propose a Bayesian multivariate mixture model for the analysis of longitudinal skewed infant motor development data with intermittent missing observations. Our approach builds on recent work on mixture models for skewed cross-sectional data. Frühwirth-Schnatter and Pyne (2010) proposed a multivariate skew normal model for high-dimensional flow cytometric data. However, their focus was on marginal inference (i.e., density estimation) rather than cluster-specific inferences, as is our focus here. More recently, Lin et al. (2018) proposed a mixture of skew-t factor analyzers for settings in which cluster-specific inference is of primary interest. However, like Frühwirth-Schnatter and Pyne (2010), their approach excluded covariates in the cluster-membership model, a focal point in our study as we expect demographics to not only play a key role in predicting cluster membership, but also help characterize developmental trajectories within clusters. Additionally, their approach, while quite flexible, relied on a computationally elaborate expectation–conditional maximization algorithm that does not enjoy the inferential benefits of a Bayesian approach. Finally, the authors adopted a single-imputation scheme for ignorable missing data that does not readily account for the uncertainty in the imputation process without additional multiple imputation steps.

Our proposed model extends these prior studies in a number of ways. First, our model enables cluster-specific inferences on longitudinal growth trajectories, while accommodating skewness patterns that may vary over time and across clusters. Second, we estimate parameters in a Bayesian framework that introduces covariates into the cluster membership model using a novel application of Pólya-Gamma data augmentation (Polson et al., 2013). Third, we accommodate intermittent missingness of longitudinal responses under a "conditional ignorability" assumption whereby the missing data mechanism is assumed to be ignorable conditional on cluster assignment. Marginally, we allow for dependence between the missing data mechanism and the missing responses, thus relaxing standard missing at random (MAR)

assumptions. We develop an online imputation procedure in which missing observations are updated iteratively conditional on cluster allocation. Finally, we propose a Bayesian inferential approach that makes use of convenient matrix skew-normal and skew-t representations. This induces closed-form full conditional updates for all model parameters, leading to efficient posterior sampling and straightforward implementation in available software.

2. Nurture Study

The Nurture study is a birth cohort of predominately black women and their infants residing in the southeastern United States from 2013 and 2017 (Benjamin-Neelon et al., 2017). The study followed mothers and infants for 12 months after birth and collected data on maternal feeding practices, infant physical activity and gross motor development, and household food security. Infant development was assessed quarterly at 3, 6, 9, and 12 months of age using the Bayley composite scale of motor development (Bayley, 2006), a standard measure of infant development ranging from 40 to 160, with higher scores indicating more advanced development compared to normally developing infants. Household food security study was assessed using the 18-item US Household Food Security Survey Module restricted to the 10 items related to household food security measured during pregnancy (USDA, 2019). Following standard protocol, a final dichotomous food security exposure was defined as "food insecure" households and "food secure" households. The Institutional Review Board of Duke University Medical Center approved this study and protocol.

Of the 666 infants who were consented into the study, 106 were missing Bayley score measurements at all timepoints. We restricted our analytic sample to the 560 remaining infants who had at least one non-missing Bayley score over the study period. Of the $560 \times 4 = 2240$ possible observations, 471 (21%) were missing, leaving an available-case sample size of 1769. Sample characteristics for the 560 participants are given in Web Table 1. In the sample, 68% of infants were black and 39% of households identified as food insecure during pregnancy.

The Bayley motor development scores ranged from 49.0 to 145.0 across visits, with a mean of 102.4 and standard deviation (SD) of 13.5. Figure 1 presents trajectory plots of the motor development scores for each infant in the available-case sample, with an overlay of the mean score at each visit. The plot indicates substantial heterogeneity in the trajectories, with a modest but statistically significant (coef $= -1.16$; $p < 0.01$) decline in scores, suggesting decreased motor development over time in the Nurture cohort relative to normally developing infants. However, because most of the literature on infant motor development has focused on the average effect over time (Shoaibi et al., 2019), little is known about trends for specific subgroups of interest – for example, among infants who may be at high-risk for delayed motor milestone achievement.

[Figure 1 about here.]

Figure 2 presents the scaled residual densities from a repeated-measures linear regression based on available cases with Bayley score as the outcome. The model included an unstructured covariance matrix and adjusted for the covariates in Web Table 1. The residuals were subset by visit to yield visit-specific residual density plots. As shown in Figure 2, the residuals are skewed at each visit, particularly at 3 and 6 months, with the direction of skewness varying over time. Shapiro-Wilk tests accounting for multiple testing rejected the null hypothesis of normality at 6 months, contravening standard assumptions. While there is a modest indication of skewness in the available-case sample, it is not clear how skewness patterns vary across latent subgroups of infants, or how missing observations impact skewness. We seek to answer these questions in subsequent analyses.

[Figure 2 about here.]

Additionally, the motor development scores are correlated over time, with pairwise correlations ranging in an unstructured pattern over time. As an illustration, we fit three repeated measures models with standardized Bayley score as the outcome and varying correlation

structures for the errors: AR1, compound symmetric and unstructured. The models included monthly visit as a categorical predictor and adjusted for the covariates in Web Table 1. The AIC values for these models were 27599, 27517, and 27478 respectively, indicating optimal fit under the unstructured pattern. We present the estimated correlation matrix from this model in Web Table 2. The pairwise correlations ranged from 0.14 to 0.27 across visits with no discernable structure. Finally, the Nurture data feature intermittent missing data, with approximately one third of the sample missing observations at each visit (Web Table 1). While it may be reasonable to assume that the missing data are missing at random (MAR), as we have no *a priori* reason to believe that the occurrence of missing observations is directly related to missing Bayley scores, we relax this assumption below by assuming ignorable missingness conditional on latent motor development cluster assignment.

3. Model

3.1 *Multivariate Skew Normal Mixture Model*

We propose a finite mixture model that accommodates relevant features of the data, namely skewness, missing values, and dependence among the responses. While alternative mixture models (e.g., Dirichlet process mixtures) provide flexibility for marginal inferences and density estimation, finite mixtures are appealing when the focus is on practical within-cluster inferences. In such cases, the primary goal is to identify a small number of clinically relevant clusters to help design targeted interventions to improve health outcomes. However, to avoid underfitting in finite mixtures, it is imperative to properly model the within-cluster distributions by accounting for important features, such as skewness or heavy tails. With this goal in mind, we present a repeated measures regression model based on a multivariate skew normal distribution — and by extension, a multivariate skew-t distribution — in which the Bayley scores across the J measurement occasions represent correlated responses.

Specifically, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$ be a $J \times 1$ vector of standardized Bayley scores for subject i $(i = 1, \ldots, n)$. We propose a mixture model of the form

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_{ki} f(\mathbf{y}_i | \boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k $(k = 1, ..., K)$ and π_{ki} is a subject-specific mixing weight representing the probability that subject i belongs to cluster k. For now we assume that K is fixed; we discuss model-selection strategies for choosing the optimal value of K in Section 3.5.3.

For posterior inference, we introduce a latent cluster indicator variable z_i taking the value $k \in \{1, ..., K\}$ with probability π_{ki}. Given $z_i = k$, we assume $\mathbf{y}_i$ is distributed according to a J-dimensional multivariate skew normal (MSN) density (Azzalini and Valle, 1996):

$$\begin{aligned} \mathbf{y}_i | (z_i = k) &\overset{ind}{\sim} \text{MSN}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k), \text{ with density} \\ f(\mathbf{y}_i | z_i = k) &= 2\phi_J(\mathbf{y}_i; \boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k)\Phi[\boldsymbol{\alpha}_k^T(\mathbf{y}_i - \boldsymbol{\zeta}_{ki})], \end{aligned} \tag{2}$$

where $\phi_J(\mathbf{y}_i; \boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k)$ denotes a J-dimensional normal density with mean $\boldsymbol{\zeta}_{ki}$ and covariance matrix $\boldsymbol{\Omega}_k$; $\Phi(\cdot)$ is the CDF of a scale standard normal random variable; $\boldsymbol{\zeta}_{ki}$ is a $J \times 1$ vector of subject- and 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 captures dependence among the J responses for subject i. The vector $\boldsymbol{\alpha}_k$ has components α_{kj}, $j = 1, ..., J$, that control the skewness of outcome j in cluster k. When $\boldsymbol{\alpha}_k = \mathbf{0}$, the MSN distribution reduces to the multivariate normal (MVN) distribution $\text{N}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k)$, where $\boldsymbol{\zeta}_{ki}$ represents a $J \times 1$ mean vector and $\boldsymbol{\Omega}_k$ is a $J \times J$ unstructured covariance matrix.

We can extend model (2) to the regression setting by modeling $\boldsymbol{\zeta}_{ki}$ as a function of covariates. Here, we adopt a convenient stochastic representation of the MSN density (Azzalini and Valle, 1996; Frühwirth-Schnatter and Pyne, 2010):

$$\mathbf{y}_i | (z_i = k, t_i) = \mathbf{X}_i \boldsymbol{\beta}_k + t_i \boldsymbol{\psi}_k + \boldsymbol{\epsilon}_{ki}, \tag{3}$$

where $\mathbf{X}_i$ is a $J \times Jp$ design matrix that includes potential time-dependent covariates; $\boldsymbol{\beta}_k = (\beta_{k11}, \ldots, \beta_{k1p}, \ldots, \beta_{kJ1}, \ldots, \beta_{kJp})^T$ is a $Jp \times 1$ vector of cluster- and outcome-specific regression coefficients; $t_i \sim \mathrm{N}_{[0,\infty)}(0,1)$ is a subject-specific standard normal random variable truncated below by zero; $\boldsymbol{\psi}_k = (\psi_{k1}, \ldots, \psi_{kJ})^T$ is a $J \times 1$ vector of cluster-specific parameters that control skewness; and $\boldsymbol{\epsilon}_{ki} \sim \mathrm{N}_J(\mathbf{0}, \boldsymbol{\Sigma}_k)$ is a $J \times 1$ vector of correlated error terms. Thus, conditional on t_i and $z_i = k$, $\boldsymbol{y}_i$ is distributed as $\mathrm{N}_J(\mathbf{X}_i\boldsymbol{\beta}_k + t_i\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$. Marginally (integrated over t_i), $\mathbf{y}_i|(z_i = k)$ is distributed $\mathrm{MSN}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$, where through back-transformation

$$\begin{aligned} \boldsymbol{\zeta}_{ki} &= \mathbf{X}_i\boldsymbol{\beta}_k, \\ \boldsymbol{\Omega}_k &= \boldsymbol{\Sigma}_k + \boldsymbol{\psi}_k\boldsymbol{\psi}_k^T, \\ \boldsymbol{\alpha}_k &= \frac{1}{\sqrt{1 - \boldsymbol{\psi}_k^T\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k}}\boldsymbol{\omega}_k\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k \quad \text{and} \\ \boldsymbol{\omega}_k &= \mathrm{Diag}(\boldsymbol{\Omega}_k)^{1/2}. \end{aligned} \tag{4}$$

