Web-based supplementary materials for "A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity"

by Brian Neelon, A. James O'Malley, and Sharon-Lise T. Normand

Web Appendix: MCMC Algorithm for Latent Class Two-Part Model

We describe the MCMC algorithm for the two-part random intercept model applied in Section 5. As in Section 5, we assume the same predictors, $\boldsymbol{X}$, for both the binomial and lognormal components.

Starting with initial values for model parameters, the algorithm iterates through the following steps until convergence (as determined by MCMC diagnostics):

1. **Update $\boldsymbol{\gamma}_k$**: The full conditional for r-dimensional vector $\boldsymbol{\gamma}_k$ $(k = 2, \dots, K)$ is given by

$$\begin{aligned} \pi(\boldsymbol{\gamma}_k|\cdot) &\propto \prod_{i=1}^{n} [\Pr(C_i = k|\boldsymbol{\gamma}_k)]^{\mathrm{I}_{(C_i=k)}} \pi(\boldsymbol{\gamma}_k) \\ &= \prod_{i:\, C_i=k} \left(\frac{e^{\boldsymbol{w}_i'\boldsymbol{\gamma}_k}}{\sum_{h=1}^{K} e^{\boldsymbol{w}_i'\boldsymbol{\gamma}_h}} \right) \mathrm{N}_r[\boldsymbol{\gamma}_k; \mathbf{0}, (9/4)\boldsymbol{I}_r], \end{aligned}$$

where $\mathrm{N}_r(\boldsymbol{\gamma}_k;\cdot)$ is an r-dimensional normal distribution evaluated at $\boldsymbol{\gamma}_k$. Since this full conditional distribution does not have a closed form, we update $\boldsymbol{\gamma}_k$ as a vector using a random walk Metropolis algorithm based on a multivariate-$t_3(s_g\boldsymbol{T}_k)$ proposal density centered at the previous value, $\boldsymbol{\gamma}_k^{old}$. To improve mixing, we apply the adaptive proposal (AP) developed by Harrio et al. (2005), which uses the empirical covariance from an extended burn-in period to tune $\boldsymbol{T}_k$ so that it emulates the true posterior covariance. The parameter s_g scales the covariance to achieve an optimal acceptance rate of approximately 20%. As a default value, we choose $s_g = 2.4/\sqrt{r}$ as recommended by Gelman, Roberts, and Gilks (1996).

2. **Update C_i**: For $i = 1, \dots, n$, draw C_i from its full conditional

$$\begin{aligned} \pi(C_i|\cdot) &= \Pr(C_i = k|\cdot) = \mathrm{Cat}(p_{ik}), \text{ where} \\ p_{ik} &= \frac{\pi_{ik}(\boldsymbol{\gamma}_k)\left[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{b}_i, \tau_k^2)\right] \mathrm{N}_2(\mathbf{b}_i; \mathbf{0}, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \pi_{ih}(\boldsymbol{\gamma}_h)\left[\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\alpha}_h, \boldsymbol{\beta}_h, \mathbf{b}_i, \tau_h^2)\right] \mathrm{N}_2(\mathbf{b}_i; \mathbf{0}, \boldsymbol{\Sigma}_h)}, \end{aligned}$$

$\pi_{ik}(\boldsymbol{\gamma}) = \Pr(C_i = k|\boldsymbol{\gamma}_k)$ as given in step (1), and $\mathbf{b}_i = (b_{1i}, b_{2i})'$. If there are no class-membership covariates [i.e., $r = 1$ in Step (1)], then update π_k directly from a Dirichlet$(n_1 + e_1, \dots, n_K + e_k)$ distribution, where $e_1, \dots, e_K$ are prior hyperparameters and $n_k = \sum_{i=1}^{n} \mathrm{I}_{(C_i=k)}$.

To avoid label switching in this case, Lenk and DeSarbo (2000) recommend sampling from an ordered Dirichlet distribution with $\pi_1 < \pi_2 < \cdots < \pi_K$. See Appendix C of their paper for details.

