Scalable Bayesian Estimation in the Multinomial Probit Model

Rubén Loaiza-Maya and Didier Nibbering

Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia

ABSTRACT

The multinomial probit (MNP) model is a popular tool for analyzing choice behavior as it allows for correlation between choice alternatives. Because current model specifications employ a full covariance matrix of the latent utilities for the choice alternatives, they are not scalable to a large number of choice alternatives. This article proposes a factor structure on the covariance matrix, which makes the model scalable to large choice sets. The main challenge in estimating this structure is that the model parameters require identifying restrictions. We identify the parameters by a trace-restriction on the covariance matrix, which is imposed through a reparameterization of the factor structure. We specify interpretable prior distributions on the model parameters and develop an MCMC sampler for parameter estimation. The proposed approach significantly improves performance in large choice sets relative to existing MNP specifications. Applications to purchase data show the economic importance of including a large number of choice alternatives in consumer choice analysis.

ARTICLE HISTORY

Received July 2020
Accepted July 2021

KEYWORDS

Factor analysis; Multinomial probit model; Parameter identification; Spherical coordinates

1. Introduction

The multinomial probit (MNP) is an important model for analyzing choice behavior, because it allows the latent utilities of the choice alternatives to be correlated. These correlations capture general substitution patterns among choice alternatives, in contrast to the case with models that impose the independence of irrelevant alternatives property (Hausman and McFadden 1984).

However, current specifications of the MNP model are not scalable to discrete choice problems with a large number of choice alternatives, as the number of parameters in the covariance matrix of the latent utilities grows quadratically in the number of choice alternatives (Burgette and Reiter 2013). This curse of dimensionality is exacerbated by the fact that, contrary to standard covariance matrix estimation settings, where multiple continuous variables are observed, all parameters in the covariance matrix have to be estimated from a single categorical variable.

Standard dimension reduction techniques, such as factor analysis, cannot straightforwardly be applied to the covariance matrix of the latent utilities. Since the scale of the latent utilities is not identified (Bunch 1991), the identification of the model parameters requires a restriction on the covariance matrix. The main challenge is to reduce the number of parameters that characterize the covariance matrix, while imposing an identifying restriction.

This article proposes an MNP model specification that is scalable to modern choice data with many choice alternatives. Specifically, we employ a factor structure on the covariance matrix, where the number of parameters scales linearly, rather than quadratically, with the number of choice alternatives. The parameters are identified by imposing a trace restriction on the covariance matrix. To impose the trace restriction on the factor representation, we transform the covariance parameters to a spherical coordinate system of angles and a spherical radius. The radius is, by construction, equal to the square root of the trace of the covariance matrix. Therefore, the trace restriction is readily imposed by setting the squared radius equal to the number of choice alternatives.

To conduct Bayesian estimation, prior densities on the angles in the reparameterization must be selected. We elicit the priors on the angles from well-understood prior assumptions popularly used in the Bayesian factor analysis literature. The process of elicitation can be performed using a fast algorithm provided in this article. The computation of the posterior distribution involves a Markov chain Monte Carlo (MCMC) sampler with Gibbs sampling steps for the coefficients and latent utilities, and a Metropolis–Hastings step for the angle parameters. A numerical experiment confirms that the MCMC sampler succeeds in accurately estimating the model parameters, in similar computation time as existing MNP specifications.

An application to real consumer choice data illustrates the empirical relevance of the scalable MNP model. We construct consumer choice data with 50 alternatives as a modern counterpart of commonly used laundry detergent and margarine purchase datasets with only six alternatives. In these large choice sets, our approach produces better predictive performance than existing methods. The model is able to identify correlations across a large set of products. The results show that limiting the analysis to only a few products may, for instance, severely bias price elasticity estimates. The proposed model has similar performance to existing MNP specifications when applied to the...
traditional laundry detergent and margarine choice data with six alternatives. This article makes three important contributions to the MNP literature. First, the proposed approach addresses the scalability of the MNP model directly. Piatek and Gensowski (2017) specified a factor structure for the covariance matrix under the assumption that the factor loadings are known. This assumption might be unrealistic in many choice problems, especially when the number of alternatives is large. Cripps, Fiebig, and Kohn (2009) proposed a covariance selection prior that permits elements of the inverse of the covariance matrix to be zero. This approach allows for a sparse representation of the model, but does not reduce the number of parameters to be estimated.

Second, this article contributes to the literature on parameter identification in MNP models. Burgette and Nordheim (2012) show that fixing the trace of the covariance matrix should be preferred over fixing a diagonal element, as in McCulloch, Polson, and Rossi (2000) and Imai and Van Dyk (2005a). Our model reparameterization satisfies the trace restriction and also naturally imposes parsimony, which is not embedded in the marginal data augmentation approach used by Burgette and Nordheim (2012).

Third, this is the first study that applies the MNP model to real choice data with a large number of choice alternatives. Empirical applications of MNP models have been limited to only a few choice alternatives. For instance, Imai and Van Dyk (2005a) considered six clothing detergent brands, McCulloch and Rossi (1994) and Burgette and Nordheim (2012) six margarine brands, Piatek and Gensowski (2017) two education levels and three occupation categories, and Cripps, Fiebig, and Kohn (2009) five tests for cervical cancer. This limitation is of particular concern today, with the widespread availability of data on large choice sets.

The outline of the remainder of this article is as follows. Section 2 discusses the model specification and Section 3 introduces a scalable covariance matrix specification. Section 4 discusses prior specifications and the MCMC sampler. Section 5 conducts a numerical experiment to evaluate estimation accuracy, and Section 6 applies the proposed methods to real consumer choice datasets. Section 7 concludes.

### 2. MNP Model

#### 2.1. Model Specification

Let \( Y_i \) be an observable unordered random categorical variable with support on the set \( A_i = \{0, 1, 2, \ldots, J\} \), with \( J + 1 \) the number of choice alternatives, and \( i = 1, \ldots, N \), with \( N \) the number of individuals. Let \( \tilde{Z}_i = (\tilde{z}_{i0}, \ldots, \tilde{z}_{iJ})^\top \) be a \((J + 1) \times 1\) vector of continuous random variables that can be interpreted as latent utilities, with

\[
Y_i(\tilde{Z}_i) = \arg \max_{j \in A_i} \tilde{z}_{ij}.
\]

(1)

The latent utilities are modeled as

\[
\tilde{Z}_i = \tilde{X}_i \tilde{\beta} + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \tilde{\Sigma}),
\]

(2)

where \( \tilde{X}_i \) is a \((J + 1) \times K\) matrix of observed regressors, \( \tilde{\beta} \) is a vector of coefficients, and \( \tilde{\epsilon}_i \) is an independent normally distributed disturbance vector with covariance matrix \( \tilde{\Sigma} \). The regressor matrix typically includes an intercept, a \( k_d \)-dimensional vector \( x_{i,d} \) of individual-specific characteristics, and a \((J + 1) \times k_d\) matrix \( x_{i,d} \) of \( k_d \) alternative-specific covariates, such that

\[
\tilde{X}_i = [I_{J+1} \quad x_{i,d}^\top \otimes I_{J+1} \quad x_{i,d}],
\]

(3)

where \( I_{J+1} \) denotes the identity matrix of dimensions \((J + 1) \times (J + 1)\).

