A symmetric prior for multinomial probit models

Lane F. Burgette
RAND Corporation

P. Richard Hahn
University of Chicago Booth School of Business

January 2013

Abstract

Under standard prior distributions, fitted probabilities from Bayesian multinomial probit models can depend strongly on the choice of a base category, which is used to identify the model. This paper proposes a novel identification strategy and prior distribution for the model parameters that makes the prior symmetric with respect to relabeling the outcome categories. Further, our new prior allows for an efficient marginal data augmentation Gibbs sampling algorithm that samples rank-deficient covariance matrices without resorting to Metropolis-Hastings updates.

Keywords: Base category, Discrete choice, Gibbs sampler, Marginal data augmentation, Sum-to-zero identification

1 Introduction

In multinomial probit (MNP) models of discrete choices, parameters are typically identified by selecting a base category relative to which the choice parameters are defined. From the point of view of identification, the choice of base category is immaterial. However, in a Bayesian framework, previously developed priors can be sensitive to the base category

specification — sometimes strongly so. Hence, when practitioners choose a base category in the MNP analysis, they are unwittingly making a decision about the prior specification for their model.

In this paper, we propose sum-to-zero restrictions on the latent utilities and regression parameters that define the MNP model. In this novel identification framework, we are able to define a prior which is symmetric with respect to relabeling of the outcome categories. Even so, this model preserves the favorable computational aspects of other, recent Bayesian MNP models (Imai and van Dyk, 2005*a*; Burgette and Nordheim, 2012).

1.1 Multinomial probit models of discrete choice

Multinomial probit (MNP) models are popular in studies involving discrete choice data (McFadden, 1974; Train, 2003). They have applications in marketing (Rossi et al., 2005), politics (Rudolph, 2003), transportation studies (McFadden, 1974; Garrido and Mahmassani, 2000), and beyond. The MNP is more flexible than standard multinomial logit models, as it need not make an assumption of independence of irrelevant alternatives (IIA). This means that the ratio of selection probabilities for two outcome categories can depend on the characteristics of another category. Further contributing to the popularity of the MNP is a series of advances in Bayesian computation, starting with Albert and Chib (1993), that has made it increasingly computationally manageable (McCulloch and Rossi, 1994; McCulloch et al., 2000; Imai and van Dyk, 2005*a*,*b*).

The MNP requires two normalizations in order to identify the model. These models can be derived through the assumption that agents construct latent Gaussian utilities and select the category that corresponds to the largest utility. Since the ordering of the utilities is maintained by an additive shift or multiplicative rescaling, identifying assumptions on the scale and location are needed.

In order to set the scale, it has been standard to fix an element on the main diagonal of the covariance matrix at one. Burgette and Nordheim (2012) demonstrated that the choice of which element one fixed could have a meaningful impact on posterior predictions, when

using the popular prior of Imai and van Dyk (2005*a*). To avoid this problem, they proposed a model that identifies the scale of the model by fixing the trace of the covariance matrix, which makes the prior covariance invariant to joint permutations of the rows and columns. This paper will build upon such a trace-restricted prior, resolving the location identification issue as well.

Previous MNP models have set the location of the latent utilities by specifying a base (or reference) category for the model. The base category's utility is then subtracted from all of the other utilities for each observation, removing the indeterminacy of the location. But, Burgette and Nordheim (2012) noted that Bayesian MNP predictions can be sensitive to the specification of the base category, though they did not provide a satisfactory solution for this issue. This problem arises because instead of specifying a prior for the original utilities and inducing a prior on the base-subtracted utilities, it has been standard to specify a prior directly on base-subtracted utilities.

Rather than selecting a reference category whose utility is assumed to be equal to zero, we enforce a sum-to-zero restriction on the latent utilities. If respondents choose from p categories, other MNP methods transform the utilities to $(p-1)$-space. Instead, we constrain our utilities to exist in a $(p-1)$-dimensional hyperplane in p-space.

We apply our new prior to two consumer choice datasets, as well as a series of simulated datasets based on the consumer choice studies. In doing so, we see that the *symmetric MNP* (sMNP) model defines a more sensible model, produces better predictions, and has favorable computational properties compared to previous MNP models.

1.2 Preliminaries

Assume that agent $i = 1, \ldots, n$ is choosing among p mutually exclusive alternatives. The MNP can be derived by assuming that there exist vectors of latent Gaussian utilities $W_i = \{w_{ij}\}$ of length p, and that each agent selects the alternative with the highest utility, so that we observe $Y_i = \arg\max_j w_{ij}$.

It is standard to assume that the utilities take the form

$$W_i = X_i\beta + \varepsilon_i. \tag{1}$$

X_i is a matrix of covariates, β is a vector of regression parameters, and $\varepsilon_i \overset{\text{iid}}{\sim} \text{normal}(0, \Sigma)$ capture variations in taste across agents. We will assume X_i contains intercept terms, k_d covariates that vary by decision-maker (e.g., a buyer's age), and k_a alternative-specific covariates (e.g., product prices). We assume the covariates are arranged in that order (from left to right) so that

$$X_i = \begin{bmatrix} I_p & x_{i,d}^\top \otimes I_p & x_{i,a} \end{bmatrix}. \tag{2}$$

The k_d-vector $x_{i,d}$ is the collection of covariates that vary by individual; $x_{i,a}$ is a $p \times k_a$ matrix whose columns contain the values of the variables that vary by alternative.

The standard identifying approach is to transform W_i to $W_i^* = T_{\text{bc}} W_i$ where

$$T_{\text{bc}} = \begin{bmatrix} -J_{p-1} & I_{p-1} \end{bmatrix} \tag{3}$$

with J_{p-1} a column vector of ones with length $p-1$. Then we would assume $W_i^* = X_i^*\beta^* + \varepsilon_i^*$ where

$$X_i^* = \begin{bmatrix} I_{p-1} & x_{i,d}^\top \otimes I_{p-1} & T_{\text{bc}} x_{i,a} \end{bmatrix} \tag{4}$$

and $\varepsilon_i^* \overset{\text{iid}}{\sim} \text{normal}(0, \Sigma^* = T_{\text{bc}} \Sigma T_{\text{bc}}^\top)$.

Albert and Chib (1993) had the key insight that data augmentation (Tanner and Wong, 1987) would greatly ease the estimation of the MNP. If we treat the latent W_i^* as parameters to be updated in the MCMC algorithm, then under a normal prior, the full conditional distribution of β^* is normal. Further, the full conditional distribution of each W_i^* is truncated multivariate normal, which can be updated one component at a time as univariate truncated normals (McCulloch and Rossi, 1994).

