Econometric Reviews

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: http://www.tandfonline.com/loi/lecr20

Parsimonious Estimation of the Covariance Matrix in Multinomial Probit Models

Edward Cripps , Denzil G. Fiebig & Robert Kohn

To cite this article: Edward Cripps , Denzil G. Fiebig & Robert Kohn (2009) Parsimonious Estimation of the Covariance Matrix in Multinomial Probit Models, Econometric Reviews, 29:2, 146-157, DOI: 10.1080/07474930903382158

To link to this article: https://doi.org/10.1080/07474930903382158

Published online: 23 Nov 2009.

Econometric Reviews, 29(2):146–157, 2010

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474930903382158

PARSIMONIOUS ESTIMATION OF THE COVARIANCE MATRIX IN MULTINOMIAL PROBIT MODELS

Edward Cripps[1], Denzil G. Fiebig[2], and Robert Kohn[3]

[1] *School of Mathematics and Statistics, University of Western Australia, Perth, Australia*
[2] *School of Economics, University of New South Wales, Sydney, Australia*
[3] *Australian School of Business, University of New South Wales, Sydney, Australia*

□ *This article presents a Bayesian analysis of a multinomial probit model by building on previous work that specified priors on identified parameters. The main contribution of our article is to propose a prior on the covariance matrix of the latent utilities that permits elements of the inverse of the covariance matrix to be identically zero. This allows a parsimonious representation of the covariance matrix when such parsimony exists. The methodology is applied to both simulated and real data, and its ability to obtain more efficient estimators of the covariance matrix and regression coefficients is assessed using simulated data.*

Keywords Bayesian analysis; Covariance selection; Data augmentation; Markov chain Monte Carlo; Multinomial probit.

JEL Classification C11; C15; C35; D03; D12.

1. INTRODUCTION

Models for consumer choice data often specify a continuous latent variable to represent an individual's derived utility from a choice alternative. Such models are described as random utility models and specify systematic and error components for each individual's choice utility. The multinomial logit (MNL) model is derived from specifying identically and independently (iid) Gumbel distributed errors for the individuals choice utilities, and these utilities are differenced against a designated base utility. The MNL model has been widely used due to its analytical

Address correspondence to Robert Kohn, Australian School of Business, University of New South Wales, Sydney, NSW 2052, Australia; E-mail: R.Kohn@unsw.edu.au

and computational tractability. However, the assumption of iid error terms, and the resulting assumption of Independence of Irrelevant Alternatives (IIA), may be too strong in practice. The multinomial (MNP) model avoids the property of IIA by allowing a more general covariance structure for the error terms that also allows for correlation among alternatives. However, MNP models are often not used as they are substantially more complex than MNL models, and their estimation is computationally more demanding.

Bayesian methods are being increasingly applied to random utility models of consumer choice. Albert and Chib (1993) present a data augmentation (Tanner and Wong, 1987) approach for estimating an MNP model, and use Markov chain Monte Carlo (MCMC) simulation to carry out the estimation. Much of the focus of subsequent Bayesian literature on the MNP model deals with issues of convergence and mixing of MCMC algorithms. See, for example, McCulloch and Rossi (1994), Chib et al. (1998), Nobile (1998), McCulloch and Rossi (2000), and Imai and Van Dyk (2005). Imai and Van Dyk (2005) include a comprehensive review of these algorithms and compare them in terms of the interpretability of the prior, the computational speed of the algorithm, and the simplicity of estimation.

The contribution of our article is to propose a prior on the covariance matrix of the latent utilities which allows the off-diagonal elements of the inverse of the covariance matrix to be identically zero. Such a prior is known as a covariance selection prior in the literature. Dempster (1972) proposed estimating the covariance matrix parsimoniously by identifying zero elements in its inverse and called models for the covariance matrix obtained in this way covariance selection models. His idea was that in many statistical problems the inverse of the covariance matrix has a large number of zeros in its off-diagonal elements, and this should be exploited in estimating the covariance matrix. In the Gaussian case, there is a natural interpretation of such zeros: the i, jth element of the inverse is zero if and only if the partial correlation between the ith and jth variables is 0. In the MNP case, zeros in the inverse mean that the latent utilities are conditionally independent. McCulloch et al. (2000) describe a parametrization of the covariance matrix for the latent variables that allows placing priors directly on the identifiable parameters of the MNP model. We extend this parametrization to include covariance selection and use simulated data to compare the estimates of the covariance matrix obtained with covariance selection with those obtained using the prior McCulloch et al. (2000). We also include an analysis of a real data set from the health economics literature.

