Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 124 (2005) 311–334
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

A Bayesian analysis of the multinomial probit model using marginal data augmentation

Kosuke Imai[a,*], David A. van Dyk[b]

[a] *Department of Politics, Princeton University, Princeton, NJ 08544, USA*
[b] *Department of Statistics, University of California, Irvine, CA 92697, USA*

Accepted 9 February 2004

Abstract

We introduce a set of new Markov chain Monte Carlo algorithms for Bayesian analysis of the multinomial probit model. Our Bayesian representation of the model places a new, and possibly improper, prior distribution directly on the identifiable parameters and thus is relatively easy to interpret and use. Our algorithms, which are based on the method of marginal data augmentation, involve only draws from standard distributions and dominate other available Bayesian methods in that they are as quick to converge as the fastest methods but with a more attractive prior specification. C-code along with an R interface for our algorithms is publicly available.[1]

JEL classification: C11; C25; C35

Keywords: Bayesian analysis; Data augmentation; Prior distributions; Probit models; Rate of convergence

1. Introduction

Discrete choice models are widely used in the social sciences and transportation studies to analyze decisions made by individuals (see, e.g., Maddala, 1983; Ben-Akiva and Lerman, 1985). Among such models, the multinomial probit model is often appealing because it lacks the unrealistic assumption of independence of irrelevant alternatives of logistic models (see, e.g. Hausman and Wise, 1978). Despite this appeal, the model is sometimes overlooked because model fitting can be computationally demanding owing

* Corresponding author. Tel.: +1-609-258-6601; fax: +1-609-258-1110.
E-mail addresses: kimai@princeton.edu (Kosuke Imai), dvd@ics.uci.edu (David A. van Dyk).

[1] R is a freely available statistical computing environment that runs on any platform. The R software that implements the algorithms introduced in this article is available from the first author's website at http://www.princeton.edu/~kimai/.

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2004.02.002

to the required high-dimensional integrations. Recent advances in Bayesian simulation, however, have shown that Gibbs sampling algorithms based on the method of data augmentation can provide reliable model fitting (Geweke et al., 1994). Hence, the development of efficient Markov chain Monte Carlo (MCMC) algorithms has been a topic of much recent work; see, e.g., McCulloch and Rossi (1994), Chib et al. (1998), Nobile (1998), McCulloch et al. (2000), Nobile (2000), and McCulloch and Rossi (2000).

The basic computational strategy of the proposed MCMC methods is to identify an underlying set of Gaussian latent variables, the relative magnitudes of which determines the choice of an individual. Because the natural parameterization of this model is unidentifiable given the observed choice data, a proper prior distribution is required to achieve posterior propriety. As proposed by McCulloch and Rossi (1994), a Monte Carlo sample of the identifiable parameters can then be recovered and be used for Monte Carlo integration in a Bayesian analysis. A complication involved in this procedure is that the prior distribution for the identifiable model parameters is determined as a byproduct. Inspection (e.g., via simulation) is therefore required to determine what prior distribution is actually being specified and how sensitive the final results are to this specification.

To improve the computational performance of McCulloch and Rossi's (1994) algorithm (but without addressing the difficulties in the prior specification), Nobile (1998) introduced a "hybrid Markov chain." This hybrid is quite similar to the original algorithm but adds an additional Metropolis step to sample the unidentifiable parameters and *appears* to dramatically improve the performance (i.e., mixing) of the resulting Markov chains. We illustrate that the improved mixing of Nobile's hybrid method seems to be primarily for the unidentifiable parameter—the gain for the identifiable model parameters is much smaller, at least in terms of the autocorrelation of their Monte Carlo draws. Nonetheless, Nobile's method has an advantage over McCulloch and Rossi (1994) in that it can be less sensitive to starting values. In addition to this clarification of the improvement offered by Nobile's method, we point out an error in Nobile's derivation which can significantly alter the stationary distribution of the resulting Markov chain and thus hamper valid inference.

A second computational innovation was introduced by McCulloch et al. (2000) and aims to address the difficulties with prior specification (but without addressing the computational speed of the algorithm). In particular, this proposal specifies a prior distribution only on the identifiable parameters and constructs a Markov chain that fixes the unidentifiable parameter. Unfortunately, as pointed out by McCulloch et al. (2000) and Nobile (2000), the resulting algorithm can be much slower to converge than either the procedure of McCulloch and Rossi (1994) or of Nobile (1998).

To clarify comparisons among existing algorithms and the algorithms we introduce, we specify three criteria: (1) the interpretability of the prior specification, (2) the computational speed of the algorithm, and (3) the simplicity of implementation. Our comparisons among the three existing algorithms appear in Table 1, which indicates that none of the algorithms dominates the others.

The primary goal of this article is to introduce new algorithms that perform better than the existing algorithms when evaluated in terms of these three criteria. That is, our

Table 1
Ranking three MCMC algorithms for fitting the multinomial probit model in terms of the interpretability of the prior specification, computational speed as measured by the autocorrelation of the Monte Carlo draws, and simplicity of implementation

Algorithm	Criteria		
	Prior	Speed	Simplicity
McCulloch and Rossi (1994)			Best[a]
Nobile (1998)		Best[b]	
McCulloch et al. (2000)	Best		

[a]The algorithms of both McCulloch and Rossi (1994) and McCulloch et al. (2000) require only draws from standard distributions; the latter is, however, more involved.

[b]Although the gain in terms of the autocorrelation of the identifiable parameters is small, the sampler in Nobile (1998) when compared with that of McCulloch and Rossi (1994), can be less sensitive to starting values.

algorithms are at least as good as the best of the three existing algorithms when measured by any of the three criteria. In particular, our algorithms are as fast as Nobile's (1998) algorithm, are not particularly sensitive to starting values, do not require a Metropolis step, directly specify the prior distribution of the identifiable regression coefficients, and can handle flat prior distributions on the coefficients.

The second goal of this article is to use the framework of conditional and marginal data augmentation (Meng and van Dyk, 1999; van Dyk and Meng, 2001) in order to illuminate the behavior of the various algorithms. In particular, using unidentifiable parameters within a Markov chain is the key to the substantial computational gains offered by marginal augmentation. Thus, it is no surprise that eliminating the unidentifiable parameters slows down the algorithm of McCulloch et al. (2000). Likewise, under this framework, Nobile's somewhat subtle error is readily apparent. It is also expected that the procedure of Nobile (1998) has better convergence properties than the procedure of McCulloch and Rossi (1994), at least for the unidentifiable parameters.

The remainder of the article is divided into five sections. In Section 2 we briefly review the multinomial probit model and introduce our prior specification. The method of marginal data augmentation is reviewed and illustrated in the context of the multinomial probit model in Section 3, which concludes with the introduction of our new algorithms. In Section 4, we present the results of both theoretical and empirical investigations, which compare our algorithms with others in the literature. More sophisticated computational examples appear in Section 5. Section 6 gives concluding remarks and two appendices present some technical details.

2. The multinomial probit model

The observed multinomial variable Y_i is modeled in terms of a latent variable $W_i = (W_{i1}, \ldots, W_{i,p-1})$ via

$$Y_i(W_i) = \begin{cases} 0 & \text{if } \max(W_i) < 0 \\ j & \text{if } \max(W_i) = W_{ij} > 0 \end{cases} \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\max(W_i)$ is the largest element of the vector W_i. The latent variables is modeled as

$$W_i = X_i\beta + e_i, \quad e_i \sim \mathrm{N}(0, \Sigma) \quad \text{for } i = 1, \ldots, n, \tag{2}$$

where X_i is a $(p-1)\times k$ matrix of observed covariates with fixed $k\times 1$ coefficients, β, and $\Sigma = (\sigma_{\ell m})$ is a positive definite $(p-1)\times(p-1)$ matrix with $\sigma_{11} = 1$. Averaging over W_i, we find $p(Y_i \mid \beta, \Sigma)$, the multiplicative contribution of Y_i to the likelihood. The constraint on σ_{11} is made to be sure the model parameters (β, Σ) are identified. In particular, consider

$$\tilde{W}_i = \alpha W_i = X_i\tilde{\beta} + \tilde{e}_i, \quad \tilde{e}_i \sim \mathrm{N}(0, \tilde{\Sigma}) \quad \text{for } i = 1, \ldots, n, \tag{3}$$

where α is a positive scalar, $\tilde{\beta} = \alpha\beta$, and $\tilde{\Sigma} = \alpha^2\Sigma$ is an unconstrained positive definite $(p-1)\times(p-1)$ matrix. Since $Y_i(W_i) = Y_i(\tilde{W}_i)$, the parameter α is unidentifiable. (Even with this constraint on σ_{11}, (β, Σ) may be unidentifiable without certain conditions on X and Y; see Keane (1992), Chib et al. (1998), and Speckman et al. (1999).)

Our analysis is based on the Bayesian posterior distribution of β and Σ resulting from the independent prior distributions

$$\beta \sim \mathrm{N}(\beta_0, A^{-1}) \quad \text{and} \quad p(\Sigma) \propto |\Sigma|^{-(\nu+p)/2}[\mathrm{trace}(S\Sigma^{-1})]^{-\nu(p-1)/2}, \tag{4}$$

subject to $\sigma_{11} = 1$, where β_0 and A^{-1} are the prior mean and variance of β, ν is the prior "degrees of freedom" for Σ, and the matrix S is the prior scale of Σ; we assume the first diagonal element of S is one.

