A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters

by

Robert E. McCulloch   Nicholas G. Polson   Peter E. Rossi [1]

April 1997

revised, May 1999

Abstract

We present a new prior and corresponding algorithm for Bayesian analysis of the multinomial probit model. Our new approach places a prior directly on the identified parameter space. The key is the specification of a prior on the covariance matrix so that the (1,1) element if fixed at 1 and it is possible to draw from the posterior using standard distributions. Analytical results are derived which can be used to aid in assessment of the prior.

---

[1]Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637. email: peter.rossi@gsb.uchicago.edu.

1 Introduction

The Multinomial Probit Model (MNP) has now become a widely accepted alternative to the multinomial logit model for situations in which one of a finite number of outcomes are observed conditional on a set of covariates. The recent popularity of the MNP model stems from the appeal of relaxing the IIA property of logit models and advances in methods for inference. Both likelihood-based and non-likelihood based methods have been advanced which make estimation of MNP models for a large number of alternatives computationally tractable. Working from the frequentist point of view, McFadden (1989) proposed the method of simulated moments and Hajivassiliou and McFadden (1990) have proposed the method of simulated scores (See also Keane (1994) and Borsch-Supan and Hajivassiliou (1993)).

A Bayesian analysis of the MNP model is given in McCulloch and Rossi (1994) (henceforth MR) (see also Nobile (1998)). The MNP model, as commonly specified, has a vector of coefficients β and a covariance matrix Σ as parameters. However, the parameters (β, Σ) are not fully identified. The model is often identified by setting the first diagonal element of the covariance matrix equal to one ($\sigma_{11} = 1$). A key feature of the MR method is the manner in which the identification issue is handled. In the MR approach, a proper prior is specified for the full set of parameters (β, Σ) and the marginal posterior of the identified parameters $(\beta/\sqrt{\sigma_{11}}, \Sigma/\sigma_{11})$ is reported. Thus, the prior on the identified parameters is the marginal prior of $(\beta/\sqrt{\sigma_{11}}, \Sigma/\sigma_{11})$ derived from the prior distribution specified for the full set of parameters (β, Σ). The approach is taken because of the difficulties associated with a Bayesian analysis of covariance matrices with first diagonal element fixed at one. Since it is impossible to specify a truly diffuse or improper prior with this approach, the induced priors must be inspected to assure that they properly represent the investigators beliefs.

In this paper, we present a new approach which places a prior directly on the set of identified parameters. We do this by specifying a prior on Σ such that the first diagonal element is one with probability one. We discuss both the assessment of the prior and the computation of the posterior. Computation of the posterior involves a simple Gibbs

sampler in which each draw is either normal, truncated univariate normal, or Wishart. The new prior can be both informative or diffuse and improper. We present analytical results that facilitate the prior specification in both the new prior and that of MR. However, we see in the examples that this simple method of achieving identification comes at a cost: the Gibbs sampler produces a Markov chain which tends to be more highly autocorrelated that the Markov chain used in the MR approach. In some extreme cases, the Markov chain for the identified parameter case will fail to converge. These cases occur in high dimensions and in situations in which the likelihood is not very informative.

In section 2, we review the multinomial probit model and discuss the identification problem. As discussed above, the way in which this issue is handled is the key difference between the method of this paper and the MR approach. Section 3 reviews the MR algorithm and section 4 presents our new prior and corresponding Gibbs sampling algorithm. For either method, we must assess the prior. In section 5 we discuss the choice of the prior and present some analytical results that aid in assessing and comparing the alternative priors. Section 6 illustrates the prior and posterior computation with some simulated examples. Section 7 discusses the implications of our new prior for the distribution of the smallest eigenvalue which explains why certain prior settings may cause convergence problems. Section 8 discusses the pros and cons of the new approach and briefly compares it with other approaches in the literature.

2   The Multinomial Probit Model

2.1   The Model

In this section we briefly review the MNP model. Let Y be a random variable such that $Y \in \{0, 1, 2, \ldots, p-1\}$ and X be a $(p-1) \times k$ matrix. The conditional distribution of $Y \mid X$ is specified as follows. First let,

$$W = X\beta + \epsilon \tag{1}$$

where ϵ is $N(0, \Sigma)$. Y is then a function of W by,

$$Y(W) = \begin{cases} 0 & \text{if } \max(W) < 0 \\ i & \text{if } \max(W) = W_i > 0. \end{cases} \quad (2)$$

Here $\max(W)$ means the maximal element of $W' = (W_1, W_2, \ldots, W_{p-1})'$. So, if all the W_i are negative then $Y = 0$ and Y equals the index of the biggest W_i if it is positive.

We have now defined $Y \mid X, \beta, \Sigma$, where X consists of observable quantities and the model parameters are β and Σ. Typically in application we observe a set of observations (Y_i, X_i) and assume that given the X's the Y's are independent. Various elaborations of the model have been considered (see for example MR sections 8 and 9).

At this point, some of the difficulties associated with the analysis of the MNP model may be appreciated. To compute the likelihood we must compute the probability of sets of the form $\{W \mid Y(W) = i\}$ where $W \sim N(X\beta, \Sigma)$. The sets are cones in R^{p-1}. Much of the research on the MNP model has been devoted to the development of efficient computational methods for computing these integrals.

2.2 Identification

In the model specified by (1) and (2) above, the parameters (β, Σ) are not identified. This is because $Y(cW) = Y(W)$ for all $c > 0$. Since $cW = c(X\beta + \epsilon) = X(c\beta) + c\epsilon$ we see that the distribution $Y \mid X, \beta, \Sigma$ is the same as the distribution $Y \mid X, c\beta, c^2\Sigma$. Given a set of observations (Y_i, X_i) the likelihood L would be such that $L(\beta, \Sigma) = L(c\beta, c^2\Sigma)$. For discussion of identification in the context of the MNP model see Dansie (1985), Bunch (1991), and Keane (1992) for example.

Let σ_{ij} be the ij^{th} element of Σ. Since σ_{11} is positive, identification may be achieved by setting σ_{11} equal to 1. This is the approach commonly adopted in frequentist methods. It is not straightforward to adopt this approach in a Bayesian analysis because of the difficulty in defining a prior on the set of covariance matrices such that the (1,1) element is one.

3 An Algorithm with Nonidentified Parameters

3.1 The Algorithm

In this section we briefly review the method developed in MR for a Bayesian analysis of the MNP model. The method uses the full set of parameters (β, Σ). Because of the model is not identified, we use a proper prior to ensure that the posterior is proper. The prior specification lets β and Σ be independent with β having a multivariate normal distribution and $G = \Sigma^{-1}$ having a Wishart distribution:

$$p(\beta \mid \bar{b}, A) \propto |A|^{\frac{1}{2}} \exp\{-\frac{1}{2}(\beta - \bar{b})'A(\beta - \bar{b})\} \tag{3}$$

$$p(G \mid \nu, V) \propto |G|^{\frac{\nu-p}{2}} \exp\{\mathrm{tr}(-\frac{1}{2}GV)\}. \tag{4}$$

Here, $\bar{b}$ and A are the parameters of the normal prior: $\beta \sim N(\bar{b}, A^{-1})$. V and ν are the parameters of the inverted Wishart prior: $\Sigma^{-1} \sim W(\nu, V)$. Note that our parameterization of the Wishart distribution is such that $E(\Sigma^{-1}) = \nu V^{-1}$. In practice we choose values for $(\bar{b}, A, \nu, V)$ and then check that the marginal prior of the identified parameters $(\beta/\sqrt{\sigma_{11}}, \Sigma/\sigma_{11})$ is reasonable. Theoretically, our marginal posterior is the same as that obtained by working with just the identified parameters $(\beta/\sqrt{\sigma_{11}}, \Sigma/\sigma_{11})$ and the marginal prior on them derived from the prior specified for the full set of parameters. MR uses the full set of parameters so that the following simple algorithm may be used to compute the posterior.

The method uses Gibbs sampling to obtain draws from the posterior distribution. Gibbs sampling is discussed in Gelfand and Smith (1990), Casella and George (1992), Smith and Roberts (1993), and Tierney (1991) among many others. In the Gibbs sampling approach, we sample the W_i in (1) above. As pointed out by Albert and Chib (1993), introducing the latent variables $\{W_i\}$ to be drawn in addition to (β, Σ) greatly simplifies the algorithm.