It then remains to sample Σ^*, the $(p-1)$-dimensional covariance over the base-subtracted utilities. Up to a constraint and the normalizing constant, the priors for both the Imai and

van Dyk and the Burgette and Nordheim models are the same:

$$p(\Sigma^*) \propto |\Sigma^*|^{-(\nu+p)/2}[\mathrm{tr}(S\Sigma^{*-1})]^{-\nu(p-1)/2}\mathbf{1}\{\mathrm{cond}\}, \tag{5}$$

where $\mathbf{1}\{\mathrm{cond}\}$ is equal to one if $\{\mathrm{cond}\}$ is a true statement, and zero otherwise. For Imai and van Dyk, this condition is $\{\sigma^*_{11} = 1\}$; for Burgette and Nordheim the condition is $\{\mathrm{tr}(\Sigma^*) = (p-1)\}$. Further, the working parameter is given the joint prior

$$p(\Sigma^*, \alpha^2) \propto |\Sigma^*|^{-(\nu+p)/2} \exp\{-1/(2\alpha^2)\,\mathrm{tr}(S\Sigma^{*-1})\}(\alpha^2)^{-[\nu(p-1)/2+1]}\mathbf{1}\{\mathrm{cond}\}, \tag{6}$$

after which Σ^* can be sampled in a Gibbs step through a draw of $\tilde{\Sigma} = \alpha^2\Sigma^*$.

1.3 Asymmetries of currently-used MNP priors

Later in this paper, we will demonstrate empirically that switching from one base category to another can result in substantial differences in estimated (posterior) purchase probabilities in marketing applications that appear elsewhere in the literature. To motivate this work, we begin by highlighting differences in the prior purchase probabilities under two base category specifications, conditional on a range of values of the structural portion of the utilities, $X_i^*\beta^*$. In what follows, we consider a simple case with $p = 3$ categories and focus on one of the three outcome categories, which we will refer to as the category of interest. First, we specify this category of interest as the base category (corresponding to $Y_i = 0$), and then we reparametrize so that the category of interest is the first non-base category (corresponding to $Y_i = 1$). Our experience indicates that sensitivity to the base category primarily comes from the prior on Σ^* (rather than β^*), so we will condition on β^* in order to clarify the issue.

We begin by considering $\Pr(Y_i = 0 \mid X_i^*\beta^* = (-v, -v)^\top)$ and $\Pr(Y_i = 1 \mid X_i^*\beta = (v, 0)^\top)$. These probabilities are marginal to the trace-restricted variant of the Imai and van Dyk prior for Σ^* with $\nu = 2$ degrees of freedom, and centered at $S = .5J_2J_2^\top + .5I_2$. These hyperparameters correspond to a default prior for a model of $p = 3$ outcome categories.

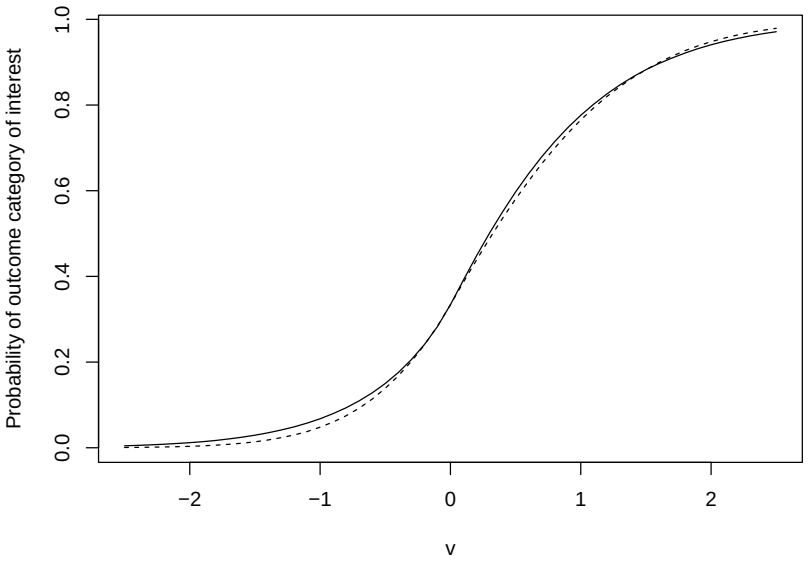

Figure 1: Plot of $\Pr(Y_i = 0 \mid X_i^*\beta^* = (-v, -v)^\top)$ (solid line) and $\Pr(Y_i = 1 \mid X_i^*\beta^* = (v, 0)^\top)$ (dotted line) marginal to Σ^* for the Imai and van Dyk prior, as a function of v. Notice that these probabilities — which correspond to a change in base category — are very similar.

(Using $S = I_2$ would result in stronger asymmetries with respect to the base category.) As seen in Figure 1, we find that the implied selection probabilities under the two base category specifications are actually very close to one another across all v.

Although Figure 1 might hint that changing from one base category to another does little to impact the model, a closer analysis displayed in Figure 2 shows that this is not the case. In short, marginalization over Σ^* is obscuring the differences induced by re-parametrizing the base category. In Figure 2, we examine the distribution $\Pr(Y_i = 0 \mid X_i^*\beta^* = (-v, -v)^\top, \Sigma^*)p(\Sigma^*)$ at $v = 1$, and the analogous distribution for the case when the category of interest corresponds to $Y_i = 1$. Stated symbolically, we find that

$$\int \Pr(Y_i = 0 \mid X_i^*\beta^* = (-1, -1)^\top, \Sigma^*)p(\Sigma^*)d\Sigma^* \approx \int \Pr(Y_i = 1 \mid X_i^*\beta^* = (1, 0)^\top, \Sigma^*)p(\Sigma^*)d\Sigma^*$$