#### 2.2. Identification

The parameters \( \tilde{\beta} \) and \( \tilde{\Sigma} \) in the MNP model specified in Equation (2) are not identified (Bunch 1991). There are two parameter identification problems. First, the location of the latent utilities is unidentified, since \( Y_i(\tilde{Z}_i) = Y_i(\tilde{Z}_i + c) \) for \( c \in \mathbb{R} \). Second, the scale of the latent utilities is also unidentified, as \( Y_i(c \tilde{Z}_i) = Y_i(\tilde{Z}_i) \) for \( c \in \mathbb{R}^+ \).

##### 2.2.1. Location

A standard solution to the identification problem in the location is to difference the utilities with respect to a baseline category. Define choice category \( j = 0 \) as the base category and define the differences in utilities as \( Z_{ij} = \tilde{z}_{ij} - \tilde{z}_{i0} \). The dependent variable \( Y_i \) equals

\[
Y_i = \begin{cases} 
0 & \text{if max}(Z_{ij}) < 0, \\
\ j & \text{if max}(Z_{ij}) > 0,
\end{cases}
\]

(4)

where max\((Z_{ij})\) is the largest element of \( Z_i \).

The utility model in Equation (2) is transformed to differences in utilities by

\[
Z_i = T\tilde{Z}_i = X_i \tilde{\beta} + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma),
\]

(5)

with transformation matrix \( T = [-I_j \quad I_j] \), the \( J \times K \) transformed regression matrix \( X_i = [I_j \quad x_{i,d}^\top \otimes I_j \quad x_{i,d}] \), and the \( J \times J \) transformed covariance matrix \( \Sigma = T\tilde{\Sigma}T^\top \). For the remainder of this article, we employ this location identification approach, and refer to \( \tilde{Z}_i \) as utilities. Moreover, we define \( Y = (Y_1, \ldots, Y_N)^\top, Z = (Z_1^0, \ldots, Z_N^0)^\top, \) and \( X = (X_1^0, \ldots, X_N^0)^\top \).

##### 2.2.2. Scale

There are multiple solutions to the unidentified scale in the latent utilities, but they all impose a constraint on the covariance matrix \( \Sigma \). McCulloch, Polson, and Rossi (2000) developed a prior that fixes the \((1,1)\) element of \( \Sigma \) to be equal to one. Burgette and Nordheim (2012) argued that the assignment of a choice alternative to the unit variance can have a large effect on the posterior choice probabilities. They propose to fix the trace of the covariance matrix instead of fixing one of its elements. We follow this approach and restrict the trace of the covariance matrix \( \Sigma \) in Equation (5) to be equal to \( J \).

### 3. Scalable Reparameterization

The total number of unique parameters in the covariance matrix, \( J(J + 1)/2 \), grows quadratically with \( J \). To reduce the dimension of the parameter space, Section 3.1 specifies a factor structure for \( \Sigma \). Section 3.2 introduces a reparameterization approach, and refer to Burgette and Nordheim (2012) for additional details.
of the factor structure that imposes a trace restriction on the covariance matrix.

### 3.1. Factor Structure

Denote as $\gamma$ a $J \times q$ matrix with $q < J$, and as $D$ a diagonal matrix with positive diagonal elements $d = (d_1, \ldots, d_J)$. We model $\Sigma$ as

$$
\Sigma = \gamma \gamma^\top + D^2. \tag{6}
$$

For the purpose of identification, the upper triangular elements of $\gamma$ are fixed at zero (Geweke and Zhou 1996). In this factor covariance structure, the total number of parameters that characterize the covariance matrix is $n = J(q+1) - q(q-1)/2$, which implies that for a given value of $q$, the number of parameters grows linearly with $J$.

### 3.2. Trace Restriction

A major challenge in the factor covariance structure is the implementation of the identifying restriction trace($\Sigma$) = $J$, which implies

$$
\text{trace}(\Sigma) = \sum_{j=1}^{l} \left( \min(q,j) \right) \left( \sum_{k=1}^{\min(q,j)} \gamma_{jk}^2 + d_j^2 \right) = J, \tag{7}
$$

where the scalar $\gamma_{jk}$ denotes the element in row $j$ and column $k$ in $\gamma$, while the scalar $d_j$ is the $j$th element in $d$.

The trace restriction implies that by construction, the elements of $\gamma$ and $d$ need to be constrained to the surface of an $n$–sphere of radius $\sqrt{J}$. To show this, define the $n$-dimensional vector $\psi$ as

$$
\psi = (\psi_1, \ldots, \psi_n)^\top = (d^\top, \text{vech}(\gamma)^\top)^\top, \tag{8}
$$

where $\text{vech}(\gamma) = (\gamma_{11}, \ldots, \gamma_{q1}, \gamma_{q2})^\top$ and $\gamma_{kl,jk} = (\gamma_{kj}, \ldots, \gamma_{jk})^\top$. The trace restriction in Equation (7) is equivalent to the spherical restriction $\sum_{l=1}^{n} \psi_l^2 = J$ on $\psi$, from which follows that the elements of $\psi$ are restricted to the $n$-dimensional spherical space $S^n = \{ \psi : \sum_{l=1}^{n} \psi_l^2 = J \}$.

As one must guarantee the spherical restriction, which involves all the elements in $\psi$, estimation of $\psi$ is challenging. To avoid direct implementation of this restriction, we exploit the fact that $\psi$ is restricted to $S^n$, which naturally allows for a spherical transformation from $\psi$ into an $(n - 1)$-dimensional vector of angles $\kappa = (\kappa_1, \ldots, \kappa_{n-1})^\top$ and the radius $\sqrt{J}$, with $\kappa \in \Lambda$ and $\Lambda = [0, \pi)^{n-2} \times [0, 2\pi)$.

The spherical transformation reparameterizes $\psi_1$ as

$$
\psi_1(\kappa, J) = \begin{cases} 
\sqrt{J} \cos \kappa_l & \text{for } l = 1, \\
\sqrt{J} \cos \kappa_l \prod_{j=l+1}^{n-1} \sin \kappa_j & \text{for } 1 < l < n, \\
\sqrt{J} \prod_{j=l}^{n-1} \sin \kappa_j & \text{for } l = n.
\end{cases} \tag{9}
$$

The main advantage of this transformation is that inference on the parameter space $\Lambda$ is a more accessible problem, as no spherical restriction is required for $\kappa$. Moreover, the inverse function of the transformation is available in closed-form,

$$
k_l(\psi) = \begin{cases} 
\arccos \left( \frac{\psi_l}{\sqrt{\sum_{j=1}^{n} \psi_j^2}} \right)^{\frac{1}{2}} & \text{for } l < n - 1, \\
\arccos \left( \frac{\psi_n}{\sqrt{\sum_{j=1}^{n} \psi_j^2}} \right)^{\frac{1}{2}} & \text{for } l = n - 1 \land \psi_n \geq 0, \\
2\pi - \arccos \left( \frac{\psi_n}{\sqrt{\sum_{j=1}^{n} \psi_j^2}} \right)^{\frac{1}{2}} & \text{for } l = n - 1 \land \psi_n < 0.\tag{10}
\end{cases}
$$

Notice from Equation (10) that $k_l(c|\psi) = k_l(\psi)$, for any positive scalar $c$. Setting $c = \sum_{l=1}^{n} \psi_l^2 = \text{trace}(\Sigma)$, shows that $k_l$ is a function of the scale of $\psi_l$ relative to the other elements in $\psi$, rather than the trace of $\Sigma$.

