BU-1313-MB

Gibbs Sampling with Diffuse Proper Priors: A Valid Approach to Data-Driven Inference?

BY RANJINI NATARAJAN

Center for Statistical Sciences,

Brown University, Providence, RI 02912

AND CHARLES E. MCCULLOCH

Biometrics Unit and Statistics Center,

Cornell University, Ithaca, NY 14853

Abstract

This paper demonstrates by example that the use of the Gibbs sampler with diffuse proper priors can lead to inaccurate posterior estimates. Our results show that such inaccuracies are not merely limited to small sample settings.

Keywords: Diffuse Priors, Impropriety, Maximum Likelihood Estimation, Non-informative Priors, Probit-Normal.

1 Introduction

In complicated statistical models (for example, hierarchical models) the elicitation and specification of subjective prior information can often be difficult. Thus, as a matter of convenience, priors based on formal rules (reference priors) are typically relied upon in practice. These priors are selected with the justification of letting the data speak for themselves. However, reference priors are generally improper (that is, do not integrate to a finite number) and do not always lead to proper posterior distributions. Some work has been done on characterizing improper priors which ensure propriety for various models (Ibrahim and Laud, 1991; Berger and Robert, 1993; Natarajan and McCulloch, 1995; Hobert and Casella, 1996; Yang and Chen, 1996), but no general rules have emerged so far. This fact is complicated further by the practical difficulty in recognizing impropriety. Monte Carlo methods such as the Gibbs sampler (Gelfand and Smith, 1990) can be implemented to sample from the posterior despite impropriety, so long as the full conditional specifications exist. Although the resulting Gibbs chains are null-recurrent and do not converge to a stationary distribution, they often move reasonably around the parameter space and provide no signs of trouble, for instance, by being trapped in absorbing states. Thus, the use of the sampler with improper priors can sometimes lead to seriously misleading results in practice.

One approach to a default Bayes analysis in this situation is to identify transformations of the parameters for which reference priors on the transformed scale ensure proper posterior distributions.h The uniform shrinkage priors of Strawdermann (1971) are based on this idea. However, in multivariate settings, suitable choices of transformations are not entirely obvious. Further, the form of the induced prior on the original parameter scale may be highly complicated and could lead to difficulties in sampling from the resulting posterior distributions. Thus, one easily-implemented practical suggestion to circumvent these concerns about impropriety has been to use very diffuse but conjugate proper priors (McCulloch and Rossi, 1994; Hamada and Wu, 1995; Hobert and Casella, 1996). These priors are flat relative to the likelihood and can be constructed by truncating the domain of an improper prior to a compact set or by using proper priors with very large variances. McCulloch and Rossi (1994) outline such an analysis for the multinomial probit model using normal priors with large variances on the fixed effects and inverted Wishart priors with small hyperparameter values for the variance components. With these conjugate distributions the full conditionals required by the Gibbs sampler have a very simple form and implementation is straightforward.

There are reasons to question whether the use of diffuse priors are a solution to this problem. First, parameters about which there is little information may suffer from sensitivity to the choice of even a diffuse prior. This point was discussed by Efron (1973) and more recently by Kass and Wasserman (1996). A second issue, which has received relatively little attention in the literature, is the computational difficulty which may accompany such an analysis. With increased prior diffusion, the convergence of Gibbs sampling methods can be extremely slow since it attempts to sample from an almost improper posterior. It may then become essential to generate a prohibitively large number of simulation samples to achieve convergence for the full vector of parameters, rendering such an approach practically infeasible.

In this note we illustrate the degree of the computational problem in using diffuse priors in a probit-normal hierarchy for clustered binary observations. A family of proper priors with limiting distributions which are improper are studied. Our focus will be on investigating the behavior of posterior modes estimated from Gibbs output for varying degrees of prior diffusion. We are specifically interested in posterior modes since diffuse Bayes procedures are often touted as a valid approach to performing exact likelihood inference. Thus, a comparison of posterior modes to the maximum likelihood estimator to evaluate these claims would be of relevance and interest. Our results show that, even with very large simulation samples, modal posterior estimates resulting from Gibbs chains with diffuse proper priors can be inaccurate.

2 Probit-Normal Hierarchy

Consider a probit-normal hierarchy with a single normally distributed random-effect u and a single fixed-effect β:

$$\begin{aligned} w_{ij} \mid u_i &\sim \text{independent Bernoulli}\,(p_{ij}), \quad i = 1, \ldots, n, \;\; j = 1, \ldots, r_i, \\ \Phi^{-1}(p_{ij}) &= x_{ij}\,\beta + u_i, \\ u_i &\sim \mathcal{N}(0,\, \theta), \end{aligned} \tag{1}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal with its inverse denoted by $\Phi^{-1}(.)$. This model corresponds to r_i binary observations drawn on each of n clusters. The random effects u provide a convenient mechanism to capture within-cluster correlation. Throughout this manuscript we consider a balanced layout, that is $r_i = r, \forall\, i$ for convenience. The parameters of interest are the fixed-effect β and component of variance θ.