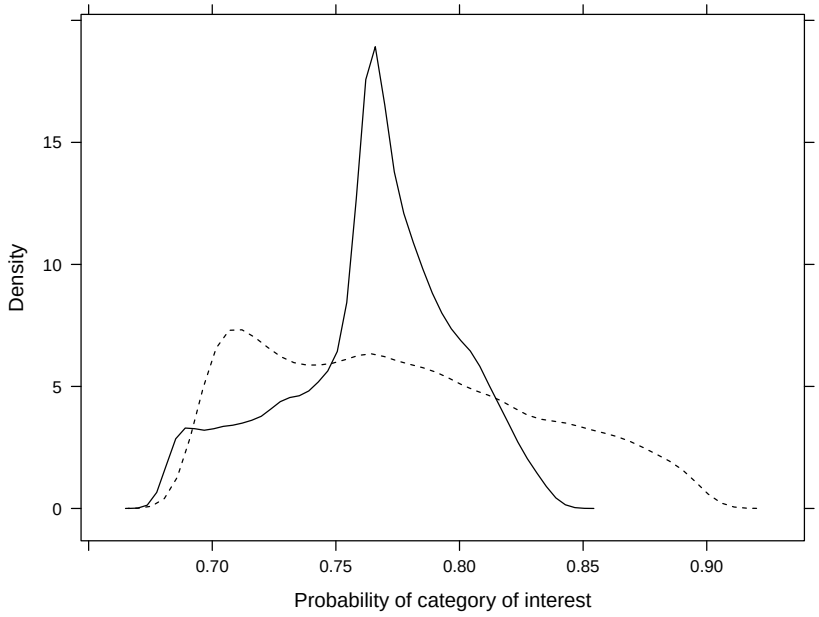

Figure 2: Density histogram of $\Pr(Y_i = 0 \mid X_i^*\beta^* = (-1,-1)^\top, \Sigma^*)p(\Sigma^*)$ (solid line) and $\Pr(Y_i = 1 \mid X_i^*\beta^* = (v,0)^\top, \Sigma^*)p(\Sigma^*)$ (dotted line). Although the means are very similar (as indicated by the $v = 1$ slice of Figure 1), the distributions themselves are quite different.

but

$$\int (\Pr(Y_i = 0 \mid X_i^*\beta^* = (-1,-1)^\top, \Sigma^*))^2 p(\Sigma^*)d\Sigma^* < \int (\Pr(Y_i = 1 \mid X_i^*\beta^* = (1,0)^\top, \Sigma^*))^2 p(\Sigma^*)d\Sigma^*.$$

Because the differences in prior probabilities appear primarily to be of second and higher moments, an ad-hoc solution to the problem of base category dependence (such as specifying alternative values of the hyperparameters, or by specifying a different $p(\beta^*|\Sigma^*)$ to compensate) may be difficult. Further, even though the prior differences are obscured by averaging over draws of Σ^*, this often is not the case after conditioning on observed data. Although we expect the impact of the prior to fade as the sample size increases, information in multinomial models accrues slowly relative to standard models of a continuous outcome, which means that asymmetries in the prior for an MNP model may persist in the posterior for sample sizes that are typical in business and economics applications. Hence, we pursue

a prior that is identically invariant to relabeling the outcome categories.

2 A symmetric prior for MNP regressions

We now propose a *symmetric MNP* (sMNP) model that is invariant under relabeling or reordering of the outcome categories. Rather than identifying the locations of the latent utilities by subtracting one from the others, we instead require that they sum to zero. (This assumes that the choice-specific covariates have mean zero for each observation, which is a convenient but inessential standardization.) Further, we assume that the regression parameters that correspond to each agent-specific covariate sum to zero, which gives the same degrees of freedom as the standard MNP, where (in a sense) the regression parameters related to the base category are set equal to zero.

With this sum-to-zero restriction on the utilities, we require a covariance for W_i that is symmetric and positive-semidefinite with $p-1$ positive eigenvalues, and constrained in some way in order to set the scale of the model. Rather than directly specifying a distribution on $p \times p$ matrices, we build it up with a mixture of trace-restricted positive-definite matrices. Conditionally, we assume that a positive-definite matrix of dimension $p-1$ describes the covariance of all but one of the dimensions of W_i. We denote the left-out category with the parameter b, and refer to it as the *faux base category* indicator. In contrast to previous MNP models, b is learned according to Bayes rule.

The proposed model is as follows:

$$
\begin{aligned}
b &\sim \text{unif}(\{1, \ldots, p\}) && (7)\\
\Sigma_{-b} &\sim p_{\text{TR}}(S_b, \nu_b) && (8)\\
R_{-b} &= [\text{chol}(\Sigma_{-b})]^\top && (9)\\
R &= \begin{bmatrix} R_{1:(b-1)} \\ R_b^* \\ R_{b:p} \end{bmatrix} && (10)\\
\beta_{-b} &\sim \text{normal}(0, B_{-b}^{-1}) && (11)\\
\beta &= f_0(\beta_{-b}) && (12)\\
W_i &\overset{\text{ind}}{\sim} \text{normal}(X_i\beta, RR^\top) && (13)\\
Y_i &= \arg\max_j W_i. && (14)
\end{aligned}
$$

In this formulation, p_{TR} is the trace-restricted variant of the Imai and van Dyk (2005*a*) prior in (5). Its hyperparameters S_b and ν_b may change with b but we recommend using common hyperparameters in most cases, since $S_b = \text{diag}\{(1+c, \ldots, 1+c)\} - cJJ^\top$ for all b and a common ν_b will yield a prior covariance structure that is symmetric with respect to the outcome categories. As a default, we recommend using $c = 1/(p-1)$. This corresponds to the first $p-1$ rows and columns of a symmetric $p \times p$ covariance matrix P with $p-1$ positive eigenvectors that is symmetric with respect to relabeling the rows and columns. This matrix has the property that vectors drawn from the normal$(0, P)$ distribution sum to zero almost surely, which is a natural center for our relabeling invariant, sum-to-zero MNP. Using $c = 0$ means roughly that we expect $p-1$ of the dimensions of the utilities to be independent, with the remaining dimension strongly anti-correlated. Using $c = 1/(p-1)$ is a more neutral prior, and seems to lead to better mixing in the MCMC.

R_{-b} is the transposed Cholesky decomposition of Σ_{-b} such that $R_{-b}R_{-b}^\top = \Sigma_{-b}$. R_b^* is a row vector inserted into R_{-b} at the bth row such that the sum of each column of R is zero. In this formulation, β_{-b} has dimension $(p-1)(k_d+1)+k_a$ (assuming that intercept terms

are included). The function f_0 acts on β_{-b} such that for each sub-vector of length $p-1$ that corresponds to an agent-specific covariate (or the intercepts), β is equal to β_{-b} with an extra dimension inserted at the bth position in the sub-vector. This inserted element is chosen so that the sub-vector sums to zero.