The prior distribution on Σ is a constrained inverse Wishart distribution. In particular, beginning with $\tilde{\Sigma} \sim \text{inv Wishart}(\nu, \tilde{S})$ and transforming to $\alpha^2 = \tilde{\sigma}_{11}$ and $\Sigma = \tilde{\Sigma}/\tilde{\sigma}_{11}$ we find

$$p(\Sigma, \alpha^2) \propto |\Sigma|^{-(\nu+p)/2} \exp\left[-\frac{\alpha_0^2}{2\alpha^2}\,\mathrm{trace}(S\Sigma^{-1})\right](\alpha^2)^{-[\nu(p-1)/2+1]}, \tag{5}$$

subject to $\sigma_{11} = 1$, where α_0^2 is a positive constant, $\tilde{S} = \alpha_0^2 S$, $\tilde{\Sigma} = (\sigma_{\ell m})$, and the Jacobian adds a factor of $\alpha^{p(p-1)-2}$. (The inverse Wishart distribution is parameterized so that $\mathrm{E}(\tilde{\Sigma}) = (\nu - p)^{-1}\tilde{S}$.) Thus, the conditional distribution of α^2 given Σ is

$$\alpha^2 \mid \Sigma \sim \alpha_0^2\,\mathrm{trace}(S\Sigma^{-1})/\chi^2_{\nu(p-1)}, \tag{6}$$

and integrating (5) over α^2 yields the marginal distribution of Σ given in (4). Thus, $p(\Sigma)$ is the distribution of $\tilde{\Sigma}/\tilde{\sigma}_{11}$, where $\tilde{\Sigma} \sim \text{inv Wishart}(\nu, \tilde{S})$. Because the inverse Wishart distribution is proper if $\nu \geqslant p-1$, $p(\Sigma)$ is proper under this same condition. To approximate $\mathrm{E}(\Sigma)$, we note

$$\mathrm{E}(\Sigma) = \mathrm{E}\left(\frac{1}{\tilde{\sigma}_{11}}\tilde{\Sigma}\right) \approx \mathrm{E}(\tilde{\Sigma})/\mathrm{E}(\tilde{\sigma}_{11}) = S, \tag{7}$$

where the approximation follows from a first-order Taylor series expansion. Finally, by construction, the prior variance of Σ decreases as ν increases, as long as the variance exists.

Combining (4) and (5) and transforming to $(\tilde{\beta}, \tilde{\Sigma})$ yields $\tilde{\beta} \mid \tilde{\Sigma} \sim \mathrm{N}(\sqrt{\tilde{\sigma}_{11}}\beta_0, \tilde{\sigma}_{11}A^{-1})$ with $\tilde{\Sigma} \sim \text{inv Wishart}(\nu, \tilde{S})$. McCulloch and Rossi (1994) on the other hand suggest

$\tilde{\beta} \mid \tilde{\Sigma} \sim \mathrm{N}(\beta_0, A^{-1})$ with $\tilde{\Sigma} \sim \text{inv Wishart}(v, \tilde{S})$. As we shall demonstrate, this seemingly minor change has important implications for the resulting algorithms.

Our choice of prior distribution is motivated by a desire to allow for both informative and diffuse prior distributions while maintaining simple and efficient algorithms. We can set $A = 0$ for a flat prior on β or choose small values of A when little prior information is available. Neither the method of McCulloch and Rossi (1994) nor of Nobile (1998) allows for a flat prior on β. A prior distribution on Σ is generally not meant to convey substantive information but rather to be weakly informative and to provide some shrinkage of the eigenvalues and correlations (McCulloch et al., 2000). The prior distribution for Σ specified by (4) should accomplish this with small degrees of freedom ($v \geqslant p - 1$ for prior propriety).

3. Conditional and marginal augmentation

3.1. Data augmentation algorithm

The data augmentation (DA) algorithm (Tanner and Wong, 1987) is designed to obtain a Monte Carlo sample from the posterior distribution $p(\theta, W \mid Y)$ by iteratively sampling from $p(\theta \mid W, Y)$ and $p(W \mid \theta, Y)$. (In this discussion, Y may be regarded as generic notation for the observed data, θ for the model parameters, and W for the latent variables.) The samples obtained with the DA algorithm form a Markov chain, which under certain regular conditions (e.g., Roberts, 1996; Tierney, 1994, 1996) has stationary distribution equal to the target posterior distribution, $p(\theta, W \mid Y)$. Thus, after a suitable burn in period (see Gelman and Rubin, 1992; Cowles and Carlin, 1996, for discussion of convergence diagnostics) the sample obtained with the DA algorithm may be regarded as a sample from $p(\theta, W \mid Y)$. The advantage of this strategy is clear when both $p(\theta \mid W, Y)$ and $p(W \mid \theta, Y)$ are easy to sample, but simulating $p(\theta, W \mid Y)$ directly is difficult or impossible.

In the context of the multinomial probit model, computation is complicated by the constraint $\sigma_{11} = 1$, i.e., $p(\beta, \Sigma \mid W, Y)$ is not particularly easy to sample directly. The methods introduced by McCulloch and Rossi (1994), Nobile (1998), and McCulloch et al. (2000) are all variations of the DA algorithm which are designed to accommodate this constraint in one way or another. In this section, we introduce the framework of conditional and marginal augmentation which generalizes the DA algorithm in order to improve its rate of convergence. From this more general perspective, we can both derive new algorithms with desirable properties for the multinomial probit model and predict the behavior of the various previously proposed algorithms.

3.2. Working parameters and working prior distributions

Conditional and marginal augmentation (Meng and van Dyk, 1999; van Dyk and Meng, 2001) take advantage of unidentifiable parameters to improve the rate of convergence of a DA algorithm. In particular, we define a *working parameter* to be a parameter that is not identified given the observed data, Y, but is identified given

(Y, W). Thus, σ_{11} is a working parameter in (2); equivalently α is a working parameter in (3). (The matrix $\tilde{\Sigma}$ is identifiable given $\tilde{W}$, as long as $n > p - 1$.)

To see how we make use of the working parameter, consider the likelihood of $\theta = (\beta, \Sigma)$,

$$L(\theta \mid Y) \propto p(Y \mid \theta) = \int p(Y, W \mid \theta)\, \mathrm{d}W. \tag{8}$$

Because there are many *augmented-data models*, $p(Y, W \mid \theta)$, that satisfy (8) the latent-variable model described in Section 2 is not a unique representation of the multinomial probit model. In principle, different augmented-data models can be used to construct different DA algorithms with the different properties. Thus, we aim to choose an augmented-data model that results in a simple and fast DA algorithm and that is formulated in terms of an easily quantifiable prior distribution.

Since α is not identifiable given Y, for any value of the working parameter, α,

$$L(\theta \mid Y) = L(\alpha, \theta \mid Y) \propto \int p(Y, W \mid \theta, \alpha)\, \mathrm{d}W, \tag{9}$$

where the equality follows because α is not identifiable and the proportionality follows from the definition of the likelihood. Thus, we may condition on any particular value of α. Such conditioning often takes the form of a constraint; e.g., setting $\sigma_{11} = 1$ in the multinomial probit model.

Alternatively, we may average (9) over any *working prior distribution* for α,

$$L(\theta \mid Y) \propto \int \left[\int p(Y, W \mid \theta, \alpha)\, p(\alpha \mid \theta)\, \mathrm{d}\alpha \right] \mathrm{d}W, \tag{10}$$

where we may change the order of integration by Fubini's theorem. Here we specify the prior distribution for α conditional on θ so we can specify a joint prior distribution on (θ, α) via the marginal prior distribution for θ. The factor in square brackets in (10) equals $p(Y, W \mid \theta)$ which makes (10) notationally equivalent to (8); i.e., (10) constitutes a legitimate augmented-data model.

The difference between (10) and (8) is that the augmented-data model, $p(Y, W \mid \theta)$, specified by (10) averages over α whereas (8) implicitly conditions on α; this is made explicit in (9). Thus, we call (9) and the resulting DA algorithms *conditional augmentation* and we call (10) and its corresponding DA algorithms *marginal augmentation* (Meng and van Dyk, 1999).

We expect the conditional augmented-data model, $p(Y, W \mid \theta, \alpha)$, to be less diffuse than the corresponding marginal model, $\int p(Y, W \mid \theta, \alpha)\, p(\alpha \mid \theta)\, \mathrm{d}\alpha$—this is the key to the computational advantage of marginal augmentation. Heuristically, we would like $p(W \mid \theta, Y)$ to be as near $p(W \mid Y)$ as possible so as to reduce the autocorrelation in the resulting Markov chain—if we could sample from $p(W \mid Y)$ and $p(\theta \mid W, Y)$ there would be no autocorrelation. Thus, $p(W \mid \theta, Y)$ should be as diffuse as possible, up to the limit of $p(W \mid Y)$. Since $p(W \mid \theta, Y) = p(Y, W \mid \theta)/p(Y \mid \theta)$ and $p(Y \mid \theta)$ remains unchanged, this is accomplished by choosing $p(Y, W \mid \theta)$ to be more diffuse, which is the case with marginal augmentation using a diffuse working prior distribution.

Formally, Meng and van Dyk (1999) proved that, starting from any augmented-data model, the following strategy can only improve the geometric rate of convergence of the DA algorithm.

Marginalization strategy

Step 1: For α in a set $\mathscr{A}$, construct a one-to-one mapping $\mathscr{D}_\alpha$ of the latent variable and define $\tilde{W} = \mathscr{D}_\alpha(W)$. The set $\mathscr{A}$ should include some $\alpha_\mathscr{I}$ such that $\mathscr{D}_{\alpha_\mathscr{I}}$ is the identity mapping.

Step 2: Choose a proper working prior distribution, $p(\alpha)$ that is independent of θ, to define an augmented-data model as defined in (10).

In the context of the multinomial probit model the mapping in Step 1 is defined in (3), i.e., $\tilde{W}_i = \mathscr{D}_\alpha(W_i) = \alpha W_i$ for each i with $\alpha_\mathscr{I} = 1$ and $\mathscr{A} = (0, +\infty)$. Thus, if we were to construct a DA algorithm using (3) in place of (2) with any proper working prior distribution (independent of θ) we would necessarily improve the geometric rate of convergence of the resulting algorithm.

The samplers introduced by McCulloch and Rossi (1994) and Nobile (1998) as well as the ones we introduce in Section 3.4 use marginal augmentation but do not fall under the Marginalization strategy because θ and α are not a priori independent. Nevertheless, marginal augmentation is an especially promising strategy because using (3) to construct a DA algorithm can be motivated by the difficulties that the constraint $\sigma_{11} = 1$ impose on computation (McCulloch and Rossi, 1994). In practice, we find that these samplers are not only much easier to implement but in many examples also converge much faster than the sampler of McCulloch et al. (2000), which does not use marginal augmentation.

3.3. Sampling schemes

In this section we describe how marginal augmentation algorithms are implemented and for clarity illustrate their use in the binomial probit model. There are two basic sampling schemes for use with marginal augmentation, which differ in how they handle the working parameter and can exhibit different convergence behavior. Starting with $\theta^{(t-1)}$ and $(\theta^{(t-1)}, \alpha^{(t-1)})$, respectively, the two schemes make the following random draws at iteration t:

Scheme 1: $\tilde{W}^{(t)} \sim p(\tilde{W} \mid \theta^{(t-1)}, Y)$ and $\theta^{(t)} \sim p(\theta \mid \tilde{W}^{(t)}, Y)$,

Scheme 2: $\tilde{W}^{(t)} \sim p(\tilde{W} \mid \theta^{(t-1)}, \alpha^{(t-1)}, Y)$ and $(\theta^{(t)}, \alpha^{(t)}) \sim p(\theta, \alpha \mid \tilde{W}^{(t)}, Y)$.