Let W_{ij} be the jth element of the ith W vector. Let $W_{i(-j)}$ be the ith W vector with W_{ij} removed. Given the data $D = \{Y_i, X_i\}$, the Gibbs sampling algorithm proceeds by drawing from the following set of conditional distributions:

$$
\begin{aligned}
\beta &\mid \Sigma, \{W_i\}, D \\
\Sigma &\mid \beta, \{W_i\}, D \\
W_{ij} &\mid \beta, \Sigma, \{W_{i(-j)}\}, D
\end{aligned}
$$

The draw of β is from a multivariate normal, the draw of Σ is from an inverted Wishart distribution, and the draw of each W_{ij} is a univariate truncated normal.

Nobile(1998) has proposed an elaboration of the algorithm above. A Metropolis step is added in which the current values of all W, β, and Σ are scaled up or down by a positive constant and the algorithm jumps to the scaled values or not in the usual Metropolis manner. Nobile provides evidence that this added step significantly improves the performance of the Markov chain.

4   A Method with Fully Identified Parameters

In this section, we describe an alternative prior and corresponding Gibbs sampling algorithm for the MNP model. The prior assigns probability one to the set $\{\Sigma \mid \sigma_{11} = 1\}$ in such a way that we are able to draw from $\Sigma \mid \beta, \{W_i\}, D$. Thus our algorithm for computing the posterior will be same as before, except for the draw of Σ. As discussed in section 2.2 above, by fixing $\sigma_{11} = 1$, we identify the parameters of the MNP model.

We define our prior by first reparametrizing Σ. Denote the ϵ in (1) by $(\epsilon_1, \epsilon_2, \ldots, \epsilon_{p-1})'$. Then let $U = \epsilon_1$ and $Z = (\epsilon_2, \epsilon_3, \ldots, \epsilon_{p-1})'$ so that $\epsilon = (U, Z')'$. Σ indexes the joint distribution of U and Z which is $N(0, \Sigma)$. We rewrite this joint distribution as the marginal distribution of U and the conditional distribution $Z \mid U$. Let $\gamma = E(UZ)$ and $\Sigma_Z = E(ZZ')$. We have $U \sim N(0, \sigma_{11})$ and $Z \mid U \sim N((\gamma/\sigma_{11})U, \Sigma_Z - \gamma\gamma'/\sigma_{11})$. Let $\Phi = \Sigma_Z - \gamma\gamma'/\sigma_{11}$. There is a one to one correspondence between Σ and $(\sigma_{11}, \gamma, \Phi)$. Hence we can put a prior

on $\{\Sigma \,|\, \sigma_{11} = 1\}$ by setting $\sigma_{11} = 1$ and putting priors on γ and Φ using the relation:

$$\Sigma = \begin{bmatrix} 1 & \gamma' \\ \gamma & \Phi + \gamma\gamma' \end{bmatrix}. \tag{5}$$

We choose the following priors:

$$\gamma \sim N(\bar{\gamma}, B^{-1}) \tag{6}$$

$$\Phi^{-1} \sim W(\kappa, C). \tag{7}$$

To obtain draws from $(\gamma, \Phi)\,|\,\beta, \{W_i\}, D$, first note that given β and $\{W_i\}$ we "observe" $\{\epsilon_i\}$. Correspondingly we have observations $(U_i, Z_i')'$. From above, γ and Φ are simply the parameters of a multivariate regression of Z on U. The draw of (γ, Φ) is easily done by again Gibbs sampling: $\gamma\,|\,\Phi, \beta, \{W_i\}, D$ and $\Phi\,|\,\gamma, \beta, \{W_i\}, D$. Our full Gibbs sampler is then:

$$
\begin{aligned}
\beta &\mid \gamma, \Phi, \{W_i\}, D \\
W_{ij} &\mid \beta, \gamma, \Phi, \{W_{i(-j)}\}, D \\
\gamma &\mid \Phi, \beta, \{W_i\}, D \\
\Phi &\mid \gamma, \beta, \{W_i\}, D.
\end{aligned}
$$

The first two draws are just as in MR with Σ obtained from γ, Φ and $\sigma_{11} = 1$. The third draw is a normal and the fourth draw is an inverted Wishart. The exact form of these last two draws is given in the appendix.

5   Analytical Results and Prior Assessment

We shall refer to the approaches of sections 3 and 4 as NID and ID (not identified and identified) respectively. In this section we present analytical results that help us understand and choose the priors in both approaches.

First some notation. In the NID approach, one dimension of the parameter is unidentified and a proper prior is used to ensure that the posterior is proper. In the ID approach

only the identified parameters are used. To clearly distinguish the identified parameters from the "full" set we shall henceforth refer to the parameters of the NID method as $\tilde{\beta}$ and $\tilde{\Sigma}$. We use β and Σ to denote the parameters identified by the restriction $\sigma_{11} = 1$: $\beta = \tilde{\beta}/\sqrt{\tilde{\sigma}_{11}}$ and $\Sigma = \tilde{\Sigma}/\tilde{\sigma}_{11}$.

Under the NID prior, $\tilde{\beta} \sim N(\bar{b}, A^{-1})$ and $\tilde{\Sigma}^{-1} \sim W(\nu, V)$ so that the set of prior parameters which must be chosen is $(\bar{b}, A, \nu, V)$. Under the ID prior we have $\beta \sim N(\bar{\beta}, D^{-1})$, $\gamma \sim N(\bar{\gamma}, B^{-1})$, and $\Phi^{-1} \sim W(\kappa, C)$ so that the set of prior parameters is $(\bar{\beta}, D, \bar{\gamma}, B, \kappa, C)$.

Note that in the ID case we have the option of using the standard improper choices: $D = 0$, $B = 0$, and $\kappa = 0$. In this case the choices of $\bar{\beta}$, $\bar{\gamma}$, and C do not matter. In the NID case we must choose a proper prior.

Both the NID prior and the ID prior have unusual forms. In the remainder of this section we derive results about these nonstandard priors and use these results to help us choose the prior parameters. We assume that the prior is assessed, at least approximately, by choosing appropriate marginals for β and Σ. For both priors, we obtain results about the marginal priors on the identified parameters β and Σ. We first explore the marginal prior of β in each prior and then that of Σ. Proofs of all results are available in the appendix.

5.1 The Prior for β

The simplest prior is that of β in the ID algorithm: $\beta \sim N(\bar{\beta}, D^{-1})$.

In the NID approach the prior for β is more complicated. Result 1 below gives the prior distribution of the identified coefficients using the NID prior. We need the following notation.

$$V^{-1} = \begin{bmatrix} v^{11} & v^{12} \\ v^{21} & V^{22} \end{bmatrix} \tag{8}$$

where v^{11} is 1×1, v^{12} is $1 \times (p-2)$, $v^{21} = (v^{12})'$, and V^{22} is $(p-2) \times (p-2)$. Let $v^{1.2} = v^{11} - v^{12}(V^{22})^{-1}v^{21}$.

Result 1:

Under the NID prior the marginal distribution is $\beta = \tilde{\beta}/\sqrt{\tilde{\sigma}_{11}} \sim \sqrt{v^{1.2}} \, \chi \, \tilde{\beta}$, where $\chi^2 \sim \chi^2_{\nu-p+2}$ independent of $\tilde{\beta} \sim N(\bar{b}, A^{-1})$.

Result 1 tells us that the form of the prior for β using the NID method is quite unusual: the square root of a chi-squared random variable times a normal. When $\bar{b}$ is 0, this will result in a distribution with heavy tails relative to the normal. When $\bar{b}$ is non-zero the distribution will be skewed.

A basic feature of the NID prior is that β and Σ are not independent. In result 1, we see that the prior for β depends on that of Σ through ν and $v^{1.2}$. The easiest thing to do in practice seems to be to choose the prior for Σ and then choose that of β given ν and $v^{1.2}$.

5.2   The Prior for Σ

In both the ID and the NID case, we assume that as a first step in choosing the prior parameters for Σ, we are able to specify its expected value.

$$E(\Sigma) = \begin{bmatrix} 1 & E(\gamma)' \\ E(\gamma) & E(\Phi) + E(\gamma\gamma') \end{bmatrix} = \begin{bmatrix} 1 & E(\gamma)' \\ E(\gamma) & \Delta + E(\gamma)E(\gamma)' \end{bmatrix}. \tag{9}$$

Where $\Delta = E(\Phi) + Var(\gamma)$.

Given $E(\Sigma)$, both $E(\gamma)$ and Δ are known.