With this model specification, we induce a prior distribution on the set of positive-semidefinite matrices of dimension p that have exactly $p-1$ positive eigenvalues. It would also be possible to work with a matrix decomposition like $\Sigma = ADA'$, where A is a $p \times (p-1)$ orthogonal matrix and D is diagonal. One could then define a prior on the Stiefel manifold that contains A (Hoff, 2009). This would be a more direct definition on positive semidefinite matrices, but inducing a prior in the manner implied by our model is conceptually simple and guarantees favorable computational properties.

To make the motivation of this new set of identifying restrictions explicit, we note that they result from transforming the unnormalized utilities not by T_{bc} as in (3), but rather multiplying them by a p-dimensional square matrix T_s that is defined to have ones on the main diagonal, and entries of $-1/(p-1)$ elsewhere. Note that $\arg\max W_i = \arg\max T_s W_i$, while the elements of $T_s W_i$ sum to zero. This transformation also induces the proposed identifying restrictions on β. If we partition $\beta = (\beta_d, \beta_a)$, where β_a corresponds to the covariates that vary by outcome category, we have

$$T_s X_i \beta = X_i \begin{bmatrix} (I \otimes T_s)\beta_d \\ \beta_a \end{bmatrix}. \tag{15}$$

This transformed version of β (i.e., the second factor on the right-hand side of the above equation) conforms to the proposed identifying restrictions. Similarly, a normal distribution with mean zero and covariance $T_s \Sigma T_s^\top$ results in draws that sum to zero almost surely. (Note that T_s is almost idempotent in the sense that $T_s T_s = c T_s$ for some scalar c. The first $p-1$ rows and columns of T_s therefore serve as our default for S_b since this corresponds to the transformed variance of ε_i if its variance in the unnormalized scale is proportional to the identity.)

We emphasize that there is nothing inherently wrong with using the asymmetric identifying transformation T_{bc}. If we do not wish for our inferences to depend on the base category, however, the prior must compensate for the asymmetries in the transformation. This seems quite difficult to achieve, especially if we hope to have a computationally tractable model. Using T_s, however, we can decouple prior specification and model identification, all while preserving the favorable computational characteristics of existant MNP models.

2.1 Model estimation

We propose a Gibbs sampler to estimate the model. As with the algorithm described by Imai and van Dyk (2005*a*), we sample the working parameter α at each step in the MCMC. Such complete marginalization over working parameters can improve the mixing of the Markov chains (van Dyk, 2010). The MCMC switches between working with parameters in the scale defined by Σ_{-b}, and the scale defined by $\tilde{\Sigma}_{-b} = \alpha^2 \Sigma_{-b}$. Any parameter topped by a tilde refers to parameters in the latter scale. For example, $\tilde{W}_{i,-b} \overset{\text{ind}}{\sim} \text{normal}(X_{i,-b}\tilde{\beta}_{-b}, \tilde{\Sigma}_{-b})$. Here, $X_{i,-b}$ indicates X_i with the bth row and the columns specific to the bth category removed.

We initialize the latent utilities W_i by sampling a standard normally-distributed vector of length p and centering it at zero. We then permute its elements so that the maximum of each W_i coincides with the observed Y_i.

The sampler then repeatedly iterates through the following steps:

1. Gibbs through the $W_{i,-b}$ elements. $w_{i,b}$ is known given b and $W_{i,-b}$. After dropping the bth element of W_i and the corresponding elements in X_i and β, the full conditionals of elements of $W_{i,-b}$ are truncated univariate normal. The conditional means and variances can be calculated as described by McCulloch and Rossi (1994). The truncations are:

    - If $Y_i = j \neq b$, sample w_{ij} from a truncated normal so that $w_{ij} > -.5 \sum_{k \notin \{j,b\}} w_{ik}$ and $w_{ij} > \max(w_{ik} : k \notin \{j, b\})$.
    - If $Y_i \neq b$ and $Y_i = k \neq j$, sample w_{ij} from a truncated normal so that $w_{ij} < w_{ik}$

and $w_{ij} > -\sum_{l \neq b} w_{il} - w_{ik}$.

- If $Y_i = b$, sample w_{ij} from a truncated univariate normal such that

$$w_{ij} < \min\{-.5 \sum_{k \notin \{b,j\}} w_{ik}, -1(\max\{W_{-\{j,b\}}\} + \sum_{k \notin \{b,j\}} w_{ik})\}.$$

2. Sample $\alpha^2|\text{all} \sim \text{tr}(S_b \Sigma_{-b}^{-1})/\chi^2_{\nu_b(p-1)}$.

3. Set $\tilde{W}_i = \alpha W_i$ for all i.

4. Sample $(\alpha, \beta_{-b}|\text{all})$ via

$$\begin{aligned}
\hat{\beta}_{-b} &= [\sum_{i=1}^{n} X_{i,-b}^\top \Sigma_{-b}^{-1} X_{i,-b} + B_{-b}]^{-1} [\sum_{i=1}^{n} X_{i,-b}^\top \Sigma_{-b}^{-1} \tilde{W}_{i,-b}], \\
\text{IP} &= \sum_{i=1}^{n} (\tilde{W}_{i,-b} - X_{i,-b}\hat{\beta}_{-b})^\top \Sigma_{-b}^{-1} (\tilde{W}_{i,-b} - X_{i,-b}\hat{\beta}_{-b}), \\
\alpha^2 &\sim \frac{\text{IP} + \hat{\beta}_{-b}^\top B_{-b} \hat{\beta}_{-b} + \text{tr}(S_b \Sigma_{-b}^{-1})}{\chi^2_{(n+\nu_b)(p-1)}}, \text{ and} \\
\tilde{\beta}_{-b} &\sim \text{normal}\Big(\hat{\beta}, \alpha^2 \big(\sum_{i=1}^{n} X_{i,-b}^\top \Sigma_{-b}^{-1} X_{i,-b} + B_{-b}\big)^{-1}\Big).
\end{aligned}$$

Record $\beta = \alpha^{-1}\tilde{\beta}$.

5. Sample $(b, \tilde{\Sigma}_{-b})$ via

$$\begin{aligned}
\text{OP}_b &= \sum_{i=1}^{n} (\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})(\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})^\top, \\
p(b|\tilde{\beta}, \tilde{W}) &\propto |S_b + \text{OP}_b|^{-(n+\nu_b)/2}, \text{and} \\
p(\tilde{\Sigma}_{-b}|b, \tilde{\beta}, \tilde{W}) &\sim \text{inv-Wishart}(n + \nu_b, S_b + \text{OP}_b).
\end{aligned}$$

(See the Appendix.)