### 4. Bayesian Estimation

This section develops a Bayesian method for estimating the identified parameters $(\beta, \kappa)$, subject to the trace restriction in Equation (7). The density of interest is the augmented posterior

$$
p(\beta, \kappa | Z | Y, X, B, \theta) \propto p(Y | Z)p(Z | X, \beta, \kappa)p(\beta | B)p(\kappa | \theta), \tag{11}
$$

where we assume the priors of $\beta$ and $\kappa$ to be independent, with corresponding hyperparameters $B$ and $\theta$. The prior on the coefficients $\beta$ is standard and specified as

$$
\beta | B \sim N(0, B^{-1}). \tag{12}
$$

We develop the prior choice for $p(\kappa | \theta)$ in the next section.

#### 4.1. Prior for Angular Coordinates

Developing prior beliefs directly on the angular coordinates $\kappa$ is challenging, because these coordinates lack interpretation in relation to the covariance matrix $\Sigma$. On the other hand, because $\psi$ directly determines the covariance matrix $\Sigma$ via (6) and (8), the implications that a prior on $\psi$ has on the prior beliefs on $\Sigma$ are well understood. At the same time, choosing a prior on $\psi$ imposes the prior on $\kappa$ required in Equation (11).

Therefore, we first select a prior $p(\psi | \theta)$. This prior implies $p(\kappa | \theta)$, which is required for parameter estimation, and $p(\Sigma | \theta)$. Because the elements of $\kappa$ and $\psi$ do not have a monotonically increasing relationship, analytical derivation of $p(\kappa | \theta)$ is challenging. Instead, we construct a parametric approximation to $p(\kappa | \theta)$ from a flexible parametric density class $\mathcal{P}$ with elements $\hat{p}(\kappa | \lambda)$ indexed by $\lambda \in \Lambda$.

The approximating prior density $\hat{p}(\kappa | \lambda)$ is calibrated by minimizing an estimate of the Kullback–Leibler divergence

$$
\text{KL} \left[ p(\kappa | \theta) || \hat{p}(\kappa | \lambda) \right], \tag{13}
$$

with respect to $\lambda$. Specifically, we minimize

$$
\frac{1}{M} \sum_{m=1}^{M} \log \left( p(\kappa^{[m]} | \theta) \right) - \frac{1}{M} \sum_{m=1}^{M} \log \left( \hat{p}(\kappa^{[m]} | \lambda) \right),
$$

where the first term can be ignored in the minimization problem, and the second term is computed using $M$ draws $\{\kappa^{[m]}\}_{m=1}^{M}$. These draws are produced by generating from the prior distribution $p(\psi | \theta)$, and then transforming into draws from the
Algorithm 1 Prior calibration on the angular coordinates

1: Set hyperparameters $\theta$ in $p(\psi|\theta)$
2: Generate $M$ draws $\psi_{[m]} \sim p(\psi|\theta)$
3: Transform $\psi_{[m]}$ to $\kappa_{[m]}$ using (10)
4: Calculate $\hat{\lambda}$ as the minimizer of (13) using $\{\kappa_{[m]}\}_{m=1}^{M}$
5: Construct $\tilde{p}(\kappa|\hat{\lambda})$

prior $p(\kappa|\theta)$. Algorithm 1 outlines the steps of the optimization process.

Once calibrated, we can use the prior $\tilde{p}(\kappa|\lambda)$ for inference, by plugging it into (11) instead of $p(\kappa|\theta)$. Key to accurate approximation of $p(\kappa|\theta)$ is the selection of a flexible parametric density class $\mathcal{P}$. Here, we use $\tilde{p}(\kappa|\lambda) = \prod_{i=1}^{n-1} \tilde{p}(\kappa(i) | \lambda(i))$ with

$$
\tilde{p}(\kappa(i) | \lambda(i)) = \phi_i \left\{ t_{n_i} \left[ \frac{G(\kappa(i)) - \mu_i}{\tau_i} \right] \right\} t_{n_i} \left[ \frac{G(\kappa(i)) - \mu_i}{\tau_i} \right] \frac{1}{\tau_i} G' (\kappa(i)),
$$

where $G(\kappa(i)) = \Phi_{1}^{-1} \left( \frac{\kappa(i)}{\lambda(i)} \right)$ for $l < n - 1$, $G(\kappa(i)) = \Phi_{1}^{-1} \left( \frac{\kappa(i)}{\lambda(i)} \right)$ for $l = n - 1$, $G'(\cdot)$ is the derivative of $G(\cdot)$, $t_{n_i}$ is the Yeo and Johnson (2000) transformation, $t_{n_i}'(\cdot)$ its first derivative, while $\phi_i()$ and $\Phi_{1}^{-1}()$ denote the density and inverse distribution function of a standard normal variable, respectively.

The density function in Equation (14) is capable of accurately approximating the marginals of the prior $p(\kappa|\theta)$, which in turn results in accurate approximation to the implied prior $p(\Sigma|\theta)$, as we will show in Section 4.2. Moreover, the evaluation of the density is computationally efficient, which increases the speed of the sampling algorithm discussed in Section 4.3. Therefore, we select this density class over generally more computationally involved nonparametric alternatives. For details on the properties and construction of this distribution, we refer to Part A of the online appendix.

4.2. Choice of Prior

To set the prior $p(\kappa|\theta)$, the practitioner needs to specify a prior on the vector $\psi$. Although simulation from $p(\psi|\theta)$ is required for Algorithm 1, this prior density does not have to be available in closed-form.

We propose the following characterization of the prior distribution $p(\psi|\theta)$:

$$
\psi = \frac{\sqrt{J}}{||\psi||} \tilde{\psi},
$$

$$
p(\tilde{\psi}|\theta) = \prod_{j=1}^{J} p(\tilde{\psi}_{j} | \nu, s) \prod_{k=1}^{\min(q) \cdot d_{j}} p(\tilde{\psi}_{j|k} | \sigma_{\gamma}^{2})
$$

$$
\tilde{\psi}_{j|k} | \sigma_{\gamma}^{2} \sim N(\mu_{\gamma}, \sigma_{\gamma}^{2}),
$$

$$
\tilde{\psi}_{j} | \nu, s \sim \text{Inverse-Gamma} (\nu, s),
$$

where $\nu$ and $s$ denote the shape and rate parameters of the Inverse-Gamma distribution.