Notice that in Scheme 1, we completely marginalize out the working parameter while in Scheme 2, the working parameter is updated in the iteration.

For the binomial model, $\tilde{\Sigma} = \alpha^2$ and $\theta = \beta$ and we use the prior distribution given in (4) with $\beta_0 = 0$ and working prior distribution given in (6). The algorithms described here for the binomial model are a slight generalization of those given by van Dyk and Meng (2001) for binomial probit regression; they assume $p(\beta) \propto 1$, while we allow $\beta \sim \mathrm{N}(0, A^{-1})$.

The first step in both sampling schemes is based on

$$W_i \mid \theta, Y_i \sim \mathrm{TN}(X_i\beta, 1, Y_i), \tag{11}$$

where $\mathrm{TN}(\mu, \sigma^2, Y_i)$ specifies a normal distribution with mean μ and variance σ^2 truncated to be positive if $Y_i = 1$ and negative if $Y_i = 0$. Since $\tilde{W}_i = \alpha W_i$, the first step of Scheme 2 is given by

$$\tilde{W}_i \mid \beta, \alpha^2, Y_i \sim \mathrm{TN}(\alpha X_i \beta, \alpha^2, Y_i). \tag{12}$$

For Scheme 1, we draw from $p(\tilde{W}, \alpha^2 \mid \beta, Y)$ and discard the draw of α^2 to obtain a draw from $p(\tilde{W} \mid \beta, Y)$. That is, we sample

$$p(\tilde{W}_i \mid \beta, Y_i) = \int p(\tilde{W}_i \mid \beta, \alpha^2, Y_i)\, p(\alpha^2 \mid \beta)\, \mathrm{d}\alpha^2 \tag{13}$$

by first drawing $\alpha^2 \sim p(\alpha^2 \mid \beta)$ and then drawing $\tilde{W}_i$ given θ, α^2, and Y_i as described in (12). In this case, $p(\alpha^2 \mid \beta) = p(\alpha^2)$, i.e., $\alpha^2 \sim \alpha_0^2/\chi_v^2$.

The second step in both sampling schemes is accomplished by sampling from $p(\beta, \alpha^2 \mid \tilde{W}, Y)$; with Scheme 1 we again discard the sampled value of α^2 to obtain a draw from $p(\beta \mid \tilde{W}, Y)$. To sample from $p(\beta, \alpha^2 \mid \tilde{W}, Y)$, we first transform to $p(\tilde{\beta}, \alpha^2 \mid \tilde{W}, Y)$, then we sample from $p(\alpha^2 \mid \tilde{W}, Y)$ and $p(\tilde{\beta} \mid \alpha^2, \tilde{W}, Y)$, and finally we set $\beta = \tilde{\beta}/\alpha$. Thus, for both sampling schemes we sample $\alpha^2 \mid \tilde{W}, Y \sim [\sum_{i=1}^n (\tilde{W}_i - X_i\hat{\beta})^2 + \alpha_0^2 + \hat{\beta}^\top A \hat{\beta}]/\chi_{n+v}^2$ and $\tilde{\beta} \mid \alpha^2, \tilde{W}, Y \sim \mathrm{N}[\hat{\beta}, \alpha^2 (A + \sum_{i=1}^n X_i^\top X_i)^{-1}]$, where $\hat{\beta} = (A + \sum_{i=1}^n X_i^\top X_i)^{-1} \sum_{i=1}^n X_i^\top \tilde{W}_i$.

Although both Schemes 1 and 2 have the same lag-1 autocorrelation for linear combinations of $\theta^{(t)}$, the geometric rate of convergence of Scheme 1 cannot be larger than that of Scheme 2 because the maximum correlation between θ and $\tilde{W}$ cannot exceed that of (θ, α) and $\tilde{W}$ (Liu et al., 1994). Thus, we generally prefer Scheme 1. As we shall see, this observation underpins the improvement of the hybrid Markov chain introduced by Nobile (1998); McCulloch and Rossi (1994) uses Scheme 2, while Nobile (1998) uses Scheme 1 with the same augmented-data model. Because we use a different prior distribution than Nobile, both sampling schemes are available without recourse to a Metropolis step; see Section 4.2 for details.

3.4. Two new algorithms for the multinomial probit model

We now generalize the samplers for the binomial model to the multinomial probit model. The resulting Gibbs samplers are somewhat more complicated and, as with other algorithms in the literature, require additional conditional draws. We introduce two algorithms, the first with two sampling schemes, which are designed to mimic Schemes 1 and 2. Because of the additional conditional draws, however, they do not technically follow the definitions given in Section 3.3; thus, we refer to them as "Scheme 1" and "Scheme 2"; see van Dyk et al. (2004) for discussion of sampling schemes in multistep marginal augmentation algorithms.

On theoretical grounds we expect, and in our numerical studies we find, that "Scheme 1" outperforms "Scheme 2." Thus, in practice, *we recommend* "Scheme 1" *of* Algorithm 1 *always be used rather than* "Scheme 2." We introduce "Scheme 2" primarily for comparison with the method of McCulloch and Rossi (1994) since both use the same sampling scheme, but with different prior distributions. Likewise, "Scheme 1" uses the same sampling scheme as Nobile's method (1998), again with a different

prior distribution; see Section 4.2 for details. In both sampling schemes of Algorithm 1, we assume $\beta_0 = 0$. We relax this constraint in Algorithm 2 but this may come at some computational cost; we expect Algorithm 1 (Scheme 1) to outperform Algorithm 2. Thus, *we recommend* Algorithm 2 *only be used when* $\beta_0 \neq 0$.

We begin with Algorithm 1, which is composed of three steps. In the first step we sample each of the components of W_i in turn, conditioning on the other components of W_i, the model parameters, and the observed data. Thus, this step is composed of $p-1$ conditional draws from truncated normal distributions. Then, we compute $\tilde{W}_i = \alpha W_i$ with α^2 drawn from its conditional prior distribution for "Scheme 1" and with the value of α^2 from the previous iteration for "Scheme 2." Step 2 samples $\beta \sim p(\beta \mid \Sigma, \tilde{W}, Y)$. Finally, Step 3 samples $(\alpha^2, \Sigma) \sim p(\alpha^2, \Sigma \mid \beta, (\tilde{W} - X_i\tilde{\beta}), Y)$ and sets $W_i = \tilde{W}_i/\alpha$ for each i. Details of Algorithm 1 appear in Appendix A.

To allow for $\beta_0 \neq 0$, we derive a second algorithm, which divides each iteration into two steps:

Step 1: Update (W, Σ) via $(W^{(t)}, \Sigma^{(t)}) \sim \mathscr{K}(W, \Sigma \mid W^{(t-1)}, \Sigma^{(t-1)}; \beta^{(t-1)})$, where $\mathscr{K}$ is the kernel of a Markov chain with stationary distribution $p(W, \Sigma \mid Y, \beta^{(t-1)})$.

Step 2: Update β via $\beta^{(t)} \sim p(\beta \mid Y, W^{(t)}, \Sigma^{(t)}, (\alpha^2)^{(t)}) = p(\beta \mid Y, W^{(t)}, \Sigma^{(t)})$.

Step 1 is made up of a number of conditional draws, which are specified in Appendix A. We construct the kernel in Step 1 using marginal augmentation with an implementation scheme in the spirit of Scheme 1; we completely marginalize out the working parameter. By construction, the conditional distribution in Step 2 does not depend on $(\alpha^2)^{(t)}$. In fact, Step 2 makes no use of marginal augmentation; it is a standard conditional draw. We introduce an alternative augmented-data model to be used in Step 1, replacing (3) with

$$\tilde{W}_i = \alpha(W_i - X_i\beta) = \tilde{e}_i, \quad \tilde{e}_i \sim \mathrm{N}(0, \tilde{\Sigma}) \quad \text{for } i = 1, \ldots, n. \tag{14}$$

Because $\tilde{W}_i$ is only used in Step 1, where β is fixed, it is permissible for $\tilde{W}_i$ to depend on β. The details of Algorithm 2 appear in Appendix A.

3.5. The choice of $p(\alpha^2|\Sigma)$

As discussed in Section 3.2, the computational gain of marginal augmentation is a result of a more diffuse augmented-data model, which allows the Gibbs sampler to move more quickly across the parameter space. Thus, we expect that the more diffuse the augmented-data model, the more computational gain that marginal augmentation will offer. This leads to a rule of thumb for selecting the prior distribution on the unidentifiable parameters—the more diffuse, the better. With the parameterization of $p(\alpha^2 \mid \Sigma)$ given in (6), only α_0^2 is completely free; changing ν or S effects the prior distribution of Σ and, thus, the fitted model. (In our numerical studies, we find that the choice of α_0^2 has little effect on the computational performance of the algorithms.) To the degree that the practitioner is indifferent to the choice of $p(\Sigma)$, values of ν and S can be chosen to increase the prior variability of α^2 and simultaneously of Σ, i.e., by choosing both ν and S small. As discussed by McCulloch et al. (2000), however, care must be taken not to push this too far because the statistical properties of the posterior distribution may suffer.

4. Comparisons with other methods

The algorithms introduced here differ from those developed by McCulloch and Rossi (1994), Nobile (1998), and McCulloch et al. (2000) in terms of their prior specification and sampling schemes. In this section we describe these differences and the advantages of our formulation; we include a number of computational comparisons involving binomial models to illustrate these advantages. More sophisticated, multinomial examples appear in Section 5.