6. Make the following transformations:

- $\alpha^2 = \text{tr}(\tilde{\Sigma}_{-b})/(p-1)$

- $\Sigma_{-b} = \alpha^{-2}\tilde{\Sigma}_{-b}$
- $\beta_{-b} = \alpha^{-1}\tilde{\beta}_{-b}$
- $W_i = \alpha^{-1}\tilde{W}_i$ for all i

Iteration for iteration, this algorithm is only slightly higher in cost computationally than the Imai and van Dyk (2005*a*) and Burgette and Nordheim (2012) algorithms. In particular, calculating the p determinants in Step 5 is necessary for the proposed sampler, but does not enter into the previous algorithms. Since the dimension of the matrices $(p-1)$ will typically be modest, this is not a great increase in computational cost. Further, we will see that the excellent mixing in the resulting Markov chains more than makes up for this extra computation, to say nothing of the desirable symmetry of the new model.

3 Demonstrations

3.1 Clothes detergent purchases

Imai and van Dyk (2005*a*,*b*) apply their methods to a consumer choice model of clothing detergent purchases. The data are available in their `MNP` package in R. We have records of purchasing decisions along with available log-prices for shoppers choosing between All, Era Plus, Solo, Surf, Tide, and Wisk brand detergents. There are 2657 observations and only six regression parameters, so we typically do not see large differences in estimated purchase probabilities based on the various base category fits. However, specifying the base category to be All — which is rarely purchased despite its low price — does give somewhat different predictions for All when its price is low. To see this, we set the prices for all other brands at their brand-specific average, and consider predicted purchase probabilities across a range of low prices for All. The predictions from five of the base categories (solid lines) are very similar. The predictions when All is the base category (dashed line) are notably higher. When we apply the sMNP to the data, we see that its predictions are intermediate to those of the various base category fits (dotted line).

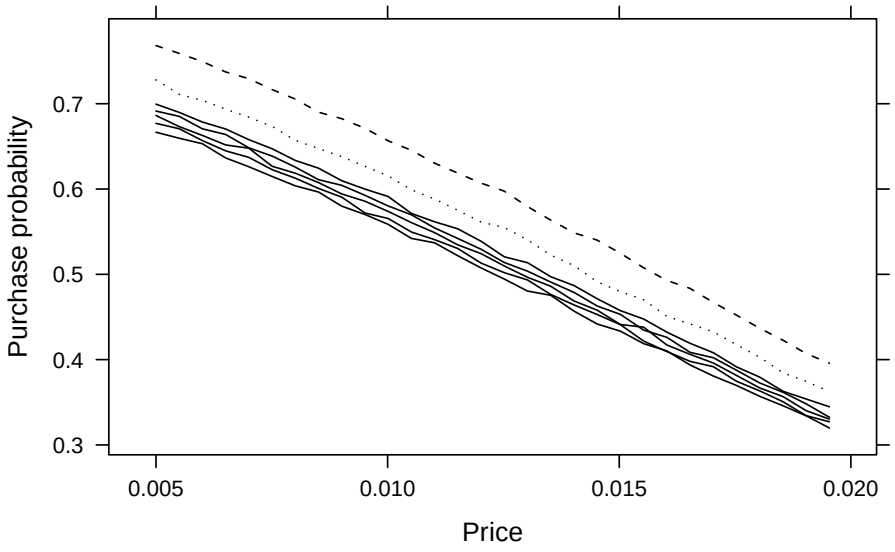

Figure 3: Estimated purchase probabilities for ALL brand detergent, with all other brands' prices fixed at the brand-specific mean observed price. The dashed line uses ALL as the base category; the solid lines use all of the other possible base categories. The dotted line results from an sMNP fit.

To interpret the β parameters, we know — by the sum-to-zero property of the intercept terms — that a brand with an intercept coefficient that is persistently negative (ALL) is less desirable than average, in a sense (Figure 5). ERAPLUS and TIDE are estimated to be more desirable. However, note that these intercepts do not reflect marginal purchase probabilities, as less desirable brands may also have lower prices. As economic theory would suggest, the price coefficient is strongly negative (Figure 6), which indicates that raising a detergent's price (relative to the competitors) will lower its estimated purchase probability.

Although these interpretations of the β parameters are accurate, we would argue that summaries of MNP results are best phrased in terms of changes in posterior predicted selection probabilities. For example, one might consider the effect of a proposed price increase on the current purchase probabilities. We advocate this because predictions take into account both β and Σ parameters, and the Σ parameters can be very difficult to interpret on their own. If only the β parameters are of interest in an application, we would

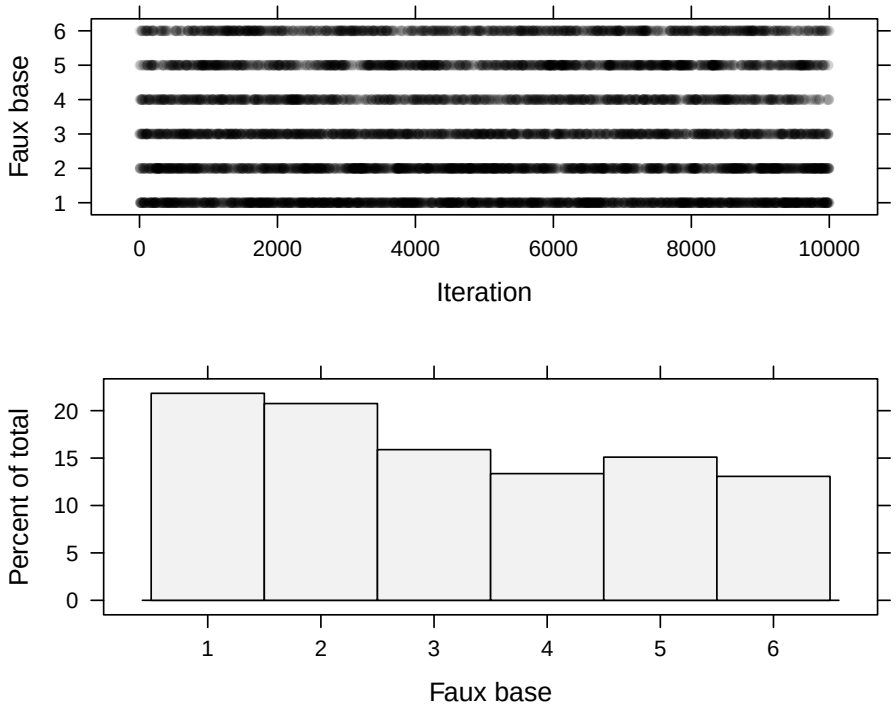

Figure 4: Trace plot and histogram of samples from the posterior distribution of faux base parameters b for the detergent data. In the upper panel, points are plotted with 10% intensity. The numbers 1 through 6 correspond to All, EraPlus, Solo, Surf, Tide, and Wisk, respectively.

argue that a model that assumes IIA may be more appropriate.