2. MODEL, PRIORS, AND SAMPLING SCHEME

2.1. Multinomial Probit Model

We now outline the MNP model for consumer choice. Let $Y \in \{0, 1, 2, \ldots, p-1\}$ denote the choice alternatives, and define the utility individual i assigns to alternative j as

$$u_{ij} = \tilde{x}'_{ij}\beta + \epsilon_{ij}, \qquad i = 1, \ldots, n, \quad j = 0, \ldots, p-1, \tag{2.1}$$

where $\tilde{x}_{ij}$ is a $k \times 1$ vector of covariates, β is a $k \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon}_i = (\epsilon_{i0}, \ldots, \epsilon_{i,p-i})$ is a Gaussian random vector with mean zero. The individual is assumed to choose the alternative with the highest utility. Since we can only make statements about the difference in utilities, we set $\epsilon_{i0} = 0$ without loss of generality and transform (2.1) by expressing the model in terms of differences in utilities relative to the base of $j = 0$.

That is,

$$u_{ij} - u_{i0} = (\tilde{x}_{ij} - \tilde{x}_{i0})'\beta + \epsilon_{ij} - \epsilon_{i0}$$

or in matrix notation as

$$W_i = X'_i\beta + \boldsymbol{\epsilon}_i, \tag{2.2}$$

where W_i is the $(p-1) \times 1$ vector of differenced utilities, X is the $(p-1) \times k$ matrix of explanatory variables, and $\boldsymbol{\epsilon}_i$ is the vector of differenced errors. Let Y_i be the choice made in the ith case, so that

$$Y_i(W_i) = \begin{cases} 0 & \text{if } \max(W_i) < 0 \\ j & \text{if } \max(W_i) = W_{ij} > 0, \end{cases} \tag{2.3}$$

such that if we know W_i, we know Y_i. Let $\Sigma = \mathrm{var}(\boldsymbol{\epsilon}_i)$. To ensure that the model is identified, we follow Train (2003, pp. 104–106), and normalize Σ by setting $\Sigma_{11} = 1$.

Following McCulloch et al. (2000), we write

$$\Sigma = \begin{bmatrix} 1 & \gamma' \\ \gamma & \Phi + \gamma\gamma' \end{bmatrix}, \tag{2.4}$$

where γ is a $(p-2) \times 1$ vector, and Φ is the $(p-2) \times (p-2)$ conditional covariance matrix of $W_1, \ldots, W_{p-1}$ given W_1. McCulloch et al. (2000) place the following priors on β, Φ, and γ:

$$\beta \sim N(\bar{b}, V_\beta), \qquad \gamma \sim N(\bar{\gamma}, V_\gamma), \qquad \text{and} \qquad \Phi^{-1} \sim W(\kappa, S),$$

where $W(\kappa, S)$ denotes a Wishart density with paramters κ and S such that $E(\Phi^{-1}) = \kappa S^{-1}$. The priors for γ and Φ are motivated by the fact that they may be viewed as the regression coefficients and covariance matrix in the multivariate regression of $(\epsilon_2, \ldots, \epsilon_{p-1})$ on ϵ_1. That is,

$$\begin{pmatrix} \epsilon_2 \\ \vdots \\ \epsilon_{p-1} \end{pmatrix} = \gamma\epsilon_1 + \eta, \qquad \eta \sim N(0, \Phi). \tag{2.5}$$

The values of $\bar{b}$, V_β, κ, and S are discussed in the analysis of the simulated and empirical data sets.

The method described in this article uses the same priors for β and γ as in McCulloch et al. (2000), but a covariance selection prior is developed for the covariance matrix Φ in (2.5). To do so, we follow Wong et al. (2003) and decompose the inverse of Φ as

$$\Phi^{-1} = TDT,$$

where T is a diagonal matrix, and D is a correlation matrix. Although Φ is somewhat removed from Σ, the elements of T and D still have a sensible interpretation in terms of the latent utilities. The ith diagonal element T_i of T is the inverse of the standard deviation of ϵ_{i+i} conditional on ϵ_j, $j = 1, \ldots, p-1, j \neq i+1$, and $-D_{ij}, i \neq j$ is the partial correlation between ϵ_{i+1} and ϵ_{j+1} given $\epsilon_k, k = 1, \ldots, p-1,\ k \neq i+1, i+j$. If $D_{ij} = 0$, then the partial correlation between ϵ_{i+1} and ϵ_{j+1} is zero, and the corresponding latent utilities are conditionally independent. We now discuss the priors on T and D.