As detailed in online supplement, conjugate full conditionals are available for all parameters in model (3), leading to straightforward Gibbs sampling. However, the Nurture analysis described in Section 5 involves no time-varying covariates, only time-varying covariate effects. In such cases, we can express the MSN density more compactly using a matrix skew normal (MatSN) representation. Structuring the data in this way greatly facilitates posterior computation by permitting low-dimensional matrix updates for the regression coefficients. For cluster k, let $\mathbf{Y}_k$ be an $n_k \times J$ matrix with rows $\mathbf{y}_i^T$ for $i = 1, ..., n_k$, where n_k is the number of subjects in cluster k. From equation (2), it follows that $\mathbf{Y}_k$ is distributed as

$$\mathbf{Y}_k \sim \mathrm{MatSN}_{n_k \times J}(\mathbf{X}_k\mathbf{B}_k, \boldsymbol{\alpha}_k, \mathbf{I}_{n_k}, \boldsymbol{\Omega}_k), \tag{5}$$

where $\mathbf{I}_{n_k}$ is the $n_k \times n_k$ identity matrix, and $\mathbf{X}_k$ and $\mathbf{B}_k$ are, respectively, $n_k \times p$ and $p \times J$ matrices of the form

$$\mathbf{X}_k = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_k 1} & \dots & x_{n_k p} \end{pmatrix} \quad \text{and} \quad \mathbf{B}_k = \begin{pmatrix} \beta_{k11} & \dots & \beta_{k1J} \\ \vdots & \ddots & \vdots \\ \beta_{kp1} & \dots & \beta_{kpJ} \end{pmatrix}.$$

Note that if we set $x_{i1} = 1$ for all i, then the first row of $\mathbf{B}_k$, $(\beta_{k11}, \dots, \beta_{k1J})$, represents time-specific intercepts that capture the time trend for the reference covariate group in cluster k. Adapting equation (7) from Chen and Gupta (2005), the density function for $\mathbf{Y}_k$ is

$$f(\mathbf{Y}_k) = 2^{n_k} \phi_{n_k \times J}(\mathbf{Y}_k; \mathbf{X}_k\mathbf{B}_k, \mathbf{I}_{n_k}, \mathbf{\Omega}_k) \Phi_{n_k}[(\mathbf{Y}_k - \mathbf{X}_k\mathbf{B}_k)\boldsymbol{\alpha}_k], \tag{6}$$

where $\phi_{n_k \times J}(\mathbf{Y}_k; \mathbf{X}_k\mathbf{B}_k, \mathbf{I}_{n_k}, \mathbf{\Omega}_k)$ is the density function for a matrix normal (MatNorm) random variable of dimension $n_k \times J$ with mean $\mathbf{X}_k\mathbf{B}_k$ and scale matrices $\mathbf{I}_{n_k}$ and $\mathbf{\Omega}_k$, and $\Phi_{n_k}(\cdot)$ denotes the CDF of an n_k-dimensional standard MVN random variable.

Further, let $\mathbf{t}_k = (t_1, ..., t_{n_k})^T$ denote the $n_k \times 1$ vector of latent variables for cluster k. By extending equation (3), it follows that the conditional distribution of $\mathbf{Y}_k$ given $\mathbf{t}_k$ is

$$\mathbf{Y}_k | \mathbf{t}_k \quad \sim \quad \text{MatNorm}_{n_k \times J}(\mathbf{X}_k^* \mathbf{B}_k^*, \mathbf{I}_{n_k}, \mathbf{\Sigma}_k) \tag{7}$$

where $\mathbf{X}_k^*$ is an $n_k \times (p+1)$ augmented design matrix formed by right column-binding $\mathbf{t}_k$ to $\mathbf{X}_k$, $\mathbf{B}_k^*$ is a $(p+1) \times J$ matrix of regression coefficients formed by lower row-binding $\boldsymbol{\psi}_k = (\psi_1, \dots, \psi_J)^T$ to $\mathbf{B}_k$, and $\mathbf{\Sigma}_k$ is the $J \times J$ covariance of $\boldsymbol{\epsilon}_{ik}$ in equation (3). This matrix normal representation admits conditionally conjugate prior distributions, which in turn leads to efficient Gibbs sampling for posterior inference. We formalize this in the following proposition, which establishes the conditional conjugacy of $\mathbf{B}_k^*$ and $\mathbf{\Sigma}_k$:

PROPOSITION 1: Let $\mathbf{B}_k^*$ and $\mathbf{\Sigma}_k$ in equation (7) have a joint Matrix Normal–Inverse Wishart (IW) prior, denoted MatNorm–IW$_{(p+1)\times J}(\mathbf{B}_{0k}^*, \mathbf{L}_{0k}, \nu_{0k}, \mathbf{V}_{0k})$, of the form

$$\begin{aligned} \pi(\mathbf{B}_k^*, \mathbf{\Sigma}_k) &= \pi(\mathbf{B}_k^* | \mathbf{\Sigma}_k) \pi(\mathbf{\Sigma}_k) \\ &= \text{MatNorm}_{(p+1)\times J}(\mathbf{B}_{0k}^*, \mathbf{L}_{0k}, \mathbf{\Sigma}_k) \text{IW}(\nu_{0k}, \mathbf{V}_{0k}), \end{aligned}$$

where $\mathbf{B}_{0k}^*$ is a $(p+1) \times J$ prior location matrix, $\mathbf{L}_{0k}$ and $\mathbf{V}_{0k}$ are, respectively, $(p+1) \times (p+1)$

and $J \times J$ prior scale matrices, and ν_{0k} denotes the prior degrees of freedom. Then, the full conditional distribution of $\mathbf{B}_k^*$ is $\text{MatNorm}_{(p+1)\times J}(\mathbb{B}_k^*, \mathbf{L}_k, \boldsymbol{\Sigma}_k)$, where

$$\begin{aligned} \mathbb{B}_k^* &= \mathbf{L}_k(\mathbf{L}_{0k}^{-1}\mathbf{B}_{0k}^* + \mathbf{X}_k^{*T}\mathbf{Y}_k) \text{ and} \\ \mathbf{L}_k &= (\mathbf{L}_{0k}^{-1} + \mathbf{X}_k^{*T}\mathbf{X}_k^*)^{-1}, \end{aligned}$$

and $\mathbf{X}_k^*$ is the augmented covariate matrix defined in equation (7). Likewise, the full conditional distribution of $\boldsymbol{\Sigma}_k$ is $\text{IW}(\nu_k, \mathbf{V}_k)$, where

$$\begin{aligned} \nu_k &= \nu_0 + n_k + p + 1 \text{ and} \\ \mathbf{V}_k &= \mathbf{V}_{0k} + (\mathbf{B}_k^* - \mathbf{B}_{0k}^*)^T\mathbf{L}_{0k}^{-1}(\mathbf{B}_k^* - \mathbf{B}_{0k}^*) + (\mathbf{Y}_k - \mathbf{X}_k^*\mathbf{B}_k^*)^T(\mathbf{Y}_k - \mathbf{X}_k^*\mathbf{B}_k^*). \end{aligned}$$

The proof is provided in Web Appendix A.

3.2 *Pólya–Gamma Multinomial Regression for Cluster Probabilities*

To accommodate heterogeneity in the cluster-membership probabilities, we model π_{ki} as a function of covariates using a multinomial logit model

$$\pi_{ki} = \Pr(z_i = k|\mathbf{w}_i) = \frac{\mathrm{e}^{\mathbf{w}_i^T\boldsymbol{\delta}_k}}{\sum_{h=1}^K \mathrm{e}^{\mathbf{w}_i^T\boldsymbol{\delta}_h}}, \ k = 1, \ldots, K, \tag{8}$$

where $\mathbf{w}_i$ is an $r \times 1$ vector of subject-level covariates, $\boldsymbol{\delta}_k$ is a $r \times 1$ vector of cluster-specific regression parameters. For identifiability, we choose category K as reference and set $\boldsymbol{\delta}_K = \mathbf{0}$. By allowing the cluster probabilities to vary across subjects, model (1) can be viewed as a *mixture of experts* model, in which π_{ki} acts as a *gating function* controlling the prior probability of membership in cluster k, and $f(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is the "expert" providing information on the within-cluster distribution of $\mathbf{y}_i$ (Bishop, 2006). An appealing feature of the mixture of experts model is the ability to discern cluster membership using information contained in the covariates $\mathbf{w}_i$. As a result, the gating functions can yield accurate cluster allocations even when the cluster-specific densities $f(\mathbf{y}_i|\boldsymbol{\theta}_k)$ are similar across clusters.

