ELSEVIER

Journal of Econometrics 99 (2000) 335–345

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

# Comment: Bayesian multinomial probit models with a normalization constraint

Agostino Nobile

*Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK*

Received 21 February 2000; received in revised form 24 April 2000; accepted 25 April 2000

**Abstract**

McCulloch, Polson and Rossi (2000), have proposed a prior for the Bayesian analysis of the multinomial probit model which incorporates the identification (or normalization) constraint $\sigma_{11} = 1$. Some empirical evidence on the performance of the prior and related sampler is provided. Direct simulation from Wishart and inverted Wishart distributions, conditional on one of the elements on the diagonal, is then considered. This suggests an alternative way of imposing the normalization constraint in a Bayesian multinomial probit model. © 2000 Elsevier Science S.A. All rights reserved.

*JEL classification:* C11; C15; C25; C35

*Keywords:* Identification; Inverted Wishart distribution; Wishart distribution

## 1. Introduction

In their very stimulating paper, McCulloch, Polson and Rossi (henceforth MPR) provide a new prior distribution for the Bayesian analysis of the multinomial probit (MNP) model, which incorporates the identification constraint $\sigma_{11} = 1$. Their approach consists of rewriting the covariance

*E-mail address:* agostino@stats.gla.ac.uk (A. Nobile).

PII: S0304-4076(00)00035-X

matrix $\Sigma$ as

$$\Sigma = \begin{pmatrix} \sigma_{11} & \gamma' \\ \gamma & \Phi + \dfrac{\gamma\gamma'}{\sigma_{11}} \end{pmatrix},$$

reparameterizing $\Sigma$ into $\{\sigma_{11}, \gamma, \Phi\}$, setting $\sigma_{11} = 1$ and then selecting independent priors for $\gamma$ and $\Phi$. In their treatment, they provide analytical results not only for this new prior (ID prior) but also for the non-identified (NID) prior introduced by McCulloch and Rossi (1994). They illustrate how simulation from the prior distribution can aid in its specification, especially for informative priors, and provide two examples of posterior simulation. They discuss some difficulties associated with the ID prior, linking them to the induced prior distribution on the smallest eigenvalue of $\Sigma$. They conclude with a survey of alternative approaches.

MPR's contribution is a welcome addition to the MNP modeling toolkit. A few short remarks on their approach are provided below, while Section 2 discusses some empirical results obtained using the ID prior. Section 3 considers direct simulation from Wishart and inverted Wishart distributions conditional on one diagonal element. The ability of generating from these distributions suggests an alternative way of imposing the normalization constraint $\sigma_{11} = 1$. The notation of MPR is used throughout, unless stated otherwise.

1. Further guidance, beyond MPR's remarks in Section 5.2.1, can be given for the important special case $E(\Sigma) = I$, $B^{-1} = \tau I$. For this case, besides requiring that all the covariances have zero expected value, it may also be reasonable to assume that their prior variability is the same:

$$Var(\gamma_i) = Var(\Phi_{jk} + \gamma_j\gamma_k) \quad i, j, k = 1, \dots, p-2; j \neq k, \tag{1}$$

   where $\gamma_i = \sigma_{1,i+1} = Cov(\varepsilon_1, \varepsilon_{i+1})$, $\Phi_{jk} + \gamma_j\gamma_k = \sigma_{j+1,k+1} = Cov(\varepsilon_{j+1}, \varepsilon_{k+1})$ and $p > 3$. This leads (details are in the appendix) to an equation that solved for $\tau$ yields

$$\tau = \frac{a}{1+a} \quad \text{where} \quad a = \frac{(\kappa - p + 1)}{(\kappa - p + 2)(\kappa - p - 1)}. \tag{2}$$

   Thus, in this case, one only needs to specify a value for $\kappa$, in order to assign the prior of $\Sigma$. Table 1 contains some values of $\tau$ for given values of $\kappa - p$.