We also highlight the mixing behavior of the sMNP algorithm. For example, the faux-base parameter b mixes extremely well, as indicated by the near constant switching between its six possible values (Figure 4). Further, the mixing of the price parameter in the symmetric MNP algorithm compares favorably to the base category MNP in fits of these data (Figure 6). Imai and van Dyk (2005a) used these data to demonstrate improved mixing performance of their model relative to earlier MNP models, so these results are a comparison against the state of the art.

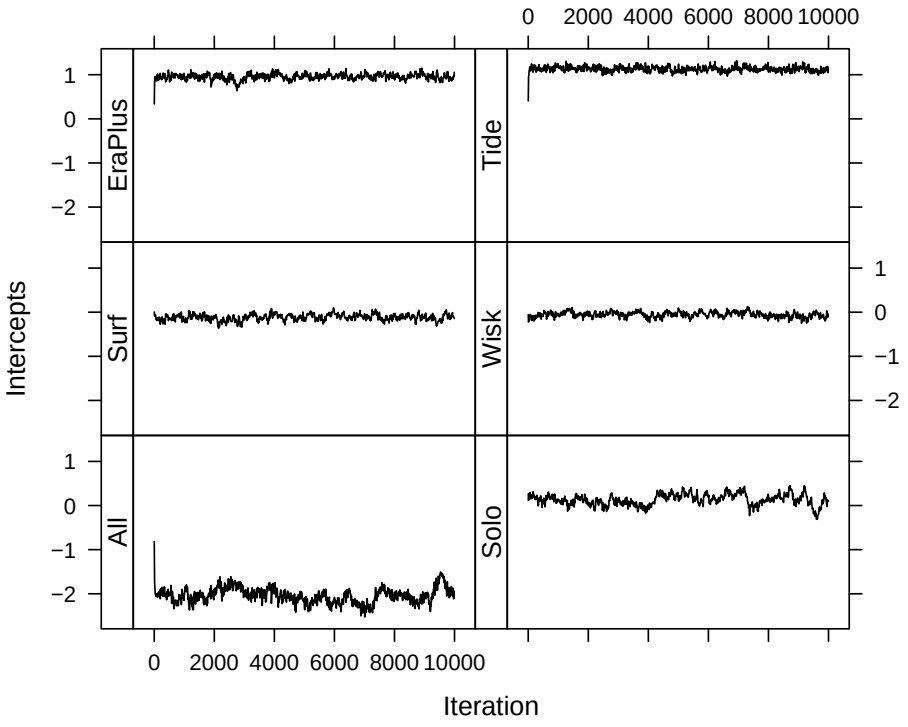

Figure 5: Trace plots of samples from the posterior distributions of the intercept terms for an sMNP fit of the detergent data.

3.2 Margarine purchases

We also consider a similar analysis of consumer purchases of margarine that are available in the `bayesm` package in R. Again, our model only has intercepts and a price coefficient. Following McCulloch and Rossi (1994), we limit our analysis to purchases of Parkay, Blue Bonnet, Fleischmanns, House brand, Generic, and Shedd Spread tub margarines. And, following Burgette and Nordheim (2012), we limit the analysis to the first purchase of one of these brands for each household. This results in a dataset with 507 observations. With the smaller sample size, there are larger differences in posterior estimated purchase probabilities when one switches from one base category to another in standard MNP fits.

In Figure 7, we see that sMNP predictions again tend to be between those of standard MNP models when we consider all possible base categories, as was the case in Figure 3.

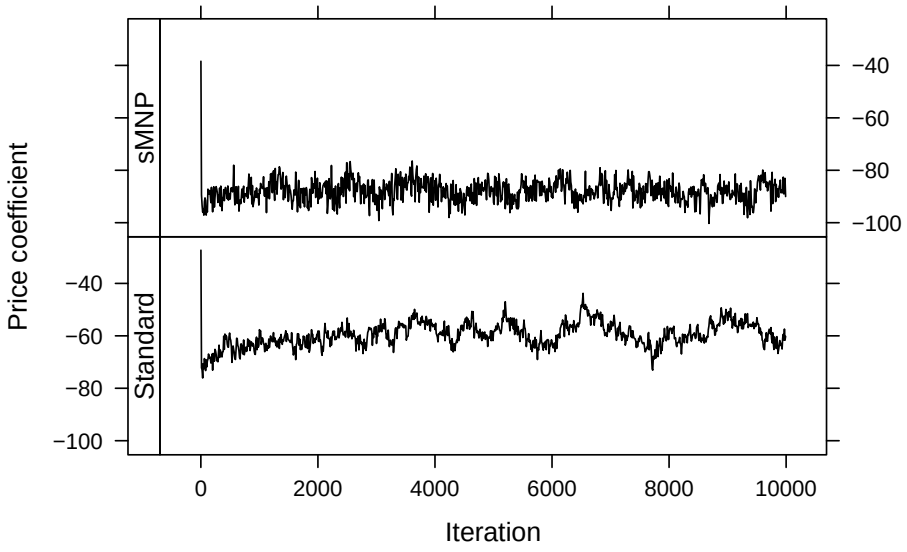

Figure 6: Trace plots of samples from the posterior distributions of the price coefficients for sMNP and standard MNP fits of the detergent data.

The observed HOUSE brand prices are between \$0.19 and \$0.64, so there is significant disagreement across nearly the entire range of observed prices for that brand. (With the larger sample size in the detergent data, we only saw meaningful differences when we extrapolated out of the observed price range.) Although there is some Monte Carlo error in the estimates, it is insignificant compared to the 19% difference between the low and high estimates of HOUSE's selection probability when it is priced at \$0.20.

Thus, in both of these examples, we see that the sMNP gives predictions that are between those of the standard MNP models that are fit alternately with each base. This is compatible with the heuristic interpretation of the sMNP as a model that averages across base categories in standard MNP models.