2.2. Prior for *T*

First, note that the squared diagonal elements of T, T_i^2, are the inverse partial variances in the regression Eq. (2.5). We specify a gamma prior for each T_i^2 such that

$$p(T_i^2) \propto (T_i^2)^{\tau-1} \exp\{-\nu T_i^2\}, \qquad \text{for } i = 1, \ldots, p-2. \tag{2.6}$$

The values of τ and ν determine how vague or informative the prior is and are discussed in the analysis of the simulated and empirical data sets. This prior is common for inverse variance parameters in Gaussian linear regression and was described by Wong et al. (2003) in the context of covariance selection. Equation (2.6) implies the prior for T_i can be written as

$$p(T_i) \propto T_i^{2\tau-1} \exp\{-\nu T_i^2\}. \tag{2.7}$$

Wong et al. (2003) show (2.7) results in a full conditional posterior distribution of T_i that is approximated by a Gaussian density as $n \to \infty$. This improves the efficiency of the MCMC scheme described later.

2.3. Covariance Selection Prior for *D*

We now describe the covariance selection prior for D proposed by Wong et al. (2003) that allows the off-diagonal elements to be identically zero subject to the constraint that D must be positive definite.

For $j = 1, \ldots, p$ and $i < j$, let the binary variable $J_{ij} = 0$ if D_{ij} is identically zero, and $J_{ij} = 1$ otherwise. Let $J = J(D) = \{J_{ij}, i < j, j = 1, \ldots, p\}$. Let

$$S(J) = \sum_{i<j} J_{ij},$$

and let $r = p(p-1)/2$ making $0 \leq S(J) \leq r$. Let D_p be the set of $p \times p$ positive definite correlation matrices. Let

$$V(J^*) = \int_{D \in D_p : J(D) = J^*} \left(\prod_{i \leq j, J_{ij}=1} dD_{ij} \right)$$

be the volume of the positive definite region for D, given the constraints imposed by J^*, and let

$$\overline{V}(l) = \binom{r}{l}^{-1} \left(\sum_{J : S(J) = l} V(J) \right)$$

be the average volume for regions with size l.

The hierarchical prior for D is given by

$$p(dD \,|\, J) = V(J)^{-1} dD_{J=1} I(D \in D_p), \tag{2.8}$$

$$p(J \mid S(J) = l) = \binom{r}{l}^{-1} \frac{V(J)}{V(l)}, \tag{2.9}$$

$$p(S(J) = l \mid \psi) = \binom{r}{l} \psi^l (1-\psi)^{r-l}, \tag{2.10}$$

where $0 \leq \psi \leq 1$, $I(D \in D_p) = 1$ if D is a correlation matrix and 0 otherwise and $D_{J=1} = \{D_{ij} : D_{ij} \neq 0\}$. The parameter ψ is the probability that $J_{ij} = 1$. In our article, we take $p(\psi) = 1$.

For further details see Wong et al. (2003).

2.4. Sampling Scheme

We generate $W_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, p-1,\ \beta, \gamma, T_i,\ i = 1, \ldots, p-2$, and $D_{ij}, i < j$ using the following MCMC scheme:

1. $W_{ij} \mid Y_{ij}, \beta, \Sigma, W_{i,-j}, X$, where $W_{i,-j}$ contains all the elements of the ith observation of W not equal to j;
2. $\beta \mid W, \Sigma, X$;
3. $\gamma \mid W, \beta, \Phi$;
4. $T_i \mid W, \beta, D, \gamma, T_{-i}$;
5. $D_{ij} \mid T, W, \beta, \gamma, D_{-ij}$.

For details on generating from the distributions contained in points 1–3, see McCulloch et al. (2000). For details on generating from the distributions contained in points 4, 5, see Wong et al. (2003).