4.1. Prior specification of McCulloch and Rossi (1994)

Rather than the prior distribution we used in (4), McCulloch and Rossi (1994) suggest

$$\tilde{\beta} \sim \mathrm{N}(\tilde{\beta}_0, A^{-1}) \quad \text{and} \quad \tilde{\Sigma} \sim \text{inv Wishart}(v, S) \tag{15}$$

which, as they noted, results in a rather cumbersome prior distribution for the identifiable parameters, (β, Σ). In particular, (15) results in the marginal prior distribution for β is given by

$$p(\beta) \propto \int |\sigma_{11} A|^{1/2} |\Sigma|^{-(v+p+1)/2} \exp\left\{ -\frac{1}{2} [(\sqrt{\sigma_{11}}\beta - \beta_0)^\top A (\sqrt{\sigma_{11}}\beta - \beta_0) + \text{trace}(S\Sigma^{-1})] \right\} d\Sigma. \tag{16}$$

Because (16) is not a standard distribution, numerical analysis is required to determine what model is actually being fit. Since (15) does not allow for an improper prior on β, proper prior information for β must be included and must be specified via (16). The motivation behind (15) is computational; the resulting model is easy to fit.

The more natural interpretation of our choice of $p(\beta, \Sigma)$ comes with no computational cost. To illustrate this, we compare the algorithm developed by McCulloch and Rossi (1994) with our Algorithm 1 using a data set generated in the same way as the data set in Example 1 of Nobile (1998). This data set, with a sample size of 2000, was generated with a single covariate drawn from a Uniform $(-0.5, 0.5)$ distribution, and $\beta = -\sqrt{2}$. Again, following Nobile (1998) we use the prior specification given in (15) with $\tilde{\beta}_0 = 0$, $A = 0.01$, $v = 3$, and $S = 3$ when running McCulloch and Rossi's algorithm. When running our algorithms, we use the prior and working prior distributions given in (4) and (6) with $\beta_0 = 0$, $A = 0.01$, $v = 3$, $S = 1$, and $\alpha_0^2 = 3$.

Fig. 1 compares both sampling schemes of Algorithm 1 with the method of McCulloch and Rossi (1994). As in Nobile (1998), we use two starting values: $(\beta, \alpha) = (-\sqrt{2}, \sqrt{2})$ and $(\beta, \alpha) = (-2, 10)$ to generate two chains of length 3000 for each of the three algorithms.[2] The contour plots represent the joint posterior distributions of the *unidentifiable* parameters, $(\tilde{\beta}, \alpha)$ and demonstrate that both the method of McCulloch and Rossi (1994) and Scheme 2 are sensitive to the starting value. Although at stationarity Schemes 1 and 2 must have the same lag-one autocorrelation for linear functions

[2] Since we aim to investigate convergence, the chains were run without burn-in in this and later examples.

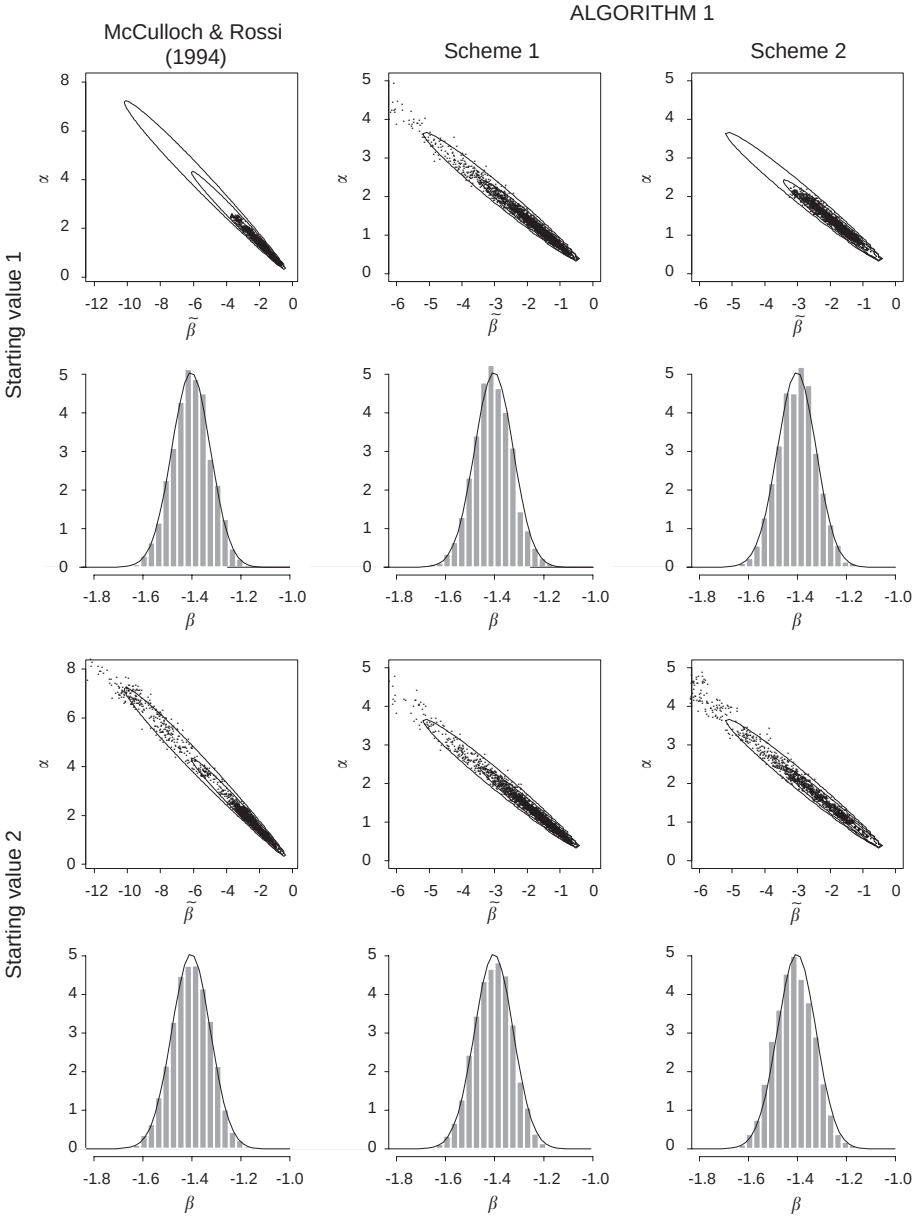

Fig. 1. Comparison of the two marginal augmentation schemes of Algorithm 1 with the algorithm of McCulloch and Rossi (1994) using the first binomial probit example of Nobile (1998). The two sets of starting values are $(\beta,\alpha)=(-\sqrt{2},\sqrt{2})$ and $(\beta,\alpha)=(-2,10)$, respectively. The plots are produced with 3000 draws for each chain and illustrate that although Scheme 1 is less sensitive to the choice of starting value, in this example, the algorithms seem equivalent for the identifiable parameter. (The scale for α in the plots in the first column is different from that in the other columns. This is due to a difference in the working prior distribution in the different samplers.)

of the parameter, an advantage of Scheme 1 is that it, like the method of Nobile (1998), can be far less sensitive to the starting values. This seems to be a general advantage of Scheme 1 over Scheme 2, in that both the method of Nobile (1998) and Algorithm 1 (Scheme 1) are less sensitive to starting values than the method of McCulloch and Rossi (1994) and Algorithm 1 (Scheme 2); in the following section, we show that the methods of Nobile (1998) and McCulloch and Rossi (1994) are examples of Scheme 1 and Scheme 2, respectively, but using the prior in (15). These differences do not persist for the *identifiable* parameter in this example; Schemes 1 and 2 appear equivalent in terms of $\beta = \tilde{\beta}/\alpha$. (In Section 4.2, however, we find that Scheme 1 is less sensitive to starting values than Scheme 2 for the *identifiable* parameters. In another example in Section 5, we find that Algorithm 1 ("Scheme 1") is less sensitive to starting values than any of the existing algorithms.)

4.2. The sampling scheme of Nobile (1998)

Nobile (1998) suggested a simple adaptation of the method of McCulloch and Rossi (1994) which aims to improve the computational performance, but uses the same prior distribution. In this section, we show that the difference between the two methods is that one uses Scheme 1 and the other uses Scheme 2. Although this explains the computational gain of Nobile's algorithm, we also find that this gain is not nearly as dramatic as it might at first seem. Finally, we point out an error in Nobile (1998) that effects the stationary distribution of the sampler and thus compromises valid inference. To avoid technical details, our discussion is in the context of the binomial model.

In the binomial case, an iteration of McCulloch and Rossi's (1994) sampler reduces to

Step 1: Draw $\tilde{W}_i \sim p(\tilde{W}_i \mid \tilde{\beta}, \tilde{\Sigma}, Y)$ independently for $i = 1, \ldots, n$.
Step 2: Draw $\tilde{\beta} \sim p(\tilde{\beta} \mid \tilde{W}, \tilde{\Sigma}, Y)$.
Step 3: Draw $\tilde{\Sigma} \sim p(\tilde{\Sigma} \mid \tilde{W}, \tilde{\beta}, Y)$.

Because of the choice of the prior distribution, the joint distribution, $p(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{W}, Y)$ is not easy to sample directly and thus it is split into Steps 2 and 3. Aside from this complication, this algorithm is an example of Scheme 2; the model and working parameters are drawn in Step 2 and Step 3 and both are conditioned on in Step 1.

Nobile (1998) modified this sampler by adding one more step:

Step 4: Sample $(\tilde{\beta}, \tilde{\Sigma})$ along the direction in which the likelihood is flat.

With the same caveat, Nobile's algorithm is an example of Scheme 1. If we could combine Steps 2 and 3 into a joint draw of $(\tilde{\beta}, \tilde{\Sigma}) \sim p(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{W}, Y)$, transforming to (β, Σ) would accomplish the second draw of Scheme 1. Secondly, Step 4 is equivalent to sampling $(\tilde{\beta}, \tilde{\Sigma}) \sim p(\tilde{\beta}, \tilde{\Sigma} \mid \Sigma, \beta, Y)$ because Σ, β, and Y determine the value of the likelihood. But this in turn is equivalent to sampling $\alpha^2 \sim p(\alpha^2 \mid \Sigma, \beta, Y)$, since given (β, Σ), α^2 is the only free parameter in $(\tilde{\beta}, \tilde{\Sigma})$. Now,

$$\begin{aligned} p(\alpha^2 \mid \Sigma, \beta, Y) &\propto p(\alpha^2, \Sigma, \beta \mid Y) \propto p(Y \mid \alpha^2, \Sigma, \beta)\, p(\alpha^2, \Sigma, \beta) \\ &= p(Y \mid \beta, \Sigma)\, p(\alpha^2, \Sigma, \beta), \end{aligned} \tag{17}$$

and so Step 4 is equivalent to drawing $\alpha^2 \sim p(\alpha^2 \mid \Sigma, \beta, Y) = p(\alpha^2 \mid \Sigma, \beta)$. Thus, Steps 4 and 1 combine into sampling $(\alpha^2, \tilde{W}) \sim p(\alpha^2, \tilde{W} \mid \beta, \Sigma, Y)$ and discarding α^2, which is equivalent to the first step of Scheme 1.