An alternative approach to handling dependence on the base category would be to fit an Imai and van Dyk-style MNP model using each base category separately, and perform a post-hoc average of the fitted probabilities. We find this to be unappealing from several perspectives. First, the computation load is p times as large as it would be; the sMNP is

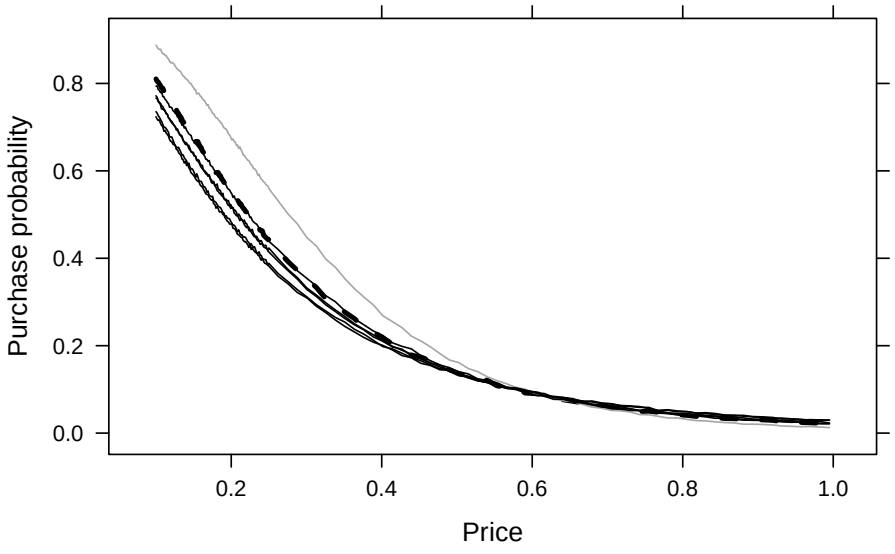

Figure 7: Estimated purchase probabilities for HOUSE brand margarine over a range of prices for that brand, with other prices fixed. The solid lines are posterior predictions from standard MNP models, and the dashed line is from the sMNP. The gray line uses HOUSE brand as the base category.

only slightly more expensive than a single base category MNP. More importantly, the sMNP constitutes a proper Bayesian procedure, which automatically confers a range of theoretical advantages.

3.3 A simulation study

Here we compare the fitted probabilities of MNP models that use each of the possible base categories and the fitted probabilities that result from the base category-free sMNP. We simulate 50 datasets that are loosely based on the consumer choice examples above. We assume that $n = 750$ consumers are choosing from $p = 6$ products. The simulated product-specific intercepts and mean prices have correlation 0.9 so that more desirable products are more expensive, as one would expect. The price coefficient was drawn uniformly from $[-1.25, -.75]$ so that if a product is relatively less expensive, it will be more popular. Finally, a $p \times p$ covariance matrix with expectation I is drawn from an inverse-Wishart distribution with 50 degrees of freedom. The simulation parameters were chosen so that each "brand" is chosen with high probability. Note that the data parameters were chosen without regard to any set of identifying restrictions.

We measure performance via the total variation between the estimated and true purchase probabilities, averaged over the first 10 sets of prices in each simulated dataset. We expect that the sMNP will be less prone to making "extreme" predictions in the sense of Figure 7. The results are summarized in Figure 8, and are consistent with this notion. The plot gives the average total variation from the true purchase probabilities for each of the base category MNP models (hollow circles) and the sMNP (solid circles). Note that the sMNP is never the worst among the various base category models. In nine of the 50 simulated cases, sMNP outperformed all of the base category models. In 43 out of 50 of the simulations, the sMNP performed better than the median base category performance.

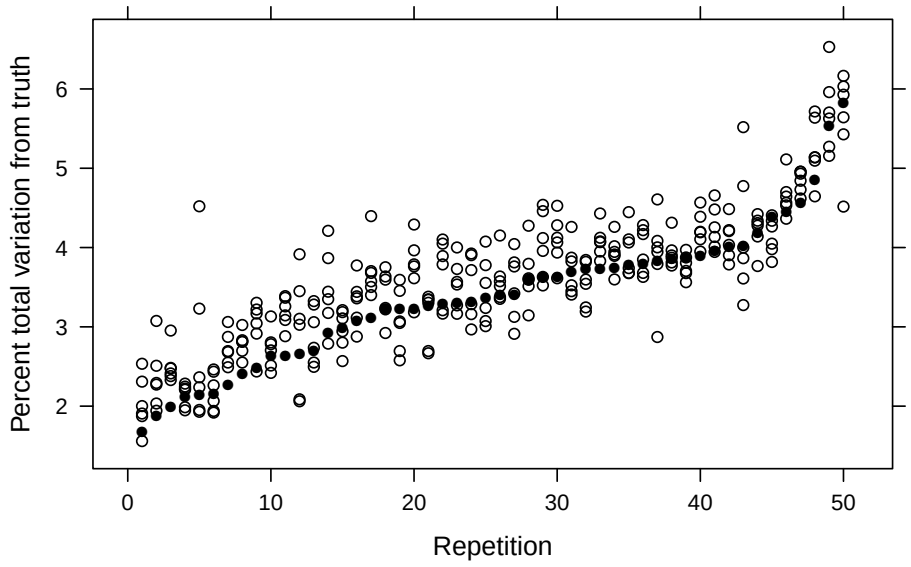

Figure 8: Simulation results. Points give the average percent total variation between true and estimated purchase probabilities. Solid circles are from the sMNP. Hollow circles are from MNP models that use each of the six possible base category identifying restrictions. The sMNP is never worse than every base category model, and in 43 out of 50 cases, it beats the median performance.

4 Identification

A potential downside to our model is that it is not formally identified. In particular, the model would be identified if we were able to restrict the trace of Σ, rather than the trace of Σ_{-b}. If one of the diagonal elements of Σ is estimated to be substantially larger than 1, then the scale of β will depend on b. Although a fully identified model may be preferable, we argue that little is lost in this case.