This choice of prior links the trace-restricted parameters in $\psi$ to an unrestricted parameter vector $\tilde{\psi} = (d^T, \text{vech}(\tilde{\psi}^T)^T)^T$. As a result, we induce a prior on the restricted parameters $\psi$ by selecting a prior for the unrestricted parameters $\tilde{\psi}$. For $\tilde{\psi}$ any prior that fits the type of factor structure in Equation (6) may be employed. We use a particularly well-established prior in the factor literature, that assumes a normal distribution on $\tilde{\psi}$ and an Inverse-Gamma distribution for $d_{j}^{2}$ (see for instance Lopes (2014) and references therein).

Since $\psi$ is a function of the scale of the elements of $\tilde{\psi}$ relative to the norm of $\tilde{\psi}$, the location of $p(\psi|\theta)$ is unidentified. To solve this, we anchor the mean of the Inverse-Gamma prior in Equation (18) at one by setting $s = \nu - 1$. Thus, the hyperparameters for the prior are $\theta = (\mu_{\gamma}, \sigma_{\gamma}, \nu)^T$.

4.2.1. Accuracy of the Approximating Prior

This section assesses the accuracy of the approximating density for the choice of prior in the previous section. We focus on the parameter space of most interest to the practitioner, the covariance matrix $\Sigma$, which is implied by $p(\kappa|\theta)$. Specifically, we focus on assessing the accuracy of $\tilde{p}(\Sigma_{2,2}|\tilde{\lambda})$ at replicating $p(\Sigma_{2,2}|\theta)$, and the accuracy of $\tilde{p}(\rho_{2,3}|\tilde{\lambda})$ at replicating $p(\rho_{2,3}|\theta)$ with $\rho_{2,3} = \frac{\Sigma_{2,2}}{\sqrt{\Sigma_{2,2} \Sigma_{3,3}}}$, for $J = 6$, $q = 1$, and $\theta = (0, 1, 5)^T$. All the results in this section also hold for the other elements in $\Sigma$ and for any number of choice alternatives $J$.

The prior $p(\Sigma_{2,2}|\theta)$ is constructed via simulation. First, we generate draws from the prior $p(\kappa|\theta)$, then transform them into draws for $\Sigma$, and finally construct a kernel density estimator of $p(\Sigma_{2,2}|\theta)$. Similarly, we use Algorithm 1 to calibrate the approximating prior $\tilde{p}(\kappa|\tilde{\lambda})$, from which we also obtain a kernel density estimator of $\tilde{p}(\Sigma_{2,2}|\tilde{\lambda})$ via simulation. We construct $p(\rho_{2,3}|\theta)$ and $\tilde{p}(\rho_{2,3}|\tilde{\lambda})$ in the same way.

In Panel (a) of Figure 1, the yellow solid line and the black dashed line represent the implied prior densities $p(\Sigma_{2,2}|\theta)$ and $\tilde{p}(\Sigma_{2,2}|\tilde{\lambda})$, respectively. The approximating prior is an accurate representation of the prior. The remaining panels in Figure 1 demonstrate that the approximating prior remains accurate for alternative hyperparameter values. A similar result is obtained when we compare the implied priors $p(\rho_{2,3}|\theta)$ and $\tilde{p}(\rho_{2,3}|\tilde{\lambda})$, as shown in Figure 2.

Additionally, Figure 3 shows that also for a larger number of factors $q = 4$, the approximating prior is accurate.

4.2.2. Hyperparameters

The user of the prior proposed in Equations (15)–(18) only has to set the hyperparameters $\theta$. Here we discuss the impact that these hyperparameters have on the implied prior for $\Sigma$. The parameters $\sigma_{\gamma}$ and $\nu$ jointly control the dispersion of the variances around their prior mean of one, and $\sigma_{\gamma}$ also governs the variance of the correlations around their prior mean.

To illustrate this, Panels (b) and (d) in Figure 1 show that, for the small value of $\sigma_{\gamma} = 0.1$, a larger value of $\nu$ makes the prior on the diagonal elements of $\Sigma$ tighter around one. In Panels (a) and (c), we observe that for large values of $\sigma_{\gamma}$, the hyperparameter $\nu$ has little effect on the prior. On the other hand, comparing Panel (a) to (b), and Panel (c) to (d), indicates that for smaller values of $\sigma_{\gamma}$, the impact of $\nu$ on the prior is more pronounced.

Panels (a) and (b) in Figure 2 show that $\sigma_{\gamma}$ governs the prior variance on the underlying correlations of $\Sigma$, with larger values for $\sigma_{\gamma}$ associated with a prior with larger variance. The parameter $\nu$ does not affect the implied prior on the correlations.
Figure 1. Approximating prior density for $\Sigma_{2,2}$.
NOTES: This figure shows the implied prior variance densities for the MNP model with a factor structure. The yellow solid line corresponds to the implied prior $p(\Sigma_{2,2}|\theta)$ and the black dashed line corresponds to its approximation $\tilde{p}(\Sigma_{2,2}|\hat{\lambda})$, and the panels correspond to different values for $\theta$ with $q = 1$.

Figure 2. Approximating prior density for $\rho_{2,3}$.
NOTES: This figure shows the implied prior correlation densities for the MNP model with a factor structure. The yellow solid line corresponds to the implied prior $p(\rho_{2,3}|\theta)$ and the black dashed line corresponds to its approximation $\tilde{p}(\rho_{2,3}|\hat{\lambda})$, and the panels correspond to different values for $\theta$ with $q = 1$.

So a small value for $\sigma_{\gamma}$ shrinks the correlations toward their prior means. A comparison of Panels (a) and (b) to Panels (c) and (d) indicates that the prior mean of $\Sigma$ equals $I_j$ for $\mu_{\gamma} = 0$, because the correlations have a prior mean of zero, and it equals an equicorrelated matrix when $\mu_{\gamma} \neq 0$.

The implied prior for $\Sigma$ is also sensitive to the number of factors considered. Figure 3 shows how these implied prior densities change when the number of factors is set to $q = 4$. Increasing the number of factors has a small effect in the priors for both the variance and correlation parameters.
For the remainder of this article, we consider two choices for \( \mu_\gamma \). The first choice, \( \mu_\gamma = 0 \), sets the prior mean of the covariance matrix \( \Sigma \) equal to the identity matrix, which is a standard MNP prior choice as we will discuss in the next section. However, \( \Sigma \) is the covariance matrix of the differences in utilities \( Z \), for which an identity covariance matrix does not necessarily imply a symmetric correlation structure for the untransformed utilities \( \tilde{Z} \). Therefore, the second choice, \( \mu_\gamma = \mu_\gamma^* \), sets the prior mean of \( \Sigma \) equal to the equicorrelated covariance matrix \( \frac{1}{2}(I_J + \mu_\gamma I_J^T) \). This follows Geweke, Keane, and Runkle (1994), who shrink the covariance matrix of the untransformed utilities \( \Sigma \) to an identity matrix, which is equivalent to shrinking \( \Sigma \) to this equicorrelated matrix. The value \( \mu_\gamma^* \) that produces the equicorrelated matrix above can be computed as in Part B of the online appendix, for any given values of \( \sigma_\gamma, \nu \) and \( q \).