Denote the vector of observed data by $\mathbf{w}$. Then the observed likelihood function $\ell(\beta, \theta\,; \mathbf{w})$ is given by

$$\ell(\beta, \theta\,; \mathbf{w}) \;=\; \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{r_i} \Phi\,(x_{ij}\beta + u_i)^{w_{ij}}\, \Phi\,(-x_{ij}\beta - u_i)^{1-w_{ij}}\; exp(-\frac{u_i^2}{2\theta})\, \frac{du_i}{\sqrt{2\pi\theta}}, \qquad (2)$$

and can be very accurately evaluated using Gauss-Hermite quadrature methods. It is therefore a good situation in which to make comparisons.

For the model in (1) Natarajan and McCulloch (1995) show that the following power family of improper priors:

$$\pi(\beta, \theta) \;\propto\; \theta^{-a-1}, \qquad -\infty < \beta < \infty,\;\; 0 < \theta < \infty,$$

where a is a pre-specified constant, lead to proper posteriors only if $-\frac{n}{2} < a < 0$ and the dimension of the polyhedral cone

$$\mathcal{C} \;=\; \{\beta, \alpha \,:\, (1 - 2w_{ij})\,(x_{ij}\beta + \alpha_i) \le 0,\;\; i \,=\, 1, \ldots, n,\;\, j \,=\, 1, \ldots, r_i\}\,,$$

is smaller than $n+1$. Thus, the classic non-informative prior, that is, when $a = 0$ (denote by $\pi^n(\beta, \theta)$) does not ever result in proper posterior distributions, while the flat prior, that is, when $a = -1$ (denoted by $\pi^f(\beta, \theta)$) fails to ensure propriety for data configurations which result in a full-dimensional $\mathcal{C}$.

3 Simulation Study

We examined the performance of the Gibbs sampler with diffuse proper priors for five data sets generated from the model in (1) corresponding to different choices of n and r. For each data set the cone $\mathcal{C}$ was full-dimensional, thereby resulting in improper posterior distributions both under the non-informative and flat prior. The true parameters used for each simulated data set were $\beta = -5$ and $\theta = 2$. The single covariate x_{ij} was generated independentally and identically from a uniform distribution on $(-0.5, 0.5)$. All programs implementing the analysis described here were written in GAUSS (Aptech Systems, 1994).

Five sample size configurations were considered. Three of the five data sets simulated had clusters of size 2. Such data arise naturally in paired samples, for example, the analysis of family psychology using married couples as the unit of analysis (Raudenbush, Brennan, and Barnett, 1995). In our simulations, the number of pairs n range from a small data set ($n = 25$), to a moderate ($n = 50$) and a large one ($n = 100$). We also considered two examples with larger cluster sizes: $n = 50$, $r = 4$ and $n = 15$, $r = 8$. The first situation may arise in a longitudinal study where repeated assessments on a subject form a cluster. The second scenario is often encountered in animal breeding and carcigenocity experiments where animals from the same litter form clusters. In this setting it is reasonable to expect a small number of groups (litters) but a moderate cluster size.

In order to implement a diffuse Bayes analysis for this model we considered proper counterparts of the flat prior and the classic non-informative prior. Proper versions of the flat prior were obtained by truncating the range to intervals of finite length. Two such distributions were considered: a uniform prior between $[-25, 25]$ for β and $[0, 25]$ for θ denoted by $\pi_1^f(\beta, \theta)$, and a uniform prior between $[-10, 10]$ for β and $[0, 10]$ for θ denoted by $\pi_2^f(\beta, \theta)$. More generally, for complicated problems where direct evaluations of the likelihood function (and consequently the ML estimate) are non-trivial, the region of truncation may be chosen to be substantially larger to avoid problems with cutting off potentially important pieces of the likelihood function. However, for the examples we consider, $\pi_1^f(\beta, \theta)$ is sufficiently diffuse. Proper versions of the classic non-informative prior were obtained by postulating a normal prior for β and inverted gamma (IG) priors for θ. (A random variable X is said to follow an IG(a,b) distribution if its probability density function is given by $f(x) = b^a\, exp(-b/x) x^{-a-1}/\Gamma(a)$, $x > 0$, $a > 0$, $b > 0$.) Three specifications were considered ranging from very diffuse to very tight:

$$\begin{aligned}
\pi_1^n(\beta, \theta) &\sim \mathcal{N}\left(-5, 10^3\right) \times IG\,(0.01, 0.1)\,, \\
\pi_2^n(\beta, \theta) &\sim \mathcal{N}\left(-5, 10^2\right) \times IG\,(0.1, 0.5)\,, \\
\pi_3^n(\beta, \theta) &\sim \mathcal{N}\left(-5, \sqrt{10}\right) \times IG\,(5, 5)\,.
\end{aligned}$$

In all cases the location of the normal prior was taken to be the true parameter value.