5.2.1   The ID Case

Our goal is to choose values for the prior parameters $(\bar{\gamma}, B, \kappa, C)$ where $\gamma \sim N(\bar{\gamma}, B^{-1})$ and $\Phi^{-1} \sim W(\kappa, C)$. We take as given $E(\gamma)$ and Δ. Clearly, $\bar{\gamma} = E(\gamma)$. Using a standard result in multivariate analysis we have, $E(\Phi) = \frac{C}{(\kappa-p+1)}$, giving,

$$\Delta = \frac{C}{(\kappa - p + 1)} + B^{-1}. \tag{10}$$

Thus,

$$C = (\kappa - p + 1)(\Delta - B^{-1}). \tag{11}$$

Consider the simple but important case where we wish to have $E(\Sigma) = I$. Then $E(\gamma) = 0$ and $\Delta = I$. If we simplify be letting $B^{-1} = \tau I$ then specification of the prior now involves only the choice of the two scalars κ and τ! Since κ is the degrees of freedom in our inverted Wishart prior, κ and τ control the tightness of the prior with large values of of κ and small values of τ giving tighter priors. The relative variability of γ and Φ determine the prior distribution of the correlations. To see this, note that in the case $p = 3$, the single correlation in Σ is $\gamma/\sqrt{(\Phi + \gamma^2)}$. If the variance of γ is small (relative to Φ) the prior distribution of the correlation will tighten up around 0.

5.2.2 The NID Case

The prior parameters of Σ in the NID case are (ν, V) where $\tilde{\Sigma} \sim W(\nu, V)$. Our goal is to have some understanding of the implications of choices for (ν, V) on the prior distribution of $\Sigma = \tilde{\Sigma}/\tilde{\sigma}_{11}$. Since (γ, Φ) constitute a one to one reparametrization of the identified Σ, our approach is to derive the marginal priors of γ and Φ given choices of ν and V.

Result 2:

Under the NID prior the marginal distribution of γ is multivariate t with $\nu - p + 3$ degrees of freedom and

$$E(\gamma) = (V^{22})^{-1}v^{21} \text{ and } \text{Var}(\gamma) = \frac{v^{1.2}}{\nu - p + 1}(V^{22})^{-1} \tag{12}$$

Result 3:

Under the NID prior the marginal distribution of Φ^{-1} is of the form

$$\Phi^{-1} = \frac{W}{\omega} \text{ where } W \sim \text{Wishart}(\nu, (V^{22})^{-1}) \text{ and } \omega \sim v^{1.2}\chi^2_{\nu-p+2} \tag{13}$$

with W and ω independent.

From result (3) we have:

$$E(\Phi) = \frac{v^{1.2}(\nu - p + 2)}{\nu - p + 1}(V^{22})^{-1}. \tag{14}$$

9

To use these results, we again suppose that we are able to specify the expected value of Σ so that $E(\gamma)$ and Δ are given. Our results give $\Delta = E(\Phi) + Var(\gamma) = v^{1.2}(V^{22})^{-1}\frac{\nu-p+3}{\nu-p+1}$. Given choices for $v^{1.2}$ and ν we can obtain $(V^{22})^{-1}$ from Δ. Given $(V^{22})^{-1}$, we can obtain v^{21} from $E(\gamma)$. Thus, given $E(\Sigma)$ we have only the two numbers ν and $v^{1.2}$ left to choose in order to specify the prior.

As a simple example we can let (as we did in the ID case) $E(\gamma) = 0$ and $\Delta = I$. This implies that $v^{12} = 0$, $v^{1.2} = v^{11}$ and

$$V^{22} = \frac{v^{11}(\nu - p + 3)}{(\nu - p + 1)}I. \tag{15}$$

Given these choices, v^{11} simply scales the distribution of $\tilde{\Sigma}$ up and down by a factor which cancels out for Σ so that the choice of v^{11} does not affect the distribution of Σ. Of course it does affect the distribution of β as a scale factor, but this can always be adjusted by choice of A. Consequently, in this simple setup, we can assume without loss of generality that $v^{11} = 1$. We are now left with only the choice of ν in order to specify the prior of Σ!

6   Simulated Examples

In this section we apply the ID approach to two simulated examples. The examples are the same as those of MR. In the first example $p = 3$ and in the second $p = 6$.

We consider two versions of the ID prior. In both cases we choose $D = 0$, an improper choice for for the prior on β. In choosing our prior for Σ, we center the prior on the I so that (as in section 5.2.1) all we must specify are values for κ and τ. In our first prior we choose $\kappa = p + 2$ and $\tau = 1/8$. In our second prior we choose $B = 0$ and $\kappa = 0$ so that the prior on Σ is improper as well.

In both priors we avoid making choices about the prior for β. In the first prior we have chosen to roughly center our prior for Σ at the identity matrix. In the second prior we avoid making choices for Σ as well.

Each run of a Gibbs sampler is started at the initial values $\beta = 0$ and $\Sigma = I$. Since we draw W first, there is no need to specify initial values.

6.1   Simulated Example with $p = 3$

With $p = 3$, Σ is 2×2 so there is just one unconstrained variance and one correlation. Figure 1 displays the marginal prior distributions of the single correlation (top panel) and the single variance given the choices of the first ID prior: $\kappa = (p+2) = 5$ and $\tau = 1/8$. The histograms are constructed from 10,000 iid draws from the prior. We see that the prior for the correlation is centered at zero but spreads out towards ± 1. The prior for the variance has mean one and a long right tail. Of course, we cannot display the any marginal priors for the second prior choice since it is improper on both β and Σ.

Data was simulated as follows. The matrix X has just one column and the corresponding true value of the single coefficient equals -1.414. Each of the X values was drawn iid from the uniform distribution on the interval $(-.5, .5)$. The true value of the correlation is .5 and the true value of the unconstrained variance is 2. 3000 observations were simulated.

Figure 2 displays the posteriors of the three parameters (β, σ_{22}, and ρ_{12}) obtained from the first choice of ID prior. In each histogram the solid line is drawn at the value of the true parameter. These histograms are based on 10,000 iterations of the Gibbs sampler outlined in section 4 (after discarding a few initial burn in draws). Figure 3 displays the results using the second prior (improper on Σ). The results displayed in figures 2 and 3 are very similar and both appear to be quite reasonable.

Figure 4 displays additional detail for the draws of the variance. Both the time series plot of the draws and the autocorrelation function are shown. Rows 1, 2, and 3 of the figure display results for the NID sampler, ID sampler with our first prior choice, and the ID sampler with the second prior choice respectively. The prior for the NID sampler was $\bar{b} = 0$, $A^{-1} = 100I$, $\nu = 6$, and $V = \nu I$. In all three cases the autocorrelation functions die off as the lag increases. Clearly, the draws from the ID samplers are more highly autocorrelated than those of the NID. The draws from the ID sampler using the

second prior are more highly autocorrelated than those of the first. For example the 10th autocorrelation is .77 for the second prior and .71 for the first.

This example illustrates a basic feature of the Gibbs sampler of section 4 for the ID prior. The more diffuse the prior on Σ is, the slower the autocorrelations die out. We have tried many simulated examples and found this to be generally true.

6.2   Simulated Example with $p = 6$

In this example there is again just one column in the X matrix and its values are iid draws from the uniform distribution on the interval $(-2, 2)$. The true value of the coefficient is .89. With $p = 6$, Σ is a 5×5 matrix so there are 4 unconstrained variances and 10 correlations. The true values of the 4 variances are .8, .6, .4, and .2 as we go down the diagonal. All true correlations are .5. 1600 observations were generated.

Figure 5 displays four of the marginal posteriors obtained from our first ID prior ($\kappa = 8$, $\tau = 1/8$). These histograms are based on 30,000 iterations of the Gibbs sampler. Again, in each case the solid line depicts the true value of the parameter. The top left panel is β, the top right is σ_{22}, the bottom left is ρ_{12} and the bottom right is ρ_{23}. As in the $p = 3$ case the sampler seems quite successful.

The sampler run using the second prior, which is improper on Σ, got stuck at a Σ which was almost singular at about the 15,000th iteration. In the next section we discuss the relationship between the performance of the sampler and the choice of prior.

7   Marginal Priors for Eigenvalues

The examples of the previous section show the choice of ID prior has an effect on the performance of the corresponding Gibbs sampler. In this section, we explain how the prior affects the performance of the sampler.

The sampler in the $p = 6$ example failed because we got stuck in a region of the parameter space where Σ was nearly singular. This suggests that the prior may be guiding

the sampler to matrices of this type. In order to get a feeling for this we use the smallest eigenvalue as a measure of how close to singularity a particular Σ may be. We now examine the prior distribution of the smallest eigenvalue.