First — from the perspective of prior specification/elicitiation — the model is identified conditional on the discrete parameter b. If the analyst wishes to specify an informative prior, this can be done conditionally for each $b = 1, \ldots, p$. If the model were only identified conditional on a continuous working parameter, this process becomes more difficult. Second — on the side of interpretation — we would argue that β parameters should be interpreted while taking Σ into account, and vice versa. Since marginal summaries do not do this, we feel that the best model summaries are changes of fitted probabilities as a function of key outcome variables such as in Figure 7, which are not impacted by this identification issue. If the analyst truly is interested in features of the marginal posterior distribution of β or Σ, it is possible to post-process the results into a single, identified scale by re-scaling the sampled values at each iteration of the MCMC such that, for example, the trace of Σ is equal to p. However, the signs of the estimated β parameters are not impacted by the under-identification of our model.

Post-processing in order to identify Bayesian MNP models was popularized by McCulloch et al. (2000), in the context of specifying a prior for $\tilde{\Sigma}^*$, rather than the identified Σ^*. As an aside, we note that a related idea for solving the base category problem would be to specify a full-rank inverse-Wishart prior for Σ, without worrying about the conditional identifying restriction on the location of the W_i. However, this approach proves to be numerically unusable. The p-dimensional inverse-Wishart prior pushes the sampled values of Σ toward the edge of the parameter space, which quickly results in numerical problems that result from sampling poorly-conditioned covariance matrices.

5 Conclusion

The analyses in this paper demonstrate that careful handling of the prior is necessary in order to obtain reliable predictions from the Bayesian MNP. As with any proper Bayesian model, our estimates are biased, but they are not biased *against* any particular outcome category in the prior. The same can not be said of previous MNP models that estimate the covariance of the utilities.

With the prior for the regression coefficients centered on zero, the sMNP estimates should be pulled toward more moderate estimates. Since multinomial data are quite coarse (in the sense that each observation contributes little information compared to a multivariate normal regression where the utilities are observed) we would argue that this prior-induced regularization toward moderate predictions is highly desirable.

When building more advanced MNP models, symmetry may take on even greater importance. For example, Cripps et al. (2010) proposed an MNP model that allows for a sparse representation of the precision matrix of the latent utilities. However, they induce sparsity in the precision of the base-subtracted utilities, not in the precision of the original utilities. This seems very likely to exacerbate the problem of posterior estimates changing across different specifications of the base category. Further, it is unclear that sparsity in the base-subtracted precision corresponds to a meaningful data-generating process. That said, it is likely that favorable bias/variance tradeoffs can be made by specifying a prior that pulls the precision toward a well-chosen, sparse structure.

More broadly, the regularizing effect of a Bayesian prior distribution is at its most powerful when the likelihood is poorly behaved in some way: when it is flat or spiky; when identification is weak; when the number of parameters is large relative to the sample size. However, in each of these situations, we should be worried that if our prior has undesirable features, they may be preserved in the posterior. For example, MNP likelihoods can be quite flat, and therefore the asymmetry of previously-proposed priors can propagate to the posterior. Data analysts may hope that such undesirable features of the prior would be

overwhelmed by the likelihood. This research suggests that while we cannot always count on the data to cover flaws of our priors, we may be able to design priors that lack the flaw in the first place, without giving up computational tractability.

Appendix: Proof

Conditional distribution of $(b, \tilde{\Sigma}_{-b})$

Computationally, the major change from the algorithm of Burgette and Nordheim (2012) is the draw from $(b, \tilde{\Sigma}_b)$. From the full conditional, we have

$$\begin{aligned} p(b, \tilde{\Sigma}_{-b}|\text{all}) &\propto \exp\{-.5 \sum (\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})^\top \tilde{\Sigma}_{-b}^{-1} (\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})\} \\ &\quad \times |\tilde{\Sigma}_{-b}|^{-n/2} p(\tilde{\Sigma}_{-b}|b) p(b) \\ &\propto \exp\{-.5 \operatorname{tr}(\tilde{\Sigma}^{-1}(S_{-b} + \sum (\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})(\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})^\top))\} \\ &\quad \times |\tilde{\Sigma}_{-b}|^{-.5(n+\nu_0+p)} \end{aligned}$$

Then

$$\begin{aligned} p(b|\tilde{\beta}, \tilde{W}) &\propto \int p(b, \tilde{\Sigma}_{-b}|\text{all}) d\tilde{\Sigma}_{-b} \\ &\propto |S_{-b} + \sum (\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})(\tilde{W}_{i,-b} - X_{i,-b}\tilde{\beta}_{-b})^\top|^{-(n+\nu_b)/2} \end{aligned}$$

from the inverse-Wishart density. Conditional on b, $\tilde{\Sigma}_{-b}$ can be sampled from an inverse-Wishart distribution as described by Imai and van Dyk (2005*a*).

References

Albert, J., and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88(422), 669–679.

Burgette, L., and Nordheim, E. (2012), "The trace restriction: An alternative identification

strategy for the Bayesian multinomial probit model," *Journal of Business and Economic Statistics*, 30(3), 404–410.

Cripps, E., Fiebig, D., and Kohn, R. (2010), "Parsimonious estimation of the covariance matrix in multinomial probit models," *Econometric Reviews*, 29(2), 146–157.

Garrido, R., and Mahmassani, H. (2000), "Forecasting freight transportation demand with the space-time multinomial probit model," *Transportation Research Part B: Methodological*, 34(5), 403–418.

Hoff, P. (2009), "Simulation of the Matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, 18(2), 438–456.

Imai, K., and van Dyk, D. (2005*a*), "A Bayesian analysis of the multinomial probit model using marginal data augmentation," *Journal of Econometrics*, 124(2), 311–334.

Imai, K., and van Dyk, D. (2005*b*), "MNP: R package for fitting the multinomial probit model," *Journal of Statistical Software*, 14(3), 1–32.

McCulloch, R., Polson, N., and Rossi, P. (2000), "A Bayesian analysis of the multinomial probit model with fully identified parameters," *Journal of Econometrics*, 99(1), 173–193.

McCulloch, R., and Rossi, P. (1994), "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64(1), 207–240.

McFadden, D. (1974), "The measurement of urban travel demand," *Journal of Public Economics*, 3(4), 303–328.

Rossi, P., Allenby, G., and McCulloch, R. (2005), *Bayesian Statistics and Marketing*, Chichester, West Sussex, England: Wiley.

Rudolph, T. (2003), "Who's responsible for the economy? The formation and consequences of responsibility attributions," *American Journal of Political Science*, 47(4), 698–713.

Tanner, M., and Wong, W. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82(398), 528–540.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press.

van Dyk, D. A. (2010), "Marginal MCMC methods," *Statistica Sinica*, Preprint No. SS-08-153, 1–33.