To facilitate sampling, we adopt the efficient data-augmentation approach introduced by Polson et al. (2013), which expresses the inverse-logit function as a scale-normal mixture

of Pólya–Gamma densities. By using Pólya–Gamma data augmentation for the multinomial model, we obtain a *Pólya–Gamma mixture of experts model* — a computationally efficient way to obtain inferences for the mixing weights in a Bayesian setting. A random variable w is said to follow a Pólya–Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$ if

$$w \sim \text{PG}(b, c) \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{s=1}^{\infty} \frac{g_s}{(s - 1/2)^2 + c^2/(4\pi^2)}, \tag{9}$$

where $g_s \stackrel{iid}{\sim} \text{Ga}(b, 1)$ for $s = 1, ..., \infty$. Polson et al. (2013) establish two key properties of the $\text{PG}(b, c)$ density. First, for $a, \eta \in \mathbb{R}$,

$$\frac{(\text{e}^{\eta})^a}{(1 + \text{e}^{\eta})^b} = 2^{-b}\text{e}^{\kappa\eta} \int_0^{\infty} \text{e}^{-w\eta^2/2} p(w|b, c = 0)dw, \tag{10}$$

where $\kappa = a - b/2$ and $p(w|b, c = 0)$ denotes a $\text{PG}(b, 0)$ density. If we choose $a = y$ and $b = 1$, then the left-hand side of equation (10) has the same functional form as a logistic regression likelihood for a binary outcome y. Note also that the right-hand side includes the kernel of a normal distribution for a random variable η with precision w. Next, the conditional distribution $p(w|b, c)$ results from "exponential tilting" the $\text{PG}(b, 0)$ density:

$$p(w|b, c) = \frac{\text{e}^{-c^2w/2} p(w|b, 0)}{\text{E}_w[\text{e}^{-c^2w/2}]} = \frac{\text{e}^{-c^2w/2} p(w|b, 0)}{\int_0^{\infty} \text{e}^{-c^2w/2} p(w|b, 0)dw}. \tag{11}$$

These results imply that, for a logistic regression model, the likelihood can be written as a scale-mixture of normal densities with Pólya-Gamma precision terms w, resulting in closed-form MVN full conditional distributions for logistic regression parameters. Details can be found in Polson et al. (2013). To extend the augmentation approach to the multinomial setting, we first introduce the binary indicators U_{ki}, such that $U_{ki} = 1_{(z_i=k)}$; that is, U_{ki} is an indicator variable taking the value 1 if subject i belongs to cluster k, and 0 otherwise. The conditional distribution of $\boldsymbol{\delta}_k$, given $\mathbf{U}_k = (U_{k1}, ..., U_{kn})^T$ and the remaining regression coefficients $\boldsymbol{\delta}_{h\neq k}$, is

$$p(\boldsymbol{\delta}_k|\mathbf{z}, \boldsymbol{\delta}_{h\neq k}) = p(\boldsymbol{\delta}_k|\mathbf{U}_k, \boldsymbol{\delta}_{h\neq k}) \propto p(\boldsymbol{\delta}_k) \prod_{i=1}^{n} \pi_{ki}^{U_{ki}} (1 - \pi_{ki})^{1-U_{ki}}, \tag{12}$$

where $p(\boldsymbol{\delta}_k)$ denotes the prior distribution of $\boldsymbol{\delta}_k$, U_{ki} is defined above, and π_{ki} is defined as

in equation (8). We can rewrite π_{ki} in terms of U_{ki} as

$$\pi_{ki} = \Pr(U_{ki} = 1) = \frac{\mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_k}}{\sum_{h=1}^{K} \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h}} = \frac{\mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_k}}{\sum_{h \neq k}^{K} \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h} + \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_k}},$$

where dividing throughout by $\sum_{h \neq k}^{K} \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$ yields

$$\pi_{ki} = \frac{\mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}}}{1 + \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}}} = \frac{\mathrm{e}^{\eta_{ki}}}{1 + \mathrm{e}^{\eta_{ki}}},$$

with $c_{ki} = \log \sum_{h \neq k} \mathrm{e}^{\mathbf{w}_i^T \boldsymbol{\delta}_h}$ and $\eta_{ki} = \mathbf{w}_i^T \boldsymbol{\delta}_k - c_{ki}$. We can use these quantities to re-express equation (12) as

$$\begin{aligned} p(\boldsymbol{\delta}_k | \mathbf{z}, \boldsymbol{\delta}_{h \neq k}) &\propto p(\boldsymbol{\delta}_k) \prod_{i=1}^{n} \left(\frac{\mathrm{e}^{\eta_{ki}}}{1 + \mathrm{e}^{\eta_{ki}}} \right)^{U_{ki}} \left(\frac{1}{1 + \mathrm{e}^{\eta_{ki}}} \right)^{1 - U_{ki}} \\ &= p(\boldsymbol{\delta}_k) \prod_{i=1}^{n} \frac{(\mathrm{e}^{\eta_{ki}})^{U_{ki}}}{1 + \mathrm{e}^{\eta_{ki}}}, \end{aligned} \tag{13}$$

where the product term denotes the likelihood from a logistic regression model. We can therefore apply the Pólya–Gamma sampler for logistic regression to update each $\boldsymbol{\delta}_k$ one at a time based on the binary indicators U_{ki}. First, we define for $k = 1, ..., K$, the $n \times 1$ vector $\mathbf{U}_k^* = \left(\frac{U_{k1} - 1/2}{w_{k1}} + c_{k1}, ..., \frac{U_{kn} - 1/2}{w_{kn}} + c_{kn} \right)^T$. As shown in Web Appendix B, the conditional distribution of $\mathbf{U}_k^*$ given $\boldsymbol{w} = (w_{k1}, ..., w_{kn})^T$ is $\mathrm{N}_n(\mathbf{W}\boldsymbol{\delta}_k, \mathbf{O}_k^{-1})$, where $\mathbf{O}_k = \mathrm{Diag}(w_{k1}, ..., w_{kn})$ and $\mathbf{W}$ is an $n \times r$ design matrix with rows $\mathbf{w}_i^T$ for $i = 1, \ldots, n$. Thus, the full conditional distribution of $\boldsymbol{\delta}_k$ is given by

$$p(\boldsymbol{\delta}_k | \mathbf{z}, \mathbf{O}_k, \boldsymbol{\delta}_{h \neq k}) \propto p(\boldsymbol{\delta}_k) \exp \left\{ -\frac{1}{2} (\mathbf{U}_k^* - \mathbf{W}\boldsymbol{\delta}_k)^T \mathbf{O}_k (\mathbf{U}_k^* - \mathbf{W}\boldsymbol{\delta}_k) \right\}. \tag{14}$$

Assuming a $\mathrm{N}_r(\mathbf{d}_{0k}, \mathbf{S}_{0k})$ prior for $\boldsymbol{\delta}_k$, we have the following Gibbs updates for the clustering model:

(1) For $k = 1, ..., K-1$ and $i = 1, \ldots, n$, update the Pólya–Gamma weight, w_{ki}, from its $\mathrm{PG}(1, \eta_{ki})$ full conditional.

(2) For $k = 1, ..., K-1$, update $\boldsymbol{\delta}_k$ from its $\mathrm{N}_r(\mathbf{d}_k, \mathbf{S}_k)$ full conditional, where

$$\begin{aligned} \mathbf{S}_k &= (\mathbf{S}_{0k}^{-1} + \mathbf{W}^T \mathbf{O}_k \mathbf{W})^{-1}, \text{ and} \\ \mathbf{d}_k &= \mathbf{S}_k (\mathbf{S}_{0k}^{-1} \mathbf{d}_{0k} + \mathbf{W}^T \mathbf{O}_k \mathbf{U}_k^*). \end{aligned}$$

(3) For $i = 1, \ldots, n$, draw the latent cluster indicator, z_i, from its multinomial full conditional as described in Web Appendix B.

3.3 *Extensions to Multivariate Skew-t Distributions*

To accommodate outliers and heavy tails, we extend equation (1) by assuming, conditional on $z_i = k$, that $\mathbf{y}_i$ follows a multivariate skew-t (MST) distribution (Gupta, 2003):

$$\begin{aligned} \mathbf{y}_i|(z_i = k) &\overset{ind}{\sim} \text{MST}_J(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k, \nu_k), \text{ with density} \\ f(\mathbf{y}_i|z_i = k) &= 2f_{t_J}(\mathbf{y}_i; \boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k, \nu_k) T_{\nu_k+J}\left(\boldsymbol{\alpha}_k^T(\mathbf{y}_i - \boldsymbol{\zeta}_{ki})\sqrt{\frac{\nu_k + J}{\nu_k + Q_{y_i}}}\right), \end{aligned} \tag{15}$$

where $f_{t_J}(\mathbf{y}_i; \boldsymbol{\zeta}_{ki}, \boldsymbol{\Omega}_k, \nu_k)$ denotes the CDF of a J-dimensional t distribution with location $\boldsymbol{\zeta}_{ki}$, covariance $\boldsymbol{\Omega}_k$, and fixed degrees of freedom ν_k that may vary across clusters; T_{ν_k+J} denotes the distribution function of the scalar standard t distribution with $\nu_k + J$ degrees of freedom; and $Q_{y_i} = (\mathbf{y}_i - \boldsymbol{\zeta}_{ki})^T \boldsymbol{\Omega}_k^{-1} (\mathbf{y}_i - \boldsymbol{\zeta}_{ki})$. As before, we adopt a stochastic representation for $\mathbf{y}_i$ to facilitate Gibbs sampling (Frühwirth-Schnatter and Pyne, 2010). Specifically, we augment the MSN stochastic representation in equation (3) by introducing subject-specific scale terms, d_i, yielding an MST regression of the form:

$$\mathbf{y}_i|(z_i = k, t_i, d_i) = \mathbf{X}_i\boldsymbol{\beta}_k + \frac{t_i}{\sqrt{d_i}}\boldsymbol{\psi}_k + \frac{1}{\sqrt{d_i}}\boldsymbol{\epsilon}_{ki}, \tag{16}$$

where $d_i \sim \text{Gamma}\left(\frac{\xi}{2}, \frac{\xi}{2}\right)$, with ξ being a pre-specified degrees of freedom parameter shared across clusters, and t_i and $\boldsymbol{\epsilon}_{ki}$ are defined as in equation (3). In principle, ξ may vary across clusters (becoming ξ_k), though we set it to be fixed across clusters here for simplicity. For posterior inference, we first draw $d_i|(z_i = k)$ from its Gamma full conditional as described in Web Appendix B. We then form cluster-specific scaled response and design matrices $\tilde{\mathbf{Y}}_k = \sqrt{\mathbf{d}_k} \circ \mathbf{Y}_k$ and $\tilde{\mathbf{X}}_k = \sqrt{\mathbf{d}_k} \circ \mathbf{X}_k^*$ from equation (7), where the $n_k \times 1$ vector $\sqrt{\mathbf{d}_k}$ has elements $\sqrt{d_i}$ for all i in cluster k and "$\circ$" denotes the Hadamard product. Finally, we fit equation (7) to the scaled data to update the remaining model parameters.