For each prior specification, posterior modes were calculated by maximizing the Rao-Blackwellized estimate of the joint posterior using a sample generated by the Gibbs sampler. (A Gibbs implementation for this model is described by McCulloch and Rossi, 1994.) This calculation was replicated across ten independent runs of the sampler. Each run generated a sample of size 10,000 collected after discarding (burning-in) the initial 1,000 variates. These numbers are significantly larger than those suggested and used in the literature for problems with only two parameters. Thus, we are in fact providing the Gibbs chains with ample opportunity to achieve convergence. The posterior modes for each of the proper priors were also calculated by maximizing a numerical integration estimate of the joint posterior distribution (direct Bayes), for comparison purposes. For small values of the shape (a) and scale (b) parameters, we experienced instabilities in numerically evaluating the denominator of the posterior, that is, the marginal distribution of the data. Thus, we ignored the normalizing constant and simply maximized the numerator. This direct approach is possible for our simple example since the likelihood function in (2) is not difficult to evaluate. However, in the more general setting we envision, direct Bayes will not be feasible due to the computational intractability of likelihood calculations.

The performance of the Gibbs sampler was investigated using a one sample t-statistic to test whether the Gibbs estimates across the ten independent runs reproduced direct Bayes on average. For those cases where we failed to reject this hypothesis of equality, we then considered how close the Gibbs estimates came to reproducing direct Bayes, by calculating two measures of closeness. The first measure was the percent increase in variance (% incr. var.) contributed because of using a simulation based approach rather than doing an exact calculation:

$$\%\text{ incr. var.} \quad = \quad 100 \times \frac{\text{variance of Gibbs}}{\text{true variance}}.$$

The second is the number of multiple runs required to approximate the direct Bayes estimate to one decimal place of accuracy with $100(1-\alpha)$ percent confidence (required runs).

We estimated "variance of Gibbs" by calculating the variability in the Gibbs modes across the ten independent runs and we estimated "true variance" by using the curvature of the logarithm posterior at the modal estimate. We judged an increase of variance of 10% or less to be acceptable. This corresponds to being willing to use a simulation-based method as long as it does not increase the variance of the estimator by 10% or more. Values greater than 10% seem like a large price to pay for noise due purely to simulation. Our experiences with simulated maximum likelihood and Monte Carlo EM for probit- and logit-normal problems such as these is that those methods typically add less than 2% variance.

The number of multiple runs for single-digit accuracy is obtained by an approximate power calculation:

$$\text{required runs} \;\approx\; \frac{z^2_{1-\frac{\alpha}{2}} \times \text{ variance of Gibbs}}{(0.05)^2},$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)$ quantile for the Gaussian distribution. In all our calculations we considered $\alpha = 0.05$. Roughly speaking, we may interpret $10,000 \times$ required runs as the simulation sample size needed to be within ± 0.05 of the direct Bayes estimate.

4 Results

Figure 1 displays the non-informative prior π^n and its diffuse proper versions. The diffuse priors π^n_1 and π^n_2 corresponding to θ are represented modulo the normalizing constant to allow representation of all priors on the same scale. The non-informative prior for θ has an infinite peak at zero, while the diffuse proper priors mimic this behavior by placing large mass for small θ and tapering off for larger values. Figure 2 displays the likelihood surface for each of the five data sets. The likelihood function is fairly flat along the θ- direction for all five data sets, which is common for parameters situated lower in hierarchical specifications (Kass and Wassermann, 1996). Table 1 a) – e) reports the results of our prior sensitivity experiment for the five data sets. The maximum likelihood estimates of β and θ and associated asymptotic standard errors are reported for comparison for each data set. The Gibbs estimates are the average of ten independent runs of the sampler with simulation standard errors in parentheses. For several runs, we monitored changes in the Rao-Blackwellized posterior density estimates. In all cases we found the changes to be minimal by 10,000 iterations. The direct Bayes estimates of β and θ are also reported alongside.

The analyses of these five data sets are interesting in several respects and reveal problems with the Gibbs implementation regardless of sample size. First, despite non-existence of the posterior distribution under the flat and non-informative prior, it is still possible to run Gibbs chains for these priors. Figure 3 displays the histogram of a sample generated for one Gibbs run with the non-informative prior for the data in Table 1 e). This histogram is an example of how reasonable some null Gibbs chains may appear. There are no obvious warnings of impropriety since the iterates appear to move reasonably around the parameter space without getting trapped in the absorbing state of zero for θ. Thus, the use of the Gibbs sampler with improper priors (without verifying propriety) is dangerous and can lead to seriously misleading results.