Figure 6 displays the marginal prior distribution of the smallest eigenvalue of Σ for two choices of the ID prior in the case $p = 6$ (so Σ is 5×5). The top panel corresponds to the choices $\kappa = p + 1$ and $\tau = 1/2$. The bottom panel corresponds to the choices $\kappa = p + 2$ and $\tau = 1/8$ used in section 6. These two priors are markedly different. In the top panel most of the mass is on values less than .1 while in the bottom panel most of the mass is on values greater than .1. Since the identity matrix has smallest eigenvalue equal to 1, it makes sense that if we tighten up our prior around $\Sigma = I$ we will move the marginal prior of the smallest eigenvalue towards 1.

Given that successively less informative proper priors put more prior weight on very small eigenvalues, one might expect that the improper prior prior on Σ might effectively put even greater weight on tiny values of the smallest eigenvalue. However, since we cannot simulate from the prior distribution of the smallest eigenvalue in the case of the improper prior, our approach is to use this prior with a very small simulated sample of data and check the posterior for influence from this prior. We simulated 7 observations from the $N_5(0, I)$ distribution and then computed the posterior of Σ given these observations. The idea is that the posterior from a small data set will largely reflect features of the prior. The data set is chosen to be large enough to turn the improper prior into a proper posterior. We then computed the marginal *posterior* distribution of the smallest eigenvalue of Σ. We did the same thing for the two priors displayed in figure 6. The results are displayed in figure 7. The top, middle, and bottom panels, correspond to the improper, $\kappa = p + 1$ and $\tau = 1/2$, and $\kappa = p + 2$ and $\tau = 1/8$ priors. The posteriors are very different. The posteriors for the proper priors are what we might expect given the priors (figure 6). The posterior obtained from the improper prior shows that it is very informative. The improper prior pushes the posterior towards Σ matrices such that the smallest eigenvalue is very small. Clearly, the improper prior is very informative about the smallest eigenvalue. This explains

why the sampler got stuck using the improper prior. Note that this can only happen if the likelihood is not sufficiently informative to overwhelm the prior. Recall that in the $p = 3$ example the sampler based on the improper prior had no problem.

8 Advantages and Disadvantages of ID Prior

The NID prior approach is most useful in situations in which a proper but fairly diffuse prior is desired. However, the results in section 5 show that it may be difficult to assess a truely informative prior on β using the NID approach. This is particularly true if a prior mean other than zero is desired. In contrast, the ID approach can be used to assess a standard normal prior directly on β. There are a number of situations in which informative priors are desireable. For example, hierarchical models for situations in which the data has a panel or grouped structure have become increasing popular. The heart of the hierarchical model is an informative prior on the coefficients. Typically, we would assume that each panel member j corresponds to a set of coefficients β_j, and use a prior,

$$\beta_j \sim N(\bar{\beta}, V_\beta). \tag{16}$$

In our view, this hierarchical model makes the most sense when applied to the identified coefficients (see Rossi et al (1996) for an example). The ID approach can be extended easily to handle a variety of hierarchical models of this sort.

Informative priors on β are also useful in situations in which the investigator has prior information from subject matter theory or experience with similar datasets. For example, a multinomial model for choices between different brands of similar products as in Nevo (1997) would feature a price coefficient which is certainly negative and never much less than -20 or so.

Note that this approach to prior specification should be viewed as a way to roughly gauge the prior since we assume that the prior is assessed by separately choosing marginals for β and Σ. Given the nature of the MNP model, prior information should involve dependence between β and Σ. For example, prior information may be about the implied

probabilities rather than directly about β and Σ which would imply dependence. For the simpler multinomial logit model (the main competitor to the MNP) it is possible to assess the prior in a more natural way (see Koop and Poirier (1993)). For the more complicated MNP model prior assessment is a difficult problem. If, for example, the researcher had prior information about the underlying utility maximization process leading to the multinomial data it might be possible to specify specify a prior on the full set of parameters ($\tilde{\beta}$ and $\tilde{\Sigma}$) and actually use that data to learn about the unidentified parameters (see Poirier (1998)).

Some investigators are uncomfortable with informative priors and would like to use improper priors. The NID approach cannot be used with improper priors on either β or Σ. The ID approach can easily handle an improper uniform priors on β. In principle, we can also used improper priors on Σ as well. However, as shown in 6 above, the improper prior on Σ is actually an extremely informative prior on the smallest eigenvalue of the sigma matrix. In some situations, the Gibbs sampler based on the ID improper prior on Σ can get stuck on a near singular matrix. This will happen with highest probabilty in high dimensional problems with a small amount of data. Note that in practice is very easy to identify when the sampler is stuck so that there is no possibility of actually reporting incorrect results.

There is an accumulating body of evidence in the statistics literature on covariance matrix estimation that a modest amount of shrinkage on the eigenvalues or correlations will produce estimators with good risk properties (see Yang and Berger (1994) and Kass and Daniels (1998)). Thus, an improper prior on Σ has at least three undesireable aspects: 1. it is actually a very informative prior on the smallest eigenvalue, 2. the sampling properties of Bayes estimators based on this prior are apt to be poor and 3. our ID Gibbs sampler may experience convergence problems with this prior. For these reasons, we advocate the use of a weakly informative default prior on Σ (centered on I) in the absense of strong prior information. The improper prior on Σ can be used as a diagnostic for prior senstivity, if desired.

Other possible approaches to assessing priors directly on the identified parameters in-

clude using the prior of Barnard, Meng and McCulloch (1996)(see, also, in McCulloch and Rossi (1996)) and the approach of Chib et al (1998). In the Barnard et al approach, Σ is written as $\text{Diag}(S)R\text{Diag}(S)$ and various priors are used on the standard deviations and correlations. This prior can be implemented in a Griddy Gibbs algorithm since the relevant range of each correlation can be expressed as function of all other correlations, allowing a one by one draw of R. The Griddy Gibbs algorithm is reliable but it can be slow and requires the choice of grid size and fineness tuning parameters. Some additional work would be required to assess truely informative priors on R. Chib et al (1998) propose using the Cholesky root parameterization with the diagonals parameterized to insure positivity and set $\sigma_{11} = 1$. As Chib et al (1998) discuss, it would be extremely difficult to assess an informative prior in this parameterization. The authors use a prior which is assessed based on preliminary estimates of the covariance matrix and asymptotic variances. A Metropolis algorithm is used with a t-style candidate sampling density. Tail and shape tuning parameters must be assessed to insure proper functioning of the Metropolis algorithm. A basic advantage of the approach presented in this paper is that we are able to obtain the analytical results of sections 5. These results help us understand the prior and guide its choice. It seems unlikely that there is any other way to specify a prior such that $\sigma_{11} = 1$ using standard distributions for which analytical results are available.

The advantages, then, of our ID approach is that we can use both truely informative or strictly improper priors on β and Σ and the MCMC algorithm can be implemented using standard conjugate normal and Wishart draws with no tuning parameters. Because standard distributions are used analytical results on the properties of the prior are available. The cost of using the ID approach (vis a vis the NID approach) is that chain defined by the ID Gibbs sampler has higher autocorrelation and is more sensitive to initial conditions than the NID Gibbs sampler or the improved hybrid NID sampler proposed by Nobile (1998). Fortunately, we have found that the ID Gibbs sampler is computationally tractible and that these problems can be avoided using longer draw sequences.

Finally, the prior developed in this paper is useful in any situation in which the marginal

prior distribution of the (1,1) element of a covariance matrix can be specified. In the case the MNP model, we focus on the special case in which this distribution is degenerate around the value 1. An earlier working paper version of this paper (McCulloch, Polson and Rossi(1994)) has already stimulated the use of this prior for switching regression models by Koop and Poirier (1997) and for strucutural equations models with limited dependent variables (Li (1996)). Jacquier, Polson and Rossi (1994) use this prior to model correlation between innovations in the level and volatility of time series. Ainslie (1998) uses the ID prior in an extension of the standard MNP model to consider purchases of the outside good.

References

Ainslie, A. (1998), "Similarities and Differences in Brand Purchase Behavior Across Categories," unpublished dissertation, University of Chicago.

Albert, J. and S. Chib (1993), "Bayesian analysis of binary and polychotomous data," *JASA* 88, 669-679.

Barnard, J., X. Meng, R. McCulloch (1996), "Modeling Covariance Matrices in terms of Standard Deviations and Correlations, with Application to Shrinkage," working paper, University of Chicago.

Borsch-Supan, A. and Hajivassiliou, V. (1993), "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models", *Journal of Econometrics*, 58, 347-368.

Bunch, D. S. (1991), "Estimability in the Multinomial Probit Model," *Transportation Research-B*, 25B, 1-12.