   If $p = 3$ one may consider that $Var(\Phi) = (2(1-\tau)^2)/(\kappa - 4)$, $\kappa > 4$. Equating this to $Var(\gamma) = \tau$, yields $\tau = (\kappa/4) - \frac{1}{2}\sqrt{(\kappa^2/4) - 4}$.
2. Notwithstanding the appeal of the ID prior, it seems to me that from a practical point of view the NID approach of McCulloch and Rossi (1994), possibly with the modification proposed in Nobile (1998), is still very much viable. In fact, MPR report that the ID Gibbs sampler tends to produce

Table 1
Values of $\tau$ corresponding to values of $\kappa - p$

| $\kappa - p$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $\tau$ | $\frac{3}{7}$ | $\frac{2}{7}$ | $\frac{5}{23}$ | $\frac{3}{17}$ | $\frac{7}{47}$ |

draws that are more highly correlated than those from the NID and hybrid samplers (more on this in Section 2). From this viewpoint, the part of MPR's contribution that may turn out to be more practically relevant is their analysis of the NID prior which, coupled with prior simulation, can form the basis for the specification of informative priors in the NID approach.

3. I wholeheartedly agree with MPR that placing an improper prior on $\Sigma$ in the ID approach has several disadvantages. One that they fail to mention is the necessity to check that the resulting posterior is proper. Improper posteriors may result while (i) the full conditional distributions are well defined and (ii) an exam of the Gibbs sampling output does not hint at posterior impropriety, see Hobert and Casella (1996).

## 2. Some simulation results

I have implemented the ID prior specification and used it to estimate a MNP model with $p = 8$ choice alternatives and $k = 10$ covariates. An artificial data set of 2000 choices $Y_i$ was generated according to models (1) and (2) in MPR, with $\tilde{\Sigma} = I + 11'$, $\tilde{\beta} = \sqrt{2}(1, 2, \dots, 10)'$, so that the identified parameters are $\beta = (1, 2, \dots, 10)'$, $\sigma_{ii} = 1$, $\sigma_{ij} = 0.5$, $i \neq j$. The covariates $X$ had values equal to either $-1$, 0, or 1. Further details on this data are in Nobile (1998), Example 3, where they were used to compare the mixing behaviour of the NID sampler and the hybrid sampler introduced there. In summary, for this data set the NID sampler required about 10,000 Gibbs sampling iterates to stabilize, while the hybrid sampler needed only about 3000 iterates, with a negligible increment in computing time per iterate. Estimates of posterior means from the hybrid sampler were closer to the true values, as the associated sampling paths were less autocorrelated thus providing more information per draw.

While estimating this model using the ID prior, the following hyperparameters were used:

$$\bar{\beta} = 0,\ D^{-1} = 100I,\ \bar{\gamma} = 0,\ \kappa = p + 3 = 11,\ B^{-1} = \tfrac{2}{7}I,$$