Second, the Gibbs modes based on proper flat priors either fail to reproduce the direct Bayes estimates on average or suffer from large amounts of simulation noise. More specifically, a 95% confidence interval for the average Gibbs mode from $\pi_2^f(\beta,\ \theta)$ does not contain the direct Bayes estimate (that is, the ML estimate) for four of the five examples considered (indicated by an asterisk). On the other hand, based on the estimates from the more diffuse $\pi_1^f(\beta,\ \theta)$ although we fail to reject the hypothesis of equality between Gibbs and direct Bayes, the proportion of simulation noise is higher than 10% in four cases and in one example it is as high as 600% which is unacceptably large. The number of required runs to achieve single digit accuracy with $\pi_1^f(\beta,\ \theta)$ is greater than one million samples for all five examples. These numbers are large even for this simple problem. For instance one run of 10,000 iterations required between 20 minutes to an hour (depending on the size of the data set) of computer time on a SUN SPARC 20 workstation. A million iterations would require upward of two days of computing time. (These estimates may be significantly larger for more realistic models with complicated design matrices and random-effect structures.) Efficient programming could undoubtedly reduce these times somewhat, but not by a substantial amount so as to make the computations more feasible. In summary, it appears that in situations where flat priors lead to improper posteriors, one should not expect that all problems can be safely avoided by simply considering priors on finite intervals. Such methods likely work well when results based on the limiting flat prior are themselves reasonable (for instance, when the likelihood is sharply peaked).

Thirdly, the non-informative prior and its diffuse counterparts clearly demonstrate problems with domination by the prior distribution, even for larger samples. The posterior mode of θ resulting from these priors is systematically smaller than the ML estimates suggesting that information contained in the data is not abundant enough to counteract the large weight placed by these priors on small θ. For example, for the moderately large data set in Table 1 e) we find the posterior mode of θ corresponding to $\pi_1^n(\beta,\ \theta)$ is 1.87 compared to the ML estimate of 3.18. Similar discrepancies are noted for all examples. The specification of these priors to approximate a likelihood inference could lead to modal estimates substantially different from the maximum likelihood estimate, which makes their use undesirable. A comparison of the Gibbs estimates with direct Bayes indicates difficulties with increased prior diffusion. As expected, the Gibbs estimates corresponding to the tight proper prior $\pi_3^n(\beta,\ \theta)$ reproduce the direct Bayes estimates very accurately (the % incr. var. is close to 0% in all cases and the number of required runs for single digit accuracy is small). However, inspection of the diffuse priors $\pi_1^n(\beta,\ \theta)$ and $\pi_2^n(\beta,\ \theta)$ reveal problems with failure to reproduce direct Bayes or large amounts of simulation variance, especially with $\pi_1^n(\beta,\ \theta)$.

5 Conclusions

This paper demonstrates by example that the use of the Gibbs sampler with diffuse priors can give inaccurate posterior estimates. Our simulations show that such inaccuracies are not necessarily limited to small samples, but in fact may be present for moderately large sample settings. By considering a simple model where direct calculations and multiple chains are feasible, we demonstrate that the Gibbs estimates either fail to reproduce direct Bayes on average, or suffer from unacceptably large simulation variances, even for reasonably large burn-in and sample choices. In practice, direct calculations are often formidable and there are no obvious methods for verifying the answers based on these Gibbs chains. In such cases, the Markov chain cannot be relied upon to signal badly-behaved problems.

Acknowledgment

We wish to thank three anonymous referees and an Associate Editor for recommendations which led to substantial improvements in our results. The second author was partially supported by NSF grant DMS-9625476.

References

APTECH SYSTEMS (1994) The GAUSS System Version 3.0, Maple Valley, Washington 98038.

Berger, J. O., and Robert, C. (1990), "Subjective Hierarchical Bayes Estimation of a Multivariate Normal Mean: On the Frequentist Interface," *The Annals of Statistics,* 18, 617–651.

Efron, B. E. (1973), Discussion of "Marginalization Paradoxes in Bayesian and Structural Inference", by A. P. Dawid, M. Stone and J. V. Zidek (1973) *Journal of the Royal Statistical Society,* Ser B, 35,219.

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.

Hamada, M., and Wu, C. F. J. (1995), "Analysis of Censored Data from Fractionated Experiments: A Bayesian Approach," *Journal of the American Statistical Association,* 90, 447–477.