Casella, G. and George E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.

Chib, S., E. Greenberg, and Y. Chen (1998), "MCMC Methods for Fitting and Comparing Multinomial Response Models," Working Paper, Olin School of Business, Washington Univerity.

Danzie, B. (1985), "Parameter Estimability in the Multinomial Probit Model," *Transportation Research-B*, 19B:6, 526-28.

Gelfand, A. and Smith A.F.M. (1990), "Sampling Based Approaches to Calculating Marginal Densities," *JASA*, 85, 972-985.

Hajivassiliou, V. and McFadden D. (1990), "The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises", working paper, Yale University.

Jacquier, E., N. Polson and P. E. Rossi (1994), "Stochastic Volatility: Univariate and Multivariate Extensions," working paper, University of Chicago.

Kass, R. and M. Daniels (1998), "Nonconjugate Bayesian Estimation of Covariance Matrices and its Use in Hierarchical Models," forthcoming *JASA*.

Keane, M. (1992), "A Note on Identification in the Multinomial Probit Model", *Journal of Business and Economics Statistics*, 10, 193-200.

Keane, M. (1994), "A Computationally Practical Simulation Estimator for Panel Data", *Econometrica*, 62, 95-116.

Li, Kai (1996), "Essays in Bayesian Financial Econometrics," PhD thesis, Department of Economics, University of Toronto.

Koop, G. and D. Poirier (1993), "Bayesian Analysis of Logit Models using Natural Conjugate Priors," *Journal of Econometrics*, 56, 323-340.

Koop, G. and D. Poirier (1997), "Learning About the Across-Regime Correlation in Swithcing Regression Models," *Journal of Econometrics*, 78, 217-228.

McCulloch, R. and Rossi, P. E. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207-240.

McCulloch, R. , N. G. Polson and P. E. Rossi (1994), "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," working paper, Graduate School of Business, University of Chicago.

McCulloch, R. and P. E. Rossi (1996), "Bayesian Analysis of the Multinomial Probit Model," forthcoming *Simulation-Based Inference in Econometrics*, (Mariano, Weeks and Schuermann, eds), Cambridge: Cambridge University.

McFadden D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57, 995-1027.

Nobile, A. (1995), "A Hybrid Markov Chain for the Bayesian Analysis of the Multinomial Probit Model," *Statistics and Computing*, 8, 229-242.

Nevo, A. (1997), "Measuring Market Power in the Ready-to-eat Cereal Industry," working paper, University of California, Berkeley.

Poirier, D. (1998), "Revising Beliefs in Nonidentifed Models," *Econometric Theory*, 14, 483-509.

Smith, A.F.M. and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *JRSS B*, 55, 3-25.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, School of Statistics, University of Minnesota.

Yang, R. and J. O. Berger (1994), "Estimation of a Covariance Matrix using the Reference Prior," *Annals of Statistics*, 22, 1195-1211.

Appendix

In this appendix, we derive the marginal distributions of γ and Φ under the prior used by MR (the NID prior). We also present the exact form of the additional conditional distributions used in the second Gibbs sampling algorithm of section 4.

Marginals of γ and Φ

Let $\tilde{\Sigma}$ denote the variance matrix of ϵ in equation(1) above and Σ denote the matrix of identified parameters. We then have:

$$\Sigma = \tilde{\Sigma}/\tilde{\sigma}_{11}. \tag{1}$$

In the first algorithm, the prior on $\tilde{\Sigma}$ is defined by $G = \tilde{\Sigma}^{-1} \sim \text{Wishart}(\nu, V)$ and the prior on Σ is then the marginal prior induced by equation(1). Note that we define ν and V to be such that $E(G) = \nu V^{-1}$.

It is useful to partition the $(p-1) \times (p-1)$ matrices Σ, G, and V as follows:

$$\Sigma = \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \Sigma_{22} \end{bmatrix} \quad G = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & G_{22} \end{bmatrix} \quad V^{-1} = \begin{bmatrix} v^{11} & v^{12} \\ v^{21} & V^{22} \end{bmatrix} \tag{2}$$

Here, σ_{12} is $1 \times (p-2)$, $\sigma_{2,1}$ is $(p-2) \times 1$, Σ_{22} is $(p-2) \times (p-2)$, and $\sigma_{12} = \sigma'_{21}$ and the partitions of the other matrices are dimensioned in the same way.

The parameters γ and Φ are defined as functions of Σ by:

$$\gamma = \sigma_{21} \quad \text{and} \quad \Phi = \Sigma_{22} - \sigma_{21}\sigma_{12}. \tag{3}$$

Equations (1) and (3) define γ and Φ as functions of G. We proceed by first deriving these functions in an explicit form and then obtaining the marginal distributions.

Using standard results on the inverse of a partitioned matrix, we have:

$$G^{-1} = \begin{bmatrix} (g_{11} - g_{12}G_{22}^{-1}g_{21})^{-1} & -g_{12}G_{22}^{-1}(g_{11} - g_{12}G_{22}^{-1}g_{21})^{-1} \\ -G_{22}^{-1}g_{21}(g_{11} - g_{12}G_{22}^{-1}g_{21})^{-1} & G_{22}^{-1} + G_{22}^{-1}g_{21}g_{12}G_{22}^{-1}(g_{11} - g_{12}G_{22}^{-1}g_{21})^{-1} \end{bmatrix}, \tag{4}$$

so that,

$$\Sigma = \begin{bmatrix} 1 & -g_{12}G_{22}^{-1} \\ -G_{22}^{-1}g_{21} & (g_{11} - g_{12}G_{22}^{-1}g_{21})G_{22}^{-1} + G_{22}^{-1}g_{21}g_{12}G_{22}^{-1} \end{bmatrix}. \tag{5}$$

From equation (3) we then have:

$$\gamma = -G_{22}^{-1}g_{21} \quad \text{and} \quad \Phi = (g_{11} - g_{12}G_{22}^{-1}g_{21})G_{22}^{-1}, \tag{6}$$

under the prior of the first algorithm.

To obtain the marginal distributions of γ and Φ write $G = \sum_{i=1}^{\nu} Z_i Z_i'$ where $Z_i \sim N(0, V^{-1})$ iid. Then let $Z_i' = (Y_i, X_i')'$ with $Y_i \in R$ and $X_i \in R^{p-2}$. With $Y' = (Y_1, Y_2, \ldots, Y_\nu)'$ and $X' = (X_1, X_2, \ldots, X_\nu)$ we have:

$$G = \begin{bmatrix} Y'Y & Y'X \\ X'Y & X'X \end{bmatrix}. \tag{7}$$

Now note that the conditional distribution of $Y_i|X_i$ is $N(X_i'(V^{22})^{-1}v^{21}, (v^{11}-v^{12}(V^{22})^{-1}v^{21}))$. Since $g_{11} - g_{12}G_{22}^{-1}g_{21}$ is the residual sum of squares from the regression of Y on X, its distribution is $(v^{11} - v^{12}(V^{22})^{-1}v^{21})\chi^2_{\nu-p+2}$ given X and hence it is independent of X, and its marginal distribution is its conditional. Clearly, $X'X = G_{22} \sim \text{Wishart}(\nu, (V^{22})^{-1})$. Thus Φ^{-1} has the distribution of a Wishart divided by an independent χ^2:

$$\Phi^{-1} = \frac{G_{22}}{g_{11} - g_{12}G_{22}^{-1}g_{21}} \sim \frac{\text{Wishart}(\nu, V^{22})^{-1})}{(v^{11} - v^{12}(V^{22})^{-1}v^{21})\chi^2_{\nu-p+2}}. \tag{8}$$

The expected value of Φ^{-1} is given by $E(\Phi^{-1}) = E(\text{Wishart}(\nu, (V^{22})^{-1}))E(1/\chi^2_{\nu-p+2})(v^{11} - v^{12}(V^{22})^{-1}v^{21})^{-1} = \nu V^{22}(\nu - p)^{-1}(v^{11} - v^{12}(V^{22})^{-1}v^{21})^{-1}$.

