# Supplementary material for A Bayesian Approach to Modeling Missing Not at Random Outcome Data: The Role of Identifying Restrictions

Antonio R. Linero*and Michael J. Daniels†

August 31, 2017

## S.1 Proof of Proposition 3.1

*Proof.* Sadinle and Reiter (2017) show that ICI implies the existence of functions $\eta_j(\bar{y}_{j-1})$ such that

$$\log \frac{p(S = j \mid Y = y)}{p(S = j - 1 \mid Y = y)} = \eta_j(\bar{y}_{j-1}). \tag{S.1}$$

We show that this occurs if and only if NCMV is satisfied. First, suppose NCMV holds. Let $f_j(y)$ denote $p(y \mid S = j)$, and, abusing notation, define $f_j(y_k \mid \bar{y}_{k-1})$ to be $p(y_k \mid S = j, \bar{Y}_{k-1} = \bar{y}_{k-1})$. Then

$$\frac{p(S = j \mid Y = y)}{p(S = j - 1 \mid Y = y)} = \frac{p(S = j)}{p(S = j - 1)} \cdot \frac{f_j(\bar{y}_{j-1}) f_j(y_j \mid \bar{y}_{j-1})}{f_{j-1}(\bar{y}_{j-1}) f_{j-1}(y_j \mid \bar{y}_{j-1})} \cdot \frac{\prod_{k>j} f_j(y_k \mid \bar{y}_{k-1})}{\prod_{k>j} f_{j-1}(y_k \mid \bar{y}_{k-1})}.$$

Applying NCMV and canceling shared terms in the numerator and denominator gives

$$\frac{p(S = j \mid Y = y)}{p(S = j - 1 \mid Y = y)} = \frac{p(S = j)}{p(S = j - 1)} \cdot \frac{f_j(\bar{y}_{j-1})}{f_{j-1}(\bar{y}_{j-1})},$$

the right-hand-side of which can be chosen for $\eta_j(\bar{y}_{j-1})$.

Conversely, suppose that (S.1) holds. We prove the converse by showing that, for $k < j$,

*Department of Statistics, Florida State University, arlinero@stat.fsu.du
†Department of Statistics, University of Florida, mjdaniels@stat.ufl.edu

$f_k(y_j \mid \bar{y}_{j-1}) = f_{k+1}(y_j \mid \bar{y}_{j-1})$. Abusing notation, by Bayes theorem we have

$$f_{k+1}(y_j \mid \bar{y}_{j-1}) = \frac{\int f(y)p(S = k+1 \mid Y = y)\ dy_{j+1}, \ldots, dy_J}{\int f(y)p(S = k+1 \mid Y = y)\ dy_j, \ldots, dy_J}$$
$$= \frac{\int f(y)p(S = k \mid Y = y)\ dy_{j+1}, \ldots, dy_J}{\int f(y)p(S = k \mid Y = y)\ dy_j, \ldots, dy_J} \cdot \frac{\exp\{\eta_{k+1}(\bar{y}_k)\}}{\exp\{\eta_{k+1}(\bar{y}_k)\}} = f_k(y_j \mid \bar{y}_k).$$

Crucially, note that this argument only holds for $k < j$; otherwise, we would not be able to factor out $\exp\{\eta_{k+1}(\bar{y}_k)\}$ in the denominator. □

## S.2 Markov chain Monte Carlo algorithm for the product-multinomial model

For completeness, we provide the MCMC algorithm used to fit the product-multinomial model used for the BCPT. For additional details, see Dunson and Xing (2009). The model

$$p^\star(y, r) = \sum_{k=1}^{\infty} \pi_k \left\{ \prod_{j=1}^{J} \gamma_{kj}^{r_j} (1 - \gamma_{kj})^{1-r_j} \right\} \left\{ \prod_{j=1}^{J} \beta_{kj}^{y_j} (1 - \beta_{kj})^{1-r_j} \right\}. \tag{S.2}$$

can be reformulated as a hierarchical model as follows. For each $i$ we have

$$\begin{aligned} Z_i &\sim \text{Categorical}(\pi) \\ Y_{ij} &\sim \text{Bernoulli}(\beta_{kj}), \qquad (\text{given } Z_i = k), \\ R_{ij} &\sim \text{Bernoulli}(\gamma_{kj}), \qquad (\text{given } Z_i = k), \end{aligned}$$

with the $Y_{ij}$'s and $R_{ij}$'s conditionally independent given $Z_i$. We use a Gibbs sampling algorithm which samples from the full conditionals.