$$C = (\kappa - p + 1)(1 - \tau)I = \tfrac{20}{7}I.$$

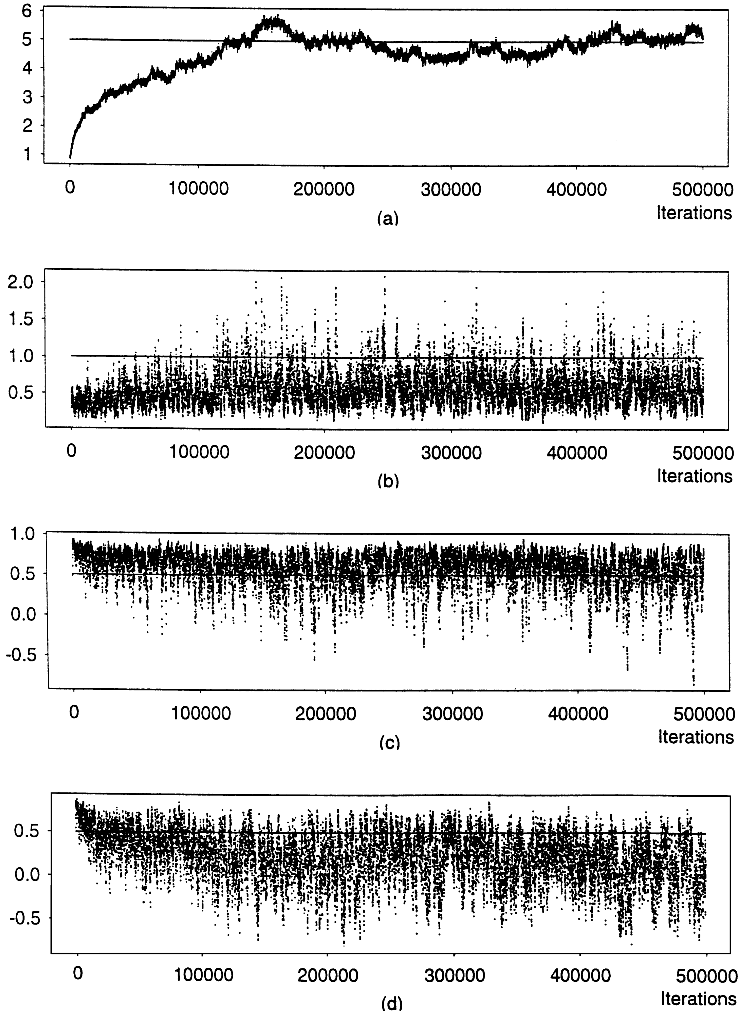

Fig. 1. ID prior Gibbs sampler, model with $p = 8$ alternatives. Time series plots of simulated values of $\beta_5$ in (a), $\sigma_{55}$ in (b), $\rho_{15} = \sigma_{15}/\sqrt{\sigma_{55}}$ in (c) and $\rho_{25} = \sigma_{25}/\sqrt{\sigma_{22}\sigma_{55}}$ in (d). Solid lines represent the true values.

The sampler was started at the same values used in Nobile (1998) for the NID and hybrid samplers: $\beta = (1, 1, \ldots, 1)$, $\gamma = 0$, $\Phi = I$. As a first run of 20,000 Gibbs sampling iterations failed to approach the true parameter values, a further run of the sampler was made with 500,000 iterations, saving only one iterate every 25. This second run required about one day on a Sun Ultra 5 workstation. Fig. 1 displays time-series plots of the simulated values of $\beta_5$, $\sigma_{55}$,

$\rho_{15} = \sigma_{15}/\sqrt{\sigma_{55}}$ and $\rho_{25} = \sigma_{25}/\sqrt{\sigma_{22}\sigma_{55}}$. The simulated paths of the other parameters exhibited similar behaviours. Sustained trends in mean and variance are evident in all plots. After the sampler has stabilized, there are still long excursions away from the true values, due to the high autocorrelation. At least on this problem, the performance of the Gibbs sampler for the ID prior was markedly inferior to both the NID sampler and the hybrid sampler. Although similar difficulties were encountered by MPR in their examples, they were not of the present gravity, perhaps due to the smaller size of the models considered.

I have re-estimated the same model with data generated using continuous covariates, uniform on the interval $(-\sqrt{1.5}, \sqrt{1.5})$ (to match the first two moments of the discrete covariates). The mixing performance of the ID sampler did not seem to improve.

This evidence is admittedly limited, but it suggests that, in medium–large problems, with several alternatives and covariates, the ID Gibbs sampler may require extremely long burn-in times. Moreover, the high autocorrelation of the sample paths may require very long runs, after the sampler has stabilized, in order to obtain a representative sample from the posterior distribution.

Perhaps some ingenuity is needed to devise a sampling Markov chain which uses the ID prior and has good mixing properties.

## 3. Generating Wishart and inverted Wishart distributions given a diagonal element

This section discusses how to generate Wishart and inverted Wishart random matrices conditional on one of the diagonal elements. This suggests an alternative way of imposing the normalization constraint in a Bayesian MNP model.

The following probability result will be used in the sequel.