For the remaining hyperparameters, we use the values \( \sigma_\gamma = 1, \nu = 5, \) and \( q = 1 \). These hyperparameter values produce prior densities for the variances that have close to zero probability mass at zero. They also imply prior densities for the correlations that have low probability mass at one and minus one. These properties make the method computationally stable, and guarantee that extreme values for the variances and correlations are the result of a strong signal in the data instead of highly uninformative priors.

### 4.2.3. Comparison to Existing MNP Prior Specifications

This section compares our prior assumptions on \( \Sigma \) to two well-known alternatives in the MNP literature. The first, from here on in referred to as MNP-MPR, is proposed by McCulloch, Polson, and Rossi (2000), who set

\[
\Sigma = \begin{bmatrix}
1 & \gamma^T \\
\gamma & \Phi + \gamma_\gamma^T
\end{bmatrix},
\gamma \sim N(0, \tau I_{J-1}), \quad \Phi^{-1} \sim \text{Wishart}(\delta, C),
\]

with degrees of freedom \( \delta = J + 3 \) and scale matrix \( C = (\delta - J)(1 - \tau)I_{J-1} \), where \( \tau = \frac{1}{8} \). The second alternative, from here on in MNP-BN, was proposed in Burgette and Nordheim (2012), who specified

\[
\Sigma = \Sigma / \text{trace}(\Sigma), \quad \Sigma \sim \text{Inverse-Wishart}(s, S),
\]

with degrees of freedom \( s = J + 3 \) and scale matrix \( S = I_J \).

Panel (a) in Figure 4 compares the implied prior on \( \Sigma_{2,2} \) for our proposed MNP model with factor structure, from here on in referred to as MNP-FS, to MNP-BN and MNP-MPR. For MNP-FS we consider \( q = 1 \) and \( \theta = (0, 1, 5)^T \). The three densities are positively asymmetric and most of their probability mass lies between zero and four. Aside from the fact that the MNP-FS density has low mass at zero, all three prior densities have a similar shape.

Panel (b) in Figure 4 compares the implied priors on the correlation element \( \rho_{2,3} \). The MNP-BN and MNP-MPR priors assign more probability mass to the edges of the support, while the MNP-FS prior is slightly tighter and has a spike at zero. Although not visible in the figure, the MNP-FS prior still assigns enough probability mass at extreme values of the support, to
allow for accurate estimation of correlations whose true parameter values are extreme.

### 4.3. Sampling Scheme

To construct the posterior in Equation (11), we employ the following MCMC sampling scheme:

**Sampling Scheme**

Step 1: Generate from $\beta | Z, \kappa, X, B$.

Step 2: Generate from $Z | \beta, \kappa, Y, X$.

Step 3: Generate from $\kappa | Z, \beta, X, \theta$.

Steps 1 and 2 are standard Gibbs sampling steps, see for instance, McCulloch and Rossi (1994). These steps require one to transform from $\kappa$ to $\Sigma$, which can be easily achieved following the instructions outlined in Table 1.

For step 3 we employ a random walk Metropolis–Hastings sampler. At the beginning of each iteration, the elements of $\kappa$ are randomly assigned to groups of five elements. The groups are then sampled, one group conditional on the other, with the 5-dimensional proposal density equal to the product of five independent truncated univariate normals. The variances of the proposal densities are set adaptively to target acceptance rates between 15% and 30%.

The random assignment into groups, plus the parameter-specific adaptive steps, allow one to target parameter-specific acceptance rates without having to sample each element of $\kappa$ one at a time. Roberts and Rosenthal (2009) provided more details on adaptive MCMC, and Smith (2015) provided an illustration on random allocation within MCMC. Part C of the online appendix discusses the sampling steps in more detail.

Although the reparameterization of $\Sigma$ into $\kappa$ results in a non-conjugate sampling step, the added computational cost is small. This is because Step 2—also required for the competing MNP estimation approaches—is the most time consuming of the sampling scheme, especially for large $J$. Section 6 illustrates this in the simulation experiment.

Potential alternatives to our reparameterization that produce a conjugate Gibbs sampling scheme may complicate the inference of the factor structure. For instance, the factor structure in Equation (6) could potentially be sampled by augmenting the parameter space with a set of latent factors as in Lopes (2014), in combination with the marginal data augmentation approach in Burgette and Nordheim (2012) to impose the trace restriction. However, as discussed in Piatek and Gensowski (2017), measurement of latent factors in an MNP model without extra data or a priori knowledge of the factor loadings $\gamma$ is challenging.

### 5. Numerical Experiment

This section presents a numerical experiment to assess the accuracy of the parameter estimates in the proposed MNP model. First we describe the simulation design, second we discuss the results, and finally we examine how sensitive the fitted probabilities are to different base category specifications.

**5.1. Design**

We generate a dataset from the data-generating process specified in Equations (4) and (5). The dataset closely matches the empirical application in Section 6.2, with the number of discrete choices $J + 1 = 50$ and number of observations $N = 5000$. The elements of the vector $x_{i,a}$ are independently generated from normal distributions with corresponding mean $\mu = 0$ and variance $\sigma^2 = 1$, and can be interpreted as the logarithm of the prices of the choice categories. We do not include individual-specific characteristics $x_{i,a}$.

The true parameter vector $\beta_0$ consists of $J$ intercepts drawn independently from normal distributions with $\mu = 0$ and $\sigma^2 = 0.5$, and the coefficient for $x_{i,a}$ which is fixed at $-0.7$. The true covariance matrix $\Sigma_0$ is set by drawing $\Sigma_0$ from the Inverse-Wishart$(S, J + 3)$ distribution, where the scale matrix $S$ has ones on the diagonal and 0.5 as the common off-diagonal element. We set $\Sigma_0 = S/\text{trace}(\Sigma_0)$.

We apply our method, MNP-FS, to the generated dataset. For the purpose of comparison we also implement the MNP-BN and MNP-MPR approaches. For our model, we set $\theta = (0, 1, 5)^T$ with $q = 1$ in the prior for $\kappa$. The implied prior densities at these parameter values are discussed in detail in Section 4.2.
Setting the number of factors to one substantially reduces the number of covariance parameters to be estimated, especially with 50 choice alternatives. Alternatively, the number of factors could be selected based on an information criterion like the BIC or a predictive performance measure like the hit rate. The model may also be extended with approaches that treat the variances and correlations of the latent utilities in the covariance matrix as an additional unknown parameter, as those proposed in Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2018).

We use the prior for the coefficients specified in Equation (12) with \( B^{-1} = 0.1I_k \) for all models. This is a rather uninformative prior, given that the covariates are scaled to a variance of one in the sampler. Since settings with large choice sets are vulnerable to numerical instabilities, we avoid the use of improper prior specifications.

The posterior results are based on 200,000 iterations of the MCMC samplers, from which the first 100,000 are discarded.

### 5.2. Results

Figure 5 compares the true parameter values of \( \beta_0 \) and \( \Sigma_0 \) (x-axis) against the corresponding posterior mean estimates (y-axis). The yellow and black circles correspond to MNP-FS and MNP-BN, respectively. The closer the circles lie to the 45° diagonal line, the closer the posterior means are to the true parameter values.