3. **Update $\boldsymbol{\alpha}_k$**: First, consider a probit link for the binomial component of equation (1). Note that $\Pr(y_{ij} > 0|C_i = k, \boldsymbol{\alpha}_k, b_{1i}) = \Phi(\boldsymbol{x}'_{ij}\boldsymbol{\alpha}_k + b_{1i})$ is equivalent to assuming $y_{ij} = \mathrm{I}_{(u_{ij}>0)}$, where $[u_{ij}|C_i = k, \boldsymbol{\alpha}_k, b_{1i}] \sim \mathrm{N}(\boldsymbol{x}'_{ij}\boldsymbol{\alpha}_k + b_{1i}, 1)$ for all i, j such that $C_i = k$. To update $\boldsymbol{\alpha}_k$, we employ the data-augmentation algorithm described in Albert and Chib (1993) by first drawing u_{ij} from its full conditional

$$\pi(u_{ij}|y_{ij}, C_i = k, \boldsymbol{\alpha}_k, b_{1i}) = \mathrm{N}(\boldsymbol{x}'_{ij}\boldsymbol{\alpha}_k + b_{1i}, 1) \text{ truncated below (above) by 0 for } y_{ij} > 0 \, (y_{ij} = 0),$$

and then drawing $\boldsymbol{\alpha}_k$ from its full conditional $\pi(\boldsymbol{\alpha}_k|\boldsymbol{u}_k, \mathbf{b}_{1k}) = \mathrm{N}_p(\boldsymbol{\eta}_{\alpha_k}, \boldsymbol{V}_{\alpha_k})$, where

$$\begin{aligned} \boldsymbol{V}_{\alpha_k} &= \left(\boldsymbol{\Sigma}_\alpha^{-1} + \boldsymbol{X}'_k\boldsymbol{X}_k\right)^{-1} \text{ and} \\ \boldsymbol{\eta}_{\alpha_k} &= \boldsymbol{V}_{\alpha_k}\left[\boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha + \boldsymbol{X}'_k(\boldsymbol{u}_k - \mathbf{b}_{1k})\right]. \end{aligned}$$

Here, $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ denote the prior mean and variance of $\boldsymbol{\alpha}_k$ (for the FEHB study, we optionally assume the same prior hyperparameters for all classes); $\boldsymbol{u}_k$ denotes an $N_k \times 1$ vector of $\{u_{ij}\}$ draws for the N_k observations in class k; $\mathbf{b}_{1k}$ denotes an $N_k \times 1$ concatenated vector of component-1 random intercepts for class k (i.e., b_{1i} is repeated n_i times for subject $i \in$ class k); and $\boldsymbol{X}_k$ is an $N_k \times p$ design matrix for class k.

For a logit link, one can use a similar data-augmentation approach, approximating the underlying logistic distribution by a mixture of normals as described in Frühwirth-Schnatter and Frühwirth (2007). Alternatively, a Metropolis-Hastings step can be used to update $\boldsymbol{\alpha}_k$.

4. **Update $\boldsymbol{\beta}_k$**: Let $\boldsymbol{y}_k^*$ denote the $M_k \times 1$ subvector of positive (i.e., nonzero) observations in class k, let $\boldsymbol{X}_k^*$ denote the corresponding design matrix, and let $\mathbf{b}_{2k}^*$ denote an $M_k \times 1$ concatenated vector of component-2 random intercepts restricted to observations greater than zero. Assuming a $\mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, the full conditional for $\boldsymbol{\beta}_k$ is

$$\pi(\boldsymbol{\beta}_k|\cdot) = \pi(\boldsymbol{\beta}_k|\boldsymbol{y}_k^*, \mathbf{b}_{2k}^*, \tau_k^2) = \mathrm{N}_p(\boldsymbol{\eta}_{\beta_k}, \boldsymbol{V}_{\beta_k}),$$

where

$$\begin{aligned} \boldsymbol{V}_{\beta_k} &= \left[\boldsymbol{\Sigma}_\beta^{-1} + \tau_k^{-2}(\boldsymbol{X}_k^{*'}\boldsymbol{X}_k^*)\right]^{-1} \text{ and} \\ \boldsymbol{\eta}_{\beta_k} &= \boldsymbol{V}_{\beta_k}\left\{\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \tau_k^{-2}\boldsymbol{X}_k^{*'}\left[\log(\boldsymbol{y}_k^*) - \mathbf{b}_{2k}^*\right]\right\}. \end{aligned}$$