The fact that Nobile's adaptation amounts to replacing Scheme 2 with Scheme 1 explains why it improves the convergence of the resulting Markov chain. (The extra conditioning required here does not effect this result; see van Dyk et al., 2004.) Unfortunately, using McCulloch and Rossi's (1994) prior specification, Nobile's extra step is not in closed form. Thus, he recommends using a Metropolis–Hastings step with jumping rule

$$J(\alpha' \mid \alpha) = \frac{1}{\alpha} \exp\left(-\frac{\alpha'}{\alpha}\right) \tag{18}$$

derived from $\alpha'/\alpha \sim \exp(1)$. The acceptance probability of the Metropolis–Hastings rule is

$$R = \frac{p(\alpha' \mid \beta, \Sigma)}{p(\alpha \mid \beta, \Sigma)} \frac{J(\alpha \mid \alpha')}{J(\alpha' \mid \alpha)}. \tag{19}$$

When deriving the explicit form of R, Nobile correctly notes that since the likelihood has the same value at α and α', only the prior needs to be included in the first factor of R. Unfortunately, he replaces (β, Σ) with $(\tilde{\beta}, \tilde{\Sigma})$ in (19). Since the likelihood need not be the same at $(\alpha, \tilde{\beta}, \tilde{\Sigma})$ and $(\alpha', \tilde{\beta}, \tilde{\Sigma})$, the resulting value of R is incorrect; a correction appears in Appendix B.

Fig. 2 illustrates the effect of this correction on posterior simulation using a data set concerning latent membranous lupus nephritis supplied by M. Haas. The data include the measurements of 55 patients of which 18 were diagnosed with the latent membranous lupus. To predict the occurrence of the disease, we fit a binomial probit model with an intercept and two covariates, which are clinical measurements related to immunoglobulin G and immunoglobulin A, two classes of antibody molecules; see Haas (1998) for scientific background. The histograms show that without the correction, Nobile's algorithm does not appropriately sample from the target distribution.

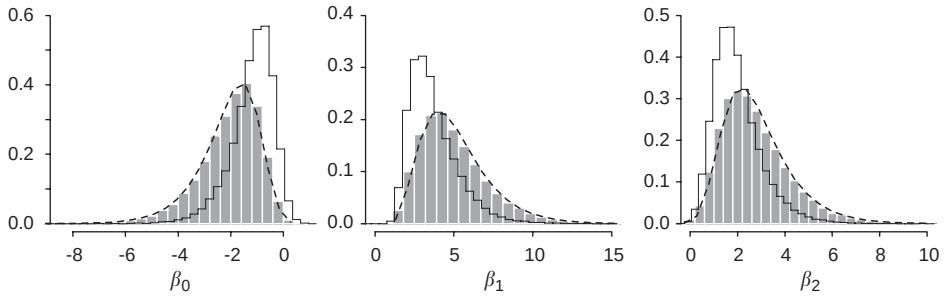

Fig. 2. Comparison of Nobile's algorithm (with and without correction) using the lupus data. The graphs compare 50,000 posterior draws of our correction (shaded histogram) with Nobile's original algorithm (outlined histogram), while the solid line represents the true posterior distribution.

In order to investigate the overall gain of adding Step 4 to the method of McCulloch and Rossi (1994), we consider the second binomial example in Nobile (1998). This example is the same as the example in Section 4.1 except that the covariate is generated from a Bernoulli(0.5) distribution and we use $\beta = 5/\sqrt{2}$ to generate the data. The same prior distributions and sample size of 2000 were used as is in Section 4.1. For this data set the likelihood is not informative, indeed the maximum likelihood estimate does not exist and a proper prior distribution for β is required for posterior propriety. Thus, this example, serves as a difficult case for testing the various algorithms.

We compare the methods of McCulloch and Rossi (1994) and Nobile (1998) (with correction) with Algorithm 1 (Scheme 1) by generating two chains of length 20,000 starting from $(\beta, \alpha) = (5/\sqrt{2}, \sqrt{2})$ and $(\beta, \alpha) = (5, 5)$. Fig. 3 compares the sample generated with the first chain of each of the three algorithms with the target posterior distribution and illustrates the autocorrelation. The time-series plots for both chains generated with each algorithm are displayed in Fig. 4. Figs. 3 and 4 illustrate that *for both identifiable and unidentifiable* parameters, the algorithm of McCulloch and Rossi

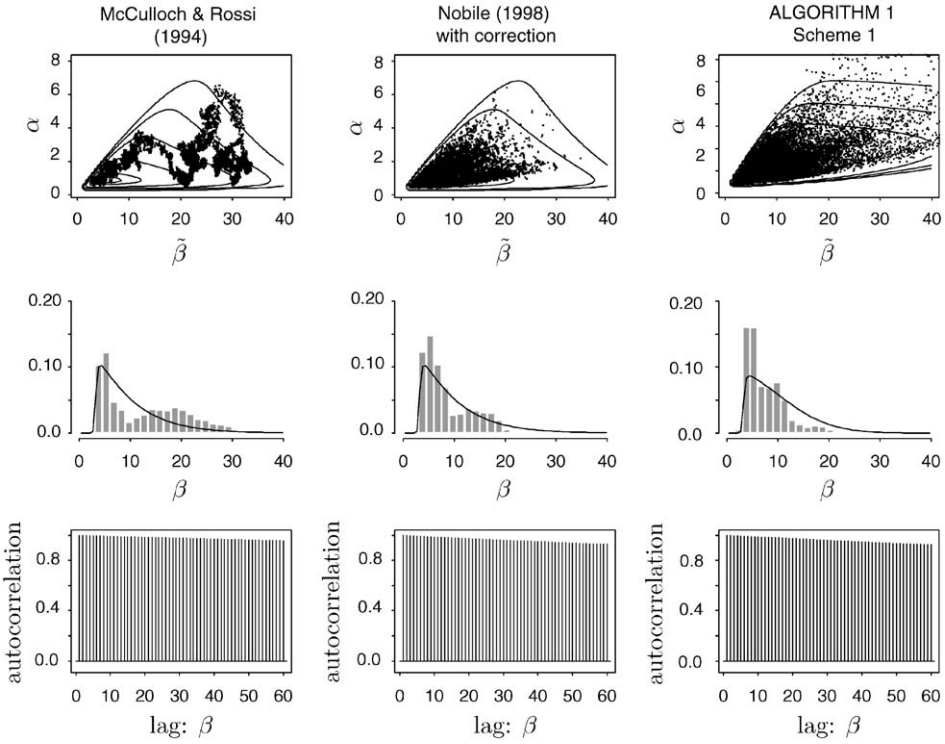

Fig. 3. Comparison of the algorithms of McCulloch and Rossi (1994) and Nobile (1998) (with correction), with Algorithm 1 (Scheme 1) using the second binomial probit example of Nobile (1998). The algorithm of McCulloch and Rossi (1994) is more sensitive to starting values than the other two algorithms; the starting value used here is $(\beta, \alpha) = (5, 5)$. However, there is little difference between the algorithms in terms of the autocorrelation of the identifiable parameter.

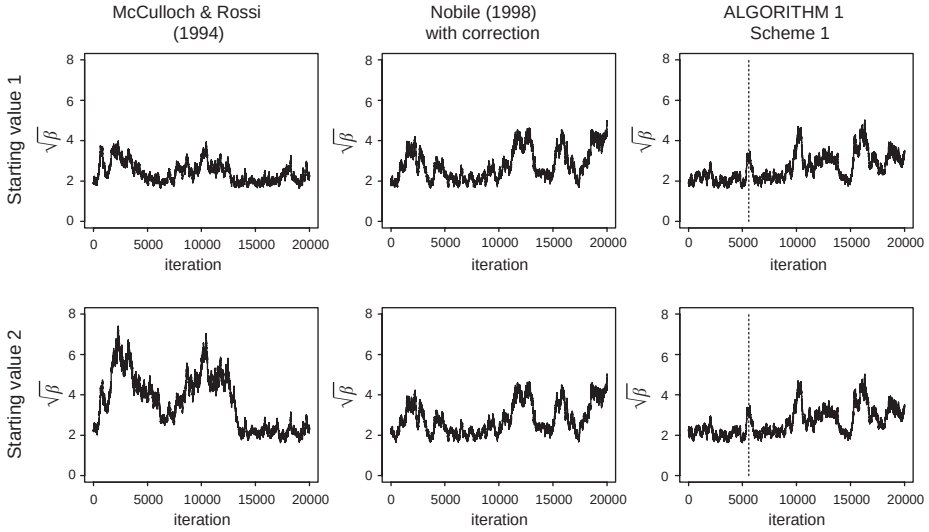

Fig. 4. Time series plot of $\sqrt{\beta}$ for the three algorithms illustrated in Fig. 3. Starting values 1 and 2 are $(\beta, \alpha) = (5/\sqrt{2}, \sqrt{2})$ and $(\beta, \alpha) = (5, 5)$, respectively. A dashed vertical line shows the iteration where the two chains of Algorithm 1 (Scheme 1) coalesced; the other two algorithms did not coalesce exactly. In this example, the algorithm of McCulloch and Rossi (1994) is more sensitive to starting values than the other two algorithms, even when measured in terms of the identifiable parameter, β.

(1994) is most sensitive to the starting value. However, in terms of the autocorrelation for the *identifiable* parameter, β, the effect of the additional Metropolis step is small. The figures also illustrate that Algorithm 1 (Scheme 1) performs at least as well as the algorithm of Nobile (1998) and outperforms that of McCulloch and Rossi (1994) with respect to both identifiable and unidentifiable parameters. With regard to the identifiable parameters, Algorithm 1 (Scheme 1) is less sensitive to the starting values; in fact, the two chains generated with this algorithm coalesced after about 5000 iterations. Unlike Nobile's algorithm, the computational gain of our algorithm comes without the computational complexity of a Metropolis step and without the inferential complexity of a prior specified on the unidentifiable parameters.