Panel (a) in Figure 5 shows that the MNP-FS approach provides more accurate estimates of \( \beta_0 \) than MNP-BN. This is also the case for the diagonal elements of \( \Sigma_0 \) in Panel (b). Finally, the results for the posterior mean estimates of the underlying correlations of \( \Sigma_0 \) are presented in panel (c) of Figure 5. The correlation estimates from MNP-FS are clearly the most tightly scattered around the diagonal line, and as such the most accurate. This result is particularly striking given the fact that the MNP-FS is the one imposing the most restrictive covariance structure. Part D of the online appendix shows that the comparison between MNP-FS and MNP-MPR, result in similar conclusions.

The error measures reported in Table 2 confirm the conclusions above. The smallest root mean squared error (RMSE) of the posterior mean estimates for the variances and correlations correspond to the MNP-FS specification. The same conclusion is reached when comparison is conducted in terms of the mean absolute error (MAE), confirming that our approach allows for more accurate estimation of the elements in the covariance matrix. For the coefficients, the RMSE and MAE values from MNP-MPR are the smallest, closely followed by MNP-FS.

### 5.3. Sensitivity Analysis Base Category Specification

Burgette, Puelz, and Hahn (2021) showed that the estimated probabilities from Bayesian MNP models can depend on the model specifications and \( \theta \) in Panel (a), and the variances and correlations of the latent utilities in \( \Sigma \) in Panel (b) and (c), respectively.

![Figure 5](image-url)

**Figure 5.** Posterior mean parameters in numerical experiment.

**NOTES:** This figure presents the estimated posterior means from the MNP-FS (yellow circles) and MNP-BN (black circles) approaches, for the coefficients \( \beta \) in Panel (a), and the variances and correlations of the latent utilities in \( \Sigma \) in Panel (b) and (c), respectively.

|            | Coefficients | Variances | Correlations |
|------------|--------------|-----------|--------------|
|            | RMSE         | MAE       | RMSE         | MAE       | RMSE         | MAE       |
| MNP-FS     | 0.095        | 0.078     | 0.231        | 0.170     | 0.221        | 0.175     |
| MNP-BN     | 0.132        | 0.111     | 0.407        | 0.308     | 0.275        | 0.212     |
| MNP-MPR    | 0.090        | 0.070     | 0.481        | 0.357     | 0.302        | 0.254     |

**NOTES:** This table reports the root-mean-squared error (RMSE) and mean absolute error (MAE) of the posterior mean estimates relative to the true parameter values. The panels for the coefficients, variances and correlations correspond to the error measures associated to the vector of coefficients \( \beta_0 \), the diagonal elements of \( \Sigma_0 \), and the implied correlations in \( \Sigma_0 \), respectively. The rows denote the alternative MNP model specifications considered.

The computational costs of the methods are similar. Since MNP-FS requires several MH steps for the generation of \( \kappa \), it takes 0.06 seconds more per sample iteration than MNP-BN and MNP-MPR, which need on average 0.36 and 0.35 sec for each sample iteration in this simulation exercise.

Burgette, Puelz, and Hahn (2021) showed that the estimated probabilities from Bayesian MNP models can depend on the base category specification. This section analyses the sensitivity of the MNP-FS results to the base category specification under the identity prior \( \theta = (0,1,5)^\top \) and the equicorrelated prior \( \theta = (1.525,1,5)^\top \) specifications and \( q = 1 \).

First, we consider the identity prior. Panel (a) in Figure 6 presents the estimated probabilities as a function of price for the least popular category, which has 16 observations. The solid black line corresponds to the true probabilities, the dashed black line is estimated using the same base category as in Section 5.2 which has 85 observations, while the yellow line corresponds to the specification that has the largest category with 440 observations as the base category. In all panels, the price of the other categories is fixed at the mean across all observations.

Panel (a) in Figure 6 shows that the yellow line and the dashed black line are different from each other. Panel (b) presents the equivalent plot for the most popular category. In this panel, there is a slightly bigger difference between the yellow and dashed black line, indicating that the estimated probabilities of...
the least popular choice are less sensitive to the base category specification than the most popular category. As discussed in Section 4.2, the identity prior does not necessarily imply an identity covariance matrix for the untransformed utilities. Since the utilities are in differences with the base category, this prior specification may result in estimates that are sensitive to the base category specification. In contrast, the equicorrelated prior follows from an identity covariance matrix for the untransformed utilities.

Panels (c) and (d) show the estimated probabilities for the least and most popular categories, respectively, when using an equicorrelated prior. Panel (c) indicates that the results for the least popular category are not as sensitive as those when using an identity prior. The estimated probabilities from the different base category specifications are now almost identical. We find the same results for the most popular category in Panel (d).

In sum, we conclude that the estimated probabilities are sensitive to the base category specification when a prior specification that shrinks \( \Sigma \) to an identity matrix is employed. However, the impact of the base category can be substantially decreased by specifying an equicorrelated prior. An alternative way to deal with the sensitivity to the base category, is to pool the estimated probabilities across all models with different base category specifications to obtain probabilities that do not depend on one base category. This approach would be computationally costly, especially when the number of choice alternatives is large.

Note that, conditional on \( q = 1 \), the estimates for \( \beta \) and \( \Sigma \) do not depend on the order of the choice alternatives in the model. With \( q > 1 \), restricting the upper triangular elements of \( \gamma \) to zero causes the estimates of \( \gamma \) to be order dependent, but has no effect on \( \Sigma \) (Lopes and West 2004). In the case where the estimates for \( \gamma \) are of particular interest, our estimation algorithm may be extended to order-independent Bayesian approaches as those proposed in Chan, Leon-Gonzalez, and Strachan (2018) and Kaufmann and Schumacher (2019).

6. Empirical Applications

This section fits the MNP model to consumer choice datasets of different dimensions. First, Section 6.1 considers a traditional consumer choice dataset on laundry detergent purchases with only six alternatives. Section 6.2 constructs a laundry detergent purchases dataset with 50 choice alternatives from a big transaction dataset. Section 6.3 considers margarine brands in both a widely used small choice set and a newly constructed large choice set.

The proposed MNP model is compared to benchmark specifications. These specifications and prior settings are discussed in Section 5.1. Moreover, we estimate an MNP model with the covariance matrix fixed to the identity matrix, referred to as MNP-I.

The in-sample and out-of-sample predictive accuracy of the models are evaluated in terms of the predictive hit-rate and the logarithmic score (log-score). The predictive probability mass function for \( Y_i \) is given by

\[
p(Y_i | X_i, Y, X, B, \theta) = \int p(Y_i | X_i, \beta, \kappa) p(\beta, \kappa | Y, X, B, \theta) d\beta d\kappa,
\]

where \( X_i \) denotes the attributes of the observation \( i \) to be predicted. For ease of notation, we refer to \( p(Y_i | X_i, Y, X, B, \theta) \) as

![Figure 6. Purchase probabilities with different base categories.](image-url)
p(Y_i|X_i). An estimate \( \hat{p}(Y_i|X_i) \) for the predictive in Equation (21) is constructed as the empirical probability mass implied by the draws \( Y_{r}^{[m]} \) obtained from \( p(Y_i|X_i, \beta^{[m]}, \kappa^{[m]}) \), where \( \beta^{[m]} \) and \( \kappa^{[m]} \) denote the MCMC draws.