*Proposition.* *Let $X_1$ and $X_2$ be independent random vectors. Let $Y_1 = h_1(X_1)$ and $Y_2 = h_2(X_1, X_2)$, where $h_1$ is one-to-one and all the functions considered are measurable. Then, $Y_1$ and $X_2$ are independent and $Y_2 = h_2(h_1^{-1}(Y_1), X_2)$.*

*Corollary. Suppose the assumptions in the above proposition hold. To generate from the conditional distribution of $Y_2$ given $Y_1 = y_1$, one can generate from the marginal distribution of $X_2$, yielding $x_2$, and then compute $y_2 = h_2(h_1^{-1}(y_1), x_2)$.*

Consider next the algorithm that generates $G$ from the Wishart distribution $W_p(v, V)$ using the Bartlett decomposition, see e.g. Smith and Hocking (1972). I assume throughout that $v > p - 1$, $V$ is positive definite, $\mathrm{E}(G) = vV^{-1}$ and that $L = ((\ell_{ij}))$ is a lower triangular matrix such that $V^{-1} = LL'$.

*Algorithm 1.* Draw $G \sim W_p(v, V)$.

1. Construct a lower triangular matrix $A$ with

(a) $a_{ii}$ equal to the square root of $\chi^2_{v+1-i}$ random variates, $i = 1, \dots, p$.
(b) $a_{ij}$ equal to N(0, 1) random variates, $i > j$.

2. Set $G = LAA'L'$.

Next note that $g_{11} = \ell_{11}^2 a_{11}^2$, a one-to-one transformation since $a_{11} > 0$. Therefore, one can use the corollary above with $X_1 = a_{11}$, $X_2 = (a_{21}, a_{22}, a_{31}, \dots, a_{pp})'$, $Y_1 = g_{11}$ and $Y_2 = (g_{21}, g_{22}, g_{31}, \dots, g_{pp})'$. It then follows that the algorithm below yields a draw from the conditional distribution of $G$ given $g_{11} = \bar{g}_{11}$.

*Algorithm 2.* Draw $G \sim W_p(v, V)$, given $g_{11} = \bar{g}_{11}$.

1. Construct a lower triangular matrix $A$ with

(a) $a_{11} = \sqrt{\bar{g}_{11}}/\ell_{11}$;
(b) $a_{ii}$ equal to the square root of $\chi^2_{v+1-i}$ random variates, $i = 2, \dots, p$.
(c) $a_{ij}$ equal to N(0, 1) random variates, $i > j$.

2. Set $G = LAA'L'$.

One can also draw from an inverted Wishart distribution, $\Sigma = G^{-1}$ with $G \sim W_p(v, V)$, conditional on one diagonal element of $\Sigma$. Consider that

$$\Sigma = G^{-1} = (L^{-1})'(A^{-1})'A^{-1}L^{-1}.$$

Since $A$ and $L$ are lower triangular, $A^{-1}$ and $L^{-1}$ are lower triangular, hence $A^{-1}L^{-1}$ is lower triangular. Denote by $a^{ij}$, $\ell^{ij}$ the $(i, j)$ elements of $A^{-1}$ and $L^{-1}$, respectively. The $p$th column of $A^{-1}L^{-1}$ is $(0, \dots, 0, a^{pp}\ell^{pp})' = (0, \dots, 0, a_{pp}^{-1}\ell_{pp}^{-1})'$, so that the $(p, p)$ element of $\Sigma$ is $\sigma_{pp} = (a_{pp}\ell_{pp})^{-2}$. Now apply again the Corollary with $X_1 = a_{pp}$, $X_2 = (a_{11}, a_{21}, a_{22}, a_{31}, \dots, a_{p,p-1})'$, $Y_1 = \sigma_{pp}$ and $Y_2 = (\sigma_{11}, \sigma_{21}, \sigma_{22}, \sigma_{31}, \dots, \sigma_{p,p-1})'$. Then, the following algorithm yields a draw $\Sigma = G^{-1}$ with $G \sim W_p(v, V)$ conditional on $\sigma_{pp} = \bar{\sigma}_{pp}$.

*Algorithm 3.* Draw $\Sigma = G^{-1}$, $G \sim W_p(v, V)$, given $\sigma_{pp} = \bar{\sigma}_{pp}$.