Hobert, J. P., and Casella, G. (1996), "The effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association,* 91, 1461-1473.

Ibrahim, J. G., and Laud, P. W. (1991), "On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior," *Journal of the American Statistical Association,* 86, 981–986.

Kass, R. E., and Wassermann, L. (1996), "Selecting Prior Distributions by Formal Rules, *Journal of the American Statistical Association,* 91, 1343-1370.

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

Natarajan, R. and McCulloch, C. E. (1995), "A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses," *Biometrika,* 82, 639–643.

Raudenbush, S.W., Brennan, R. T., and Barnett,R. C. (1995), "A Multivariate Hierarchical Model for Studying Psychological Change Within Married Couples," *Journal-of-Family-Psychology,* 9, 161-174.

Strawdermann, W.E. (1971), "Proper Bayes Minimax Estimators of the Multivariate Normal Mean," *Annals of Mathematical Statistics,* 42, 385-388.

Yang, R., and Chen, M-H. (1995), "Bayesian Analysis for Random Coefficient Regression Models using Non-informative Priors," *Journal of Multivariate Analysis,* 55, 283-311.

Caption for Table 1

Estimates of posterior modes of β and θ for varying prior specifications in a probit-normal hierarchy for five data sets with varying sample sizes. Posterior estimates are calculated by both a Gibbs sampler and by direct maximization of the numerator of the posterior distribution. MLEs and their asymptotic standard errors are given for reference. The average Gibbs mode derived from ten replicate runs of 10,000 variates each is reported (simulation standard errors in parentheses). A statistically significant difference ($\alpha = 0.05$) between the average Gibbs mode and the direct estimate is indicated by an asterisk. The % incr. var. (rounded to nearest integer) and required runs (rounded to two significant digits) are also reported for each prior specification. Priors with a superscript f (π^f) are flat priors with subscripts denoting truncated versions; priors with a superscript n (π^n) are priors which approach the classic noninformative prior as the subscripts decrease.

Footnotes for Table 1

[1] Since the joint posterior distribution does not exist under π^f and π^n, the calculation of % incr. var. and required runs are not applicable. Further, the reported direct Bayes estimates are also not theoretically applicable.

[2] Since the Gibbs mode does not reproduce the direct Bayes estimate on average, the calculation of % incr. var. and required runs are not applicable.

[3] Optimization routine to compute the direct Bayes estimate failed to converge.

Table 1

a) $n = 25$ clusters and $r = 2$ observations per cluster; MLE(β) = -6.24 with asymptotic standard error of 3.36, MLE(θ) = 1.85 with an asymptotic standard error of 2.90

	β				θ			
Prior	Direct	Gibbs	% incr. var.	Required runs	Direct	Gibbs	% incr. var.	Required runs
π^f	-6.24[1]	-9.09 (1.56)	—[1]	—[1]	1.85[1]	7.08 (2.59)	—[1]	—[1]
π^f_1	-6.24	-6.89 (1.55)	200	37,000	1.89	3.58 (2.26)	600	79,000
π^f_2	-6.24	-5.70* (0.19)	—[2]	—[2]	1.85	1.38* (0.15)	—[2]	—[2]
π^n	—[1,3]	-4.25 (0.22)	—[1]	—[1]	—[1,3]	0.19 (0.31)	—[1]	—[1]
π^n_1	-4.17	-4.50 (0.17)	23	440	0.13	0.24 (0.08)	200	96
π^n_2	-4.91	-4.89 (0.11)	5	180	0.66	0.66 (0.05)	6	40
π^n_3	-5.11	-5.10 (0.02)	0	5	0.87	0.88 (0.00)	0	1

b) n = 50 clusters and r = 2 observations per cluster; MLE(β) = -5.93 with asymptotic standard error of 2.46, MLE(θ) = 2.90 with an asymptotic standard error of 2.83

	β				θ			
Prior	Direct	Gibbs	% incr. var.	Required runs	Direct	Gibbs	% incr. var.	Required runs
π^f	-5.93[1]	-6.21 (0.49)	—[1]	—[1]	2.90[1]	3.55 (0.75)	—[1]	—[1]
π_1^f	-5.93	-6.00 (0.35)	20	1,896	2.90	3.14 (0.42)	20	2,750
π_2^f	-5.93	-5.33* (0.22)	—[2]	—[2]	2.90	2.22* (0.23)	—[2]	—[2]
π^n	-4.30[1]	-3.70* (0.16)	—[1]	—[1]	1.08[1]	0.59* (0.13)	—[1]	—[1]
π_1^n	-4.41	-4.21 (0.23)	20	810	1.19	1.06 (0.19)	20	540
π_2^n	-4.66	-4.60 (0.16)	10	390	1.43	1.36 (0.13)	10	260
π_3^n	-4.40	-4.42 (0.01)	0	1	1.06	1.06 (0.01)	0	1

c) $n = 100$ clusters and $r = 2$ observations per cluster; MLE(β) = -5.83 with asymptotic standard error of 1.93, MLE(θ) = 2.58 with an asymptotic standard error of 2.19