5. **Update τ_k^{-2}**: Assuming a $\mathrm{Ga}(\lambda, \delta)$ prior for τ_k^{-2}, draw τ_k^{-2} from its full conditional

$$\begin{aligned} \pi(\tau_k^{-2}|\cdot) &= \pi(\tau_k^{-2}|\boldsymbol{y}_k^*, \boldsymbol{\beta}_k, \mathbf{b}_{2k}^*) \\ &= \mathrm{Ga}\left(\lambda + M_k/2, \delta + \frac{1}{2}\left[\log(\boldsymbol{y}_k^*) - \boldsymbol{X}_k^{*'}\boldsymbol{\beta}_k - \mathbf{b}_{2k}^*\right]'\left[\log(\boldsymbol{y}_k^*) - \boldsymbol{X}_k^{*'}\boldsymbol{\beta}_k - \mathbf{b}_{2k}^*\right]\right), \end{aligned}$$

where M_k, $\boldsymbol{y}_k^*$, $\boldsymbol{X}_k^*$, and $\mathbf{b}_{2k}^*$ are defined in step (4).

6. **Update $\boldsymbol{\Sigma}_k$**:

Assuming an $\text{IW}(\nu_0, \boldsymbol{D}_0)$ prior, draw $\boldsymbol{\Sigma}_k$ from its full conditional

$$\pi(\boldsymbol{\Sigma}_k|\cdot) = \pi(\boldsymbol{\Sigma}_k|\mathbf{b}_k) = \text{IW}(n_k + \nu_0, \boldsymbol{D}_0 + \mathbf{b}_k'\mathbf{b}_k),$$

where $n_k = \sum_{i=1}^{n} \text{I}_{(C_i=k)}$ denotes the number of subjects currently assigned to class k, and $\mathbf{b}_k$ is an $n_k \times 2$ matrix with the first column containing the component-1 random intercepts and the second column containing the component-2 random intercepts for subjects in class k.

Repeat steps (3) - (6) for classes $k = 1, \ldots, K$.

7. **Update $\mathbf{b}_i$**: The full conditional for $\mathbf{b}_i = (b_{1i}, b_{2i})'$ is

$$\pi(\mathbf{b}_i|\cdot) \propto f(\boldsymbol{y}_i|C_i = k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{b}_i, \tau_k^2)\text{N}_2(\mathbf{b}_i; \mathbf{0}, \boldsymbol{\Sigma}_k).$$

Conditional on $C_i = k$, we update $\mathbf{b}_i$ using a random walk Metropolis algorithm with a bivariate $t_3(s_b\boldsymbol{R}_k)$ proposal centered at the previous value, $\mathbf{b}_i^{old}$. The scale matrix $\boldsymbol{R}_k$ can be estimated using the inverse information matrix obtained from a frequentist fit of the model, and s_b is a scaling factor used to achieve optimal acceptance rates. Note that, given $C_i = k$, the acceptance ratio for updating $\mathbf{b}_i$ is a function of the class-k parameters only.

Web Tables

Web Table 1

Model Comparison statistics for simulation study.