3.4 *Cluster-Specific Imputation under Conditional Ignorability*

To accommodate intermittent missing data, we propose a convenient imputation algorithm in which we assume that the missingness mechanism is conditionally ignorable given the cluster indicators z_i, extending recent work on latent class pattern mixture models for informative dropout (Roy, 2007). Here, z_i functions as discrete shared parameter that induces unobserved association between the missingness process and the missing data. Suppose $\mathbf{y}_i$ has $q_i \in (1, \ldots, J)$ observed values, denoted $\mathbf{y}_i^{obs}$, and $J - q_i$ intermittent missing values, denoted $\mathbf{y}_i^{miss}$. Let $\mathbf{R}_i = (R_{i1}, \ldots, R_{iJ})^T$ be a $J \times 1$ vector of binary response indicators, such that $R_{ij} = 1$ if infant i has a Bayley measurement at visit j. Under conditional ignorability, the conditional distribution of $\mathbf{R}_i$ given $(z_i, \mathbf{y}_i^{obs}, \mathbf{y}_i^{miss})$ is

$$f(\mathbf{R}_i | z_i = k, \mathbf{y}_i^{obs}, \mathbf{y}_i^{miss}, \mathbf{X}_i, \boldsymbol{\gamma}_k) = f(\mathbf{R}_i | z_i = k, \mathbf{y}_i^{obs}, \mathbf{X}_i, \boldsymbol{\gamma}_k) \tag{17}$$

where, in this context, $\mathbf{X}_i$ is a $J \times m$ design matrix and $\boldsymbol{\gamma}_k$ is an $m \times 1$ vector of cluster-specific parameters related to the missing data mechanism.

Under conditional ignorability, conditioning on z_i ensures that $\mathbf{R}_i$ does not depend on the missing observations $\mathbf{y}_i^{miss}$. We can therefore impute $\mathbf{y}_i^{miss}$ from its conditional MVN distribution given $(z_i, t_i, \mathbf{y}_i^{obs})$ as follows. Let $\boldsymbol{\mu}_{ki} = \mathbf{X}_i \boldsymbol{\beta}_k + t_i \boldsymbol{\psi}_k$ from equation (3). We first partition $\boldsymbol{\mu}_{ki}$ and $\boldsymbol{\Sigma}_k$, the variance of $\boldsymbol{\epsilon}_{ki}$ in equation (3), as

$$\boldsymbol{\mu}_{ki} = \begin{pmatrix} \boldsymbol{\mu}_{ki}^{miss} \\ \boldsymbol{\mu}_{ki}^{obs} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_k = \begin{pmatrix} \boldsymbol{\Sigma}_{k11} & \boldsymbol{\Sigma}_{k12} \\ \boldsymbol{\Sigma}_{k21} & \boldsymbol{\Sigma}_{k22} \end{pmatrix};$$

we then impute $\mathbf{y}_i^{miss}$ according to

$$\begin{aligned} \mathbf{y}_i^{miss} | (z_i = k, t_i, \mathbf{y}_i^{obs}) &\sim \mathrm{N}_{J-q_i}(\boldsymbol{\mu}_{ki}^{cond}, \boldsymbol{\Sigma}_k^{cond}), \text{ where} \\ \boldsymbol{\mu}_{ki}^{cond} &= \boldsymbol{\mu}_{ki}^{miss} + \boldsymbol{\Sigma}_{k12} \boldsymbol{\Sigma}_{k22}^{-1} (\mathbf{y}_i^{obs} - \boldsymbol{\mu}_{ki}^{obs}) \text{ and} \\ \boldsymbol{\Sigma}_k^{cond} &= \boldsymbol{\Sigma}_{k11} - \boldsymbol{\Sigma}_{k12} \boldsymbol{\Sigma}_{k22}^{-1} \boldsymbol{\Sigma}_{k21}. \end{aligned} \tag{18}$$

These results follow from conventional multivariate normal theory. Note that, while the complete data vector $\mathbf{y}_i = \{\mathbf{y}_i^{obs}, \mathbf{y}_i^{miss}\}$ follows a MVN distribution conditional on t_i, after

marginalizing over t_i, $\mathbf{y}_i$ follows a joint MSN distribution. Thus, the proposed conditional imputation procedure provides a convenient way of imputing missing MSN responses using samples from more standard densities.

Finally, given $z_i = k$, we independently model the J response indicators for infant i as

$$\begin{aligned}(R_{ij}|z_i = k, b_{ik}) &\overset{ind}{\sim} \text{Bern}(\phi_{ijk}), \quad j = 1, \ldots, J \\ \text{logit}(\phi_{ijk}) &= \mathbf{x}_{ij}^T \boldsymbol{\gamma}_k + b_{ki}, \end{aligned} \tag{19}$$

where $\mathbf{x}_{ij}$ is an $m \times 1$ vector of covariates, and $\boldsymbol{\gamma}_k$ is the $m \times 1$ vector of cluster-specific regression parameters from equation (17). Equation (19) may additionally include terms for the observed outcomes $\mathbf{y}_i^{obs}$ (e.g., baseline Bayley score). Because the response indicators may be correlated over time, we also include a subject-level random intercept b_{ki} conditionally distributed as $\text{N}(0, \sigma_k^2)$ given $z_i = k$. Although we assume conditional ignorability of $\mathbf{R}_i$ and $\mathbf{y}_i^{miss}$ given z_i, because the ϕ_{ijk} terms from model (19) appear in the full conditional update for z_i (Web Appendix B), $\mathbf{R}_i$ and $\mathbf{y}_i^{miss}$ are marginally correlated, resulting in a marginal missing not at random (MNAR) mechanism.

3.5 *Bayesian Inference*

3.5.1 *Prior Specification.* We adopt a Bayesian approach and assign prior distributions to all model parameters. For designs not involving time-dependent covariates, we assign a joint $\text{MatNorm–IW}_{(p+1)\times J}(\mathbf{B}_{0k}^*, \mathbf{L}_{0k}, \nu_{0k}, \mathbf{V}_{0k})$ to $(\mathbf{B}_k^*, \boldsymbol{\Sigma}_k)$ as described in Proposition 1. For time-varying designs, we assign independent MVN priors to $\boldsymbol{\beta}_k$ and $\boldsymbol{\psi}_k$ from equation (3); details are provided in Step 5(b) of Web Appendix B. For the multinomial logit model, the regression parameters $\boldsymbol{\delta}_k = (\delta_{k1}, ..., \delta_{kr})^T$ are assigned a $\text{N}_r(\mathbf{d}_{0k}, \mathbf{S}_{0k})$ prior for $k = 1, ..., K-1$, which is conditionally conjugate under the Pólya–Gamma sampling scheme described in Section 3.2. Finally, from equation (19), we assume a $\text{N}_m(\mathbf{g}_{0k}, \mathbf{G}_{0k})$ prior for $\boldsymbol{\gamma}_k$ and an inverse-gamma $\text{IG}(\lambda_{1k}, \lambda_{2k})$ prior for σ_k^2, where λ_{2k} is a scale parameter. In general, hyperparameters can

vary across clusters, though they may be shared across clusters in practice. For the skew-t model, we assume $d_i \sim \text{Gamma}\left(\frac{\xi}{2}, \frac{\xi}{2}\right)$, where ξ is a pre-specified value. More generally, one can place a Gamma prior on ξ and use Metropolis-Hastings for posterior updating.

3.5.2 *Posterior Computation.* The above prior specification induces closed-form full conditionals for all model parameters, which can be efficiently updated as part of the Gibbs sampler outlined below. Details of the MCMC sampler are provided in Web Appendix B.

(1) For $i = 1, \ldots, n$, impute $\mathbf{y}_i^{miss}$ from its MVN full conditional as described in equation (18). Conclude by constructing a complete outcome vector $\mathbf{y}_i$.

(2) Using Pólya–Gamma augmentation, update the parameters for the missing data model in equation (19) from their full conditionals, as described in Web Appendix B.

(3) Draw the cluster-membership parameters, w_{ki}, $\boldsymbol{\delta}_k$ and z_i, from their full conditionals, as described at the end of Section 3.2.

(4) For $i = 1, \ldots, n$, update the latent truncated normal variable, t_i, from its cluster-specific truncated normal full conditional given $z_i = k$.

(5) For analysis involving no time-varying covariates, update $\mathbf{B}_k^*$ and $\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, K)$ using the results from Proposition 1. For the MST model, we additionally draw latent scale terms d_i $(i = 1, \ldots, n)$ from their Gamma full conditionals, construct the augmented data $\tilde{\mathbf{Y}}_k$ and $\tilde{\mathbf{X}}_k$, and use these in the remaining updates. We back-transform using equation (4) to recover the original MSN and MST parameters. For the skew-t model, the back-transformations will employ $\tilde{\mathbf{Y}}_k$ and $\tilde{\mathbf{X}}_k$. For designs with time-varying covariates, we update $\boldsymbol{\beta}_k$ and $\boldsymbol{\psi}_k$ from equation (3) using MSN or MST updates as described in Steps 5(b) and 6(b) of Appendix B.

3.5.3 *Assessment of MCMC Convergence, Label Switching, and Model Selection.* We monitor MCMC convergence through standard diagnostics, such trace plots and effective

sample sizes. To address label switching, a common issue for Bayesian mixture models, we implemented the iterative ECR relabeling algorithm included in the `label.switching` package in `R` (Papastamoulis, 2016). In our simulation studies and application, we observed immediate convergence of the ECR algorithm, indicating no evidence of label switching in our analyses. Because our primary objective is to identify a small number of clinically meaningful motor development clusters, we adopt the widely applicable information criterion (WAIC) to select the number of clusters K (Watanabe, 2010). In Section 4.3, we demonstrate that this measure accurately recovers the true number of clusters under realistic parameter settings.