Following Si and Reiter (2013), we truncate the infinite mixture model at some fixed $K < \infty$; unlike the usual latent-class setup, more mixture components are better. Let

$M_k = \sum_{i=1}^{N} I(Z_i = k)$ and $M_{>k} = \sum_{i=1}^{N} I(Z_i > k)$. Define

$$\begin{aligned} N_{\beta kj} &= \sum_{i:R_{ij}=1,Z_i=k} 1, & T_{\beta kj} &= \sum_{i:R_{ij}=1,Z_i=k} Y_{ij}, \\ N_{\gamma kj} &= \sum_{i:Z_i=k} 1, & T_{\gamma kj} &= \sum_{i:Z_i=k} R_{ij}. \end{aligned}$$

Then the full conditionals are given by

$$\begin{aligned} V_k &\overset{\text{indep}}{\sim} \text{Beta}(1 + M_k, \alpha + M_{>k}), && (1 \le k \le K-1) \\ \beta_{kj} &\overset{\text{indep}}{\sim} \text{Beta}\{a_{\beta j}\rho_{\beta j} + T_{\beta kj}, a_{\beta j}(1-\rho_{\beta j}) + N_{\beta kj} - T_{\beta kj}\} && (1 \le k \le K, 1 \le j \le J) \\ \gamma_{kj} &\overset{\text{indep}}{\sim} \text{Beta}\{a_{\gamma j}\rho_{\gamma j} + T_{\gamma kj}, a_{\gamma j}(1-\rho_{\gamma j}) + N_{\gamma kj} - T_{\gamma kj}\} && (1 \le k \le K, 1 \le j \le J). \end{aligned}$$

Each distribution above is understood to be the conditional distribution given both the observed data $(Y_{iR_i}, R_i : 1 \le i \le N)$ and the variables given on the other rows. We must also update the hyperparameters $a_{\beta j}, \rho_{\beta j}, a_{\gamma j}, \rho_{\gamma j}$ for $j = 1, \ldots, J$. The priors we used for these parameters are not conditionally conjugate. We use Metropolis-within-Gibbs for these parameters. The density of the full conditionals are (up-to a normalizing constant),

$$\begin{aligned} a_{\beta j} &\overset{\text{indep}}{\sim} \frac{\sigma}{(\sigma + a_{\beta j})^2} \prod_{k=1}^{K} \text{Beta}\{\beta_{kj} \mid a_{\beta j}\rho_{\beta j} + T_{\beta kj}, a_{\beta j}(1-\rho_{\beta j}) + N_{\beta kj} - T_{\beta kj}\}, \\ \rho_{\beta j} &\overset{\text{indep}}{\sim} I(0 \le \rho_{\beta j} \le 1) \prod_{k=1}^{K} \text{Beta}\{\beta_{kj} \mid a_{\beta j}\rho_{\beta j} + T_{\beta kj}, a_{\beta j}(1-\rho_{\beta j}) + N_{\beta kj} - T_{\beta kj}\}, \\ a_{\gamma j} &\overset{\text{indep}}{\sim} \frac{\sigma}{(\sigma + a_{\gamma j})^2} \prod_{k=1}^{K} \text{Beta}\{\gamma_{kj} \mid a_{\gamma j}\rho_{\gamma j} + T_{\gamma kj}, a_{\gamma j}(1-\rho_{\gamma j}) + N_{\gamma kj} - T_{\gamma kj}\}, \\ \rho_{\gamma j} &\overset{\text{indep}}{\sim} I(0 \le \rho_{\gamma j} \le 1) \prod_{k=1}^{K} \text{Beta}\{\gamma_{kj} \mid a_{\gamma j}\rho_{\gamma j} + T_{\gamma kj}, a_{\gamma j}(1-\rho_{\gamma j}) + N_{\gamma kj} - T_{\gamma kj}\}. \end{aligned}$$

Slice sampling (Neal, 2000) was used for the Metropolis-within-Gibbs steps. To improve mixing on the hyperparameters, it is also valid to ignore any mixture components with $M_k = 0$; that

**Algorithm 1** Monte Carlo integration for PMAR

**procedure** GCOMP($\theta, T, j$) ▷ Approximates $\mu_j$ by simulating $T$ samples from $p_\theta(y)$
  **for** $t = 1, \ldots, T$ **do**
    Sample $(Y^{(t)}_{R^{(t)}}, R^{(t)}) \sim p_\theta(y_r, r)$.
    Sample $Y^{(t)}_{-R^{(t)}} \sim p_\theta(Y_{s(-r)} \mid Y_r = y_r, R = \mathbb{1})$.
  **end for**
  Set $\mu_j = T^{-1} \sum_{s=1}^{T} Y_{sj}$.
  **return** $\mu_j$
**end procedure**

**Algorithm 2** Monte Carlo integration for tilted-last-occasion

**procedure** GCOMP($\theta, T, \xi$) ▷ Approximates $\mu_J$ by simulating $T$ samples from $p_\theta(y_J)$
  **for** $t = 1, \ldots, T$ **do**
    Sample $(Y^{(t)}_{R^{(t)}}, R^{(t)}) \sim p_\theta(y_r, r)$.
    **if** $R^{(t)}_J = 0$ **then**
      Sample

$$Y^{(t)}_J \sim \frac{p_\theta(y_J \mid y_r, R^{(t)} = \mathbb{1}) e^{\xi y_J}}{E_\theta(e^{\xi Y_J} \mid Y_r = y_r, R^{(t)} = \mathbb{1})}.$$

    **end if**
  **end for**
  Set $\mu_J = T^{-1} \sum_{s=1}^{T} Y_{sJ}$.
  **return** $\mu_J$
**end procedure**

is, in each of the products above, we take the product over $\{k : M_k > 0\}$.

