NISS

A Hybrid Markov Chain for the Bayesian Analysis of the Multinomial Probit Model

Agostino Nobile

Technical Report Number 36
August, 1995

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

A Hybrid Markov Chain for the Bayesian Analysis of the Multinomial Probit Model

Agostino Nobile[1]

National Institute of Statistical Sciences

and

Institute of Statistics and Decision Sciences, Duke University

August 16, 1995

[1] Research supported by NSF grants DMS9208758 and DMS9313013 to NISS. Comments from the participants in the NISS Transportation Project are gratefully acknowledged.

Abstract

Bayesian inference for the Multinomial probit model, using the Gibbs sampler with data augmentation, has been recently considered by some authors. The present paper introduces a modification of the sampling technique, by defining a hybrid Markov chain in which, after each Gibbs sampling cycle, a Metropolis step is carried out along a direction of constant likelihood. Several candidate distributions for the Metropolis step are considered. Examples with two simulated and one real data sets motivate and illustrate the new technique. A proof of the ergodicity of the hybrid Markov chain is also given.

Keywords: Multinomial probit model, Gibbs sampling, Metropolis algorithm, Bayesian analysis.

1 Introduction

The multinomial probit (MNP) model belongs to the wider class of discrete choice models. In these models, see e.g. Ben-Akiva and Lerman (1985) and Anderson, de Palma and Thisse (1992), it is assumed that n agents (individuals, households, etc.) choose between p alternatives in a way to maximize their utility, which is modeled as some function (usually linear) of covariates and noise. MNP occurs when the noise is additive and multivariate normal.

The appeal of the MNP model is that it does not imply the independence of irrelevant alternatives (IIA) property, unlike other models, such as the multinomial logit. The IIA property says that, for any two alternatives in the set of alternatives, the ratio between their choice probabilities is left unaffected by adding some alternatives to (removing some alternatives from) the set. When it applies, IIA affords useful restrictions on the structure of the model, see, e.g., McFadden (1984). The property fails to hold when some alternatives are substitutes of others, thus, in many applications, IIA is considered as an unrealistic assumption.

Despite its appeal, MNP has been used relatively little because of the difficulties associated with its estimation, when the number of alternatives is not very small. Recently several simulation methods have been proposed to carry out the computations (McFadden 1989, Geweke 1991, Hajivassilou and McFadden 1990, Keane 1994, Albert and Chib 1993, McCulloch and Rossi 1994), thus rekindling the interest in MNP. This paper concerns a modification of the Gibbs sampling with data augmentation scheme advocated by McCulloch and Rossi (1994) to perform a Bayesian analysis of the MNP model. Section 2 reviews the MNP model and some issues related to its use. Section 3 describes the Bayesian approach, giving the prior distributions of the parameters and the full conditional distributions used by the Gibbs sampler. Some binomial probit examples motivate the proposed modification, illustrated in Section 4, and consisting in performing, after each Gibbs cycle, a Metropolis step along a subset of constant likelihood. Proof that the resulting Markov chain is ergodic is given in the appendix.

2 The Multinomial Probit Model

Let $y_i = (y_{i1}, \ldots, y_{ip})^T$ be a multinomial vector, with $y_{ij} = 1$ if agent i chooses alternative j, $y_{ij} = 0$ otherwise. A more compact representation of the choices is afforded by a vector $d = (d_1, \ldots, d_n)^T$ containing the indexes of the chosen alternatives: $d_i = j$ if $y_{ij} = 1$. Agent i is assumed to maximize its (unobserved) utility z_{ij} over the alternatives' set, so that

$$d_i = j \Longleftrightarrow z_{ij} \geq \max_{1 \leq b \leq p} z_{ib}. \tag{1}$$

The vector of utilities z_i of agent i satisfies:

$$z_i = R_i\beta + u_i \qquad i = 1, \ldots, n; \tag{2}$$

where R_i is a $p \times k$ matrix of covariates and

$$u_i \overset{i.i.d.}{\sim} \mathrm{N}(0, V). \tag{3}$$

Equations (1), (2), (3) specify the multinomial probit model. More general specifications are possible, see, e.g., Geweke, Keane and Runkle (1994a, 1994b). The one given above suffices for the purpose of the present paper, the extension of the proposed technique to more general models being straightforward.

Note that one can add an arbitrary constant to both sides of (2) while leaving the distribution of the data vector d unchanged. This identification problem (Dansie 1985, Bunch 1991) is commonly dealt with by subtracting the p-th equation in (2) from the first $p-1$, obtaining:

$$w_i = X_i\beta + \epsilon_i \qquad i = 1, \ldots, n; \tag{4}$$

$$\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \Sigma), \tag{5}$$

where $w_{ij} = z_{ij} - z_{ip}$, $X_{ijh} = R_{ijh} - R_{iph}$, $\epsilon_{ij} = u_{ij} - u_{ip}$ and the covariance matrix of the new error term satisfies:

$$\Sigma = [I_{p-1}, -1_{p-1}]\, V\, [I_{p-1}, -1_{p-1}]^T,$$

with I denoting the identity matrix and 1 a vector of 1's. The choice vector d can be re-expressed in terms of the utility differentials w_{ij} as follows:

$$d_i = \begin{cases} j & \text{if } w_{ij} = \max\limits_{1 \leq b \leq p-1} w_{ib} > 0 \\ 0 & \text{if } \max\limits_{1 \leq b \leq p-1} w_{ib} < 0 \end{cases} . \tag{6}$$

Yet, the model given by (4), (5) and (6) is still lacking identification since multiplication of both sides of (4) by a positive constant leaves unaltered the distribution of d. The usual way this problem is solved consists in restricting the $(1,1)$ element of Σ to be unity: $\sigma_{1,1} = 1$, thus implicitly assigning the arbitrary multiplicative constant.

The multinomial choice probability vector of agent i is easily shown to be

$$P_{ij} = \Pr[d_i = j] = \int_{E_j} \mathrm{N}(\epsilon; 0, \Sigma)\, d\epsilon \qquad 1 \leq j < p \tag{7}$$

where the sets E_j in the above $(p-1)$-dimensional multivariate normal integrals are given by

$$E_j = \bigcap_{b \neq j} \left\{ \epsilon_{ij} - \epsilon_{ib} > (x_{ib} - x_{ij})^T \beta \right\} \cap \left\{ \epsilon_{ij} > x_{ij}^T \beta \right\}$$

and x_{ij}^T denotes the $j-th$ row of X_i. Unless the number of alternatives is very small, computation by quadrature of the integrals in (7) is difficult. Since the likelihood function of (β, Σ) is

$$\ell(\beta, \Sigma) = \prod_{i=1}^{n} \prod_{j=1}^{p} [P_{ij}(\beta, \Sigma)]^{y_{ij}} \tag{8}$$

the method of maximum likelihood requires very accurate estimates of the choice probability, generally unavailable by quadrature.

Lerman and Manski (1981) suggested the method of simulated maximum likelihood (SML), where Monte Carlo estimates of the choice probabilities, obtained from a relative frequency estimator, are used. However, they found that a very large simulation sample size is needed when P_{ij} is small. McFadden (1989) introduced the method of simulated moments (MSM), based on the solution of some moments conditions involving the choice probabilities, which he suggested to estimate using a smoothed version of the frequency estimator. Both SML and MSM have greatly benefited by the development of the GHK probability simulator (Geweke 1991, Hajivassilou and McFadden 1990, Keane 1994). McCulloch and Rossi (1994) have developed a Bayesian approach using Gibbs sampling with data augmentation, expanding upon earlier work of Albert and Chib (1993). Geweke, Keane and Runkle (1994a) compare the performance of SML using the GHK simulator, MSM using the GHK simulator and the Bayesian approach using Gibbs sampling, by means of some Monte Carlo experiments. The overall conclusion is that the Bayesian method seems to have a clear edge on the other methods, especially when covariates are correlated and error variances vary across alternatives. Similar conclusions are reached by Geweke, Keane and Runkle (1994b) in comparing the relative performance of the above methods when estimating the multinomial multiperiod probit model.

3 The Bayesian approach

This section summarizes the Bayesian approach of McCulloch and Rossi (1994): the reader is referred to their paper for more details as well as extensions, such as hierarchical models, of the basic MNP.

In a Bayesian approach, the specification of a model is complete only after the assignment of a prior distribution for the parameter. A convenient prior specification for the MNP is as follows: $G = \Sigma^{-1}$ is assumed to have Wishart distribution with $\nu > p$ degrees of freedom and precision matrix P (see e.g. DeGroot (1970)):

$$\Sigma^{-1} = G \sim \mathrm{W}_{p-1}(\nu, P). \tag{9}$$

Indepedently of Σ, β is assumed multivariate normal with mean vector $\beta_{(0)}$ and covariance matrix $A_{(0)}^{-1}$:

$$\beta|\Sigma \sim \mathrm{N}_k(\beta_{(0)}, A_{(0)}^{-1}). \tag{10}$$

On occasion the parameter vector will be denoted by θ and the prior (9)-(10) by $\pi(\theta)$.