The point forecast \( \hat{Y}_i \) for \( Y_i \) is constructed as the mode of \( \hat{p}(Y_i|X_i) \). The hit-rate is defined as

\[
\text{hit-rate} = \frac{1}{N} \sum_{i=1}^{N} I[\hat{Y}_i = Y_i],
\]

where \( I[A] \) is an indicator function. The log-score is defined as follows:

\[
\text{log-score} = \frac{1}{N} \sum_{i=1}^{N} \ln(\hat{p}(Y_i|X_i)).
\]

For both the hit-rate and the log-score large values are preferred. As a rough measure of statistical significance, we test the difference of the hit-rates between MNP-FS and the benchmark models by a normal test and the difference of the log-scores by the Giaco mini and White’s (2006) test.

We randomly allocate 80% of the observations for estimation of the model, and the remaining 20% are employed for out-of-sample evaluation.

### 6.1. Traditional Dataset With Six Choice Alternatives

Imai and Van Dyk (2005a) and Burgette, Puelz, and Hahn (2021) fitted MNP models to purchase data of laundry detergents. This dataset contains purchases of 2657 households out of six brands of laundry detergents, and the log price of each brand. The dataset is described in detail by Chintagunta and Prasad (1998) and available in Imai and Van Dyk (2005b). We follow Imai and Van Dyk (2005a) and Burgette, Puelz, and Hahn (2021) and fit the MNP models with an intercept and the log price for each brand.

Table 3 shows that for a small choice set the hit-rates and log-scores of our proposed model are similar to the ones of the benchmark methods. The symbol \( (+) \) indicates that MNP-FS significantly outperforms the benchmark method, while the symbol \( (-) \) indicates that the benchmark significantly outperforms MNP-FS. Only the in-sample log-score of MNP-BN significantly improves upon MNP-FS. All models show substantial improvements over the naive forecast, which is constructed as the observed in-sample frequency of the choice alternatives. However, MNP-FS, MNP-BN, and MNP-MPR do not have a better in-sample hit-rate or out-of-sample log-score than the MNP-I.

The posterior parameter estimates from MNP-FS, MNP-BN, and MNP-MPR are also similar. For instance, the posterior mean of the covariance matrix of the latent utilities presented in Figure 7, has similar patterns across all three methods. The small differences in the posterior estimates are also observed for the price coefficient. Figure 8 shows the posterior densities for the different models. The densities are concentrated around similar values.

### 6.2. Modern Dataset With 50 Choice Alternatives

Nowadays, almost all real-life consumer choice sets contain many more choice alternatives than six. To illustrate the importance of a scalable MNP model in these settings, we analyze a laundry detergent purchase dataset with 50 choice alternatives.

We use the Complete Journey dataset published by Dunnhumby.¹ This dataset contains all purchases of 92,339 products over two years from a group of 2500 households at a retailer. We filter the purchases of products with the description “Laundry Detergents”, which results in 300 unique products with different brands, sizes and variants such as liquid or powder detergents. Since the unique products with a small purchase volume are of less interest to a marketing manager, we focus on the 50 top-selling products.

We define the log price of each brand in the same way as, for instance, Allenby and Rossi (1991) and Van et al. (2017). The Dunnhumby dataset only contains records of shelf prices at purchase dates. We impute the prices for products that are not sold on a certain purchase date by taking the mean of the observed prices of a specific product on the nearest date in the same week. In cases where there is no purchase record in the same week, we take the most recent observed price. We remove the observations for which we cannot impute a price for each product.

The final sample contains 4839 observations on 50 categories, which contains 64% of the laundry detergent purchases and the purchase frequency varies from 30 to 274 per category. We fit the same models as in the exercise with six choice alternatives, also considering an intercept and log price coefficient.

For this large choice set, Table 3 shows that the in-sample and out-of-sample log-score of our proposed model are larger than those of the benchmark models, with the out-of-sample improvement also being statistically significant. The hit-rates

---

¹ https://www.dunnhumby.com/sourcefiles
are not statistically different from those of MNP-BN and MNP-MPR. However, the MNP-FS reports both hit-rates and log-scores significantly larger than that of MNP-I. Comparing this result to the relative performance of MNP-FS and MNP-I with six choice alternatives, suggests that accounting for correlations across utilities is especially important when the choice set is large.

For the large laundry detergent choice set, the posterior parameter estimates show differences across the different models. Figure 9 presents the posterior means of the elements of the correlation matrix of the latent utilities. While some general patterns are common to all three methods, the posterior correlations of MNP-BN and MNP-MPR show more variation. The factor structure in MNP-FS has fewer parameters, which restrict the patterns in the correlation matrix. Parsimony may become more important in this large choice set, as the number of estimated covariance parameters by MNP-BN and MNP-MPR equals $49 \times 50/2 = 1225$ relative to 4839 observations. This might explain the fact that the predictive accuracy of the MNP-FS, which only estimates $49 \times 2 = 98$ covariance parameters, significantly improves upon benchmark models in the large choice set, while the differences are not significant in the small choice set.

Figure 10 reports the posterior densities for the price coefficient. We find that the differences in the specifications of the covariance matrix are also reflected in the posterior of the price coefficient. The posterior mean of the price coefficient in the MNP-FS model equals $-0.262$, compared to $-0.372$ and $-0.958$ in the MNP-BN and MNP-MPR, respectively. The corresponding posterior standard deviation is, respectively, 0.018, 0.025, and 0.031. Hence, we conclude that the benchmark specifications result in a larger price effect estimate with more posterior uncertainty than the MNP-FS model.

One might argue that in most practical settings only a small set of high volume products are of interest. To examine the importance of considering a large choice set, we compare the
effect of price on the purchase probability of the six most popular products between an MNP model estimated on only those six products and an MNP model estimated on the total choice set. Figure 11 suggests that, even when only the six top selling products are of interest, including the other products in the model is important for an effective pricing strategy.

Panel (a) of Figure 11 shows the probability of buying one of the top six selling products as a function of the price of the most popular product. The dashed black line corresponds to MNP-FS with only the top six products included as choice alternatives, and the solid yellow line to MNP-FS with 50 products included. The model that includes all 50 products indicates that the probability of buying in the top six decreases when the price of the top product increases. In other words, consumers substitute away from the top product to products outside the top six. This effect cannot be captured by the model that only includes the top six products, which sets the probability of buying in the top six equal to one by construction.

Panel (b) of Figure 11 shows the probability of buying the most popular product as a function of its price. The dashed black line shows that demand for laundry detergent is inelastic between a price of zero and four, and highly elastic for prices higher than four. However, the solid yellow line shows that including the purchases of all 50 products, results in purchase probabilities that change smoothly with price. This result suggests that excluding purchases of low volume products from the analysis can bias the estimated price effect of high volume products.