For the distribution of γ we have $-\gamma = G_{22}^{-1}g_{21} = (X'X)^{-1}X'Y$. Given X we have $-\gamma \sim N((V^{22})^{-1}v^{21}, (v^{11} - v^{12}(V^{22})^{-1}v^{21})(X'X)^{-1})$. The joint distribution of (γ, G_{22}) is now seen to be of the same form as that of the conjugate prior for the mean and covariance matrix in the analysis of iid samples from the multivariate normal distribution. Hence the marginal distribution of γ is multivariate t with $\nu - p + 3$ degrees of freedom and moments $E(E(-\gamma \mid X)) = E((V^{22})^{-1}v^{21}) = (V^{22})^{-1}v^{21}$ and $Var(\gamma) = E(Var(\gamma \mid X)) = E((v^{11} - v^{12}(V^{22})^{-1}v^{21})(X'X)^{-1}) = (v^{11} - v^{12}(V^{22})^{-1}v^{21})E((X'X)^{-1}) = (v^{11} - v^{12}(V^{22})^{-1}v^{21})(V^{22})^{-1}(\nu - p + 1)^{-1}$. The last equality follows from the fact that if W is $p \times p$ and $W \sim \text{Wishart}(\nu, A)$ then $E(W^{-1}) = A(\nu - p - 1)^{-1}$.

Conditional Distributions for γ and Phi Draws

As discussed in section 4, drawing γ and Φ is like drawing from the posterior distribution of the multivariate regression of the last $p - 2$ ϵ's on the first ϵ. In the notation of section 4, let U_i be the ith observation of the first ϵ and Z_i the the ith observation of last $p - 2$ ϵ's. Let $U' = (U_1, U_2, \ldots U_n)'$ and $Z' = (Z_1, Z_2, \ldots, Z_n)'$. Then we have the multivariate regression $Z = U\gamma' + \varepsilon$ where the rows of ε are $N(0, \Phi)$ iid.

Given γ and the conjugate prior $\Phi^{-1} \sim \text{Wishart}(\kappa, C)$, the posterior for Φ^{-1} is $\text{Wishart}(\kappa + n, C + (Z - U\gamma')'(Z - U\gamma'))$.

Given Φ we have $Vec(Z') = (U \otimes I)\gamma + Vec(\varepsilon')$, with $Vec(\varepsilon') \sim N(0, I \otimes \Phi)$. If we write $(I \otimes \Phi^{-1/2})Vec(Z') = (I \otimes \Phi^{-1/2})(U \otimes I)\gamma + (I \otimes \Phi^{-1/2})Vec(\varepsilon')$ and $\gamma \sim N(\bar{\gamma}, B^{-1})$ we then have a standard univariate regression with conjugate prior. From this we obtain $\gamma \sim N(A_\gamma(Vec(\Phi^{-1}Z'U) + B\bar{\gamma}), A_\gamma)$ where $A_\gamma = (U'U\Phi^{-1} + B)^{-1}$.