The hyperparameters ν and $A_{(0)}$ can be chosen so that the prior distribution is proper and at the same time rather diffuse. This will yield a posterior distribution mostly reflecting the shape of the likelihood and not depending much on the prior location parameters P and $\beta_{(0)}$. Though convenient, the above prior specification does not impose the identifying constraint $\sigma_{11} = 1$. Therefore the sampling part of the model is still lacking identification, which is achieved only by the use of a proper prior. This allows one to sample from the posterior distribution of (β, Σ) and then make inference about some "identified" functionals, such as $\beta/\sqrt{\sigma_{11}}$, σ_{ii}/σ_{11} and $\rho_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$.

Sampling from the posterior distribution of (β, Σ) is done using the Gibbs sampler with data augmentation (Gelfand and Smith 1990, Tanner and Wong 1987 are basic references on these methods). The general idea of Gibbs sampling is that if the parameter vector θ can be partitioned as $\theta = \{\theta_1, \ldots, \theta_m\}$ and the full conditional distributions

$$[\theta_i | \{\theta_j, j \neq i\}, Data] \tag{11}$$

are available, then successively drawing from these distributions will asymptotically yield a draw from the posterior of θ. In some cases the distributions (11) are available only conditional on some vector W of latent variables. Then, one augments the available data with W and also simulates from the conditional distribution of W, thus employing, in place of (11),

$$[\theta_i | \{\theta_j, j \neq i\}, W, Data]$$

$$[W | \theta, Data] .$$

The first application of these sampling methods to the MNP model is in Albert and Chib (1993), who propose a rejection technique to sample from the truncated multivariate normal distribution of the utility differentials w_i. McCulloch and Rossi (1994), instead, use the following partitioning of the variables to be successively simulated using the Gibbs sampler:

$$\{w_{11}, w_{12}, ..., w_{n,p-1}, \beta, \Sigma\}.$$

This amounts to replacing the difficult draw from $[w_i | \beta, \Sigma, d]$ with simpler univariate draws that will converge to a draw from it. The full conditional distributions in the McCulloch and Rossi scheme are reported below.

The distribution of the utility differential w_{ij} given the other w's, Σ , β, and the data d, is a truncated normal distribution with truncation point depending on whether the alternative j is the one selected by agent i. To write it down explicitely, some notation is needed. Let

$$\begin{array}{rl} x_{ij}^T & \text{be the } j\text{-th row of } X_i; \\ X_{i,-j} & \text{be } X_i \text{ with the } j\text{-th row deleted;} \\ w_{i,-j} & \text{be } w_i \text{ with } w_{ij} \text{ deleted;} \\ \sigma_{j,-j} & \text{be the } j\text{-th column of } \Sigma \text{ with } \sigma_{jj} \text{ deleted;} \\ \Sigma_{-j,-j} & \text{be } \Sigma \text{ with the } j\text{-th row and column deleted;} \end{array}$$

also let

$$\begin{aligned} g_{jj} &= \left[\sigma_{jj} - \sigma_{j,-j}^T \Sigma_{-j,-j}^{-1} \sigma_{j,-j}\right]^{-1}, \\ g_{j,-j} &= -\Sigma_{-j,-j}^{-1} \sigma_{j,-j} g_{jj}. \end{aligned}$$

Then

$$w_{ij}|w_{i,-j},\Sigma,\beta,d \sim \mathrm{N}(m_{ij},\tau_{ij}^2)\left[I_{\{d_i=j\}}I_{[\max(w_{i,-j},0),\infty)}(w_{ij}) + I_{\{d_i\neq j\}}I_{(-\infty,\max(w_{i,-j},0)]}(w_{ij})\right], \tag{12}$$

where $\tau_{ij}^2 = g_{jj}^{-1}$ and $m_{ij} = x_{ij}^T\beta - g_{jj}^{-1}g_{j,-j}^T(w_{i,-j} - X_{i,-j}\beta)$.

The distribution of G given the w's, β and the data is Wishart:

$$G|W,\beta,d \sim \mathrm{W}_{p-1}\left(\nu+n,\ P+\sum_{i=1}^n \epsilon_i\epsilon_i^T\right), \tag{13}$$

where $W = \{w_1,\ldots,w_n\}$.

Finally, the full conditional of β is multivariate normal:

$$\beta|W,\Sigma,d \sim \mathrm{N}_k(\beta_{(1)},A_{(1)}^{-1}), \tag{14}$$

where

$$\beta_{(1)} = A_{(1)}^{-1}\left(A_{(0)}\beta_{(0)} + \sum_{i=1}^n X_i^{*T}w_i^*\right);$$

$$
\begin{aligned}
A_{(1)} &= A_{(0)} + \sum_{i=1}^{n} X_i^{*T} X_i^*; \\
X_i^* &= L^T X_i; \\
w_i^* &= L^T w_i;
\end{aligned}
$$

and $G = LL^T$ is the Cholesky decomposition of G.

Example 1. As an illustration, $n = 2000$ observation were generated from the model (1)-(3) with $p = 2$, $\beta = -2$, the covariate R sampled from a Uniform(-0.5,0.5) distribution and V equal to I_2. Then the error variance in the model (4)-(6) is $\sigma = 2$ and one is interested in making inference about the "identified" parameter $\beta/\sqrt{\sigma}$, with true value $-\sqrt{2}$. This special case (binomial probit with only two parameters) was chosen for two reasons: first, in a binomial probit model choice probabilities are readily computed, involving only a univariate normal distribution; second, with only two parameters, prior, likelihood and posterior are easily displayed. Figure 1 (a) is a contour plot of the prior distribution, with β on the horizontal axis and $\sqrt{\sigma}$ on the vertical axis. The following values of the hyperparameters were employed: $\nu = 3$, $P = 3$, $\beta_{(0)} = 0$, $A_{(0)} = 0.01$. The ratio between function values on adjacent contour lines is 1/4. Figure 1 (b) contains a contour plot of the likelihood function: it is constant over lines out of the origin: the likelihood is only informative about the ratio $\beta/\sqrt{\sigma}$. Part (c) of Figure 1 is a countour plot of the posterior distribution, which, although a proper distribution, is seen to be rather concentrated about the line $\beta/\sqrt{\sigma} = -\sqrt{2}$, where the likelihood is largest.

Figure 2 reproposes the contour plots of the posterior distribution of $(\beta, \sqrt{\sigma})$, using a doubly logarithmic scale on the axes, for better resolution. Superimposed on the posterior contour plots are points representing the pairs $(\beta, \sqrt{\sigma})$ yielded by the first 2000 draws from two Gibbs sampling with data augmentation chains, one started at the true parameter values $(-2, \sqrt{2})$ (part (a)), the other started relatively far fom it, at $(-20, 10)$ (part (b)). The same stream of random numbers was used in both runs. One may note that the second chain quickly moves close to $\beta/\sqrt{\sigma} = -\sqrt{2}$, yet finds it somewhat difficult to move along this line to the region where the posterior is highest. Since one is only interested in the posterior of the identified paramater $\beta/\sqrt{\sigma}$, one may wonder about the relevance of the above observation, and in fact in the present example the estimates from the two chains are practically the same. This need not be true in general, though, since, unlike the likelihood, the prior $\pi(\beta, \sqrt{\sigma}) = \pi(\theta)$ does not posses the following property:

$$\forall\, c, c' > 0, \quad \forall\, \theta, \theta' \qquad \frac{\pi(c\theta)}{\pi(c\theta')} = \frac{\pi(c'\theta)}{\pi(c'\theta')}.$$

Example 2. To illustrate these difficulties, consider a binomial probit model with one covariate, $\beta = 5$, $\sigma = 2$, the covariates R were generated using draws from a Binomial(1/2) distribution. As in the previous example, 2000 observations were generated from this model and the same prior for β, σ was used. Prior, likelihood and posterior contour plots are displayed in Figure 3. Since in this example the magnitute of the systematic part of the utility is much larger than the error term's ($\beta x \in \{-5, 0, 5\}$), the likelihood function, for the given sample, is not very informative. It is approximately constant for $\beta/\sqrt{\sigma}$ above a certain value and decreases very fast for smaller values, thus in practice only ruling out a region of the parameter space. As a result, the posterior is much more widespread than in the former example, as it is evident from part (c) of Figure 3. Figure 4 displays again contour plots of the posterior distribution, with superimposed the first 20000 pairs $(\beta, \sqrt{\sigma})$ obtained from two sampling chains run with the same stream of random numbers and different starting points: $(5, \sqrt{2})$ and $(25, 5)$. It is clear that, if started away from the region of high posterior, the chain finds it difficult to move towards it. This time, moreover, the inferences from the two chains are rather different, as attested by the histograms of the sampled $\beta/\sqrt{\sigma}$ reported in Figure 5.

This second example is perhaps extreme, in that the likelihood only gives a lower bound to $\beta/\sqrt{\sigma}$. However one should note that it was produced using a very simple model: it would not be wise to rule out similar situations to arise in models with many more alternatives and covariates, espacially when the latter comprise dummy variables. These remarks somewhat echo the concers of Keane (1992) about the "extremely tenuous" parameter identification in the MNP, meaning that an identified model may have a likelihood which does not vary much over a wide range of parameters including the maximizer. These are cases where the prior distribution may matter, e.g. a more diffuse prior in the precedeing example would lead to a posterior giving more mass to higher values of $\beta/\sqrt{\sigma}$. Still, once the prior is assigned, one would expect the simulation procedure to yield, after discarding enough observations from the sampling chain (burn-in), a sample representative of the posterior, irrespective of the starting values. It just seems that, using the procedure described in this section, burn-in times may be very long.