4. Simulation Studies

4.1 *Simulation to Compare the MSN Model to the MVN Model*

Our first simulation compared MSN and MVN mixture models to investigate whether ignoring skewness leads to poor inferences in a setting resembling the Nurture study. To emulate the Nurture study, we simulated $n = 1000$ subjects from the following model

$$f(\mathbf{y}_i) = \sum_{k=1}^{3} \pi_{ki} f(\mathbf{y}_i|\boldsymbol{\theta}_k), \tag{20}$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{i4})^T$ to conform to the $J = 4$ measurement occasions in the Nurture study; $\boldsymbol{\theta}_k$ is the set of parameters specific to cluster k ($k = 1, 2, 3$), and $f(\mathbf{y}_i|\boldsymbol{\theta}_k) \overset{\text{d}}{=} \text{MSN}_4(\boldsymbol{\zeta}_{ki}, \boldsymbol{\alpha}_k, \boldsymbol{\Omega}_k)$; $\boldsymbol{\zeta}_{ki} = (\zeta_{ki1}, ..., \zeta_{ki4})^T$, $\zeta_{ki1} = \beta_{kj1} + \beta_{kj2} x_i$, and x_i is a $\text{N}(0, 1)$ covariate whose effect varies across the J measurement occasions. We modeled the cluster probabilities in equation (8) as a function of an intercept and one baseline covariate, w_{i1}, implying that $r = 2$. We did not introduce missing data into this simulation, as we address missing data in the second simulation study. As a result, the total number of complete measurements was $N = n \times J = 4000$. The generated data included $n_1 = 318$ infants in cluster 1, $n_2 = 288$ in cluster 2, and $n_3 = 394$ in cluster 3.

Because the model included no time-varying covariates — only time-varying effects — we

used the matrix normal formulation given in Proposition 1, yielding a $(p+1) \times J = 3 \times 4$ matrix $\mathbf{B}_k^*$. We chose the matrix normal hyperparameters described in Section 3.5.1 to be homogeneous across the three clusters by setting, for $k = 1, 2, 3$, $\mathbf{B}_{0k}^* = \mathbf{0}_{3\times 4}$, $\mathbf{L}_{0k} = \mathbf{I}_3$, $\mathbf{V}_{0k} = \mathbf{I}_4$, and $\nu_{0k} = J + 2 = 6$, which gives $\mathrm{E}(\boldsymbol{\Sigma}_k) = \mathbf{I}_4$. Similarly, for the clustering model, we set $\mathbf{d}_{01} = \mathbf{d}_{02} = (0, 0)^T$ and $\mathbf{S}_{01} = \mathbf{S}_{02} = \mathbf{I}_2$, noting that $k = 3$ is the reference cluster. To investigate the effect of ignoring skewness, we allowed the vector of skewness parameters, $\boldsymbol{\alpha}_k$, to vary across clusters; for cluster 3, we assumed no skewness ($\boldsymbol{\alpha}_k = \mathbf{0}$), implying MVN data for this cluster. We then fit both the MSN and MVN mixture models to data generated from model (20). We ran the MCMC for 10000 iterations with a burn-in of 1000. MCMC diagnostics indicated rapid convergence and excellent mixing (Web Figure 1).

The WAIC values for the MSN and MVN mixture models were 12112 and 17499, respectively, indicating better fit for the MSN model, as expected. Table 1 presents posterior mean estimates and 95% credible intervals for cluster 1 from the MSN and MVN models. Web Table 3 presents the results for the other two clusters. As expected, the MSN model provided accurate estimates throughout, whereas the MVN model consistently produced incorrect estimates with poor coverage when data were skewed, as in clusters 1 and 2. In particular, ignoring skewness inflated the variance estimates under the MVN model as a way to compensate for the skewness in the data. However, when data were not skewed, as in cluster 3, both models performed similarly (Web Table 3). Thus, the MSN model can be reliably used in place of the MVN model even when data are not overtly skewed.

[Table 1 about here.]

4.2 *Simulation to Compare Imputation Methods*

Next, we compared three competing methods for imputing missing response values. The first method was the MNAR online imputation approach proposed in Section 3.4. We also fit MAR online imputation and Bayesian multiple imputation (MI). We defined MAR online

imputation to be imputation of $\mathbf{y}_i^{miss}$ as in equation (18) under a global MAR assumption. Thus, we ignored the missing data model (19) when updating the cluster indicators z_i, implying marginal as well as conditional ignorability. Finally, we implemented Bayesian MI, the standard approach for handling missing data whereby many imputed data sets are generated via MVN imputation of missing responses, MCMC is performed on each imputed data set, and the posterior samples are pooled for final inference. Here, the missing responses are imputed under a marginal MVN assumption prior to modeling, and hence this imputation method ignored skewness, clustering, and the missing data model.

To compare the three methods, we generated $n = 1000$ observations from a three-cluster ($K = 3$) MSN mixture model similar in design to simulation 1. We then removed observations intermittently across the 4 measurement occasions according to model (19), which included three continuous covariates but no fixed intercept, implying $m = 3$ from equation (19). The model also included a random intercept with a common variance of $\sigma_k^2 = 1$ across clusters. After removing missing data, the number of available measurements in each cluster was $N_1 = 1463$, $N_2 = 819$ and $N_3 = 1209$. We ran each model for 10000 iterations with a burn-in of 1000. MCMC diagnostics showed rapid convergence (see Web Figure 2 for diagnostics for the MNAR online model).

As described by Zhou and Reiter (2010), Bayesian MI requires a large number of imputations. Accordingly, we generated 1000 imputed data sets by assuming a MVN distribution for $\mathbf{y}_i$, where the mean and covariance parameters were estimated from the observed responses. The MI procedure required 36 hours of computation utilizing parallelization across 8 cores, compared to 17 minutes for non-parallelized versions of the online MNAR and MAR methods, illustrating the computational gains of online imputation. Results for cluster 1 are presented in Table 2. Results for clusters 2 and 3 are given in Web Table 4.

[Table 2 about here.]

As shown in Table 2, the online MNAR imputation method is the only method of the three that recovered true parameter values. Ignoring or making incorrect assumptions regarding the missing data mechanism resulted in incorrect inference. The poor performance of Bayesian MI method is likely due to the fact that it ignores clustering during the imputation phase and instead imputes missing responses under a marginal MVN distribution. Thus, it ignores important modeling considerations such as skewness and clustering. By comparison, MAR imputation does make cluster-specific inferences; however, it does not take into account the missing data model, which again results in poor inference. Results for the other two clusters show similar behavior (Web Table 4). Taken together, these results suggest that under the conditional ignorability assumption described in Section 3.4, ignoring the missing data model (19) can lead to highly biased estimates.

4.3 *Simulation to Validate Choice of K*

We conducted a final simulation to validate the use of WAIC for determining the number of clusters, K. We generated data sets from MSN models with $K = \{2, 3, 4, 5\}$. For each simulated data set, we fit the proposed Bayesian MSN model with $K = \{2, 3, 4, 5\}$ and computed WAIC in each case. For each scenario, we ran the MCMC algorithms for 10000 iterations with a burn-in of 1000. MCMC diagnostics indicated rapid convergence for all models (Web Figure 3). As shown in Web Table 5, the WAIC measure recovered the true value of K in all cases. For some simulations (e.g., true $K = 2$), we were unable to fit the MSN model when the fitted K was large due to the occurrence of vacant clusters during MCMC sampling. We have found that this generally occurs when the data do not support large values of K.

5. Application to Nurture Study

We applied our proposed model to the Nurture data by fitting an MSN mixture model that included standardized Bayley scores as the response, indicators for the four study visits, and binary food security status as the exposure of interest. The model also included time-invariant birth weight for gestational age z-score, number of children in the household, and an indicator for breastfeeding, as these likely impact infant development within each cluster. We excluded an intercept, but allowed the covariate effects to vary over time, resulting in a parameter dimension of $p = 20$ for this component of the model (Table 3). For the multinomial logit cluster-membership model, we included an intercept, birth weight for gestational age z-score, infant race, and infant gender as covariates, as these variables are believed to affect the placement of infants into latent development clusters. The 471 missing measurements were imputed using the online MNAR imputation method described in Section 3.4. The missing data model (19) included a fixed intercept, birth weight for gestational age z-score, infant gender, infant race, and a random intercept. To select the number of clusters, we fit several MSN models with varying specifications for K, and used WAIC to chose the best fitting model. The WAIC values were 9141, 10088, 11203, and 11410 for $K = 2, 3, 4, 5$, respectively. We also fit 3-df MST models with $2, 3, 4$, and 5 clusters; these yielded WAIC values of 13228, 13934, 14002, and 14356 respectively, suggesting that the 2-cluster MSN model provided best fit among all models considered. We ran each model for 10000 MCMC iterations, with a burn in of 1000. We observed fast MCMC convergence in all cases with no evidence of label switching. MCMC diagnostics for the 2-cluster MSN model are presented in Web Figure 4.

Table 3 presents posterior means and 95% credible intervals (CrIs) for the 2-cluster model.

[Table 3 about here.]

In cluster 1, we observed a significant detrimental effect of food insecurity at each timepoint. However, in cluster 2, we only observed a significant detrimental effect of food insecurity at

months 9 and 12, though the effect sizes were more modest than in cluster 1. These trends are also displayed in Figure 3. We observed a significant positive effect of breastfeeding in cluster 1, but not in cluster 2, suggesting that breastfeeding may especially benefit infants exhibiting delayed motor development. We did not observe a significant effect of either birth weight z-score or number of children in the household. From the Pólya–Gamma multinomial logit component, we found that female infants were more likely to belong to cluster 1. From the missing data model, the intercepts suggest more missing observations for infants in cluster 1 compared to those in cluster 2, at least for the reference covariate group. Moreover, female infants in cluster 1 had significantly higher log-odds of missing a measurement compared to male infants in cluster 1, while black infants in cluster 2 had significantly lower log-odds of missing a measurement compared to other infants.

[Figure 3 about here.]