Figure 1.
ID Prior Distributions

Correlation
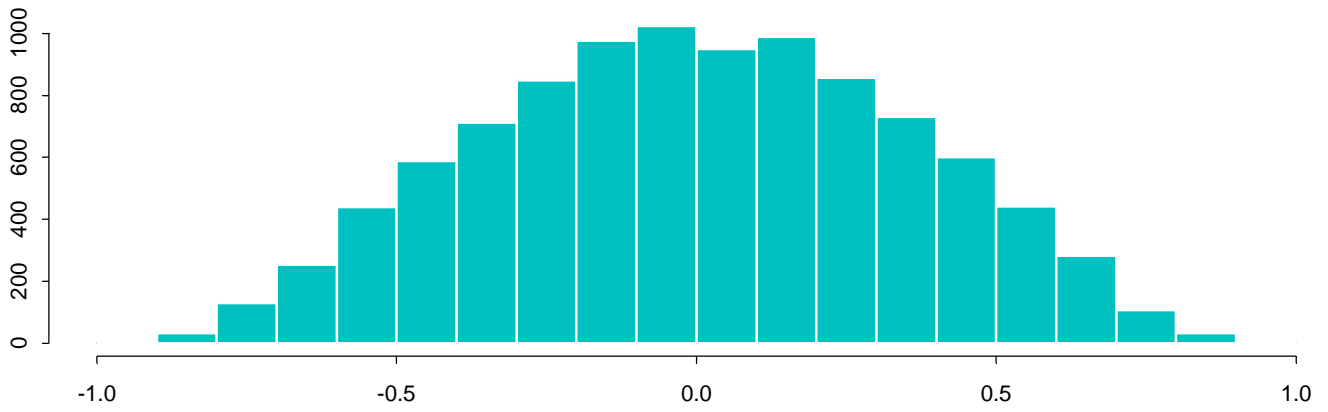
Variance (σ_{22})
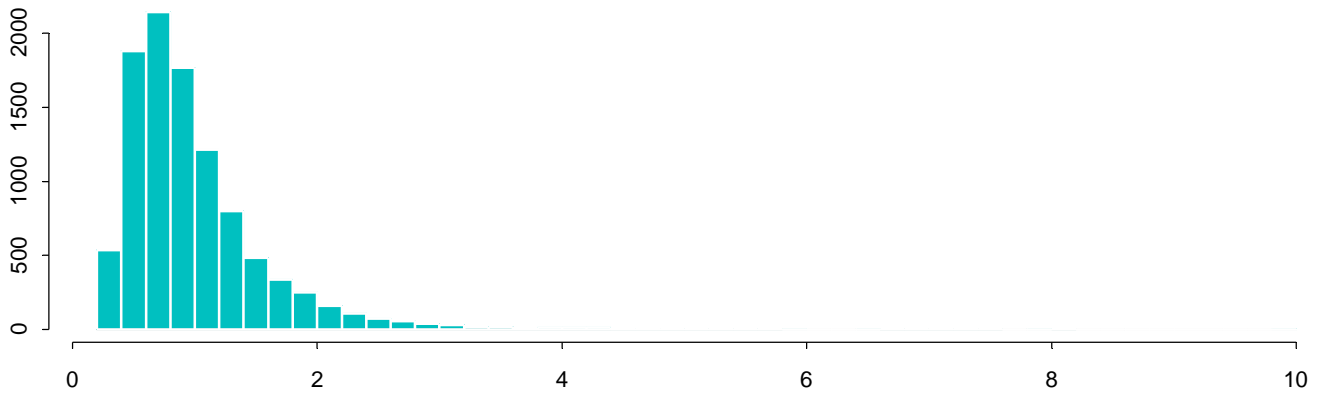
Figure 2.
Posterior Distributions - Proper ID prior
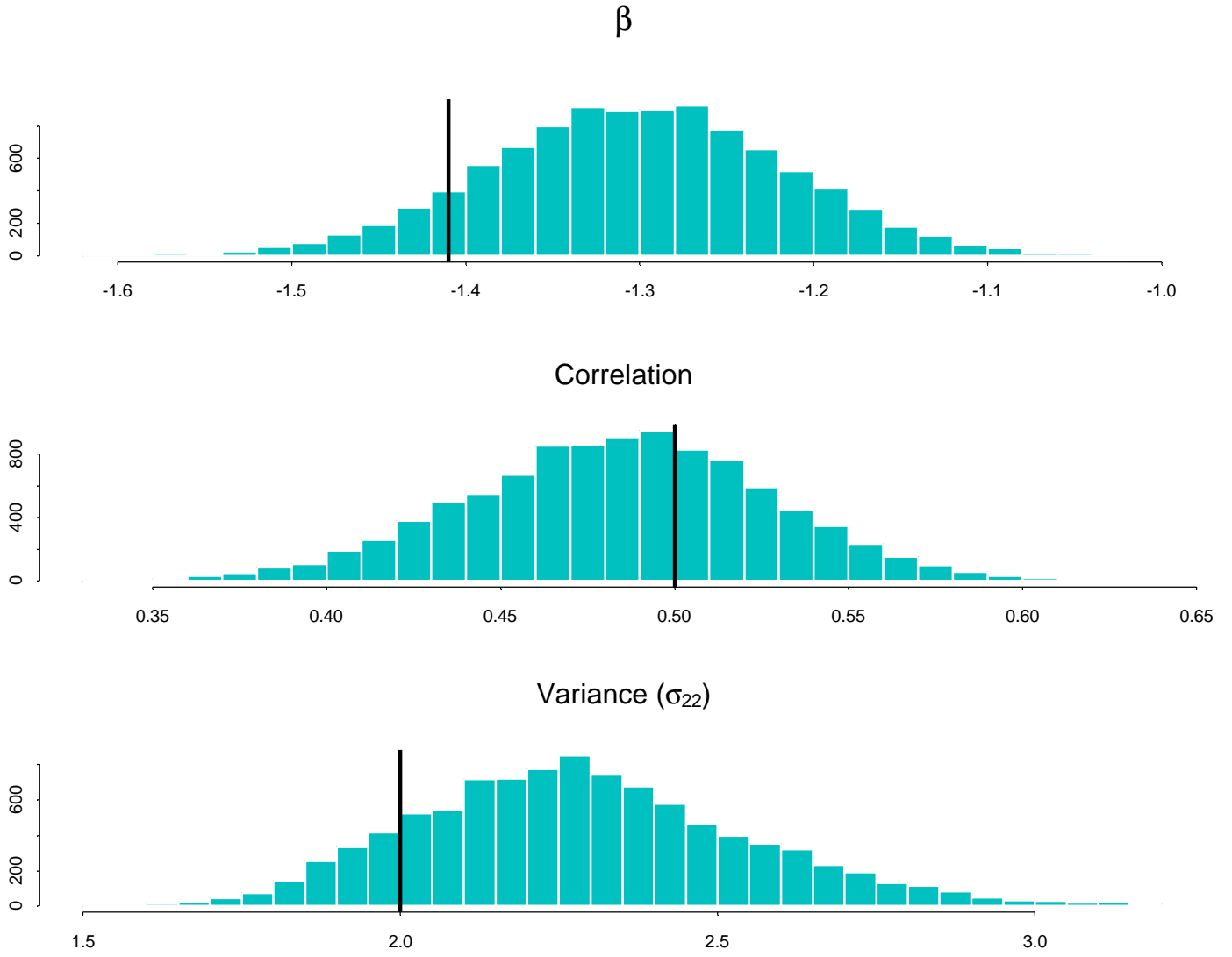
Figure 3.
Posterior Distributions - Improper ID Prior
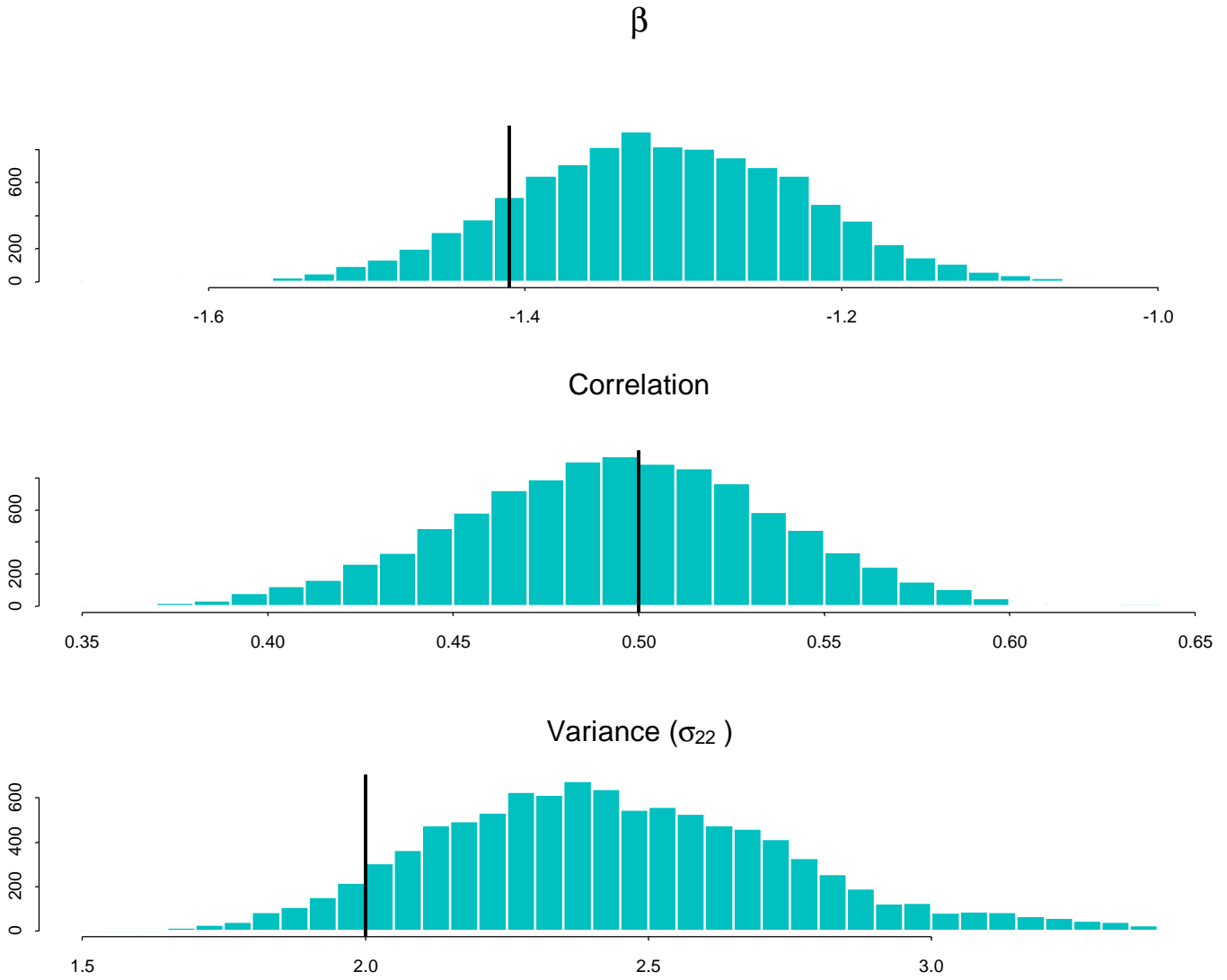
Figure 4.
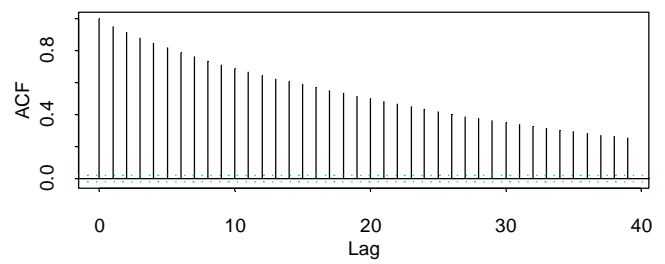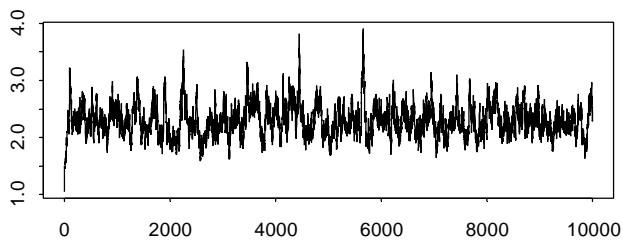## Time Series Properties of NID, ID proper and ID Improper Samplers

NID Sampler

Series : s22
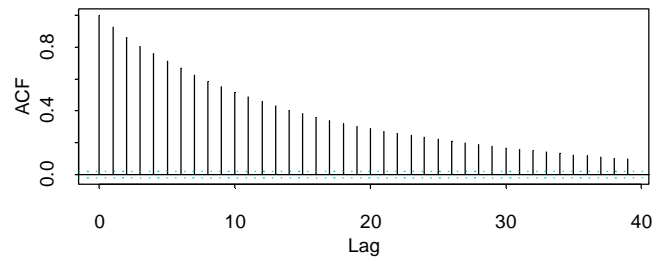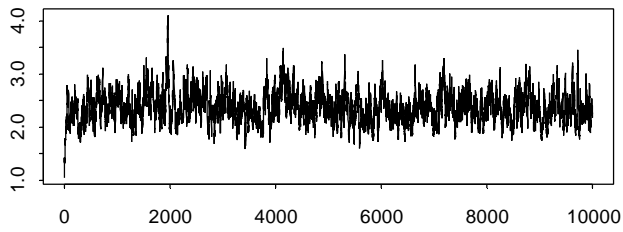
ID Proper

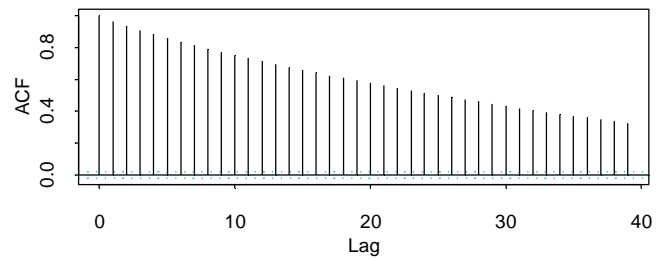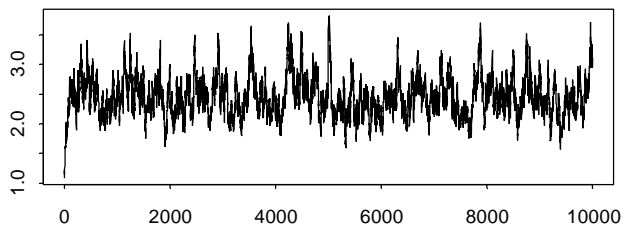Series : s22