4.3. Prior specification of McCulloch et al. (2000)

A second adaptation of the method of McCulloch and Rossi (1994) was introduced by McCulloch et al. (2000). Rather than focusing on computational performance, McCulloch et al. (2000) aimed to introduce a more natural prior specification. In particular, they formulated a prior distribution on the identifiable parameters as

$$\beta \sim \mathrm{N}(\beta_0, A^{-1}), \quad \gamma \sim \mathrm{N}(\gamma_0, B^{-1}) \quad \text{and} \quad \Phi \sim \text{inv Wishart}(\kappa, C), \tag{20}$$

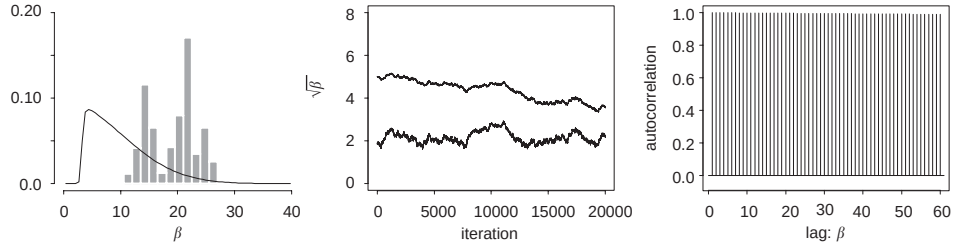

Fig. 5. The convergence of the algorithm of McCulloch et al. (2000) using the same binomial example and starting values as in Fig. 3. The time-series plot displays the chains generated using both starting values as in Fig. 4. The convergence and autocorrelation of this algorithm are less desirable than the algorithms described in Figs. 3 and 4.

where γ, $(p-2)\times 1$ and Φ, $(p-2)\times(p-2)$ are submatrices of Σ defined by

$$\Sigma = \begin{pmatrix} 1 & \gamma^\top \\ \gamma & \Phi + \gamma\gamma^\top \end{pmatrix}.$$

This prior specification is useful in practice because like our specification, it allows analysts to specify an informative or flat prior distribution on the identifiable parameter, β. As noted by McCulloch et al. (2000) and Nobile (2000), however, the resulting sampler can be inefficient. In contrast, our algorithms allow the same flexibility in the prior specification on the identifiable parameter of interest, β, without incurring computational cost.

To illustrate the computational advantage of our algorithms over that of McCulloch et al. (2000), we use the same data set, prior distributions, and sampling strategy as in Section 4.2. Fig. 5 shows the marginal posterior distribution, time-series plot, and the autocorrelation plot for the regression coefficient as in Fig. 4. The figure supports our theoretical expectation and empirical findings in the literature that this algorithm can be very sensitive to starting values and can exhibit very high autocorrelation.

5. Multinomial examples

In this section, we compare the computational performance of Algorithm 1 (using both sampling schemes) and Algorithm 2 with the algorithms of McCulloch and Rossi (1994), Nobile (1998) (with correction), and McCulloch et al. (2000) using four multinomial examples including two simulations and data from two empirical studies. Although the computational performance of the algorithms we propose is clearly an important consideration, we emphasize that other considerations are at least as important. We have demonstrated that, like the methods of McCulloch et al. (2000), our algorithms allow the practitioner to directly specify the prior distribution of the identifiable regression coefficients, can handle flat prior distributions on the regression coefficients, and do not require a Metropolis step. Even so, our algorithms are at least as good as any of the available algorithms in terms of their computational properties. In the

many examples that we have considered none of the other algorithms exhibit better autocorrelation or are less sensitive to starting values than our proposals, and, as we shall soon illustrate, in some cases our algorithms exhibit much better computational properties than their competitors.

Our first multinomial example is taken from McCulloch et al. (2000) and is also used in McCulloch and Rossi (1994). A three choice model with one covariate is used to generate a data set of size 100 with parameters values $(\beta, \sigma_{22}, \rho) = (-\sqrt{2}, 2, \frac{1}{2})$; the covariate is generated from a Uniform $(-0.5, 0.5)$ distribution. We follow the prior specification of McCulloch et al. (2000) whenever possible. For our three samplers, we use (4) and (6) with $\beta_0 = 0$, $A = 0.01$, $\nu = 6$, $S = I$ and $\alpha_0^2 = \nu$. For the algorithms of McCulloch and Rossi (1994) and Nobile (1998), we use (15) with $\tilde{\beta}_0 = 0$, $A = 0.01$, $\nu = 6$, and $S = \nu I$. Finally, for the algorithm of McCulloch et al. (2000), we use (20) with $\beta_0 = 0$, $A = 0.01$, $\gamma_0 = 0$, $B^{-1} = \frac{1}{8}$, $\kappa = 5$, and $C = \frac{21}{8}$.

We generate a single chain of length 50,000 starting from $(\beta, \sigma_{22}, \rho) = (1, 1, 1)$, for each of the six algorithms. To ensure the accuracy of our computer code, we compare draws from each sampler with bivariate marginal posterior distributions computed using direct numerical integration.[3] Autocorrelation plots for each model parameter using each of the six samplers appear in Fig. 6. As expected, the algorithm of McCulloch et al. (2000) is the slowest, while the algorithm of Nobile (1998) performs slightly better than that of McCulloch and Rossi (1994). Our algorithms perform as well computationally as any other available method. As expected, Algorithm 1 (Scheme 1) exhibits the best convergence of the methods we propose.

Although the results are not reported here due to space limitations, we replicated the six choice simulation of McCulloch and Rossi (1994) using Algorithm 1 (Scheme 1) and the samplers of McCulloch and Rossi (1994) and Nobile (1998). The results regarding the relative performance of the samplers is confirmed in this simulation. Although the lag-1 autocorrelations of the algorithm of Nobile (1998) are generally lower than those of Algorithm 1 (Scheme 1), the autocorrelations of the latter sampler diminish more quickly than those of the former. The algorithm of McCulloch and Rossi (1994) is slower than both Algorithm 1 (Scheme 1) and the algorithm of Nobile (1998).

We also studied computational performance using two data examples. In the first, we fit the multinomial probit model to survey data on voter choice in Dutch parliamentary elections using Algorithm 1 (Scheme 1) and the samplers of McCulloch and Rossi (1994), Nobile (1998), and McCulloch et al. (2000). The data include four choices and 20 covariates (with 3 intercepts); this data set is described in detail by Quinn et al. (1999). When using our sampler we specify the prior distribution with $\beta_0 = \mathbf{0}$, $A = 0.01I$, $\nu = 6$, $S = I$ and $\alpha_0^2 = \nu$; when using the algorithms of McCulloch and Rossi (1994) and Nobile (1998), we use $\tilde{\beta}_0 = \mathbf{0}$, $A = 0.02I$, $\nu = 6$, and $S = \nu I$. These two prior distributions specify the same distribution on $\tilde{\Sigma}$ and roughly the same prior mean and variance on β. Finally, when using the algorithm of McCulloch et al. (2000), we use $\beta_0 = \mathbf{0}$, $A = 0.01I$, $\nu = 6$, $\gamma_0 = 0$, and $B^{-1} = I$. This prior specification is the same for β as the prior distribution used with our algorithm and similar for Σ.

[3] These figures are available from the authors upon request.

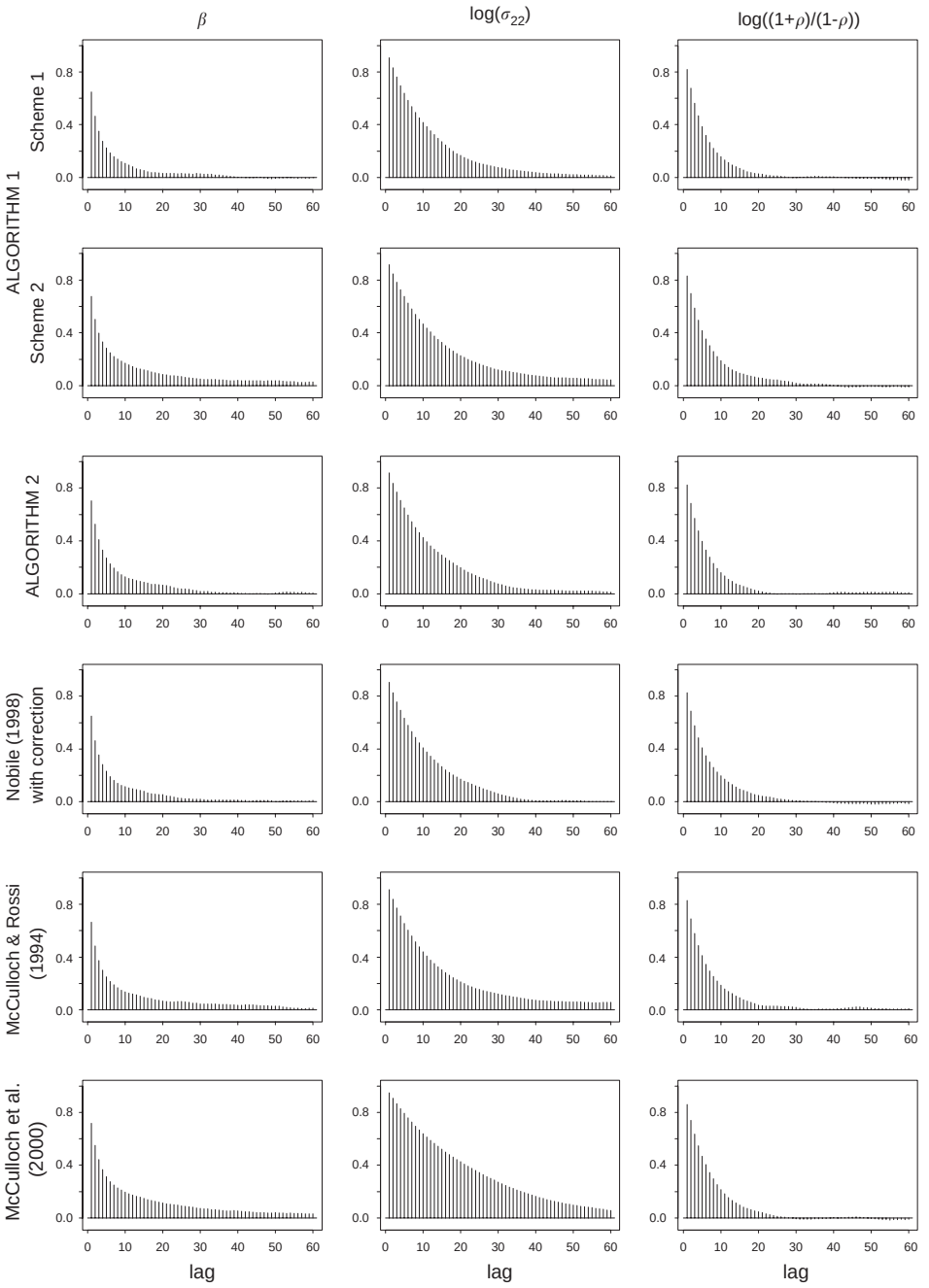

Fig. 6. The autocorrelation functions of the three model parameters using all six algorithms with the simulated trinomial example. The two schemes of Algorithm 1 perform as well as those of Nobile (1998) (with correction) and McCulloch and Rossi (1994). Algorithm 2 performs nearly as well as Scheme 1 of Algorithm 1.