As shown in Table 3, the skewness estimates for cluster 1 indicate little evidence of skewness, as all associated 95% credible intervals contained zero. However, in cluster 2, the predicted Bayley scores were negatively skewed at 6 months, in agreement with the preliminary analysis presented in Section 2. This suggests that the skewness observed in the data was driven primarily by the healthy-developing class, highlighting the model's ability to discern different skewness patterns across clusters. Finally, the estimated covariance matrix (Web Table 6) indicated an unstructured pattern for both clusters, with greater variability in cluster 2.

6. Discussion

We have developed Bayesian multivariate skew-normal and skew-t mixture of experts models for skewed longitudinal data that feature intermittent missingness. The model has many appealing features: it accounts for skewness in the infant development scores, associations

among repeated measures, and efficient inference for the cluster assignment probabilities. Additionally, the model handles missing data under a conditional ignorability assumption that relaxes standard MAR assumptions. Here, we proposed an online procedure to impute missing responses from conditional MSN distributions. The model can incorporate time-varying and time-invariant covariates. While both designs admit closed-form full conditionals, the latter enables matrix skew-normal updates that enhance computational efficiency.

Through simulations, we showed that ignoring skewness in even moderately skewed data results in incorrect inference, whereas the MSN mixture model recovers the true parameter values when the data are skewed. Furthermore, we showed that failing to account for conditional ignorability results in biased estimates when the response mechanism depends on cluster assignment. Finally, we conducted simulations to validate the use of WAIC, supporting the use of this measure in practice.

We applied our method to the Nurture data to assess the effect of household food security during pregnancy on motor development scores and to investigate possible clustering of infant development trajectories. We identified two distinct clusters of infants: one with delayed motor development that was significantly impaired by food insecurity, and a second that exhibited healthy motor development but was only modestly affected by food insecurity toward the end of infancy. This suggests that household food insecurity may compound the negative impacts of delayed motor development. On the other hand, we found that breastfeeding improved motor development among infants with delayed development. Thus, health providers might work to encourage breastfeeding for infants showing signs of delayed motor achievement. These results add to the growing body of literature on the effect of household food security on infant outcomes, and provide potential targets for intervention during this critical developmental window.

There are a number of possible extensions of this work. The model could accommodate

dropout in addition to intermittent missingness by incorporating a cluster-specific discrete time-to-event model. Additionally, cluster-specific shared parameters could be used to link the outcome and missing data models, thus relaxing the conditional ignorability assumption. More broadly, the method should prove useful in a wide range of settings involving multivariate skew data with informative missing responses.

References

Aaltonen, S., Latvala, A., Rose, R. J., Pulkkinen, L., Kujala, U. M., Kaprio, J., and Silventoinen, K. (2015). Motor development and physical activity: a longitudinal discordant twin-pair study. *Medicine and Science in Sports and Exercise* **47,** 2111–2118.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika* **83,** 715–726.

Bayley, N. (2006). *Bayley-III: Bayley Scales of Infant and Toddler Development.* San Antonio, Texas, USA.

Benjamin-Neelon, S. E., Østbye, T., Bennett, G. G., Kravitz, R. M., Clancy, S. M., Stroo, M., Iversen, E., and Hoyo, C. (2017). Cohort profile for the nurture observational study examining associations of multiple caregivers on infant growth in the Southeastern USA. *BMJ Open* **7,** e013939.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer, New York.

Filatova, S., Koivumaa-Honkanen, H., Hirvonen, N., Freeman, A., Ivandic, I., Hurtig, T., Khandaker, G., Jones, P., Moilanen, K., and Miettunen, J. (2017). Early motor developmental milestones and schizophrenia: a systematic review and meta-analysis. *Schizophrenia Research* **188,** 13–20.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11,** 317–336.

Gupta, A. (2003). Multivariate skew t-distribution. *Statistics: A Journal of Theoretical and Applied Statistics* **37,** 359–363.

Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software* **69,** 1–24.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association* **108,** 1339–1349.

Roy, J. (2007). Latent class models and their application to missing-data patterns in longitudinal studies. *Statistical Methods in Medical Research* **16,** 441–456.

Sánchez, G. F. L., Williams, G., Aggio, D., Vicinanza, D., Stubbs, B., Kerr, C., Johnstone, J., Roberts, J., and Smith, L. (2017). Prospective associations between measures of gross and fine motor coordination in infants and objectively measured physical activity and sedentary behavior in childhood. *Medicine* **96,** e8424.

Shoaibi, A., Neelon, B., Østbye, T., and Benjamin-Neelon, S. E. (2019). Longitudinal associations of gross motor development, motor milestone achievement and weight-for-length z score in a racially diverse cohort of us infants. *BMJ Open* **9,** e024440.

Taanila, A., Murray, G. K., Jokelainen, J., Isohanni, M., and Rantakallio, P. (2005). Infant developmental milestones: a 31-year follow-up. *Developmental Medicine and Child Neurology* **47,** 581–586.

USDA (2019). Food security in the U.S.: Measurement. `https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/measurement.aspx`. Accessed: 2020-01-11.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11,** 3571–3594.

Zhou, X. and Reiter, J. P. (2010). A note on Bayesian inference after multiple imputation. *The American Statistician* **64,** 159–163.

Acknowledgements

This work was supported by NIH grants R21 LM012866 and R01DK094841.

Supporting Information

The proof of Proposition 1, the MCMC algorithm referenced in Section 3, Web Tables 1 and 2 referenced in Section 2, Web Figures 1-3 and Web Tables 3-5 referenced in Section 4, and Web Figure 4 and Web Table 6 referenced in Section 5 may be found in the online version of this article at the publisher's web site. Data used for the simulation studies are available from the first author upon reasonable request. The data from the Nurture Study are not publicly available due to privacy or ethical restrictions.

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

Figure 1. Longitudinal profile plot of infant development trajectories, with mean Bayley motor development score shown in black. Plot is based on the $N = 1769$ available measurements for $n = 560$ infants.

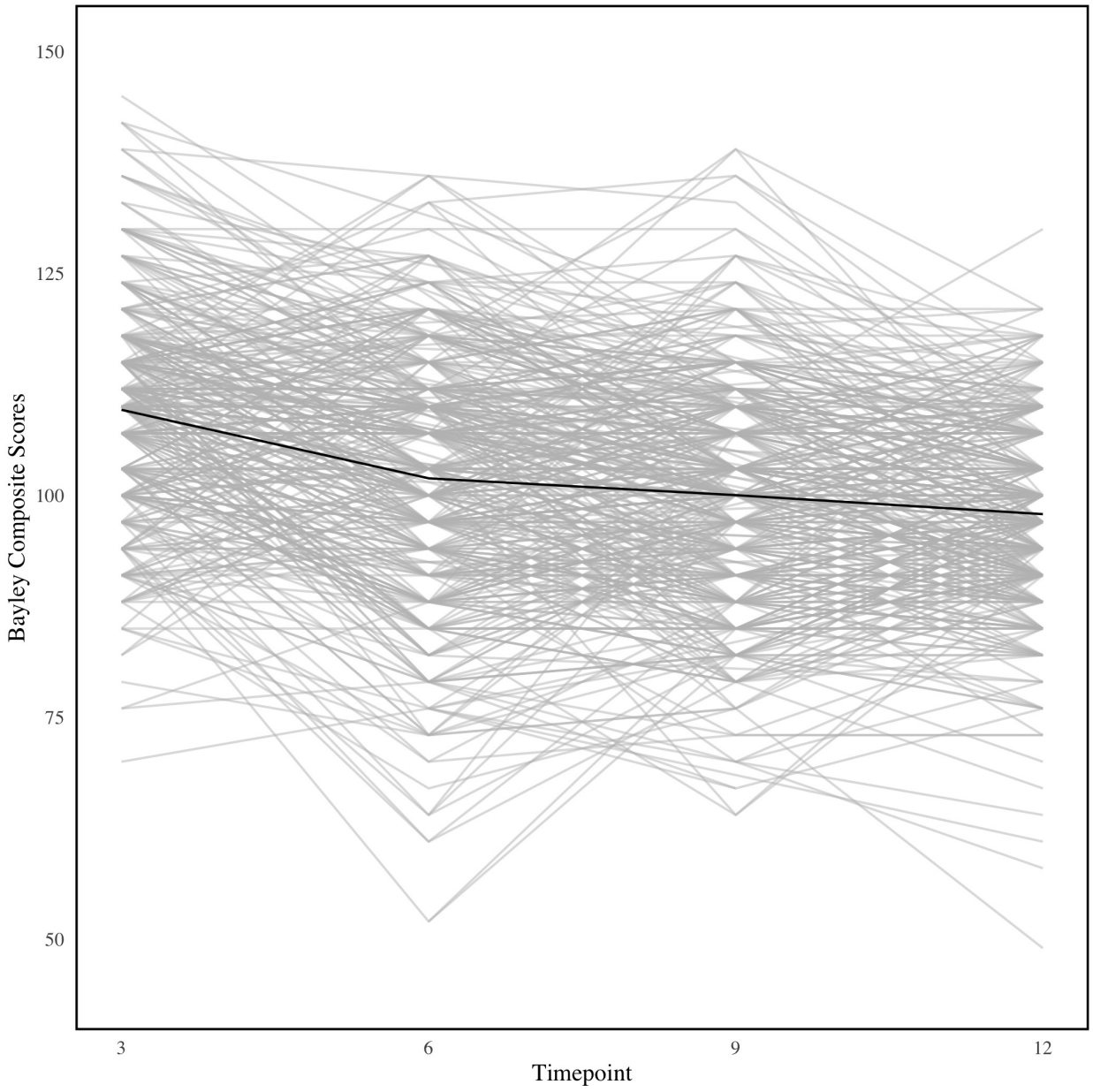

Figure 2. Scaled residual plots at each visit based on a repeated-measures linear regression model with Bayley score as the outcome. Sample skewness statistics and p-values from Shapiro-Wilk (SW) tests are provide in the legends. Plots are based on the $N = 1769$ available measurements for $n = 560$ infants.