	β				θ			
Prior	Direct	Gibbs	% incr. var.	Required runs	Direct	Gibbs	% incr. var.	Required runs
π^f	-5.83[1]	-5.36 (0.44)	–[1]	–[1]	2.58[1]	3.55 (0.62)	–[1]	–[1]
π_1^f	-5.83	-5.65 (0.21)	10	692	2.58	2.38 (0.24)	10	853
π_2^f	-5.83	-5.18* (0.13)	–[2]	–[2]	2.58	1.90* (0.12)	–[2]	–[2]
π^n	-4.68[1]	-4.19* (0.12)	–[1]	–[1]	1.30[1]	0.91* (0.14)	–[1]	–[1]
π_1^n	-4.74	-4.03* (0.16)	–[2]	–[2]	1.36	0.78* (0.13)	–[2]	–[2]
π_2^n	-4.90	-4.47* (0.14)	–[2]	–[2]	1.53	1.16* (0.18)	–[2]	–[2]
π_3^n	-4.51	-4.51 (0.02)	0	4	1.09	1.09 (0.01)	0	1

d) $n = 50$ clusters and $r = 4$ observations per cluster; MLE(β) = -9.45 with asymptotic standard error of 2.62, MLE(θ) = 4.30 with an asymptotic standard error of 2.74

	β				θ			
Prior	Direct	Gibbs	% incr. var.	Required runs	Direct	Gibbs	% incr. var.	Required runs
π^f	-9.45[1]	-7.31* (0.64)	–[1]	–[1]	4.30[1]	2.86 (0.76)	–[1]	–[1]
π_1^f	-9.45	-8.14 (0.74)	80	8,490	4.30	3.48 (0.70)	65	7,500
π_2^f	-9.45	-8.95 (0.34)	17	1,830	4.30	4.00 (0.36)	17	2,050
π^n	-8.12[1]	-7.40 (0.39)	–[1]	–[1]	2.83[1]	2.25 (0.28)	–[1]	–[1]
π_1^n	-8.15	-7.57* (0.20)	–[2]	–[2]	2.85	2.36* (0.17)	–[2]	–[2]
π_2^n	-8.10	-7.20 (0.52)	73	4,190	2.82	2.17 (0.43)	60	2,860
π_3^n	-6.23	-6.24 (0.02)	0	9	1.29	1.30 (0.01)	0	2

e) n = 15 clusters and r = 8 observations per cluster; MLE(β) = -9.38 with asymptotic standard error of 2.79, MLE(θ) = 3.18 with an asymptotic standard error of 2.36

	β				θ			
Prior	Direct	Gibbs	% incr. var.	Required runs	Direct	Gibbs	% incr. var.	Required runs
π^f	-9.38[1]	-9.95 (0.91)	–[1]	–[1]	3.18[1]	4.02 (0.87)	–[1]	–[1]
π_1^f	-9.38	-8.62 (0.40)	20	2,480	3.18	2.71 (0.30)	16	1,370
π_2^f	-9.38	-9.75* (0.09)	–[2]	–[2]	3.18	3.58* (0.09)	–[2]	–[2]
π^n	-7.96[1]	-7.52 (0.29)	–[1]	–[1]	1.83[1]	1.62 (0.14)	–[1]	–[1]
π_1^n	-8.00	-7.68 (0.33)	26	1,720	1.87	1.70 (0.17)	15	430
π_2^n	-8.02	-8.12 (0.14)	5	300	1.93	1.99 (0.07)	3	85
π_3^n	-6.45	-6.43 (0.01)	0	3	1.06	1.06 (0.00)	0	1