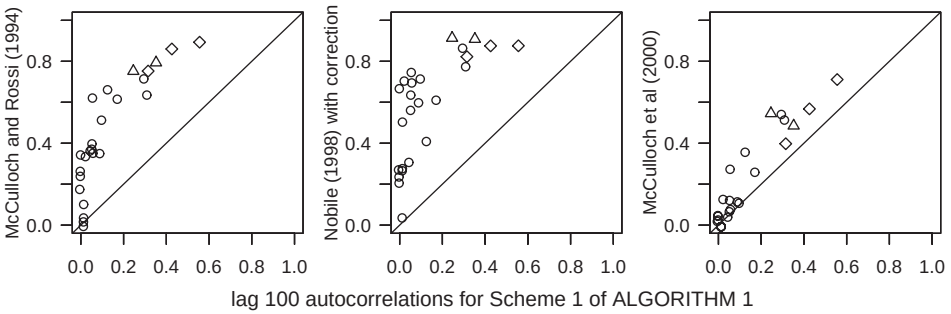

Fig. 7. Comparing the autocorrelations resulting from the four algorithms using the Dutch election data. The figures compare "Scheme 1" of Algorithm 1 (horizontal axis) with the algorithms of McCulloch and Rossi (1994), Nobile (1998), and McCulloch et al. (2000) (vertical axis), respectively. The lag 100 autocorrelations of all model parameters are plotted. Circles, triangles, and rectangles represent regression coefficients, variances, and correlation parameters, respectively. (We use the standard normalizing transformations when computing the autocorrelations.) Algorithm 1 (Scheme 1) clearly outperforms the other algorithms.

Using each of the three samplers, we generate a single chain of length 50,000 starting from $\beta=\mathbf{0}$ and $\Sigma=I$. The resulting posterior distributions agree substantively with those results reported in Quinn et al. (1999). Fig. 7 compares the lag 100 autocorrelations of all the model parameters. Judging from this figure, Algorithm 1 (Scheme 1) exhibits smaller autocorrelation than the other three algorithms. In this example, the algorithms of McCulloch and Rossi (1994) and McCulloch et al. (2000) outperform that of Nobile (1998), although this pattern reverses if we place a more informative prior distribution on the coefficients.

In the final example, we fit the multinomial probit model to data on consumer choice in market research. In particular, we use data on purchases of liquid laundry detergents by 2657 households in the Sioux Falls, South Dakota market that are analyzed by Chintagunta and Prasad (1998). In this data set, we have six national brands (Tide, Wisk, EraPlus, Surf, Solo, and All) and hence $p=6$. The data also include the log price for each brand, which we use as a covariate. The data set is described in more detail by Chintagunta and Prasad (1998). We also estimate the intercept for each brand separately, yielding the total of 6 coefficients to be estimated. We use the same prior specification as in the previous example.

For each of the three samplers, we generate three chains of length 10,000 with the same three sets of starting values for β. For Σ, we use the identity matrix as the starting value for all chains. Fig. 8 presents time series plots of selected parameters for all four algorithms. The figure shows that Algorithm 1 converges much quicker than the other three algorithms do. For example, the price coefficient, the parameter of substantive interest, converges after 1000 iterations for Algorithm 1 whereas the same parameter can require 4000–8000 draws to converge with the other algorithms. We originally ran the four algorithms with another starting value. The resulting chains do not appear in Fig. 8 because the algorithms of McCulloch and Rossi (1994) and Nobile

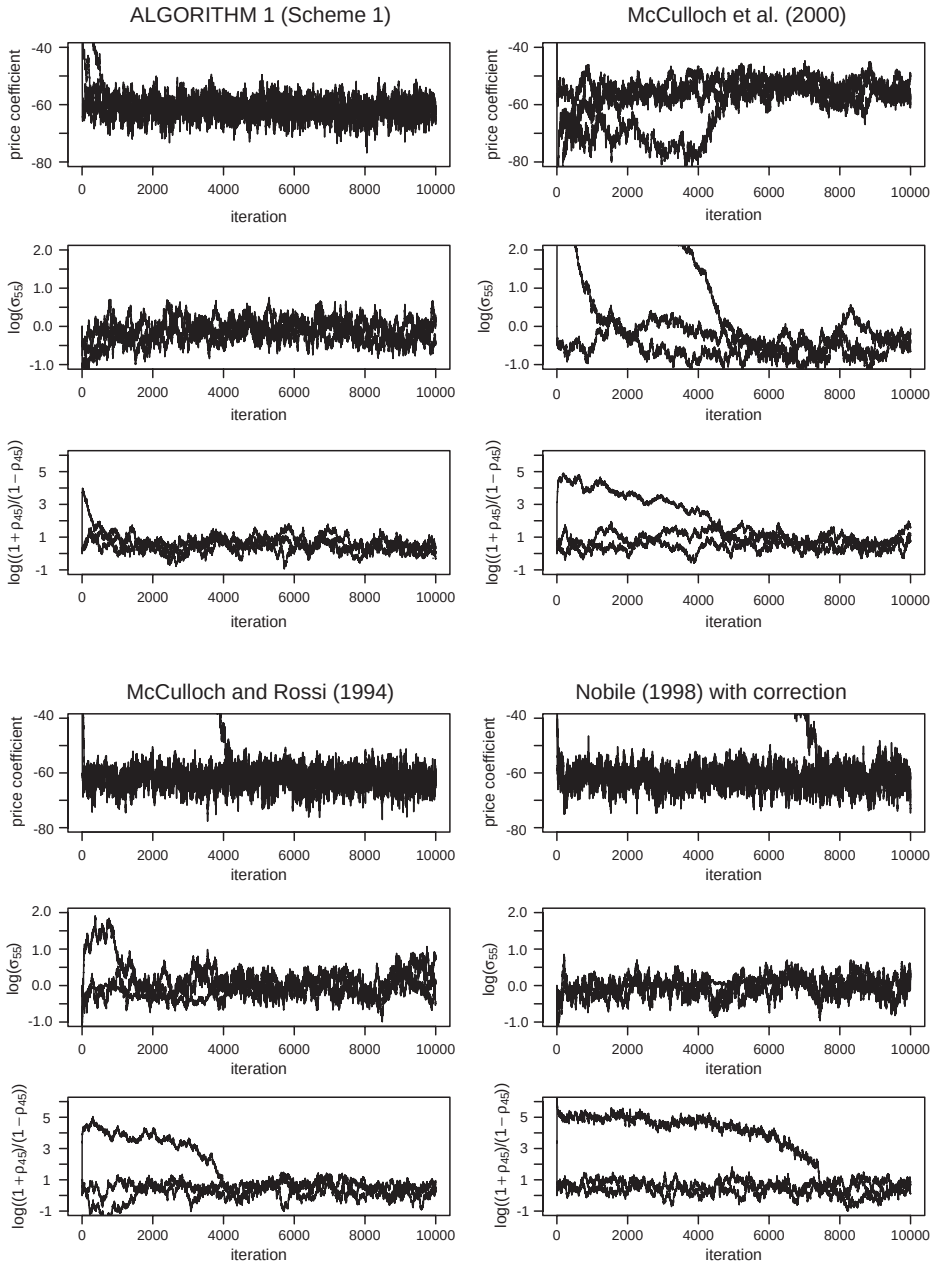

Fig. 8. Time series plots of the draws of selected parameters using the four algorithms to fit the detergent brand choice data. The figure shows the three chains for the price coefficient, a variance parameter, and a correlation coefficient. Algorithm 1 converges much faster than the algorithms of Nobile (1998) (with correction), McCulloch and Rossi (1994), and McCulloch et al. (2000).

(1998) were extremely slow to converge, taking about 60,000 and 80,000 draws, respectively, to converge. Even with this starting value, our algorithm again converged in under 1000 draws. In terms of the autocorrelations of the draws upon convergence in this example, Algorithm 1 performs better than the algorithm of McCulloch et al. (2000) and about the same as the algorithms of McCulloch and Rossi (1994) and Nobile (1998).

6. Concluding remarks

When comparing statistical methods, computational performance should nearly always be a less important criterion than the choice of model specification. In the context of the multinomial probit model, we expect methods which place a prior distribution directly on the identifiable parameters to be easier to use and more readily able to reflect available prior information in practice. Thus, we agree completely with the approach of McCulloch et al. (2000), which sacrifices computational efficiency in order to accommodate a more attractive model specification. Of course, as some of the examples is McCulloch et al. (2000) illustrate, in extreme cases computational concerns can render a model practically untenable. In some fortunate situations, however, algorithms exist that allow for attractive model specifications while also maintaining efficient computation. As has been demonstrated in the literature, the method of marginal augmentation is a useful tool for finding such algorithms. Thus, this article uses marginal augmentation to accomplish what Nobile (2000) hoped for when he remarked,

> Perhaps some ingenuity is needed to devise a sampling Markov chain which uses [the prior of McCulloch et al. (2000)] and has good mixing properties.

Although our prior distribution is not that of McCulloch et al. (2000), in terms of each of the three criteria used to compare the algorithms in Table 1, we believe the algorithms we recommend are equal to the best available methods. We specify an easy-to-use prior distribution on the identifiable model parameters and our algorithms require only draws from standard distributions. In our numerical comparisons, these advantages come without computational cost; overall the performance of the methods we recommend is better than any other method in terms of both autocorrelation and insensitivity to starting values. It would seem that sometimes there is such a thing as a free lunch!