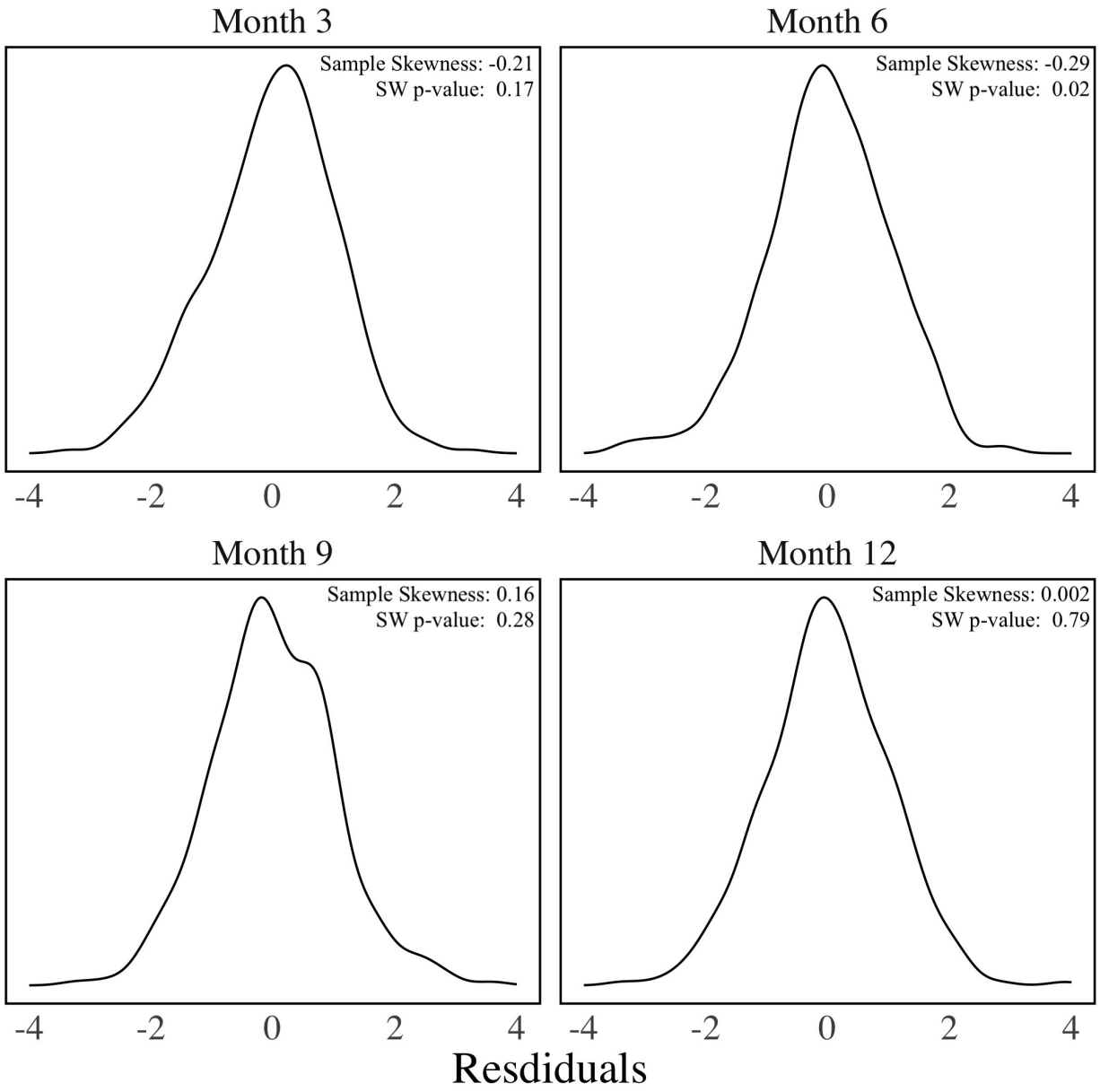

Figure 3. Predicted motor development trajectories for each cluster and food security group in the Nurture analysis. Estimated trajectories are given for a typical infant with a birth weight for gestational age z-score of 0, who was not breastfed, and who had 2.5 other children in the household. Solid lines indicate Cluster 1 and dashed lines indicate Cluster 2. Light shading represents food-secure infants, while dark shading represents food-insecure infants.

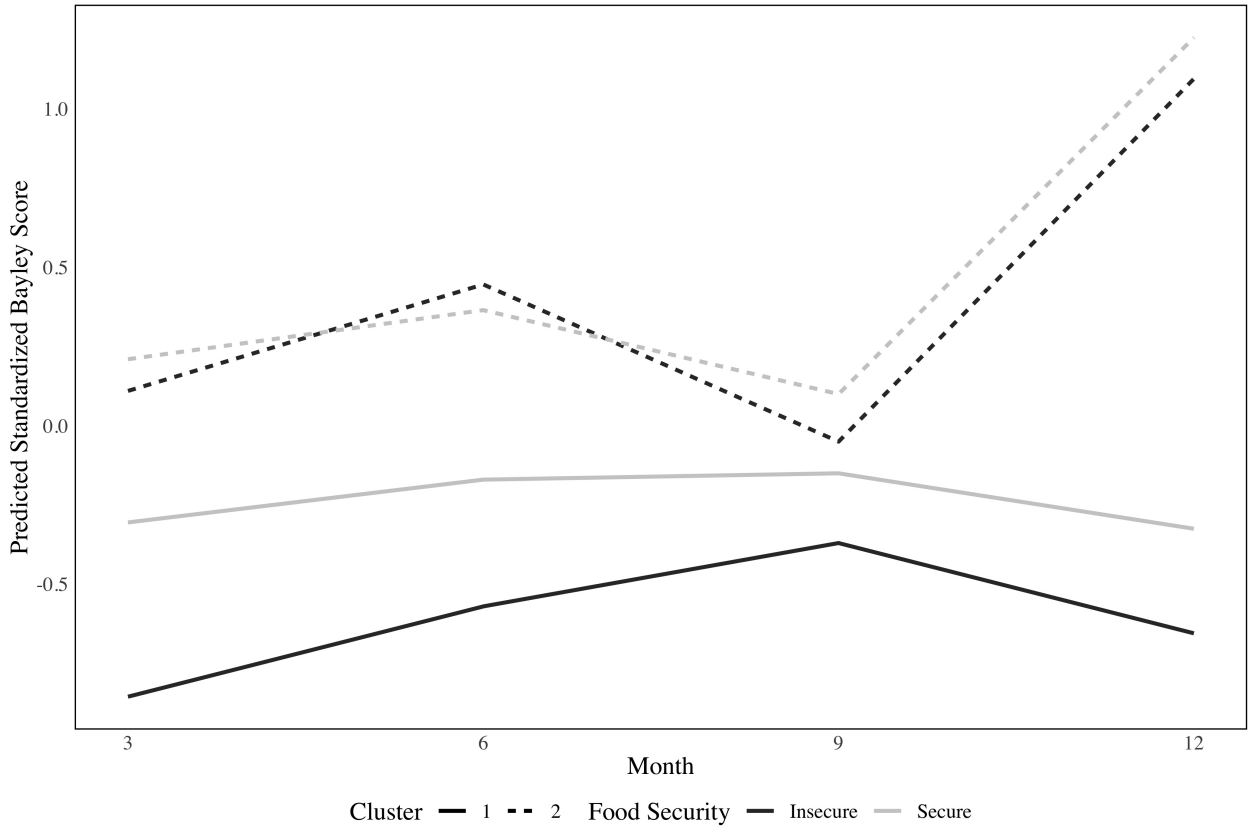

Table 1

Results for cluster 1 from Simulation Study 1 with n = 1000, J = 4, p = 2, K = 3, r = 2. 10000 iterations were run with a burn in of 1000. Posterior means (95% CrIs) are presented for the multivariate skew normal (MSN) and multivariate normal (MVN) mixtures. No missing data were introduced.

Component	**Parameter**	**True Value**	MSN Est. (95% CrI)	MVN Est. (95% CrI)
MSN	β_{111}	110.00	110.20 (109.97, 110.41)	106.36 (105.97, 108.71)
Regression	β_{121}	115.00	115.13 (114.91, 115.33)	104.17 (103.93, 104.44)
	β_{131}	120.00	120.08 (119.83, 120.49)	128.02 (128.57, 129.08)
	β_{141}	125.00	125.15 (124.86, 125.49)	126.67 (126.31, 127.05)
	β_{112}	1.00	0.97 (0.84, 1.11)	0.90 (0.74, 1.08)
	β_{122}	1.50	1.51 (1.40, 1.62)	1.53 (1.41, 1.66)
	β_{132}	2.00	2.01 (1.89, 2.14)	2.20 (2.08, 2.33)
	β_{142}	2.50	2.50 (2.35, 2.66)	2.46 (2.28, 2.64)
	Σ_{111}	1.00	0.96 (0.77, 1.14)	2.42 (2.06, 2.84)
	Σ_{112}	0.50	0.47 (0.34, 0.61)	1.20 (0.99, 1.48)
	Σ_{113}	0.25	0.25 (0.04, 0.40)	-0.54 (-0.75, -0.34)
	Σ_{114}	0.12	0.11 (-0.02, 0.30)	-1.35 (-1.67, -1.06)
	Σ_{122}	1.00	0.99 (0.74, 1.19)	1.20 (0.99, 1.48)
	Σ_{123}	0.50	0.49 (0.26, 0.66)	1.24 (1.06, 1.46)
	Σ_{124}	0.25	0.24 (0.10, 0.43)	0.08 (-0.06, 0.21)
	Σ_{133}	1.00	0.99 (0.77, 1.09)	1.24 (1.06, 1.46)
	Σ_{134}	0.50	0.47 (0.22, 0.65)	1.15 (0.93, 1.40)
	Σ_{144}	1.00	1.01 (0.63, 1.23)	2.48 (2.15, 2.91)
	α_{11}	-2.00	-2.05 (-2.28, -1.66)	/
	α_{12}	-1.00	-1.01 (-1.30, -0.75)	/
	α_{13}	1.00	0.97 (0.65, 1.28)	/
	α_{14}	2.00	1.97 (1.67, 2.28)	/
Multinomial	δ_{11}	-0.27	-0.23 (-0.47, -0.09)	-0.14 (-0.35, 0.08)
Logit†	δ_{12}	0.07	0.07 (-0.24, 0.37)	0.08 (-0.24, 0.38)
Missing	γ_{11}	-0.82	-0.84 (-0.96, -0.73)	-1.08 (-1.19, -0.99)
Data	γ_{12}	-1.08	-1.01 (-1.20, -0.91)	-1.80 (-1.96, -1.64)
	γ_{13}	-1.12	-1.08 (-1.20, -1.00)	-0.90 (-1.00, -0.80)
	σ_1^2	1.00	1.07 (0.92, 1.28)	0.89 (0.76, 1.07)
Estimated proportion‡	π_1	0.32	0.32 (0.31, 0.33)	0.32 (0.30, 0.34)

† Multinomial logit parameters comparing cluster 1 to cluster 3 (reference cluster).