1. Construct a lower triangular matrix $A$ with

(a) $a_{ii}$ equal to the square root of $\chi^2_{v+1-i}$ random variates, $i = 1, \dots, p-1$;
(b) $a_{pp} = (\sqrt{\bar{\sigma}_{pp}}\ell_{pp})^{-1}$;
(c) $a_{ij}$ equal to N(0, 1) random variates, $i > j$.

2. Set $\Sigma = (L^{-1})'(A^{-1})'A^{-1}L^{-1}$.

To draw $\Sigma$ conditional on $\sigma_{11} = \bar{\sigma}_{11}$, exchange rows/columns 1 and $p$ in the matrix $V$ before applying Algorithm 3, then exchange them again in the draw $\Sigma$.

The ability of generating from Wishart and inverted Wishart distributions conditional of $\sigma_{11}$ allows one to achieve identification of the NID prior in a manner alternative to MPR's. The prior on ($\beta$, $\Sigma$) and the Gibbs sampler remain as in the original formulation of McCulloch and Rossi (1994), described in Section 3.1 of MPR, but all distributions are conditional on $\sigma_{11} = 1$ and $\Sigma$ is drawn from an inverted Wishart distribution conditional on its (1, 1) element. For lack of a better name I will call this the NIW (normalized inverted Wishart) prior/sampler.

Both the ID and NIW priors are fully identified. To appreciate the difference between them, consider the joint distribution of ($\sigma_{11}$, $\gamma$, $\Phi$) under the NID prior. This is basically provided by MPR in their appendix, or can be obtained from the joint distribution of the blocks in a partitioned inverted Wishart matrix. One has that

$$\sigma_{11}^{-1} \sim v_{11}^{-1}\chi^2_{\nu-p+2} \quad \text{independent of } \Phi \text{ and } \gamma/\sigma_{11},$$

$$\Phi^{-1} \sim W_{p-2}(\nu,\ V_{22} - V_{21}V_{12}/v_{11}),$$

$$\frac{\gamma}{\sigma_{11}} \mid \Phi \sim N_{p-2}(V_{12}/v_{11},\ \Phi/v_{11}).$$

Therefore, under the NIW prior the joint distribution of $\gamma$ and $\Phi$ is

$$\Phi^{-1} \sim W_{p-2}(\nu,\ V_{22} - V_{21}V_{12}/v_{11}),$$

$$\gamma|\Phi \sim N_{p-2}(V_{12}/v_{11},\ \Phi/v_{11}).$$

While independent distributions are placed on $\gamma$ and $\Phi$ in the ID prior, in the NIW prior they are not independent.

I have used the NIW prior/sampler to estimate the model discussed in Section 6.1 of MPR, using the hyperparameter values $\bar{b} = 0$, $A^{-1} = 100$, $\nu = 6$, $V = \nu I$. Fig. 2 displays time-series plots and histograms for the three identified parameters $\beta$, $\rho_{12}$ and $\sigma_{22}$. Results are comparable to those reported in MPR for the ID and NID samplers.

I have also used the NIW prior/sampler to re-estimate the model with 8 alternatives and 10 covariates of Section 2, with hyperparameters $\bar{b} = 0$, $A = 100I$, $\nu = 11$, $V = 3I$, so that a priori $\mathrm{E}(\tilde{\Sigma}) = I$. Fig. 3 contains time-series plots of the sampled values of $\beta_5$, $\sigma_{55}$, $\rho_{15}$ and $\rho_{25}$. Comparing these plots with those in Fig. 1, it appears that the mixing behaviours of the NIW and ID samplers are comparable in this problem, thus the remarks addressed to the latter in Section 2 also apply to the former.