			DIC Value‡			
Number of Classes	**Average DIC* (SD)**	$\Delta^\dagger$	Lowest	2nd Lowest	3rd Lowest	Highest
One Class	14205.12 (388.84)	—	0	3	82	15
Two Class	13641.52 (516.62)	563.60	0	97	3	0
Three Class (True Model)	**13330.95** (372.33)	310.57	100	0	0	0
Four Class	14572.18 (694.65)	−1241.23	0	0	15	85

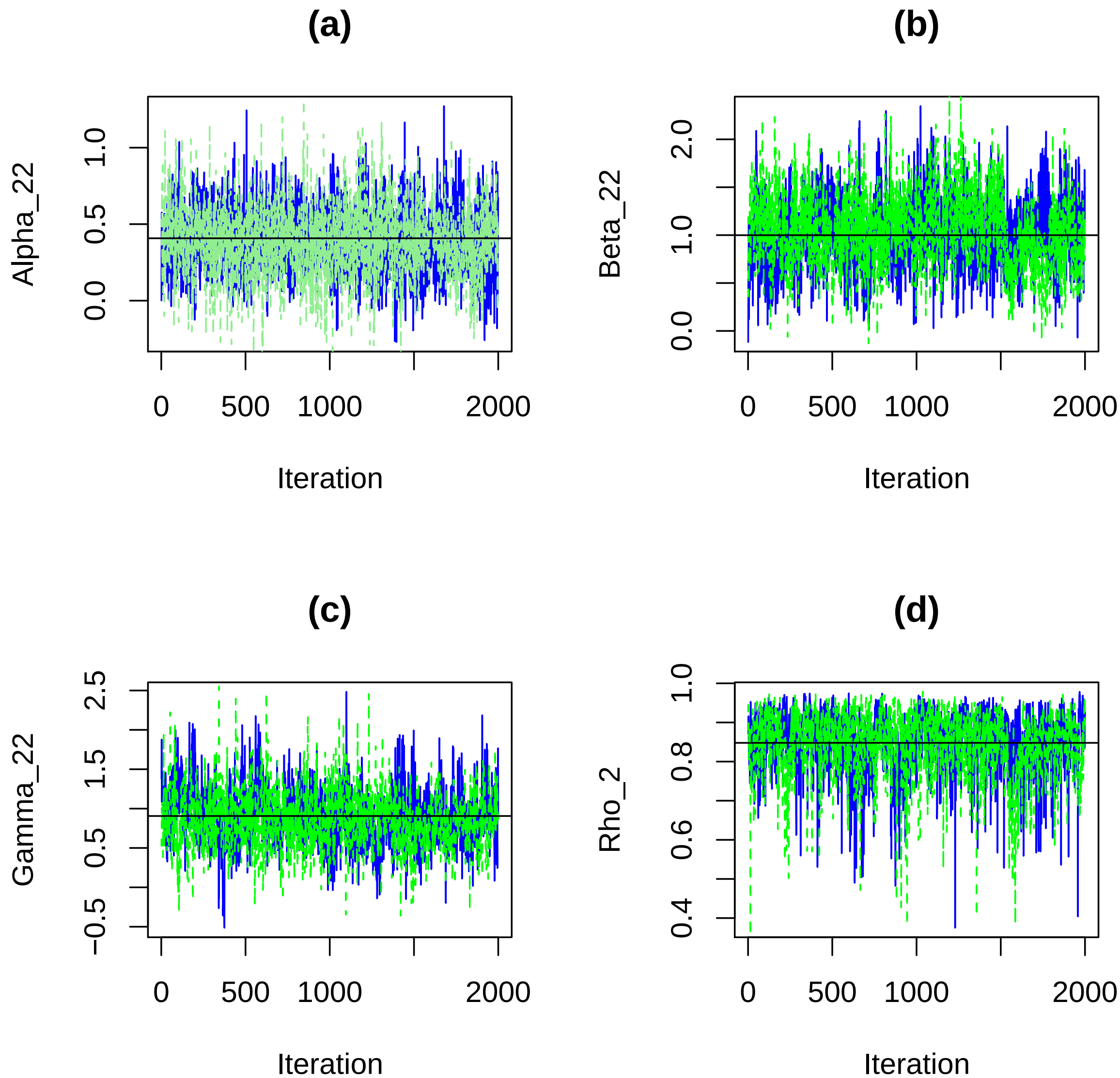
* Average DIC across 100 simulated datasets.
† Change in average DIC from previous model.
‡ Number of simulations in which model had lowest (most preferred) to highest (least preferred) DIC value.
Bold = lowest average DIC.