3. SIMULATION EXPERIMENTS

In the following simulations, call the model outlined by McCulloch et al. (2000) Model *NCS* and the covariance selection model presented in our article Model *CS*. Data is simulated for $p = 5, 6$, and 7, and we consider three different structures for Σ. For each of the 9 combinations of p and a structure of Σ, we carried out 50 replications. For all cases, we took the sample size as $n = 1600$. The data is generated from

$$W_i = X_i\beta + \epsilon, \qquad \epsilon_i \sim N(0, \Sigma), \quad \text{for } i = 1, \ldots, n,$$

where W_i is a $(p-1) \times 1$ response vector of latent variables, X is a $(p-1) \times 1$ vector of explanatory variables generated from a uniform distribution on the interval $(-0.5, 0.5)$ and $\beta = 0.89$. The three different covariance matrices are taken to be (a) a full covariance matrix, (b) a tridiagonal covariance matrix, and (c) a diagonal covariance matrix. When the covariance matrix is full or tridiagonal, the nonzero off-diagonal elements are set to 0.5. Since Model *CS* attempts to exploit parsimony in D, we expect Model *CS* to better estimate the covariance matrix than Model *NCS* when D is sparse. Therefore, when estimating the covariance matrix, we expect Model *CS* to perform best when the covariance matrix is diagonal. Note also that in the simulations described above the full covariance matrix results in a sparser D matrix than the tridiagonal covariance matrix, and hence we expect to find Model *CS* performs better when estimating the covariance matrix that is full rather than tridiagonal.

As in McCulloch et al. (2000), the hyperparameters are set to $\bar{b} = 0$, $V_\beta = 0$, $\bar{\gamma} = 0$, and $V_\gamma = 8I_{(p-2)}$ for both Model *CS* and Model *NCS*. For

Model *NCS*, when specifying hyperparameters for Φ we follow McCulloch et al. (2000) and set $\kappa = p + 2$ and $S = (\kappa - p + 1)(I_{(p-2)} - V_\gamma)$.

For Model *CS*, the hyperparameters for $T_i, i = 1, \ldots, p-2$, are $\tau = 10^{-10}$ and $v = 10^{-8}$. This prior is uninformative for the diagonal elements of T. For each simulation the algorithm is initialized by generating W_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p-1$ from independent standard normal densities. Before reporting the simulated data results, we note that although the MCMC output exhibits substantial autocorrelation, the algorithms were run for long periods such that initializing from different starting values yielded similar results. In all the simulations reported below, the MCMC algorithm was allowed 80,000 iterations to achieve convergence and a further 80,000 iterations to obtain draws from the posterior distribution.

3.1. Performance of the Estimators

Performance is assessed using two loss functions, one for the covariance matrix and one for the regression coefficients. While covariance estimation is the primary goal of this article, we also include results for the regression coefficients for completeness. For the estimators of the covariance matrix, we use the L_1 loss function described in Yang and Berger (1994). Let $\widehat{\Sigma}$ be an estimator of the $(p-1) \times (p-1)$ covariance matrix Σ. Then the L_1 loss function is defined as

$$L_1(\widehat{\Sigma}, \Sigma) = \mathrm{tr}(\widehat{\Sigma}\Sigma^{-1}) - \log|\widehat{\Sigma}\Sigma^{-1}| - (p-1).$$

For the estimates of the regression coefficients a root squared error loss (RSE) is used. That is, if $\hat{\beta}$ is an estimate of the true regression coefficient, β, then

$$\mathrm{RSE} = \sqrt{(\hat{\beta} - \beta)^2}.$$

For both loss functions, we report the percentage change in going from Model *CS* to Model *NCS*. That is, for a given loss *LOSS*, we calculate

$$\%L(CS, NCS) = \frac{LOSS(NCS) - LOSS(CS)}{LOSS(CS)} \times 100, \tag{3.11}$$

and if $\%L > 0$, then Model *CS* outperforms Model *NCS*.

3.2. Simulation Results

When $p = 5$, Σ is a 4×4 covariance matrix with variances, σ_{ii}, $i = 1, \ldots, 4$, equal to 1, 0.8, 0.6, and 0.4. Table 1 contains percentiles of

TABLE 1 Percentiles of $\%L$ for the L_1 loss function for 50 replications for $p = 5, 6$, and 7 choices with three different correlation matrices

p	Corr	10th	25th	50th	75th	90th
5	Full	−68.35	−40.75	1.37	157.05	219.84
	Tridiagonal	4.11	22.05	46.17	91.56	119.86
	Diagonal	37.18	68.05	139.75	251.24	373.91
6	Full	−49.52	−18.34	29.83	116.56	298.20
	Tridiagonal	−17.57	−3.44	12.32	38.88	48.63
	Diagonal	−11.70	51.57	78.95	143.87	246.73
7	Full	−31.00	2.40	129.82	294.81	518.08
	Tridiagonal	12.51	21.97	41.30	64.62	88.75
	Diagonal	48.84	99.04	159.22	236.16	355.24