ID Improper

Series : s22

Figure 5.
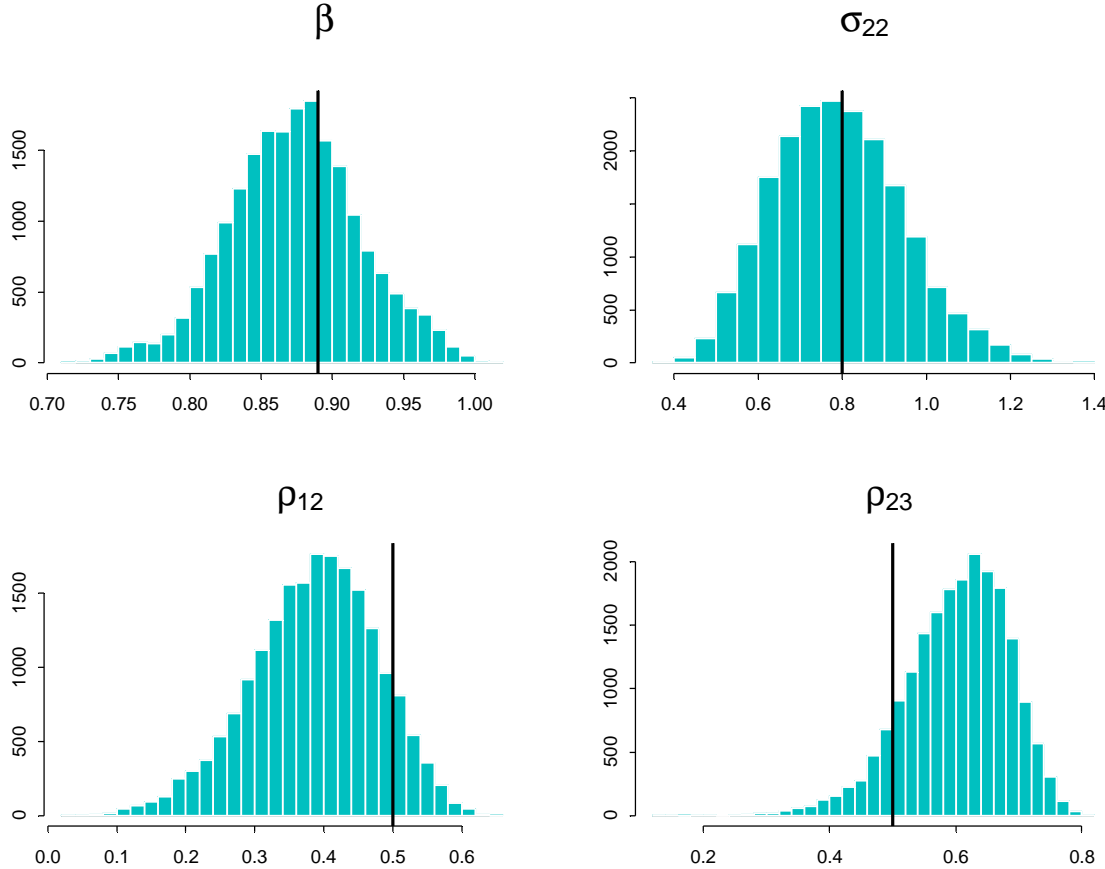Posterior distributions of Model Parameters: p = 6 example
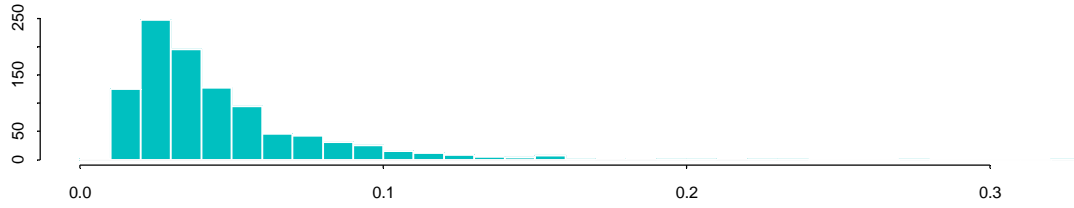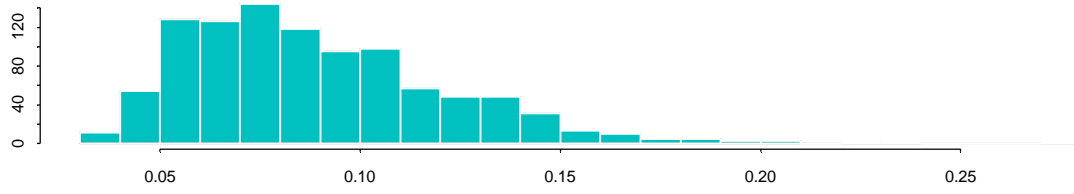
Figure 6.
Prior Distributions of Smallest Eigenvalue
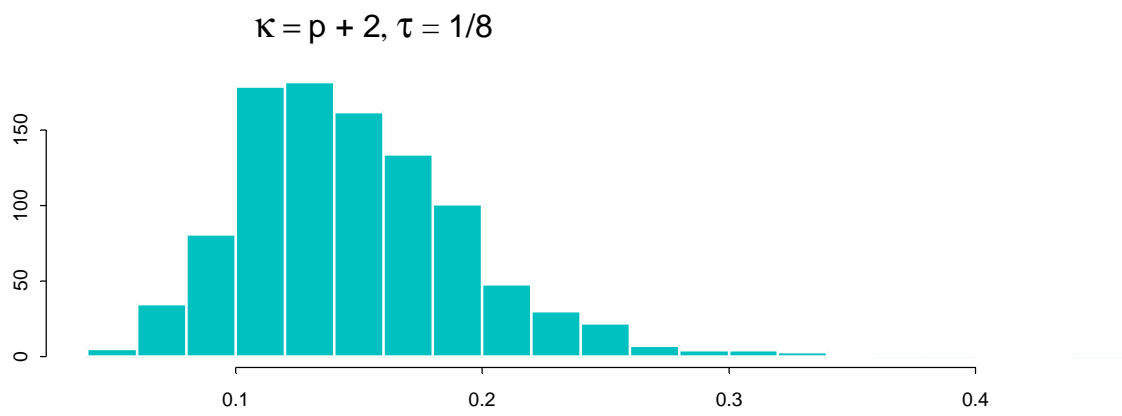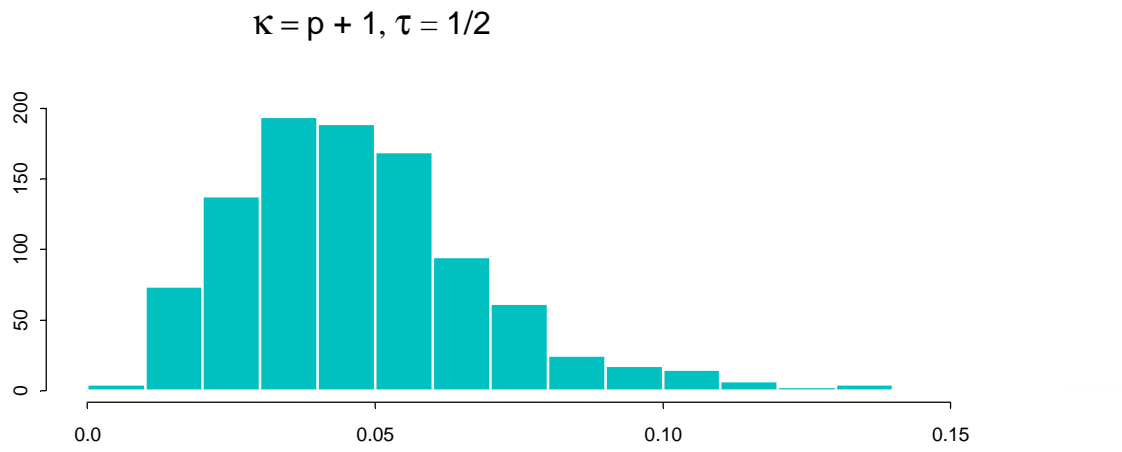
$\kappa = p + 1, \tau = 1/2$

$\kappa = p + 2, \tau = 1/8$

Figure 7.
Posterior Distribution of Smallest Eigenvalue - Small Dataset