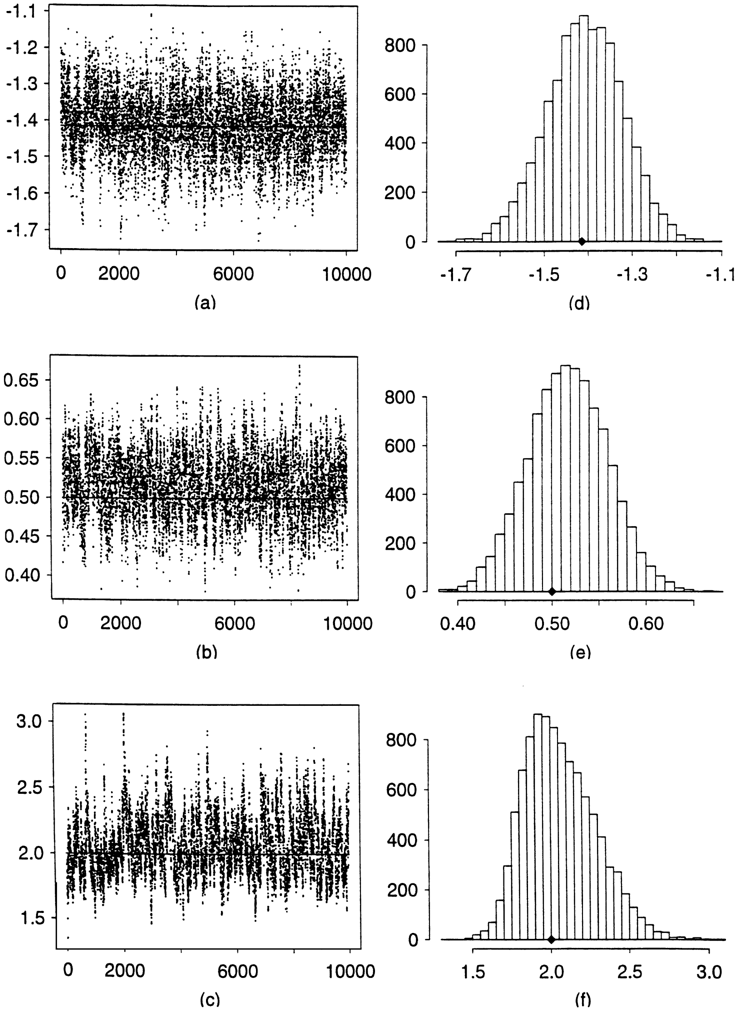

Fig. 2. NIW prior Gibbs sampler, MPR Example 1. Time series plots of simulated values of $\beta$, $\rho_{12}$ and $\sigma_{22}$ are in panels (a), (b) and (c), respectively. Panels (d), (e) and (f) contain histograms of the simulated values, in the same order. Solid lines and black diamonds represent the true values.

Based on the available evidence, it would seem to me that further research is needed to establish an approach that directly incorporates the normalization constraint in the prior distribution and provides accurate estimates in reasonable time for medium–large problems.