$\%L$ for the L_1 loss function for 50 replications for each correlation structure. Table 2 contains similar percentiles for the *RSE* loss function for 50 replications for each correlation structure. Tables 1 and 2 show that when the correlation matrix is full Model *NCS* and Model *CS* perform similarly when estimating both Σ and β. In both cases, $\%L$ is positively skewed suggesting Model *CS* may be preferred. When the correlation matrix is tridiagonal or diagonal Table 1 reports 10th percentiles that are both above zero and medians equal to 46.17% and 139.75%, respectively, suggesting that Model *CS* substantially improves upon Model *NCS* when estimating Σ. Table 2 reports a median of 134.74% for the tridiagonal correlation matrix and 61.22% for diagonal correlation matrix with 90th percentiles equal to 841.14% and 889.91%, respectively. That is, Model *CS* outperforms Model *NCS* when estimating β in both cases and the improvement can be large. We conclude that in the 5 choice simulation data, when the correlation matrix is full Model *CS* and Model *NCS* perform similarly when estimating Σ and β and when the correlation

TABLE 2 Percentiles of $\%L$ for the *RSE* loss function for 50 replications for $p = 5, 6$, and 7 choices with three different correlation matrices

p	Corr	10th	25th	50th	75th	90th
5	Full	−89.69	−76.06	−3.61	213.43	529.74
	Tridiagonal	−46.49	42.02	134.74	442.80	841.14
	Diagonal	−74.41	−30.89	61.22	142.51	889.91
6	Full	−91.87	−72.06	−49.58	50.92	1331.9
	Tridiagonal	−37.68	−13.98	8.30	79.71	304.89
	Diagonal	−63.37	−47.42	−22.14	35.10	465.40
7	Full	−58.13	−14.27	27.64	133.98	726.4
	Tridiagonal	−24.12	33.21	142.08	426.35	1180.7
	Diagonal	−81.48	−54.71	−2.63	53.20	216.52

matrix is tridiagonal or diagonal Model *CS* outperforms Model *NCS* when estimating Σ and β.

When $p = 6$, Σ is a 5×5 covariance matrix with variances, σ_{ii}, $i = 1, \ldots, 5$, equal to 1, 0.8, 0.6, 0.4, and 0.2. Table 1 shows that Model *CS* outperforms Model *NCS* when estimating Σ when the correlation matrix is diagonal (a median improvement of 78.95%) and when the correlation matrix is full (a median improvement is 29.83%). The 10th percentile is 51.57% for the diagonal correlation matrix and −18.34% for the full correlation matrix implying the improvement is most apparent for the diagonal correlation matrix. Table 2 reports a median of −22.14% for the diagonal correlation matrix and −49.58% for the full correlation matrix. That is, Model *NCS* outperforms Model *CS* when estimating β and when the correlation matrix is diagonal or full. However, it is noted that the 90th percentiles for the diagonal and full correlation matrices are 465.40% and 1331.9% indicating large improvements are obtained by model *CS* over Model *NCS* when estimating β. Table 1 shows that when the correlation matrix is tridiagonal Model *CS* improves marginally over Model *NCS* with a median improvement of 12.32% and 25th and 75th percentiles equal to −3.44% and 38.88%. Table 2 reports a similar pattern when estimating β with a median improvement of 8.30% and 25th and 75th percentiles equal to −13.98% and 79.71%. We conclude that when there are 6 choices, Model *CS* outperforms Model *NCS* for the full and diagonal correlation matrices when estimating Σ. When estimating β and the correlation matrix is full or diagonal, Model *NCS* outperforms Model *CS* (although there are occasions when Model *CS* greatly improves over Model *NCS*). When the correlation matrix is tridiagonal Model *CS* and Model *NCS* perform similarly when estimating both Σ and β.