Caption for Figure 1

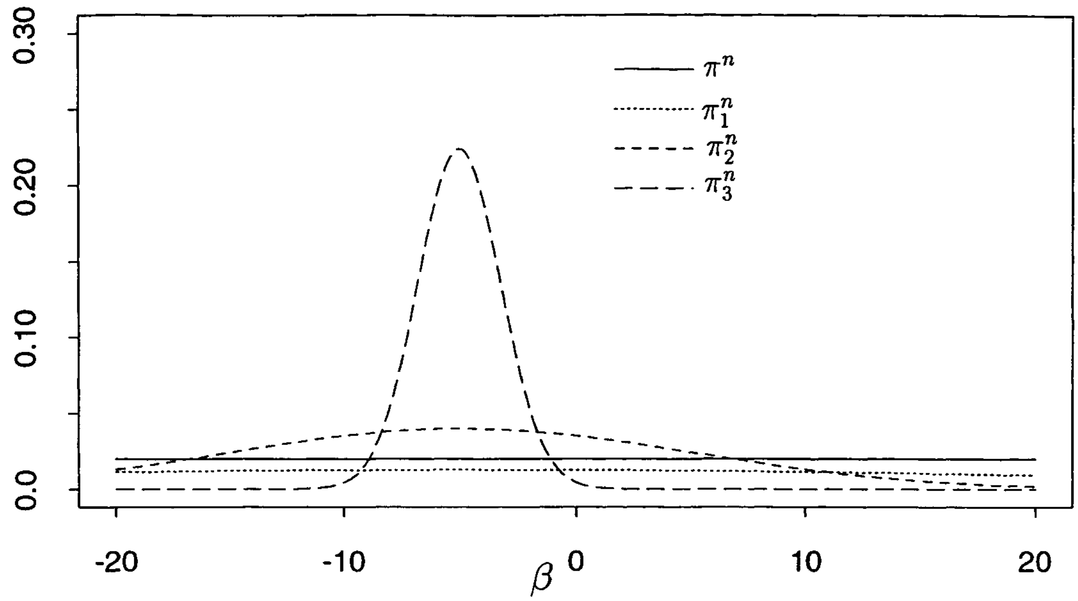
Four priors for β and θ. The non-informative prior is represented by a solid line; the broken lines are diffuse proper priors in decreasing order of diffusion, with the dotted line representing the tightest specification.

Caption for Figure 2

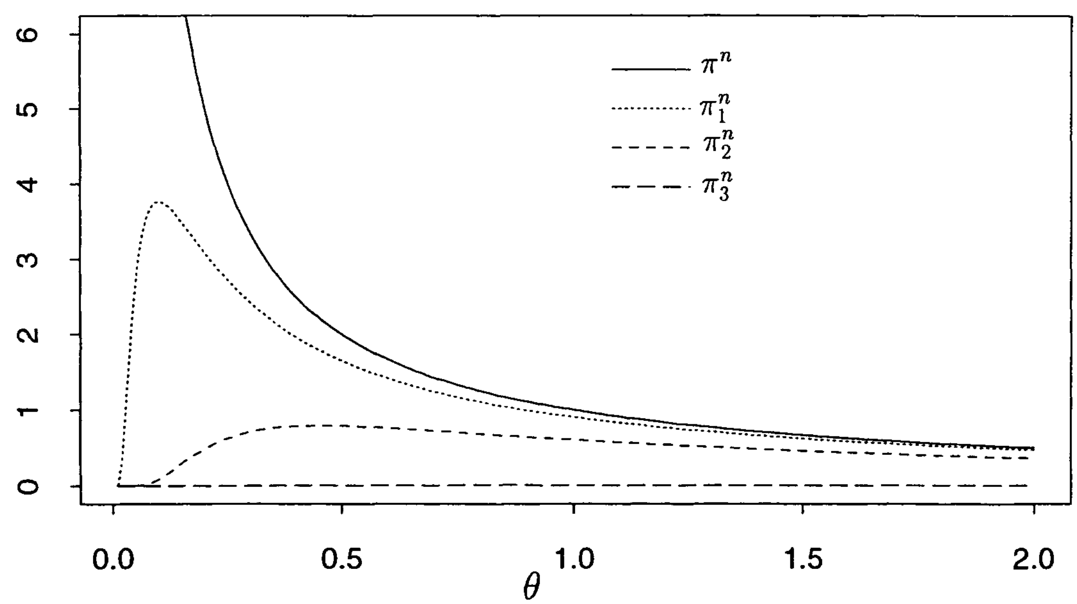
Likelihood surface for five data sets simulated from a probit-normal hierarchy with (a) $n = 25, r = 2$, (b) $n = 50, r = 2$, (c) $n = 100, r = 2$, (d) $n = 50, r = 4$, (e) $n = 15, r = 8$. True values of β and θ for all five data sets are -5 and 2 respectively.