4 A hybrid chain

In this section a modification of the sampling scheme described in Section 3 is proposed and its performance in the examples therein is examined. The idea is very simple: after each Gibbs cycle through the full conditional distributions of W and θ, perform a Metropolis step to change the scale of the current state. This allows the chain to move faster across the parameter/latent data space. The additional computational cost is minimal since the Metropolis candidate is selected in a direction of constant likelihood, so that one needs only to evaluate the prior density.

The Metropolis algorithm was introduced by Metropolis et al. (1953) and generalized by Hastings (1970). Tierney (1994) illustrates the flexibility of the algorithm and its use in combined simulation strategies.

Suppose one wants to sample from a distribution with density $f(\psi)$, with respect to some σ-finite measure μ, and let $\psi^{(0)}$ be an initial state vector. The Metropolis algorithm moves from the current state $\psi^{(n)}$ to the next one $\psi^{(n+1)}$, by first selecting a candidate state ψ^* according to some distribution $H(\psi^*|\psi^{(n)})$, with density $h(\psi^*|\psi^{(n)})$ with respect to μ. The candidate is accepted, and $\psi^{(n+1)} = \psi^*$, with probability

$$a(\psi^{(n)}, \psi^*) = \min\left[\frac{f(\psi^*)h(\psi^{(n)}|\psi^*)}{f(\psi^{(n)})h(\psi^*|\psi^{(n)})}, 1\right]. \tag{15}$$

If the candidate is rejected, the chain remains at the current state: $\psi^{(n+1)} = \psi^{(n)}$. Note that f enters in the acceptance probability only as the ratio $f(\psi^*)/f(\psi^{(n)})$, so that one need only be able to evaluate f up to a proportionaly constant.

Turning to sampling for the MNP model, let's redefine, for the sake of simplicity, the parameter vector as $\theta = (vec_L(S), \beta)$, where S is the lower triangular Cholesky factor of Σ and vec_L stacks in a vector s_{ij}, $j \leq i$. Since one needs to sample from the joint posterior distribution of θ and W, it is $\psi = (\theta, W)$ and

$$f(\psi) \propto \pi(\theta) \cdot g(W|\theta) \cdot m(d|\theta, W),$$

where π is the prior and g and m have obvious meaning. Consider the Markov chain which progresses from the current state $\psi^{(n)} = \{\theta^{(n)}, W^{(n)}\}$ to the next one by first using the Gibbs sampling procedure of Section 3 and then rescaling the resulting state $\psi' = \{\theta', W'\}$ as follows. A Metropolis step is performed with the candidate $\psi^* = \{\theta^*, W^*\}$ sampled from some distribution $H_{\Psi^*|\Psi'}(\psi^*|\psi')$ with support

$$S_{\psi'} = \{\psi^* : \psi^* = c\psi', c > 0\}.$$

With probability $a(\psi', \psi^*)$ the candidate is accepted, in which case $\psi^{(n+1)} = \psi^*$; if it is rejected then the next state is the outcome of the Gibbs cycle: $\psi^{(n+1)} = \psi'$.

Because of the particular form of the support $S_{\psi'}$, the term $f(\psi^*)/f(\psi')$ in the acceptance probability reduces to $\pi(\theta^*)/\pi(\theta')$, due to the lack of identification noted following (6). Therefore one needs only evaluate the ratio of prior densities at the parameter yielded by the Gibbs cycle and at the parameter sampled from the candidate distribution:

$$\begin{aligned}
&\frac{\pi(\theta^*)}{\pi(\theta')} = \\
&= \frac{\exp\{-\frac{1}{2}(\beta^* - \beta_{(0)})^T A_{(0)}(\beta^* - \beta_{(0)})\} \, |S^*|^{-(\nu+p)} \prod_{j=1}^{p-1} {s^*_{jj}}^{p-j} \exp\{-\frac{1}{2} \operatorname{tr} P(S^* S^{*T})^{-1}\}}{\exp\{-\frac{1}{2}(\beta' - \beta_{(0)})^T A_{(0)}(\beta' - \beta_{(0)})\} \, |S'|^{-(\nu+p)} \prod_{j=1}^{p-1} {s'_{jj}}^{p-j} \exp\{-\frac{1}{2} \operatorname{tr} P(S' S'^T)^{-1}\}} \\
&= c^{-(p-1)(\nu+\frac{p}{2})} \exp\left\{-\frac{1}{2}\left[(c^2-1){\beta'}^T A_{(0)}\beta' - 2(c-1){\beta'}^T A_{(0)}\beta_{(0)} + (c^{-2}-1) \operatorname{tr} P(S'S'^T)^{-1}\right]\right\},
\end{aligned}$$

where $\theta^* = c\theta'$ and, in the first equation, use was made of the prior distribution of S, $\pi(S) \propto |P|^{\nu/2}|S|^{-(\nu+p)} \prod_{j=1}^{p-1} s_{jj}^{p-j} \exp\{-\frac{1}{2} \operatorname{tr} P(SS^T)^{-1}\}$.

Concerning the candidate distribution H, it is assumed that $h(\psi^*|\psi') > 0 \iff h(\psi'|\psi^*) > 0$, so that a certain move is allowed if and only if the move in the reverse direction is also allowed.

Because of the form of $S_{\psi'}$, sampling of the candidate ψ^* is done easiest by sampling the scale factor C according to some univariate distribution F_C and then setting $\Psi^* = C\Psi'$. Then

$$\begin{aligned} H_{\Psi^*|\Psi'}(\psi^*|\psi') &= \Pr[\Psi^* \le \psi^*|\Psi' = \psi'] \\ &= \Pr[C\Psi' \le \psi^*|\Psi' = \psi'] \\ &= F_C\left(\frac{\psi_1^*}{\psi_1'}\right) \end{aligned}$$

where, in the first two lines, the inequalities are componentwise and the last passage can be done because of the special form of $S_{\psi'}$. If μ is Lebesgue measure and F_C has density f_C with respect to μ, then

$$h(\psi^*|\psi') = f_C\left(\frac{\psi_1^*}{\psi_1'}\right)\frac{1}{\psi_1'}.$$

Some possible choices for F_C are reported below, along with the corresponding values of the ratio $h(\psi'|\psi^*)/h(\psi^*|\psi')$ needed to compute the acceptance probability.

(a) F_C is a mixture of a point mass at γ and a point mass at γ^{-1}. The density, with respect to counting measure, is:

$$f_C(c) = \lambda I_{\{\gamma\}}(c) + (1-\lambda)I_{\{\gamma^{-1}\}}(c) \qquad \gamma \in (1,\infty),\ \lambda \in (0,1);$$

$$\frac{h(\psi'|\psi^*)}{h(\psi^*|\psi')} = \begin{cases} (1-\lambda)/\lambda & \text{if } \psi_1^*/\psi_1' = \gamma \\ \lambda/(1-\lambda) & \text{if } \psi_1^*/\psi_1' = \gamma^{-1} \end{cases}.$$

(b) F_C is a finite mixture of two uniform distributions on the intervals $(1,\gamma)$ and $(\gamma^{-1},1)$:

$$f_C(c) = \lambda\frac{1}{(\gamma-1)}I_{(1,\gamma)}(c) + (1-\lambda)\frac{1}{(1-\gamma^{-1})}I_{(\gamma^{-1},1)}(c) \quad \gamma \in (1,\infty),\ \lambda \in (0,1);$$

$$\frac{h(\psi'|\psi^*)}{h(\psi^*|\psi')} = \begin{cases} \gamma\dfrac{1-\lambda}{\lambda}\dfrac{\psi_1'}{\psi_1^*} & \text{if } \dfrac{\psi_1^*}{\psi_1'} \in (1,\gamma) \\ \gamma^{-1}\dfrac{\lambda}{1-\lambda}\dfrac{\psi_1'}{\psi_1^*} & \text{if } \dfrac{\psi_1^*}{\psi_1'} \in (\gamma^{-1},1) \end{cases}.$$

(c) C equals U with probability λ and U^{-1} with probability $1-\lambda$, where U is uniform on $(1,\gamma)$:

$$f_C(c) = \lambda\frac{1}{\gamma-1}I_{(1,\gamma)}(c) + (1-\lambda)\frac{1}{\gamma-1}\frac{1}{c^2}I_{(\gamma^{-1},1)}(c) \quad \gamma \in (1,\infty),\ \lambda \in (0,1);$$

$$\frac{h(\psi'|\psi^*)}{h(\psi^*|\psi')} = \begin{cases} \dfrac{1-\lambda}{\lambda}\dfrac{\psi_1^*}{\psi_1'} & \text{if } \dfrac{\psi_1^*}{\psi_1'} \in (1,\gamma) \\ \dfrac{\lambda}{1-\lambda}\dfrac{\psi_1^*}{\psi_1'} & \text{if } \dfrac{\psi_1^*}{\psi_1'} \in (\gamma^{-1},1) \end{cases}.$$

(d) F_C is a gamma distribution with parameters δ and γ:

$$f_C(c) = \frac{\gamma^\delta}{\Gamma(\delta)} c^{\delta-1} e^{-\gamma c};$$

$$\frac{h(\psi'|\psi^*)}{h(\psi^*|\psi')} = \left(\frac{\psi_1'}{\psi_1^*}\right)^{2\delta-1} \exp\left\{-\gamma\left[\frac{\psi_1'}{\psi_1^*} - \frac{\psi_1^*}{\psi_1'}\right]\right\}.$$

In cases (a), (b) and (c), values of γ very close to 1 are likely to result in a high acceptance rates, yet the chain will not move much, in any step, from the output of the Gibbs cycle. Values of γ very large will instead result in very low acceptance rates, since most of the proposals will fall in a region of relatively low prior probability. Extreme values of λ should be avoided. In case (d), one may recommend using a gamma distribution with unit mean, for example the $Exp(1)$ distribution.