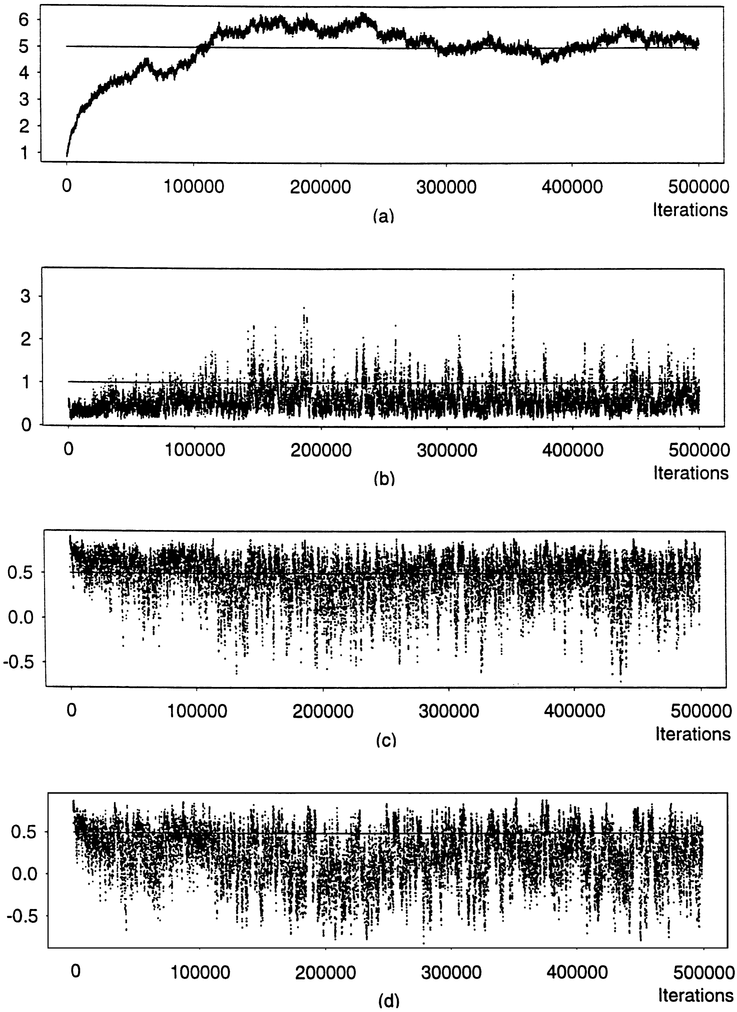

Fig. 3. NIW prior Gibbs sampler, model with $p = 8$ alternatives. Time series plots of simulated values of $\beta_5$ in (a), $\sigma_{55}$ in (b), $\rho_{15} = \sigma_{15}/\sqrt{\sigma_{55}}$ in (c) and $\rho_{25} = \sigma_{25}/\sqrt{\sigma_{22}\sigma_{55}}$ in (d). Solid lines represent the true values.

## Note added in proofs

After this note was accepted for publication I became aware that the idea of drawing from the conditional distribution of $\Sigma$ given one diagonal element had been independently used by Linardakis and Dellaportas (1999).

## Acknowledgements

The author would like to thank the Editor and an Associate Editor for inviting him to write the present note and for providing comments leading to an improved Section 3.

## Appendix

Here, I derive Eq. (2). Recall that $Var(\gamma_i) = \tau$. Also,

$$Var(\Phi_{jk} + \gamma_j \gamma_k) = Var(\Phi_{jk}) + Var(\gamma_j \gamma_k)$$

since $\Phi$ and $\gamma$ are a priori independent. Now, since the components of $\gamma$ are a priori independent and with zero mean, $Var(\gamma_j \gamma_k) = Var(\gamma_j) Var(\gamma_k) = \tau^2$. The second moments of the entries in $\Phi$, with $\Phi^{-1} \sim W_{p-2}(\kappa, C)$, have been computed, see e.g., Press (1972), Theorem 5.2.2. (However, note that the parameterization of the Wishart distribution used by Press differs from the one employed by MPR). One has

$$Var(\Phi_{jk}) = \frac{c_{jj}c_{kk} + c_{jk}^2(\kappa - p + 3)/(\kappa - p + 1)}{(\kappa - p + 2)(\kappa - p + 1)(\kappa - p - 1)} \quad \kappa > p + 1, j \neq k. \tag{3}$$

In the case under consideration, $C = (\kappa - p + 1)(\Delta - B^{-1}) = (\kappa - p + 1)(1 - \tau)I$ so that Eq. (3) reduces to

$$Var(\Phi_{jk}) = \frac{(\kappa - p + 1)(1 - \tau)^2}{(\kappa - p + 2)(\kappa - p - 1)} \quad \kappa > p + 1, j \neq k.$$

Therefore Eq. (1) can be rewritten as

$$\tau = \frac{(\kappa - p + 1)(1 - \tau)^2}{(\kappa - p + 2)(\kappa - p - 1)} + \tau^2,$$

which when solved for $\tau$ yields (2).

## References

Hobert, J.P., Casella, G., 1996. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. Journal of the American Statistical Association 91, 1461–1473.

Linardakis, M., Dellaportas, P., 1999. Bayesian Analysis of latent utilities for transportation services via extensions of the multinomial probit model. Working paper, Athens University of Economics and Business.

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.

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

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

Press, S.J., 1972. Applied Multivariate Analysis, Holt, Rinehart & Winston, New York.

Smith, W.B., Hocking, R.R., 1972. Algorithm AS 53: Wishart variate generator, Applied Statistics 21, 341–345.