6.3. Margarine Purchases

We compare the hit-rates and log-scores of the MNP-FS to the benchmark methods on another commonly used choice dataset. McCulloch and Rossi (1994), Burgette and Nordheim (2012), and Burgette, Puelz, and Hahn (2021) fitted MNP models to panel data on purchases of margarine by 516 households. They estimate an intercept and a log price coefficient for six brands. The data are described in detail by Allenby and Rossi (1991) and available in Rossi, Allenby, and McCulloch (2012). We follow Burgette and Nordheim (2012) and Burgette, Puelz, and Hahn (2021) and fit the MNP models to the first purchase of each household.

As an alternative for the small margarine choice set, we construct a large margarine choice dataset in the same way as for laundry detergents in Section 6.2. We filter the purchases of products with the description “Margarines” from the Complete Journey dataset which results in 178 unique products. The final sample contains 11754 observations on 50 categories, which contains 96% of the margarine purchases and the purchase frequency varies from 32 to 1206 per category.

Table 4 shows that for the margarine data our proposed model has higher in-sample and out-of-sample hit-rates and log-scores than the benchmark models when the number of choice alternatives is large. The log-scores of the MNP-FS are significantly larger than the log-scores of all the benchmark models. The MNP-FS model does not show significantly different performance from the MNP-I in the small choice set,
but shows significant improvements in the large choice set on all metrics. This result is in line with the laundry detergent application and supports the claim that it is important to take correlations into account in large choice sets.

Table 1 in Part E of the online appendix reports the hit-rates and log-scores for MNP-FS when the equicorrelated prior is used. For all metrics, the hit-rates and log-scores for MNP-FS are higher than those for the other methods, indicating that the equicorrelated prior improves predictive performance.

### 7. Conclusion

This article proposes a factor structure on the covariance matrix in the MNP model that makes the model scalable to the dimensions of modern choice sets. The model parameters are identified by a reparameterization of the factor structure that imposes a tract-restriction on the covariance matrix.

A numerical experiment shows that the model parameters can be accurately estimated on a choice set with 50 alternatives in the proposed MNP specification. On a real dataset with 50 choice alternatives, the hit-rates and log-scores demonstrate significant predictive improvements relative to benchmark approaches.

The large size of modern assortments and the increasing amount of product differentiation makes the scalable choice model that accounts for correlations across choice alternatives of managerial relevance. The empirical application to retail data of 50 laundry detergents suggests that managers may overestimate the price effect when only analyzing top-selling products, relative to including all products in the MNP model.

### Acknowledgments

We would like to thank an editor, an associate editor, and two anonymous referees for very constructive comments. We would like to thank Gail M. Martin, David T. Frazier, Richard Paap, and Michael S. Smith for helpful discussions.

### References

Allenby, G. M., and Rossi, P. E. (1991), "Quality Perceptions and Asymmetric Switching Between Brands," *Marketing Science*, 10, 185–204. [1687,1689]  
Bunch, D. S. (1991), "Estimability in the Multinomial Probit Model," *Transportation Research Part B: Methodological*, 25, 1–12. [1678,1679]  
Burgette, L. F., and Nordeim, E. V. (2012), "The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model," *Journal of Business & Economic Statistics*, 30, 404–410. [1679,1683,1684,1689]  
Burgette, L. F., Puelz, D., and Hahn, P. R. (2021), "A Symmetric Prior for Multinomial Probit Models," *Bayesian Analysis*, 1–18. [1685,1687,1689]  
Burgette, L. F., and Reiter, J. P. (2013), "Multiple-Shrinkage Multinomial Probit Models With Applications to Simulating Geographies in Public Use Data," *Bayesian Analysis*, 8, 453–478. [1678]  
Chan, J., Leon-Gonzalez, R., and Strachan, R. W. (2018), "Invariant Inference and Efficient Computation in the Static Factor Model," *Journal of the American Statistical Association*, 113, 819–828. [1686]  
Chintagunta, P. K., and Prasad, A. R. (1998), "An Empirical Investigation of the "Dynamic McFadden" Model of Purchase Timing and Brand Choice: Implications for Market Structure," *Journal of Business & Economic Statistics*, 16, 2–12. [1687]  
Cripps, E., Fiebig, D. G., and Kohn, R. (2009), "Parsimonious Estimation of the Covariance Matrix in Multinomial Probit Models," *Econometric Reviews*, 29, 146–157. [1679]  
Frühwirth-Schnatter, S., and Lopes, H. F. (2018), "Sparse Bayesian Factor Analysis When the Number of Factors is Unknown," arXiv: 1804.04231. [1685]  
Geweke, J., Keane, M., and Runkle, D. (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *The Review of Economics and Statistics*, 76, 609–632. [1683]  
Geweke, J., and Zhou, G. (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory," *The Review of Financial Studies*, 9, 557–587. [1680]  
Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578. [1687]  
Hausman, J., and McFadden, D. (1984), "Specification Tests for the Multinomial Logit Model," *Econometrica: Journal of the Econometric Society*, 52, 1219–1240. [1678]  
Imai, K., and Van Dyk, D. A. (2005a), "A Bayesian Analysis of the Multinomial Probit Model Using Marginal Data Augmentation," *Journal of Econometrics*, 124, 311–334. [1679,1687]  
Imai, K., and Van Dyk, D. A. (2005b), "MNP: R Package for Fitting the Multinomial Probit Model," *Journal of Statistical Software*, 14, 1–32. [1687]  
Kaufmann, S., and Schumacher, C. (2019), "Bayesian Estimation of Sparse Dynamic Factor Models With Order-Independent and Ex-Post Mode Identification," *Journal of Econometrics*, 210, 116–134. [1686]  
Lopes, H. F. (2014), "Modern Bayesian Factor Analysis," *Bayesian Inference in the Social Sciences*, 115–153. [1681,1684]  
Lopes, H. F., and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, pages 41–67. [1685,1686]  
McCulloch, R., and Rossi, P. E. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240. [1679,1684,1689]  
McCulloch, R. E., Polson, N. G., and Rossi, P. P. (2000), "A Bayesian Analysis of the Multinomial Probit Model With Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193. [1679,1683]  
Piatek, R., and Ghensowski, M. (2017), "A Multinomial Probit Model With Latent Factors: Identification and Interpretation Without a Measurement System," IZA Discussion Paper. [1679,1684]  
Roberts, G. O., and Rosenthal, J. S. (2009), "Examples of Adaptive MCMC," *Journal of Computational and Graphical Statistics*, 18, 349–367. [1684]  
Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012), *Bayesian Statistics and Marketing*, Wiley. [1689]  
Smith, M. S. (2015), "Copula Modelling of Dependence in Multivariate Time Series," *International Journal of Forecasting*, 31, 815–833. [1684]  
Wan, M., Wang, D., Goldman, M., Taddy, M., Rao, J., Liu, J., Lynamopoulos, D., and McAuley, J. (2017), "Modeling Consumer Preferences and Price Sensitivities From Large-Scale Grocery Shopping Transaction Logs," *Proceedings of the 26th International Conference on World Wide Web*, pp. 1103–1112. [1687]  
Yeo, I. K., and Johnson, R. A. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika*, 87, 954–959. [1681]