The appendix contains a proof that the hybrid Markov chain, consisting of the combination of the Gibbs cycle of Section 3 and the rescaling Metropolis step, is ergodic (positive Harris recurrent and aperiodic).

The simulation technique described above, using F_C equal to the $Exp(1)$ distribution, was employed with the two artificial data sets of Section 3. Figure 6 displays two sets of 2000 simulated pairs $(\beta, \sqrt{\sigma})$, superimposed on a contour plot of the posterior distribution, for the first example. The simulation was carried out twice, using different starting points, the same as used in the simulations displayed in Figure 2. Comparing the panels in Figure 6 and Figure 2 suggests that the influence of the starting point is greatly reduced, moreover the distributions of sampled points in Figure 6 seem to better match the posterior distribution, as described by the countour lines.

Figure 7 displays the results for the second example. It contains plots of two sets of 20000 pairs $(\beta, \sqrt{\sigma})$ obtained by running the hybrid sampling chain using the two different starting points already used in Section 3. Compared with the ones in Figure 4, the results of the new technique appear much less dependent on the simulation starting point and the overall fit between the posterior contours and the sampled points is greatly improved. Figure 8 contains histograms of the "identified" parameter $\beta/\sqrt{\sigma}$ obtained from the two simulations. There is a great improvement in the agreement between the results of the two runs, with respect to the results reported in Figure 5.

Example 3. I next illustrate the relevance of the new technique using a MNP model with three alternatives estimated on a real data set. Data from wave 7 of the Dutch mobility panel (see, e.g., Van Wissen and Meurs 1989), was used to model car ownership level (1, 2 or more, 0 cars) in terms of socio-demographic characteristics of the households. Details on the specification of the model will be reported elsewhere, along with estimates based on several waves in the panel. Here I concentrate on the differences between the posterior samples obtained using the Gibbs sampler and the hybrid Markov chain. Two runs, each consisting of 20000 simulation cycles and with the same stream of random numbers, were performed using the two methods. The $Exp(1)$ distribution was used as candidate distribution for the scale factor C in the hybrid chain. Some of the results are reported in Figure 9, which displays time series plots of some identified parameters in the model. Each panel contains two series: larger dots denote the output of the Gibbs sampler, smaller dots the output of the hybrid chain. Panel (a) refers to the constant term in the equation for 2 or more cars, while panel (b) refers to the ratio between the variances and panel (c) to the correlation coefficient in the bivariate normal error term. In all panels the time series produced by the hybrid chain reaches rather quickly its average level and exhibits a stationary behavior. The Gibbs sampler output seems to require longer times to reach the steady state and in all panels some evidence of trend, either in mean or in variance, is apparent. Using the simulation sample size of 20000, inferences from the two simulation techniques are not very different, for the quantities in panels (a) and (b). However, inferences about the correlation coefficient differ greatly.

5 Conclusions

In this paper I described a modification of the Gibbs sampler used for the Bayesian analysis of the multinomial probit model. The modification consists in performing, after each Gibbs cycle, a

Metropolis step along a direction of constant likelihood. It is relatively inexpensive and, in some examples with artificial and real data sets, seemed to improve considerably the sampler's ability of exploring the posterior distribution.

Appendix: convergence of the hybrid chain

Let E be the state space of the sampling chain: E is the direct product of the parameter space of $\theta = (vec_L(S), \beta)$ and the subset of $\mathbb{R}^{n(p-1)}$, determined by the observed choices, where the utilities differentials w_i live. The letters x, y, z will be used to denote elements of E. Let $\mathcal{E}$ be the Borel σ-field on E.

Denote by $P_G : E \times \mathcal{E} \to [0,1]$ the transition probability kernel on $(E; \mathcal{E})$ associated with the Gibbs sampler: $P_G(x, A)$ is the probability of making a transition from $x \in E$ to $A \in \mathcal{E}$ by means of a Gibbs cycle. It is assumed that $P_G(\cdot, A)$ is an $\mathcal{E}$-measurable function for any $A \in \mathcal{E}$, and that $P_G(x, \cdot)$ is a probability on $(E, \mathcal{E})$ for any $x \in E$.

Similar definitions apply to P_M, the transition kernel associated with the Metropolis step. Denote by $Q(y, A)$ the candidate transition kernel (in terms of the notation of Section 4, $Q(y, A) = \int_A H(dx|y))$ and let $a(y, z)$ be the probability of accepting the candidate z sampled according to $Q(y, dz)$. Then P_M can be written as

$$P_M(y, A) = \int_A a(y, z)Q(y, dz) + r(y)\delta_y(A)$$

where δ_y is a point mass at y and $r(y)$ is the marginal probability of rejecting the candidate:

$$r(y) = 1 - \int a(y, z)Q(y, dz).$$

The transition kernel of the hybrid chain is

$$(P_G P_M)(x, A) = \int P_G(x, dy)P_M(y, A). \tag{16}$$

Let φ denote the posterior distribution and μ the Lebesgue measure, mutually absolutely continuous, on E.

The following results hold:

(i) φ is invariant for $(P_G P_M)$.

It follows from the fact that φ is invariant for both P_G and P_M.

(ii) For every $x \in E$, for every $A \in \mathcal{E}$ with $\varphi(A) > 0$, $(P_G P_M)(x, A) > 0$.

Consider

$$B = \{y : P_M(y, A) > 0\} = \{y : y = cx,\ c \in \{1\} \cup D,\ x \in A\} \tag{17}$$

where $D \subseteq \mathrm{I\!R}^+$ (for the candidate distributions reported in Section 4, one has: (a) $D = \{\gamma^{-1}, \gamma\}$; (b and c) $D = (\gamma^{-1}, 1) \cup (1, \gamma)$; (d) $D = \mathrm{I\!R}^+$). Clearly $B \supset A$ so that $\varphi(B) > 0$. Now the Gibbs transition kernel P_G is defined in terms of strictly positive densities with respect to μ (see McCulloch and Rossi 1994). Therefore for all $x \in E$, $P_G(x, B) > 0$. It then follows, from the definition (16), that $(P_G P_M)(x, A) > 0$.

(iii) $(P_G P_M)$ is φ-irreducible.

It follows from (ii).

(iv) $(P_G P_M)$ is aperiodic.

It follows from (ii).

(v) $(P_G P_M)$ is absolutely continuous with respect to φ: $(P_G P_M)(x, \cdot) \ll \varphi$, for all $x \in E$.

Let $A \in \mathcal{E}$ with $\mu(A) = 0$. We show that $(P_G P_M)(x, A) = 0$ for all $x \in E$, separately for the case of $Q \ll \mu$ and Q discrete.

Consider first the case where the candidate transition kernel Q is absolutely continuous with respect to μ, as in (b), (c) and (d) of Section 4. Then $\int_A a(y, z) Q(y, dz) \leq Q(y, A) = 0$, for all $y \in E$. Therefore $P_M(y, A) = r(y)\delta_y(A)$, so that

$$(P_G P_M)(x, A) = \int_A r(y) P_G(x, dy) \leq P_G(x, A) = 0 \qquad \forall x \in E$$

since $P_G(x, \cdot) \ll \mu$.

If the candidate transition kernel Q is discrete, as in (a) of Section 4, then P_M is discrete and D in (17) is a countable set. Then $\mu(A) = 0$ implies $\mu(B) = 0$, so that $P_G(x, B) = 0$, for all

$x \in E$, follows. Hence $(P_G P_M)(x, A) = 0$, $\forall x \in E$.

(vi) $(P_G P_M)$ is Harris recurrent.

It follows from (i), (iii), (v) and Corollary 1 in Tierney (1994).

Finally, using (i), (iii), (iv) and (vi) as the conditions of Theorem 1 in Tierney (1994), one can conclude that $(P_G P_M)$ is positive Harris recurrent, φ is the unique invariant distribution and

$$||(P_G P_M)^n(x, \cdot) - \varphi|| \to 0 \qquad \forall x \in E,$$

where $(P_G P_M)^n$ is the n-th iterate of the transition kernel and $|| \cdot ||$ denotes the total variation distance.

References

[1] Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association*, 88, 669-679.

[2] Anderson, S. P., de Palma, A. and Thisse, J.-F. (1992). *Discrete Choice Theory of Product Differentiation*, The MIT Press.

[3] Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, The MIT Press.

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

[5] Dansie, B. R. (1985). Parameter Estimability in the Multinomial Probit Model. *Transportation Research*, 19B, 526-528.

[6] DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill.

[7] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398-409.

[8] Geweke, J. (1991). Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 571-578.