Web Table 2

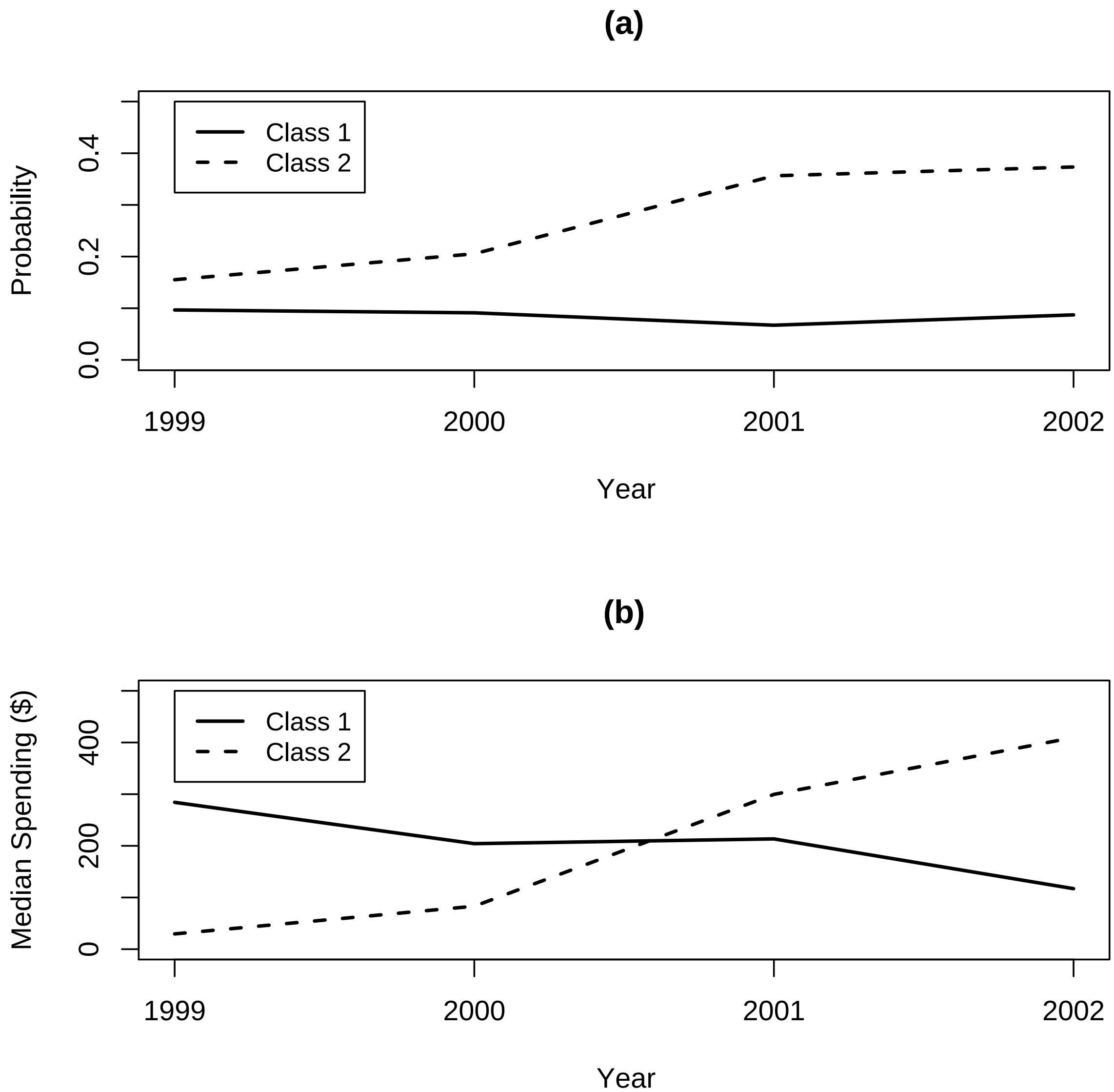
Summary statistics for three-class model based on 100 simulated datasets.