Acknowledgements

The authors gratefully acknowledge funding for this project partially provided by NSF Grant DMS-01-04129, the U.S. Census Bureau, and by the Princeton University Committee on Research in the Humanities and Social Sciences. The first author thanks Gary King for his continuous encouragement.

Appendix A. Details of two new algorithms

In this Appendix, we give details of Algorithms 1 and 2 for the multinomial probit model. We begin with Algorithm 1 with starting value $\theta^{(0)} = (\beta^{(0)}, \Sigma^{(0)})$ and for "Scheme 2," $\alpha^{(0)} = 1$ and proceed via the following three steps at iteration t:

Algorithm 1. *Step* 1: For each i and each $j = 1, \ldots, p-1$ draw W_{ij}^* given $W_{i,-j}^*, \theta^{(t-1)}$, and Y, where

$$W_{i,-j}^* = (W_{i1}^*, \ldots, W_{i,j-1}^*, W_{i,j+1}^{(t-1)}, \ldots, W_{i,p-1}^{(t-1)}) \tag{A.1}$$

exactly as described in Section 3 of McCulloch and Rossi (1994). (McCulloch and Rossi (1994) use the symbols β and Σ for the *unidentified* parameters, i.e., our $\tilde{\beta}$ and $\tilde{\Sigma}$. In this step W^* should be generated using the *identified parameters*, so notationally this is equivalent to McCulloch and Rossi (1994).) For "Scheme 1" draw $(\alpha^2)^*$ from $p(\alpha^2 \mid \beta, \Sigma) = p(\alpha^2 \mid \Sigma)$ as given in (6) and set $\tilde{W}_{ij}^* = \alpha^* W_{ij}^*$. For "Scheme 2" set $\tilde{W}_{ij}^* = \alpha^{(t-1)} W_{ij}^*$.

Step 2: Draw $\tilde{\beta}^*$ and $(\alpha^2)^*$ given $\tilde{W}^*$ and $\Sigma^{(t-1)}$,

$$(\alpha^2)^* \sim \frac{\{\sum_{i=1}^n (\tilde{W}_i^* - X_i\hat{\beta})^\top (\Sigma^{(t-1)})^{-1} (\tilde{W}_i^* - X_i\hat{\beta}) + \hat{\beta}^\top A \hat{\beta} + \text{trace}[\tilde{S}(\Sigma^{(t-1)})^{-1}]\}}{\chi^2_{(n+v)(p-1)}}, \tag{A.2}$$

where

$$\hat{\beta} = \left[\sum_{i=1}^n X_i^\top (\Sigma^{(t-1)})^{-1} X_i + A\right]^{-1} \left[\sum_{i=1}^n X_i^\top (\Sigma^{(t-1)})^{-1} \tilde{W}_i^*\right],$$

$$\tilde{\beta}^* \sim \text{N}\left[\hat{\beta}, (\alpha^2)^* \left(\sum_{i=1}^n X_i^\top (\Sigma^{(t-1)})^{-1} X_i + A\right)^{-1}\right], \tag{A.3}$$

and set $\beta^{(t)} = \tilde{\beta}^*/\alpha^*$.

Step 3: Finally, draw $\tilde{\Sigma}^{(t)}$ given $\beta^{(t)}$ and $(\tilde{W}_i^* - X_i\tilde{\beta}^*)$, for $i = 1, \ldots, n$ via

$$\tilde{\Sigma}^* \sim \text{Inv Wishart}\left[n + v, \tilde{S} + \sum_{i=1}^n (\tilde{W}_i^* - X_i\tilde{\beta}^*)(\tilde{W}_i^* - X_i\tilde{\beta}^*)^\top\right], \tag{A.4}$$

set $\Sigma^{(t)} = \tilde{\Sigma}^*/\tilde{\sigma}_{11}^*$ and $W_{ij}^{(t)} = \tilde{W}_{ij}^*/\sqrt{\tilde{\sigma}_{11}^*}$ for each i and j, and for "Scheme 2" set $(\alpha^2)^{(t)} = \tilde{\sigma}_{11}^*$.

Here, superscript stars indicate quantities that are intermediate and not part of the Markov chain, $(W^{(t)}, \beta^{(t)}, \Sigma^{(t)})$ and $(W^{(t)}, \beta^{(t)}, \Sigma^{(t)}, \alpha^{(t)})$ for "Scheme 1" and "Scheme 2", respectively.

We now give the details of Algorithm 2.

Algorithm 2. *Step* 1a: Draw W_{ij}^* just as in Step 1 of Algorithm 1. Then draw $(\alpha^2)^*$ from $p(\alpha^2 \mid \beta, \Sigma) = p(\alpha^2 \mid \Sigma)$ as given in (6) and set $\tilde{W}_{ij}^* = \alpha^*(W_{ij}^* - X_i\beta^{(t-1)})$.

Step 1b: Draw $\tilde{\Sigma}^*$ given $\tilde{W}^*$ and $\beta^{(t-1)}$,

$$\tilde{\Sigma}^* \sim \text{Inv Wishart}\left[n + v, \tilde{S} + \sum_{i=1}^{n} \tilde{W}_i^* (\tilde{W}_i^*)^\top\right], \tag{A.5}$$

and set $\Sigma^{(t)} = \tilde{\Sigma}^*/\tilde{\sigma}_{11}^*$.

Step 1c: Set

$$W_i^{(t)} = \frac{1}{\sqrt{\tilde{\sigma}_{11}^*}} \tilde{W}_i^* + X_i\beta^{(t-1)} \quad \text{for each } i. \tag{A.6}$$

Step 2: Finally, draw $\beta^{(t)}$ given $\Sigma^{(t)}$ and $W^{(t)}$

$$\beta^{(t)} \sim \text{N}\left[\hat{\beta}, \left(\sum_{i=1}^{n} X_i^\top (\Sigma^{(t)})^{-1} X_i + A\right)^{-1}\right], \tag{A.7}$$

where

$$\hat{\beta} = \left[\sum_{i=1}^{n} X_i^\top (\Sigma^{(t)})^{-1} X_i + A\right]^{-1} \left[\sum_{i=1}^{n} X_i^\top (\Sigma^{(t)})^{-1} W_i^{(t)} + A\beta_0\right].$$

(Again superscript stars are used to indicate intermediate quantities.)

Appendix B. Nobile's Metropolis–Hastings acceptance rate

To compute the correct acceptance rate for Nobile's (1998) Metropolis–Hastings rule, we first derive the target distribution,

$$p(\alpha \mid \beta, \Sigma) \propto p(\beta, \Sigma, \alpha) \tag{B.1}$$

$$\propto \alpha^{-(v(p-1)-k+1)} \exp\left\{-\frac{1}{2}\left[(\alpha\beta - \tilde{\beta}_0)^\top A(\alpha\beta - \tilde{\beta}_0) + \frac{1}{\alpha^2}\,\text{trace}(S\Sigma^{-1})\right]\right\}. \tag{B.2}$$

Thus,

$$R = \frac{p(\alpha' \mid \beta, \Sigma)}{p(\alpha \mid \beta, \Sigma)} \frac{J(\alpha \mid \alpha')}{J(\alpha' \mid \alpha)}$$

$$= \exp\left\{-\frac{1}{2}\left[\alpha^2\beta^\top A\beta(c^2-1) - 2\alpha\tilde{\beta}_0^\top A\beta(c-1) + \frac{1}{\alpha^2}\,\text{trace}(S\Sigma^{-1})\left(\frac{1}{c^2}-1\right)\right]\right\}$$

$$\times c^{-(v(p-1)-k+2)} \exp\left(c - \frac{1}{c}\right), \tag{B.3}$$

where $c = \alpha'/\alpha$. This differs from the value computed by Nobile (1998) by a factor of $c^{-(2-k-p(p-1)/2)}$.

References

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand.. MIT Press, Cambridge, MA.

Chib, S., Greenberg, E., Chen, Y., 1998. MCMC methods for fitting and comparing multinomial response models. Economics Working Paper Archive, Econometrics, No. 9802001, Washington University at St. Louis.

Chintagunta, P.K., Prasad, A.R., 1998. An empirical investigation of the "Dynamic McFadden" model of purchase timing and brand choice: implications for market structure. Journal of Business and Economic Statistics 16 (1), 2–12.

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91, 883–904.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulations using multiple sequences (with discussion). Statistical Science 7, 457–472.

Geweke, J., Keane, M., Runkle, D., 1994. Alternative computational approaches to inference in the multinomial probit model. The Review of Economics and Statistics 76, 609–632.

Haas, M., 1998. Value of IgG subclasses and ultrastructural markers in predicting latent membranous lupus nephritis. Modern Pathology 11, 147A.

Hausman, J.A., Wise, D.A., 1978. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. Econometrica 46, 403–426.

Keane, M.P., 1992. A note on identification in the multinomial probit model. Journal of Business and Economic Statistics 10, 193–200.

Liu, J.S., Wong, W.H., Kong, A., 1994. Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. Biometrika 81, 27–40.

Maddala, G., 1983. Limited-dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

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

McCulloch, R., Rossi, P., 2000. Reply to Nobile. Journal of Econometrics 99, 347–348.

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

Meng, X.-L., van Dyk, D.A., 1999. Seeking efficient data augmentation schemes via conditional and marginal augmentation. Biometrika 86, 301–320.

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

Nobile, A., 2000. Comment: Bayesian multinomial probit models with normalization constraint. Journal of Econometrics 99, 335–345.

Quinn, K.M., Martin, A.D., Whitford, A.B., 1999. Voter choice in multi-party democracies: a test of competing theories and models. American Journal of Political Science 43, 1231–1247.

Roberts, G.O., 1996. Markov chain concepts related to sampling algorithms. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

Speckman, P.L., Lee, J., Sun, D., 1999. Existence of the MLE and propriety of posteriors for a general multinomial choice model. Technical Report, Department of Statistics, University of Missouri at Columbia.

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

Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). The Annals of Statistics 22, 1701–1762.

Tierney, L., 1996. Introduction to general state-space markov chain theory. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, London.

van Dyk, D.A., Meng, X.-L., 2001. The art of data augmentation (with discussions). The Journal of Computational and Graphical Statistics 10, 1–111.

van Dyk, D.A., Kang, H., Meng, X.-L., 2004. Using a marginal gibbs sampler to improve bayesian computation for the glmm. Technical Report.