When $p = 7$, Σ is a 6×6 covariance matrix with the variances, σ_{ii}, $i = 1, \ldots, 6$, equal to 1, 0.8, 0.6, 0.4, 0.2, and 0.2. Table 1 shows Model *CS* outperforms Model *NCS* for the full, tridiagonal and diagonal correlation matrices. The median improvement for the diagonal correlation matrix is 159.22%, the median improvement when the correlation matrix is full equals 129.82%, and the median improvement when the correlation matrix is tridiagonal equals 41.30%. Similar to the 6 choice simulation, the diagonal correlation case exhibits the most improvement, and the tridiagonal correlation case has the lowest improvement, but the improvement for each correlation matrix is substantially larger in the 7 choice simulation than in the 6 choice simulation. Table 2 reports a median of 27.64% when the correlation matrix is full, and 142.08% when the correlation matrix is tridiagonal. Again, we note that the 90th percentiles for the full correlation matrix (726.4%) and tridiagonal correlation matrix (1180.7%) imply Model *CS* can substantially improve over Model *NCS* at times when estimating β. When the correlation matrix is diagonal Model *CS* and Model *NCS* perform similarly when estimating β.

We conclude that when there are 7 choices Model *CS* outperforms Model *NCS* for the full, tridiagonal and diagonal correlation matrices when estimating Σ. When estimating β Model *CS* performs similarly to Model *NCS*, when the correlation matrix is diagonal, and Model *CS* outperforms Model *NCS*, when the correlation matrix is full or tridiagonal.

4. REAL DATA: CHOICE OF CERVICAL CANCER TEST

The real data set analyzed is a health economics data set for women's choice of test for cervical cancer. When analyzing this data sets, it was found that if the priors were too uninformative the MCMC algorithm experienced numerical and convergence difficulties. The hyperparameters are set equal to $\bar{b} = 0$, $V_\beta = 100 \times I$, $\bar{\gamma} = 0$, $V_\gamma = I$, $\tau = 3$, and $v = 1$. We believe the priors are still sufficiently uninformative to allow a reasonable range of values for the parameters. For each data set the MCMC algorithm was run for 30,000 iterates to achieve convergence, and a further 30,000 iterates were used to obtain draws from the posterior distribution.

The data set contains 2,273 observations of women's choice of test for cervical cancer and human papillomavirus (HPV) and is described in Fiebig and Hall (2005). The women have 5 choices for the type of test: a standard Pap test with an additional HPV test, a Pap test without an HPV test, a liquid-based Pap test without an HPV test, a liquid-based Pap test with no HPV test, or no test. The utilities of the first four tests listed above are differenced against no test and are denoted *ph*, *pnh*, *lh*, and *lnh*. Also recorded in the data set and used as covariates are the sex of the GP, the cost of the visit, the probability of a false negative from the Pap test, and a false positive from the Pap test to be used as covariates. After differencing the utilities against no test, we denote these covariates as *sexgp*, *cost*, *fneg*, and *fpos*, respectively. All predictor variables are scaled to lie between -0.5 and 0.5. In our model given by (5.1) W_i is now a 4×1 vector containing the latent variables for *ph*, *pnh*, *lh*, and *lnh*. X_i is a 4×8 matrix with the first four columns containing the intercepts for each of the responses and the final 4 columns represent *sexgp*, *cost*, *fneg*, and *fpos*.

Table 3 contains the 10th, 50th, and 90th percentiles, as well as the mean, of the estimated posterior distribution of the regression coefficients, standard deviations, and correlations obtained by Model *CS*. Similarly, Table 4 reports the 10th, 50th, and 90th percentiles, as well as the mean, of the negative of the partial correlations between *pnh*, *lh*, and *lnh*. Table 4 also contains the posterior means of the probabilities that those partial correlations are nonzero. Analogous to the notation in Section 2.3 denote $D_{(A,B)}$ as the negative partial correlations between two variables, A and B, and the probability that this negative partial correlation is nonzero as $J_{(A,B)}$. Recall that Model *CS* attempts to identify parsimony in D, such that if D

TABLE 3 Results for the choice of cervical cancer data set. Each row report the 10th percentile, mean, median, and 90th percentile of the posterior distribution of the regression coefficients, standard deviations, and correlations