Class (%)*	Model Component	Parameter (Variable Name)	True Value	Mean Posterior Estimate[†] (SD)	95% Coverage
1 (31%)	**Binomial**	α_{11} (Intercept)	0.25	0.20 (0.21)	0.94
		α_{12} (Linear Time)	0.50	0.51 (0.07)	0.95
	Lognormal	β_{11} (Intercept)	3.25	3.23 (0.14)	0.94
		β_{12} (Linear Time)	−0.75	−0.75 (0.02)	0.94
	Variance Components	τ_1^2 (Lognormal Variance)	0.50	0.50 (0.02)	0.96
		σ_{11}^2 (Var[b_{1i}])	2.00	2.27 (0.86)	0.96
		σ_{12}^2 (Var[b_{2i}])	2.00	2.03 (0.25)	0.94
		ρ_1 (Corr[b_{1i}, b_{2i}])	0.25	0.26 (0.15)	0.91
2 (26%)	**Binomial**	α_{21} (Intercept)	0.50	0.57 (0.16)	0.95
		α_{22} (Linear Time)	−0.50	−0.51 (0.05)	0.99
	Lognormal	β_{21} (Intercept)	2.25	2.25 (0.16)	0.96
		β_{22} (Linear Time)	2.25	1.25 (0.06)	0.95
	Variance Components	τ_2^2 (Lognormal Variance)	1.00	1.01 (0.06)	0.96
		σ_{21}^2 (Var[b_{1i}])	1.00	0.95 (0.30)	0.93
		σ_{22}^2 (Var[b_{2i}])	1.00	1.01 (0.25)	0.94
		ρ_2 (Corr[b_{1i}, b_{2i}])	−0.25	−0.22 (0.18)	0.98
3 (43%)	**Binomial**	α_{31} (Intercept)	−0.50	−0.51 (0.16)	0.95
		α_{32} (Linear Time)	0.50	0.50 (0.05)	0.93
	Lognormal	β_{31} (Intercept)	0.50	0.49 (0.15)	0.98
		β_{32} (Linear Time)	0.50	0.50 (0.05)	0.99
	Variance Components	τ_3^2 (Lognormal Variance)	1.50	1.50 (0.05)	0.95
		σ_{31}^2 (Var[b_{1i}])	1.50	1.58 (0.26)	0.94
		σ_{32}^2 (Var[b_{2i}])	0.50	0.52 (0.03)	0.93
		ρ_3 (Corr[b_{1i}, b_{2i}])	0.00	0.03 (0.20)	0.94
	Class Membership Parameters	γ_{21} (Class 2 Intercept)	−0.50	−0.55 (0.28)	0.96
		γ_{22} (Class 2 Covariate, w)	0.50	0.52 (0.32)	0.94
		γ_{31} (Class 3 Intercept)	0.75	0.82 (0.24)	0.92
		γ_{32} (Class 3 Covariate, w)	−0.75	−0.78 (0.28)	0.92