[9] Geweke, J., Keane, M. and Runkle, D. (1994a). Alternative Computational Approaches to Inference in the Multinomial Probit Model. *The Review of Economics and Statistics*, 76, 609-632.

[10] Geweke, J., Keane, M. and Runkle, D. (1994b). Statistical Inference in the Multinomial Multiperiod Probit Model. Federal Reserve Bank of Minneapolis, Research Department, Staff Report 177.

[11] Hajivassiliou, V. and McFadden, D. (1990). The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises. Cowles Foundation Discussion Paper 967, Yale University.

[12] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

[13] Keane, M. P. (1992). A Note on Identification in the Multinomial Probit Model. *Journal of Business & Economic Statistics*, 10, 193-200.

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

[15] Lerman, S. and Manski, C. (1981). On the use of simulated frequencies to approximate choice probabilities. In C. Manski and D. McFadden (eds.), *Structural analysis of discrete data with econometric applications*, 305-319, The MIT Press.

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

[17] McFadden, D. (1984). Econometric analysis of qualitative response models. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, vol. II, 1395-1457, North-Holland.

[18] McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica*, 57, 995-1026.

[19] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 27, 1087-1092.

[20] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-540.

[21] Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics*, 22, 1701-1762.

[22] Wissen, L. van and Meurs, H. (1989). The Dutch mobility panel: Experiences and evaluation. *Transportation*, 16, 99-119.

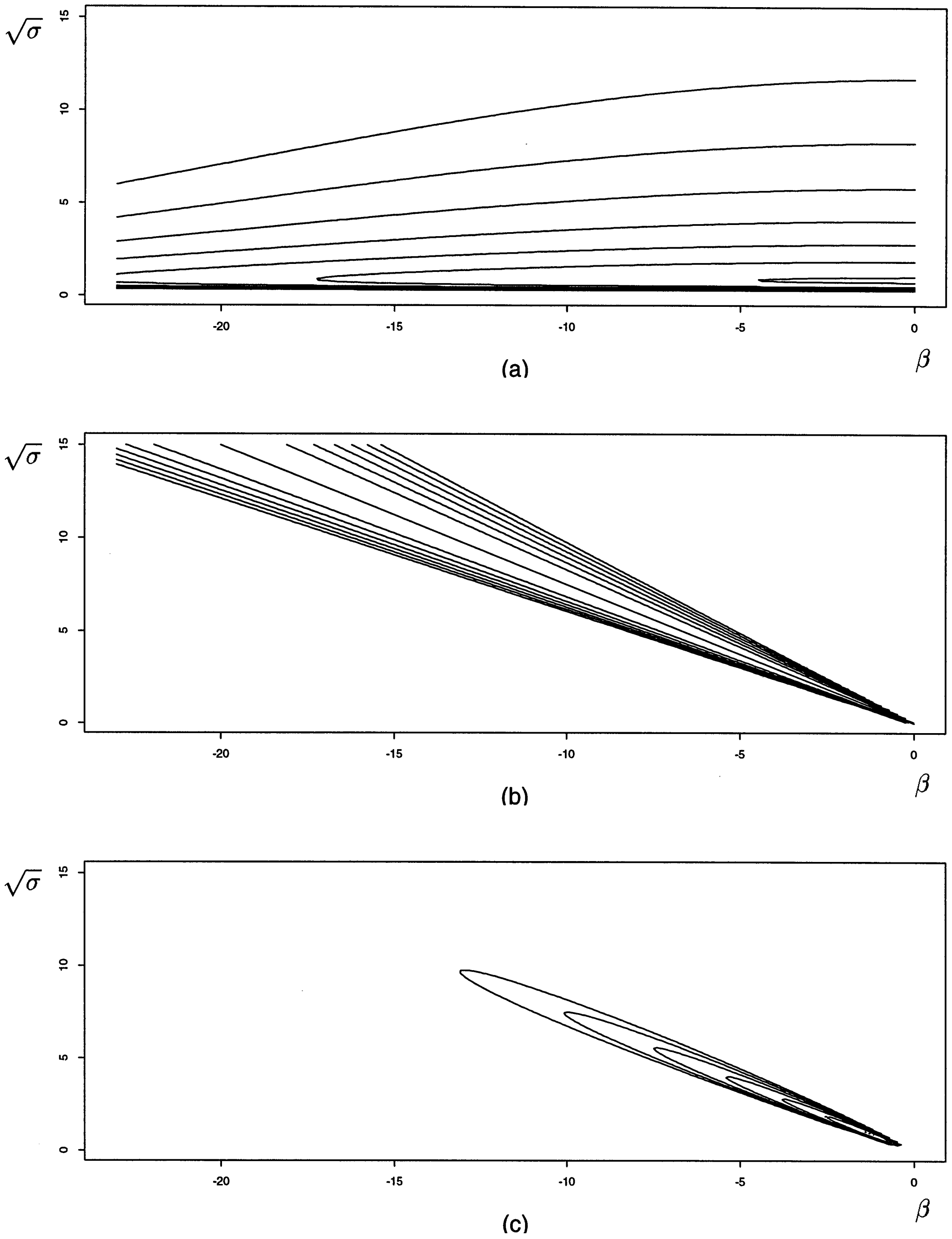


Figure 1: Example 1. Contour plots of: (a) the prior; (b) the likelihood; (c) the posterior.

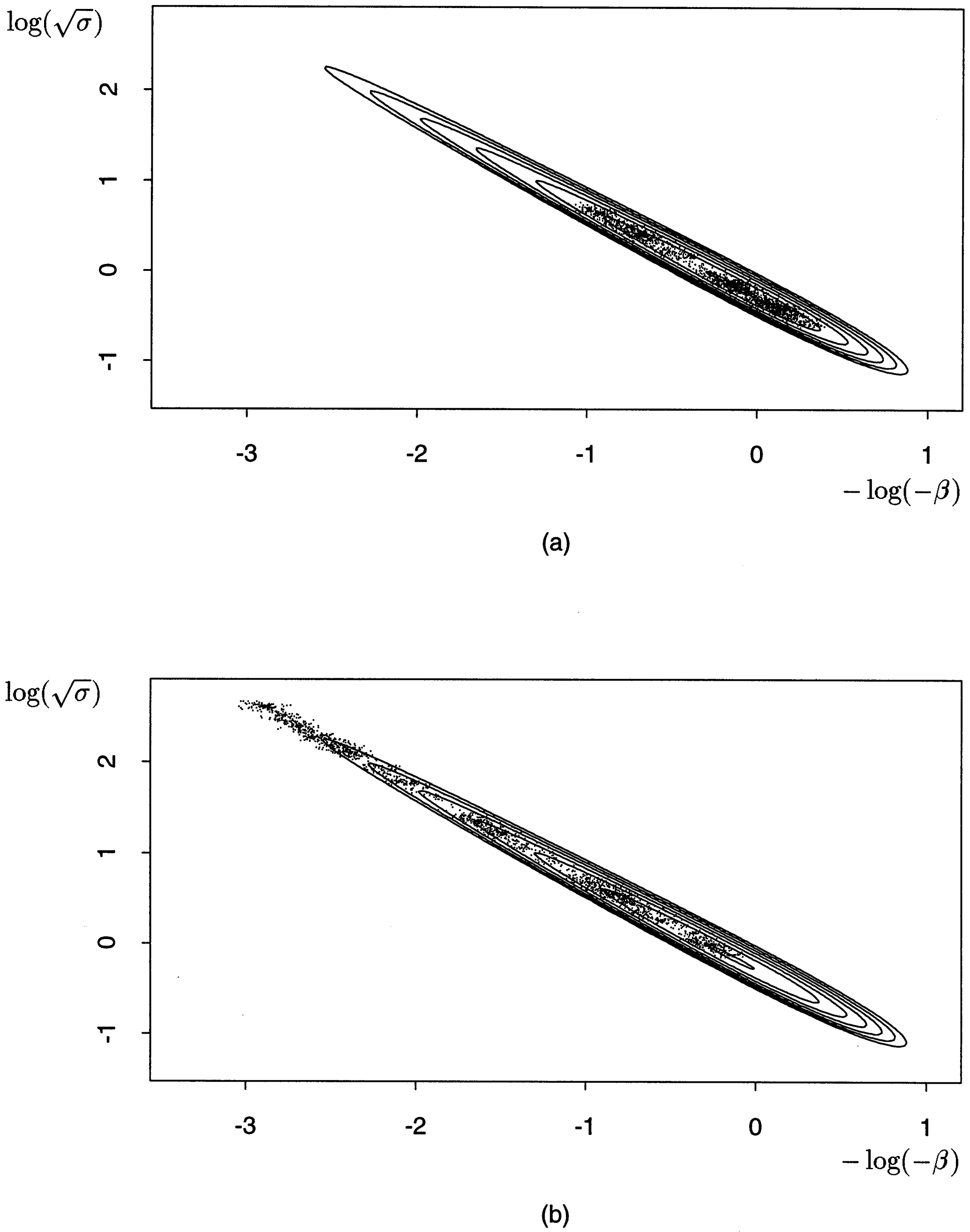
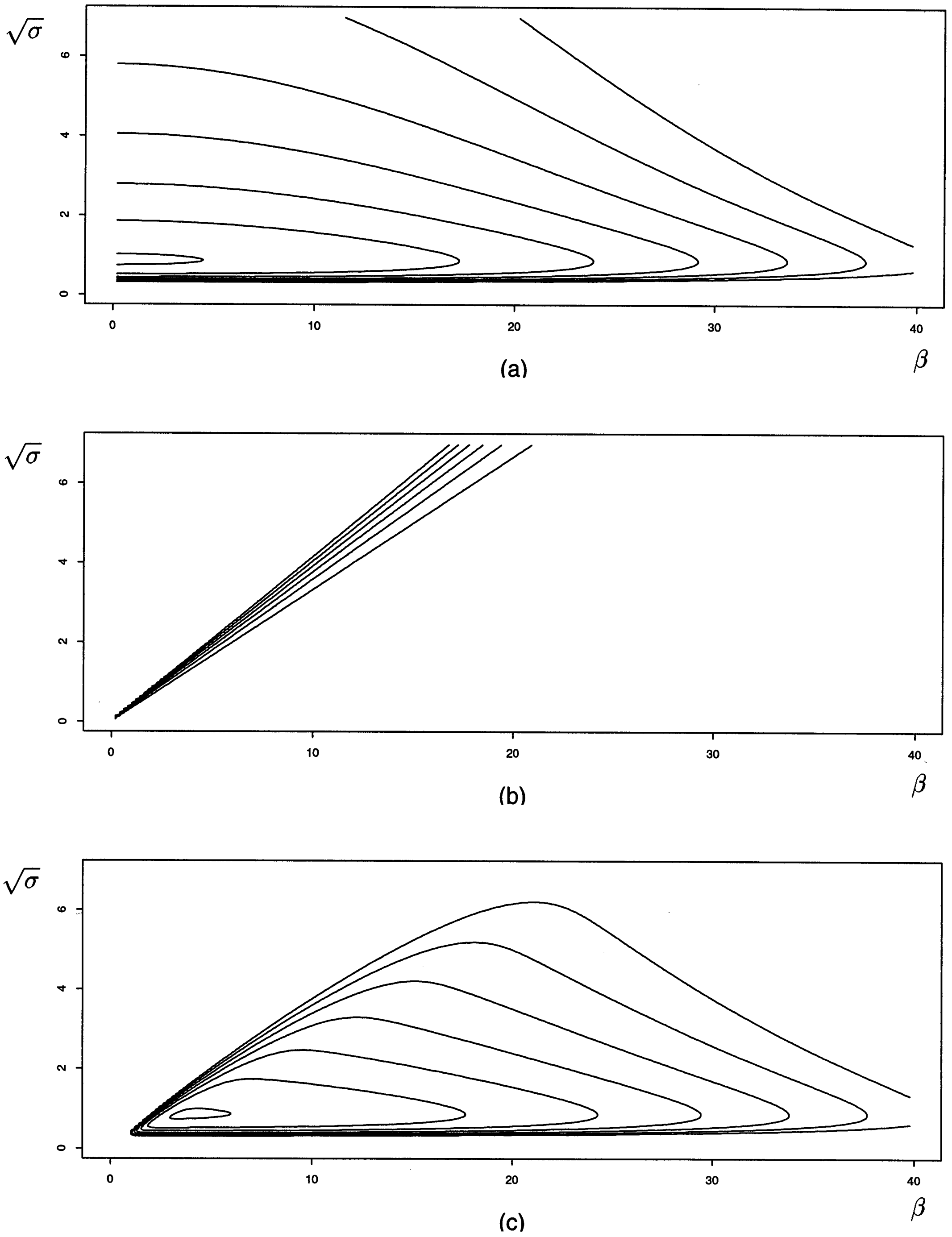