	$\hat{\beta}_{pnh}$	$\hat{\beta}_{lh}$	$\hat{\beta}_{lnh}$	$\hat{\beta}_{ph}$	$\hat{\beta}_{sexpg}$	$\hat{\beta}_{cost}$	$\hat{\beta}_{fneg}$	$\hat{\beta}_{fpos}$	
10th perc.	−0.4	−1.74	−1.02	−1.14	−0.44	−1.48	−0.26	−0.16	
Mean	−0.21	−1.24	−0.72	−0.73	−0.36	−1.32	−0.18	−0.11	
Median	−0.21	−1.17	−0.7	−0.66	−0.36	−1.31	−0.17	−0.11	
90th perc.	−0.01	−0.78	−0.45	−0.46	−0.29	−1.12	−0.1	−0.05	
	$\hat{\sigma}_{pnh}$	$\hat{\sigma}_{lh}$	$\hat{\sigma}_{lnh}$	$\hat{\rho}_{(ph,pnh)}$	$\hat{\rho}_{(ph,lh)}$	$\hat{\rho}_{(ph,lnh)}$	$\hat{\rho}_{(pnh,lh)}$	$\hat{\rho}_{(pnh,lnh)}$	$\hat{\rho}_{(lh,lnh)}$
10th perc.	1.5	0.9	0.9	0.07	0.23	0.23	0.3	−0.25	−0.13
Mean	1.87	1.15	1.12	0.41	0.47	0.44	0.61	0.36	0.44
Median	1.82	1.14	1.06	0.45	0.47	0.44	0.64	0.47	0.58
90th perc.	2.33	1.4	1.32	0.67	0.7	0.66	0.84	0.73	0.75

is a sparse matrix Model *CS* maybe a sensible model. From Tables 3 and 4, the estimated posterior mean of the correlation between *pnh* and *lnh* is 0.36, and the estimated posterior mean of the negative partial correlation is −0.08 with a posterior probability of being nonzero equal to 0.67, suggesting it may be advantageous to analyze this data set using Model *CS*. The 10th percentile and 90th percentile of the posterior distribution for the negative partial correlation between *pnh* and *lnh* are −0.57 and 0.3, showing it is indeed not significantly different from zero. Similarly, the estimated posterior means of the correlations between *pnh* and *lh* and *lh* and *lnh* are 0.61 and 0.38, respectively with negative partial correlations estimated as −0.4 and −0.25. Notice though that the evidence is strong that the partial correlation between *pnh* and *lh* is nonzero with a posterior probability of being nonzero of 0.84 and 10th and 90th percentiles equal to −0.74 and 0, respectively.

TABLE 4 Results for the choice of cervical cancer data set. The first four rows contain the 10th percentile, mean, median, and 90th percentile of the posterior distribution of the negative partial correlations. The final row contains the posterior probabilities of these negative partial correlations being nonzero

	$\widehat{D}_{(pnh,lh)}$	$\widehat{D}_{(pnh,lnh)}$	$\widehat{D}_{(lh,lnh)}$
10th perc.	−0.74	−0.57	−0.54
Mean	−0.40	−0.08	−0.25
Median	−0.44	0.0	−0.33
90th perc.	0	0.3	0.09
Post. prob.	0.84	0.67	0.87

5. CONCLUSION

A new Bayesian covariance selection method for estimating the multinomial probit model is proposed. We ran simulations comparing this method with another method previously proposed in the literature and showed that our method compares favorably when estimating the covariance matrix in terms of the L_1 loss function, particularly when the partial correlations of the utilities are negligible and when the number of choices increases. When estimating the regression coefficients, performance is judged using a root squared error loss function. While results for estimating the regression coefficients do not conclusively suggest our method is preferred, there is substantial improvement on occasions, and any loss is comparatively negligible. We have also analyzed a real consumer choice data set from the health economics literature and reported the results obtained by estimating the covariance selection model, paying attention to the partial correlations that our covariance selection can directly estimate.

ACKNOWLEDGMENT

This project was supported in part by an NHMRC Program Grant and an ARC Discovery Grant.

REFERENCES

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

Chib, S., Greenberg, E., Chen, Y. (1998). MCMC methods for fitting and comparing multinomial response models. Economics Working Paper Archive, Econometrics, No. 9802001.

Dempster, A. (1972). Covariance selection. *Biometrics* 28:157–175.

Fiebig, D. G., Hall, J. (2005). Productivity Commission Conference, November, 2004. Discrete choice experiments in the analysis of health policy. Chapter 6, pp. 119–136.

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

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

McCulloch, R. E., Rossi, P. E. (2000). Reply to Nobile. *Journal of Econometrics* 99:347–348.

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

Nobile, A. (1998) A hybrid Markov chain for the Bayesian analysis of the multinomial probit model. *Statistics and Computing* 8:229–242.

Tanner, M. A., Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82:528–550.

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

Wong, F., Carter, C. K., Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* 90:809–830.

Yang, R., Berger, J. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics* 22:1195–1211.