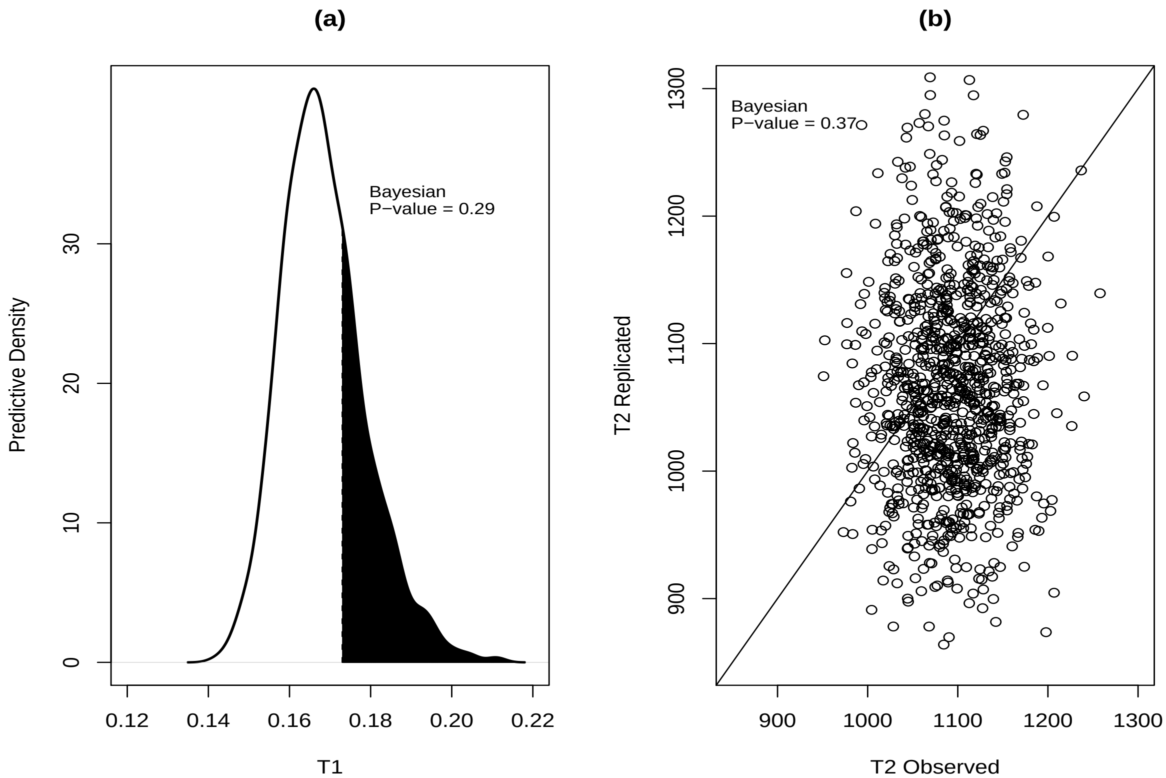
* Estimated class proportions averaged across the 100 simulated datasets. True proportions are 0.31, 0.26 and 0.43.
† Posterior means averaged across the 100 simulated datasets.



Web Figure 1. Trace plots based on two MCMC chains for four representative parameters from the three-class correlated model: (a) $\boldsymbol{\alpha}_{22}$ (log odds use for year 2, class 2); (b) $\boldsymbol{\beta}_{22}$ (log-spending for year 2, class 2); (c) $\boldsymbol{\gamma}_{22}$ (log-odds class-2 membership, females vs. males); (d) ρ_2 (random effect correlation, class 2). Horizontal lines denote posterior means.



Web Figure 2. Enlarged view of classes 1 and 2 posterior trajectories.



Web Figure 3. Results from posterior predictive checks: (a) posterior distribution of the proportion of nonzero observations; (b) scatterplot of predicted versus observed chi-square discrepancy measure across MCMC samples. In figure (a), the Bayesian predictive p-value (0.29) represents the area in the shaded region; in figure (b), the Bayesian predictive p-value (0.37) represents the proportion of samples above the diagonal.

References

[1] Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88,** 669–679.

[2] Frühwirth-Schnatter, S. and Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, **51,** 3509–3528.

[3] Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient metropolis jumping rules. In J.M. Beruardo, J.O. Berger, A.E. Dawid, and A.F.M. Smith (eds), *Bayesian Statistics,* Volume 5, 599–607. Cambridge: Oxford University Press.

[4] Harrio, H., Saksman, E., and Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics* **20,** 165–273.

[5] Lenk, P.J. and DeSarbo, W.S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* **65,** 93–119.