‡ Estimated proportion of infants in cluster 1. True proportion is 0.32.

Table 2

Results for cluster 1 from Simulation Study 2. Posterior means (95% CrIs) are presented under online MNAR imputation (MNAR), online MAR imputation (MAR), and Bayesian multiple imputation (MI) as described in Section 3.4.

Model Component	**Parameter**	**True Value**	**MNAR**	**MAR**	**MI**
k = 1	β_{111}	-2.90	-3.02 (-3.62, -2.56)	-4.10 (-5.70, -2.10)	-4.23 (-14.03, -1.37)
	β_{121}	-2.70	-2.85 (-2.99, -2.71)	-2.81 (-3.84, -2.19)	-2.97 (-13.68, -1.84)
MSN	β_{131}	-2.92	-2.84 (-3.53, -2.38)	-4.25 (-4.95, -2.10)	-4.21 (-4.91, -1.54)
Regression	β_{141}	-3.68	-3.88 (-4.03, -3.65)	-3.67 (-3.98, -3.08)	-3.57 (-3.95, -2.88)
	β_{112}	-2.78	-2.68 (-3.40, -2.23)	-4.52 (-5.39, -2.13)	-4.31 (-4.86, -1.92)
	β_{122}	-2.59	-2.83 (-2.96, -2.70)	-2.68 (-3.23, -2.19)	-2.53 (-2.95, -1.78)
	β_{132}	-2.71	-2.46 (-3.38, -2.14)	-4.14 (-4.80, -1.51)	-4.18 (-4.98, -1.46)
	β_{142}	-2.79	-2.98 (-3.11, -2.84)	-2.45 (-3.02, -1.88)	-2.33 (-2.99, -1.51)
	Σ_{111}	1.00	1.14 (0.70, 1.64)	1.18 (0.64, 2.25)	2.30 (1.41, 4.02)
	Σ_{112}	0.50	0.58 (0.21, 1.17)	0.81 (0.29, 1.82)	1.45 (0.64, 3.04)
	Σ_{113}	0.25	0.25 (0.14, 0.34)	0.38 (0.01, 1.23)	0.80 (0.16, 2.09)
	Σ_{114}	0.12	0.14 (0.06, 0.23)	0.19 (-0.13, 0.91)	0.91 (0.20, 2.18)
	Σ_{122}	1.00	0.97 (0.58, 1.70)	1.33 (0.73, 2.32)	1.61 (0.86, 3.10)
	Σ_{123}	0.50	0.52 (0.14, 1.04)	0.69 (0.23, 1.51)	0.67 (0.16, 1.85)
	Σ_{124}	0.25	0.24 (0.06, 0.39)	0.38 (-0.03, 1.08)	0.43 (-0.14, 1.65)
	Σ_{133}	1.00	1.32 (0.81, 1.89)	0.99 (0.57, 1.78)	0.91 (0.20, 2.18)
	Σ_{134}	0.50	0.55 (0.26, 0.83)	0.45 (0.09, 1.07)	0.52 (0.11, 1.49)
	Σ_{144}	1.00	0.98 (0.60, 1.57)	0.95 (0.50, 1.60)	1.32 (0.78, 2.34)
	α_{11}	-1.00	-0.81 (-1.38, -0.05)	1.98 (-1.17, 4.04)	1.67 (-4.53, 4.64)
	α_{12}	-1.00	-1.10 (-1.63, -0.23)	2.12 (-1.25, 3.46)	2.23 (-1.43, 3.41)
	α_{13}	-1.00	-1.08 (-1.7, -0.14)	2.50 (-1.25, 3.40)	2.37 (-1.04, 3.57)
	α_{14}	-1.00	-1.31 (-1.66, -0.73)	2.43 (-1.35, 3.50)	2.43 (-0.98, 3.98)
Multinomial	δ_{11}	-0.54	-0.54 (-0.73, -0.33)	-0.54 (-0.75, -0.33)	-0.52 (-0.72, -0.3)
Logit†	δ_{12}	-0.01	-0.01 (-0.31, 0.28)	-0.02 (-0.33, 0.28)	-0.02 (-0.33, 0.28)
Missing Data	γ_{11}	-1.10	-1.09 (-1.36, -0.84)	/	/
	γ_{12}	-1.27	-1.04 (-1.33, -0.81)	/	/
	γ_{13}	-1.07	-1.06 (-1.33, -0.80)	/	/
	σ_1^2	1.00	1.01 (0.86, 1.15)	/	/

† Multinomial logit parameters comparing cluster 1 to cluster 3 (reference cluster).

Table 3

Results from the 2-cluster model applied to the Nurture data. Posterior means (95% CrI) are presented in each cluster for the effects of time, food security during pregnancy (FS), birth weight for gestational age z-score (BW), any breastfeeding throughout the study period (BF), and total number of children in the household (TC). The effects of time, FS, BW, BF and TC were allowed to vary over time, yielding separate estimates for each 3-month visit. Posterior means (95% CrI) are also given for effects of birth weight for gestational age z-score, race, and gender in the multinomial logit clustering and missing data models.

Model Component	**Parameter**	**Variable**	**Cluster 1 (37.0%)**† Est. (95% CrI)	**Cluster 2 (63.0%)**† Est. (95% CrI)
MSN Reg.	β_{k11}	3 mo.	-0.33 (-0.48, -0.18)	0.26 (-0.04, 0.53)
	β_{k21}	6 mo.	-0.22 (-0.37, -0.05)	0.54 (0.17, 0.86)
	β_{k31}	9 mo.	-0.20 (-0.52, 0.11)	0.10 (-0.47, 0.56)
	β_{k41}	12 mo.	-0.35 (-0.45, -0.27)	0.80 (0.37, 1.11)
	β_{k12}	FS (3 mo.)	-0.55 (-0.68, -0.40)	-0.10 (-0.28, 0.12)
	β_{k22}	FS (6 mo.)	-0.40 (-0.56, -0.23)	0.08 (-0.08, 0.32)
	β_{k32}	FS (9 mo.)	-0.22 (-0.41, -0.03)	-0.15 (-0.27, -0.02)
	β_{k42}	FS (12 mo.)	-0.33 (-0.50, -0.12)	-0.13 (-0.26, -0.06)
	β_{k13}	BW (3 mo.)	-0.02 (-0.09, 0.06)	0.07 (-0.05, 0.16)
	β_{k23}	BW (6 mo.)	-0.03 (-0.11, 0.04)	0.03 (-0.09, 0.11)
	β_{k33}	BW (9 mo.)	-0.03 (-0.13, 0.06)	0.11 (-0.07, 0.29)
	β_{k43}	BW (12 mo.)	-0.03 (-0.11, 0.04)	0.06 (-0.05, 0.14)
	β_{k14}	BF (3 mo.)	0.41 (0.29, 0.51)	0.07 (-0.15, 0.22)
	β_{k24}	BF (6 mo.)	0.46 (0.36, 0.55)	0.04 (-0.14, 0.20)
	β_{k34}	BF (9 mo.)	0.62 (0.30, 0.91)	0.03 (-0.05, 0.12)
	β_{k44}	BF (12 mo.)	0.17 (-0.21, 0.55)	0.04 (-0.12, 0.24)
	β_{k15}	TC (3 mo.)	0.01 (-0.03, 0.06)	-0.02 (-0.09, 0.05)
	β_{k25}	TC (6 mo.)	0.02 (-0.02, 0.06)	-0.07 (-0.13, -0.02)
	β_{k35}	TC (9 mo.)	0.02 (-0.03, 0.07)	0.00 (-0.06, 0.06)
	β_{k45}	TC (12 mo.)	0.01 (-0.03, 0.06)	0.17 (-0.01, 0.35)
	α_{k1}	Skewness (3 mo.)	0.00 (-0.12, 0.11)	0.16 (-0.23, 0.41)
	α_{k2}	Skewness (6 mo.)	-0.02 (-0.15, 0.1)	-0.53 (-0.80, -0.17)
	α_{k3}	Skewness (9 mo.)	-0.02 (-0.16, 0.13)	0.05 (-0.32, 0.44)
	α_{k4}	Skewness (12 mo.)	-0.03 (-0.16, 0.10)	-0.07 (-0.41, 0.28)
Multinomial Logit‡	δ_{k1}	Intercept	1.03 (0.79, 1.25)	Ref.
	δ_{k2}	BW	0.03 (-0.09, 0.15)	Ref.
	δ_{k3}	Race (Black)	-0.02 (-0.29, 0.27)	Ref.
	δ_{k4}	Gender (Female)	0.90 (0.65, 1.27)	Ref.
Missing Data	γ_{k1}	Intercept	0.37 (0.32, 0.41)	-0.16 (-0.19, -0.14)
	γ_{k2}	BW	0.05 (-0.51, 0.59)	0.03 (-0.14, 0.19)
	γ_{k3}	Gender (Female)	0.80 (0.25, 1.57)	-0.04 (-0.41, 0.30)
	γ_{k4}	Race (Black)	0.35 (-0.56, 1.37)	-0.60 (-1.02, -0.20)
	σ_k^2	Random Int. Var.	1.34 (0.86, 1.74)	1.11 (0.79, 1.43)

† Posterior mean percent in each cluster.

‡ With only 2 clusters, this reduces to a conventional logistic model.