Figure 2: Example 1. Contour plots of the posterior distribution with superimposed 2000 draws from the Gibbs sampler started at: (a) $(\beta, \sqrt{\sigma}) = (-2, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (-20, 10)$.



(a)

(b)

(c)

Figure 3: Example 2. Contour plots of: (a) the prior; (b) the likelihood; (c) the posterior.

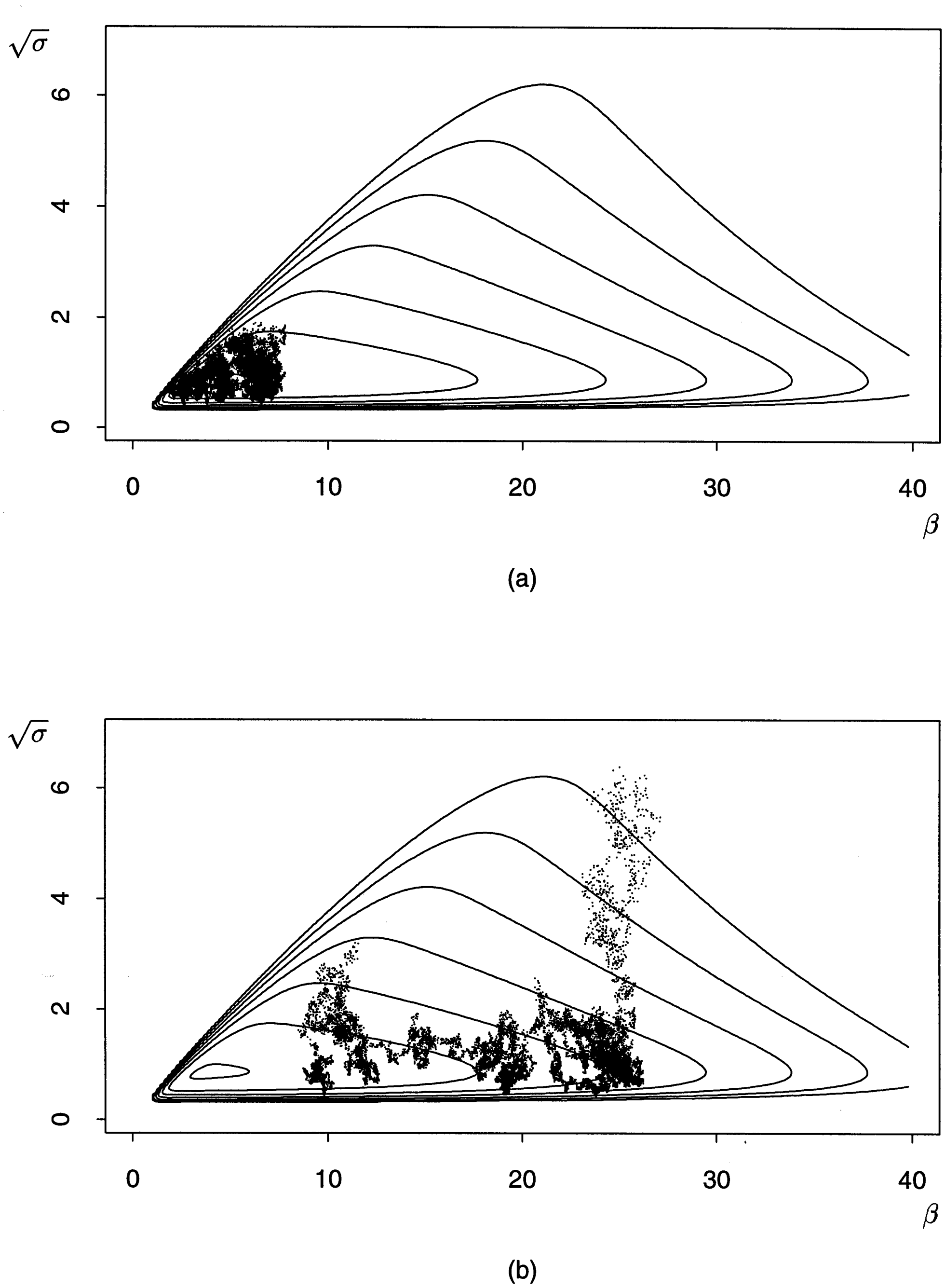


Figure 4: Example 2.Contour plots of the posterior distribution with superimposed 20000 draws from the Gibbs sampler started at: (a) $(\beta, \sqrt{\sigma}) = (5, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (25, 5)$.