## S.3 Monte Carlo algorithms

In Algorithm 1 and Algorithm 2 we give the Monte Carlo integration algorithms used for PMAR and the tilted-last-occasion restriction in the analysis of the BCPT data in Section 5.

# S.4 Conditional distributions for the product-multinomial mixture model

For the model given in Section 5, we need to sample from the conditional distributions of (S.2). This task is straight-forward due to the within-class independence. For example, the conditional density $p_\theta(y_{-r} \mid y_r, R = \mathbb{1})$ is given by

$$
\begin{aligned}
p^\star_\theta(y_{-r} \mid y_r, \mathbb{1}) &= \frac{\sum_{k=1}^K \pi_k \left\{\prod_{j=1}^J \gamma_{kj}\right\} \left\{\prod_{j=1}^J \beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j}\right\}}{\sum_{k=1}^K \pi_k \left\{\prod_{j=1}^J \gamma_{kj}\right\} \left\{\prod_{j:r_j=1} \beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j}\right\}} \\
&= \sum_{k=1}^K \vartheta_k(y_r) \prod_{j:r_j=0} \beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j},
\end{aligned}
$$

where $\sum_{k=1}^K \vartheta_k(y_r) = 1$ and

$$
\vartheta_k(y_r) \propto \pi_k \left\{\prod_{j=1}^J \gamma_{kj}\right\} \left\{\prod_{j:r_j=1} \beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j}\right\}.
$$

Hence sampling from this conditional can be accomplished by choosing a mixture component $k$ with probability $\vartheta_k(y_r)$ and then sampling independently $Y_j \sim \text{Bernoulli}(\beta_{kj})$. The density of $p_\theta(y_j \mid \bar{o}_{j-1})$ is given by

$$
\frac{\sum_{k=1}^K \pi_k \left\{\prod_{\ell=1}^{j-1} \gamma_{k\ell}^{r_\ell}(1-\gamma_{k\ell})^{1-r_\ell}\right\} \left\{\prod_{\ell<j,r_\ell=1} \beta_{k\ell}^{y_\ell}(1-\beta_{k\ell})^{1-y_\ell}\right\} \beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j}}{\sum_{k=1}^K \pi_k \left\{\prod_{\ell=1}^{j-1} \gamma_{k\ell}^{r_\ell}(1-\gamma_{k\ell})^{1-r_\ell}\right\} \left\{\prod_{\ell<j,r_\ell=1} \beta_{k\ell}^{y_\ell}(1-\beta_{k\ell})^{1-y_\ell}\right\}}.
$$

As before, this can be written as $\sum_{k=1}^K \vartheta_k(y_r, r)\beta_{kj}^{y_j}(1-\beta_{kj})^{1-y_j}$ where

$$
\vartheta_k(y_r, r) \propto \pi_k \left\{\prod_{\ell=1}^{j-1} \gamma_{k\ell}^{r_\ell}(1-\gamma_{k\ell})^{1-r_\ell}\right\} \left\{\prod_{\ell<j,r_\ell=1} \beta_{k\ell}^{y_\ell}(1-\beta_{k\ell})^{1-y_\ell}\right\}.
$$

| | | ATE | Lower limit | Upper limit | $P$-value |
|---|---|---|---|---|---|
| MAR | FB | 0.752 | -0.693 | 2.233 | 0.168 |
| | MI | 0.638 | -0.874 | 2.150 | 0.204 |
| PMAR | FB | 0.497 | -0.992 | 2.005 | 0.258 |
| | MI | 0.468 | -1.008 | 1.944 | 0.267 |
| SE | FB | 4.091 | 0.841 | 7.410 | 0.006 |
| | MI | 3.850 | 0.329 | 7.372 | 0.016 |
| TLO | FB | 0.292 | -2.536 | 3.151 | 0.420 |
| | MI | 0.331 | -2.557 | 3.218 | 0.411 |

Table S.2: Results for the BCPT study. MAR, PMAR, SE, and TLO denote the missing at random, pairwise missing at random, sequential explainability, and tilted-last-occasion models. FB denotes inference obtained using the fully-Bayes procedure while MI denotes inference obtained using multiple imputation. ATE denotes the point estimate of the average treatment effect, with lower and upper limits associated to a 95% credible interval. The posterior probability is associated to a one-sided test for $\psi > 0$.

## S.5 Results of BCPT

We provide the results for the BCPT from Figure 2 in tabular form in Table S.2. Results are given using both fully-Bayesian inference with the G-computation algorithm and multiple imputation with $M = 50$ imputed datasets. We see that the inferences using either fully-Bayesian inference or multiple imputation largely agree.

## References

Dunson, D. B. and Xing, C. (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265.

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika*, 104(1):207–220.

Si, Y. and Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521.