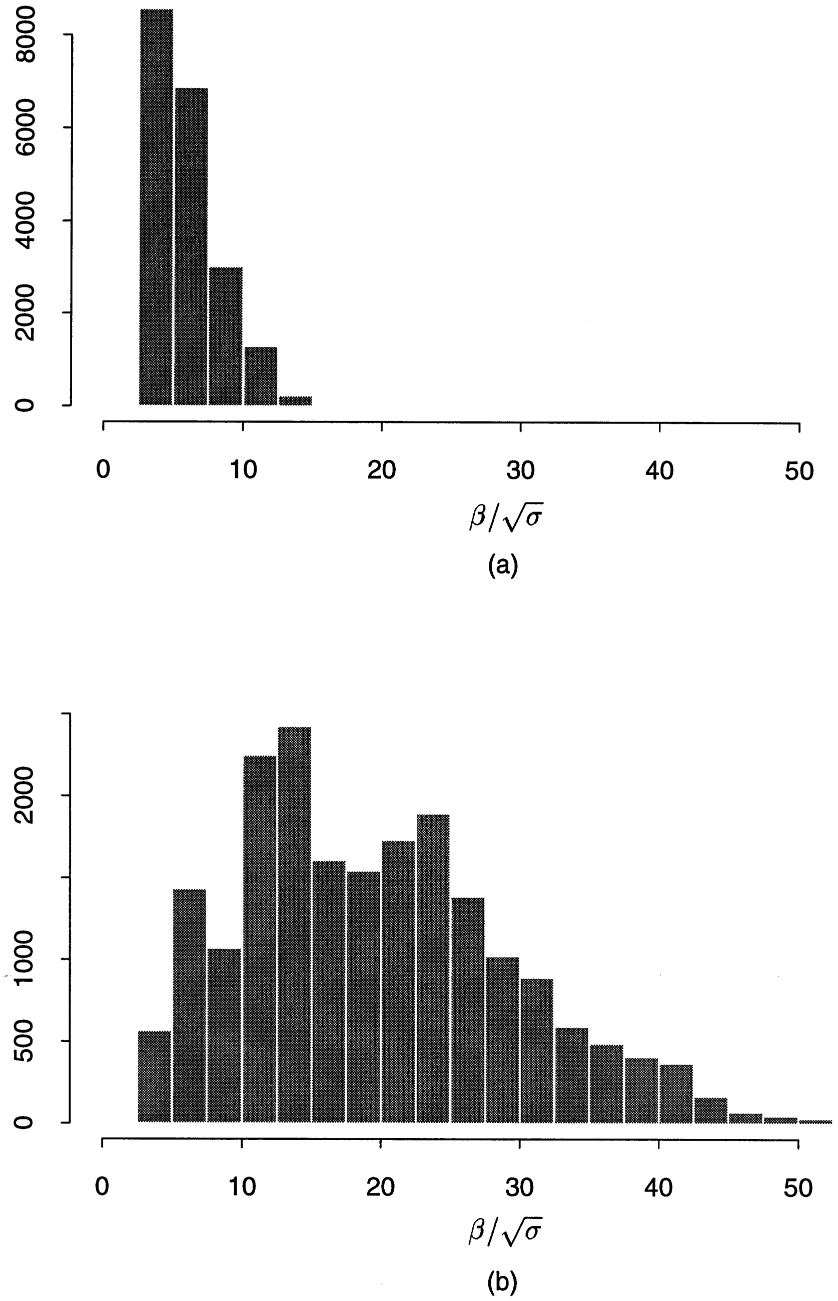
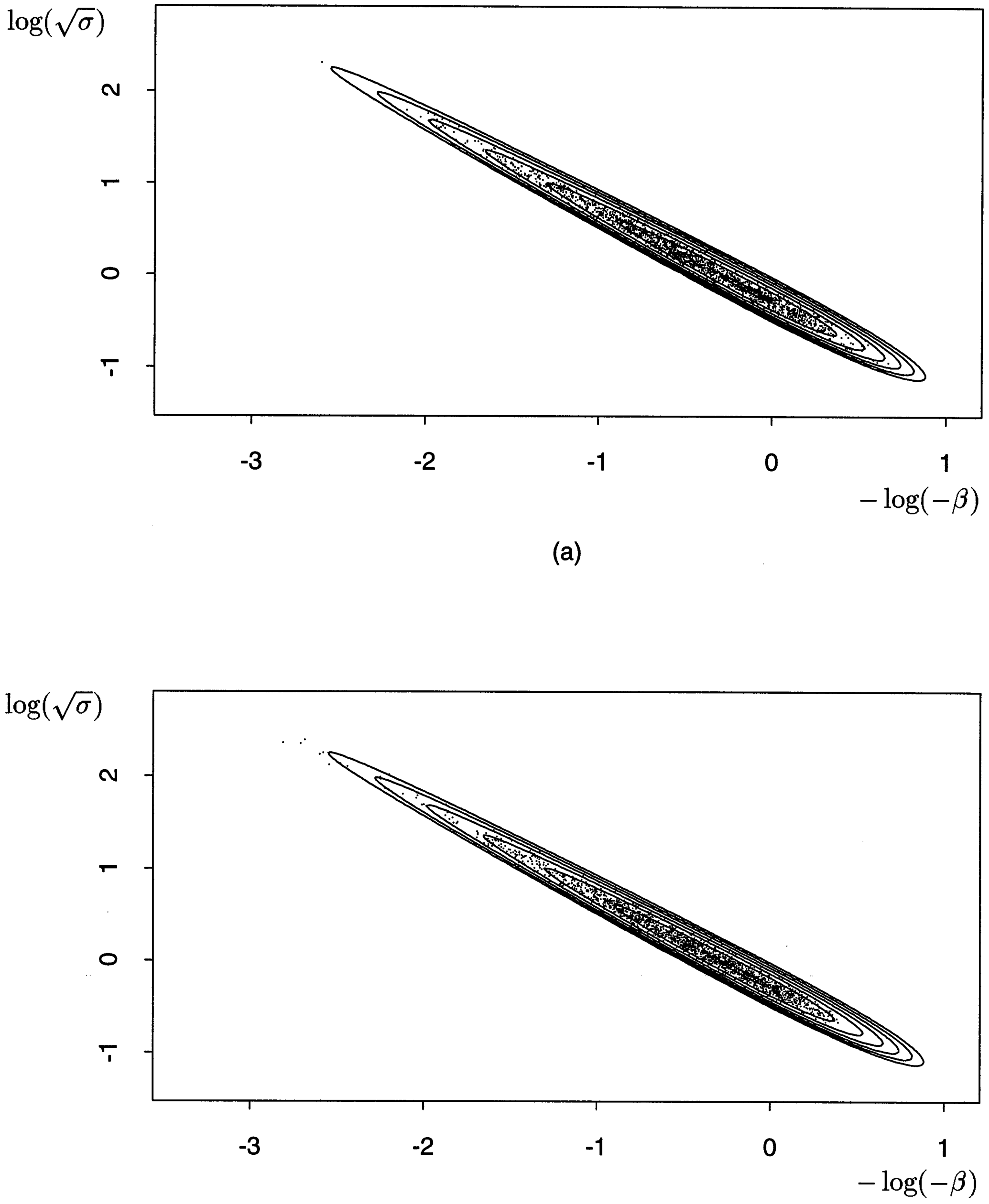


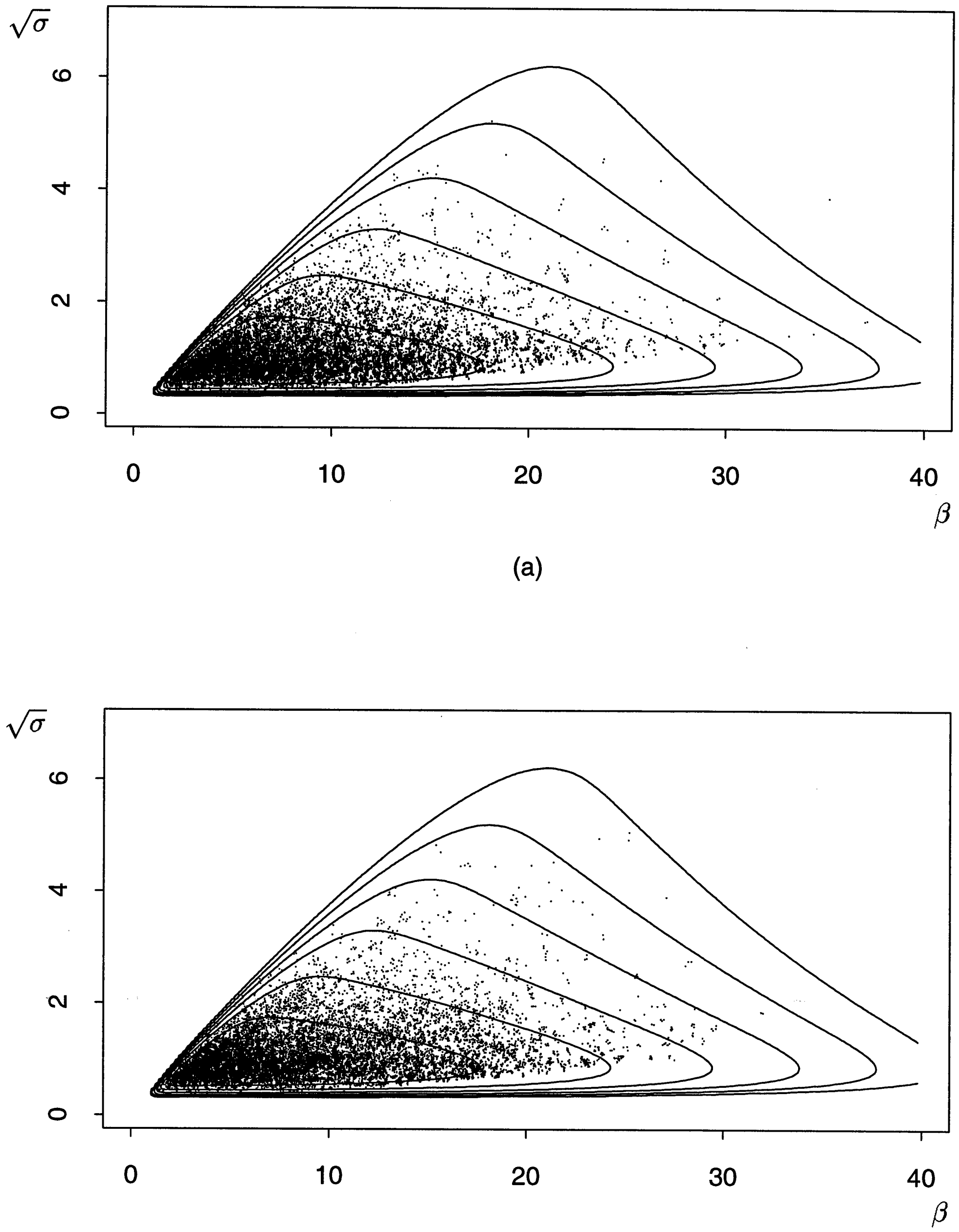
Figure 5: Example 2. Histograms of $\beta/\sqrt{\sigma}$ computed on 20000 draws from the Gibbs sampler started at: (a) $(\beta, \sqrt{\sigma}) = (5, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (25, 5)$.



(a)

(b)

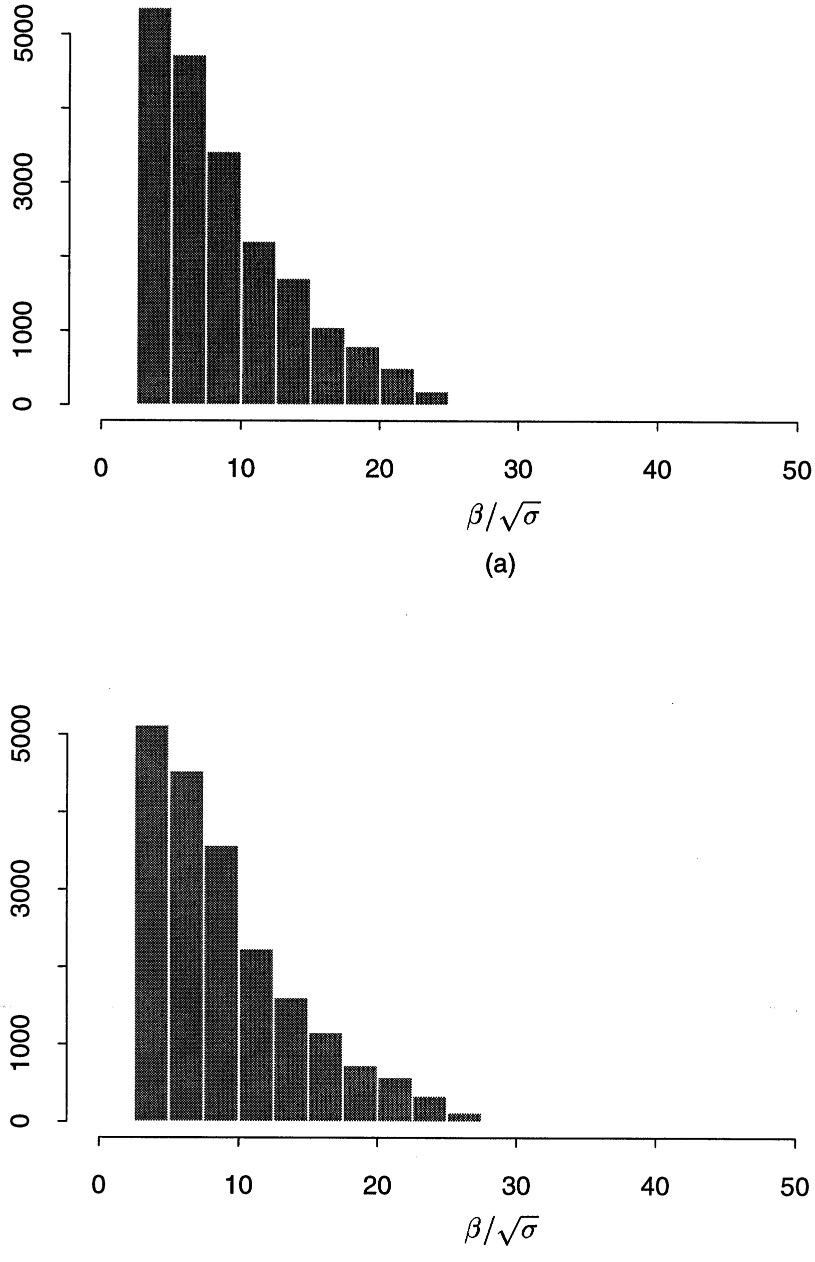
Figure 6: Example 1. Contour plots of the posterior distribution with superimposed 2000 draws from the hybrid chain with exponential candidate distribution in the Metropolis step, started at: (a) $(\beta, \sqrt{\sigma}) = (-2, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (-20, 10)$.



(a)

(b)

Figure 7: Example 2.Contour plots of the posterior distribution with superimposed 20000 draws from the hybrid chain with exponential candidate distribution in the Metropolis step, started at: (a) $(\beta, \sqrt{\sigma}) = (5, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (25, 5)$.



(a)

(b)

Figure 8: Example 2. Histograms of $\beta/\sqrt{\sigma}$ computed on 20000 draws from hybrid chain with exponential candidate distribution in the Metropolis step, started at: (a) $(\beta, \sqrt{\sigma}) = (5, \sqrt{2})$; (b) $(\beta, \sqrt{\sigma}) = (25, 5)$.

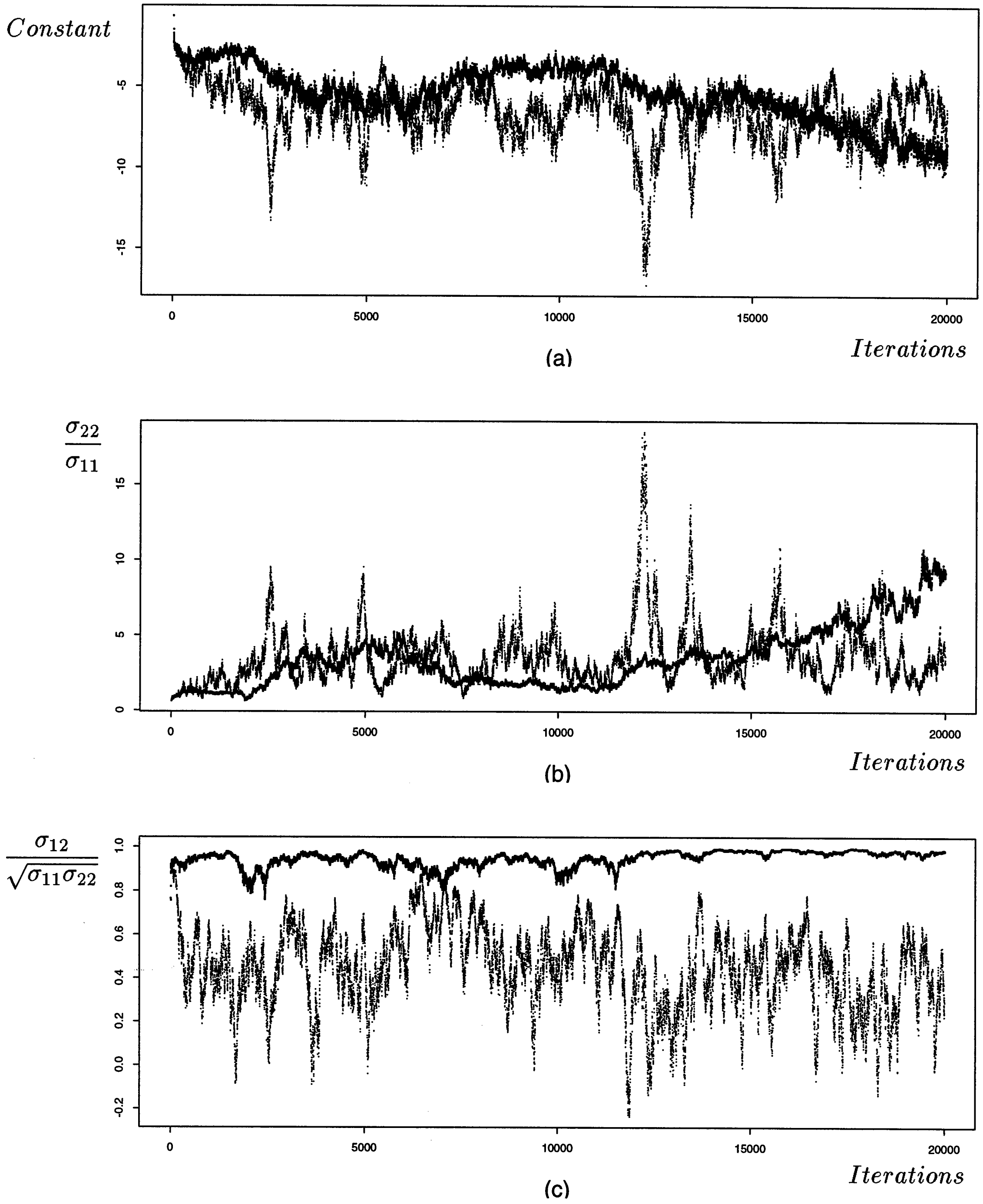


Figure 9: Example 3. Time series plots of some identified parameters in the car ownership model: (a) constant in the equation for 2 or more cars; (b) ratio of the variances of the error term; (c) correlation coefficient of the error term. Larger dots denote the output of the Gibbs sampler with data augmentation, smaller dots the output of the hybrid chain.