Caption for Figure 3

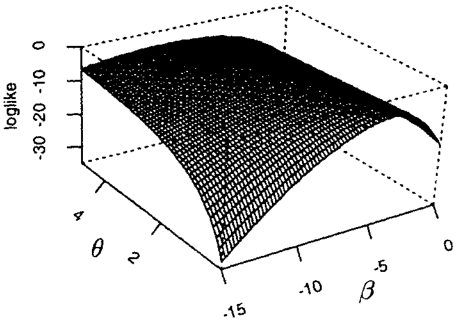
Histogram of β and θ iterates for one run of 10,000 iterates using the Gibbs sampler with a non-informative prior for data from a probit-normal hierarchy with $n = 15$, $r = 8$. (Note that 10% of the generated θ values were larger than 21 but are not included in the histogram)



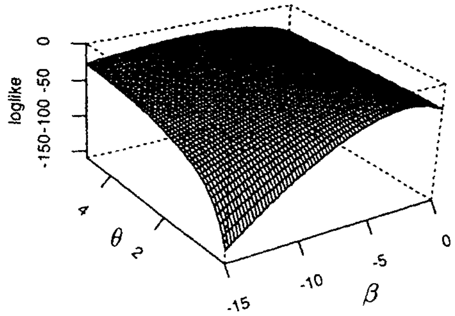
0.0
0.10
0.20
0.30
-20
-10
0
10
20
β
π^n
π_1^n
π_2^n
π_3^n



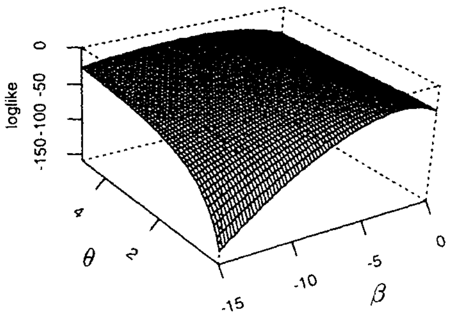
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5
2.0
θ
π^n
π_1^n
π_2^n
π_3^n


loglike
0
-10
-20
-30
θ
4
2
β
-15
-10
-5
0

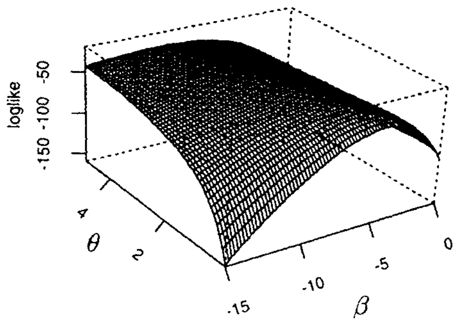
(a)


loglike
0
-50
-100
-150
θ
4
2
β
-15
-10
-5
0

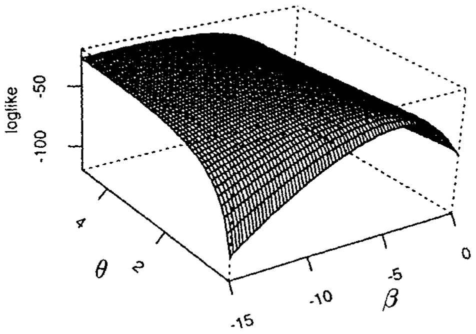
(b)


loglike
0
-50
-100
-150
θ
4
2
β
-15
-10
-5
0

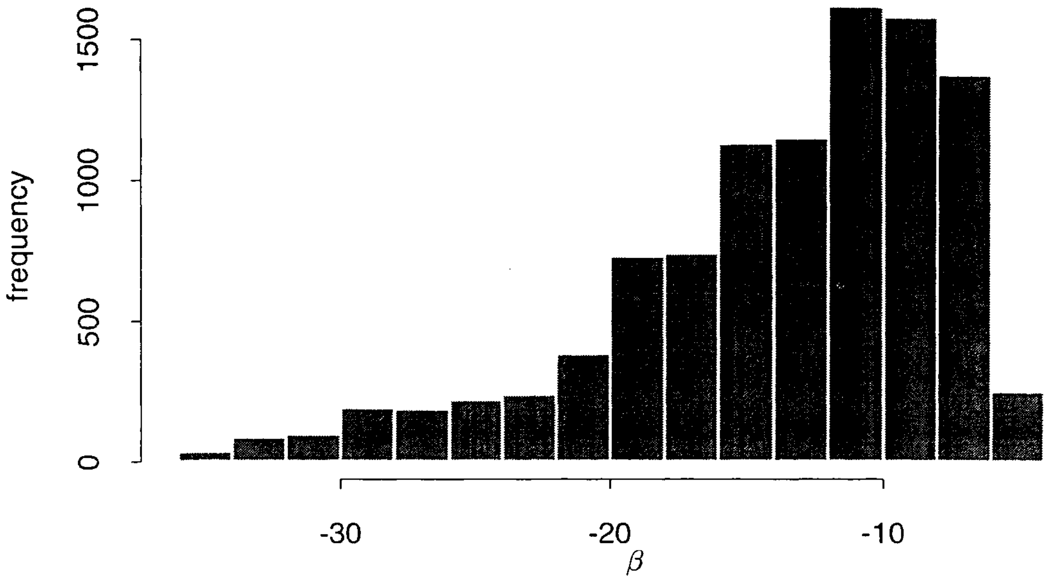
(c)

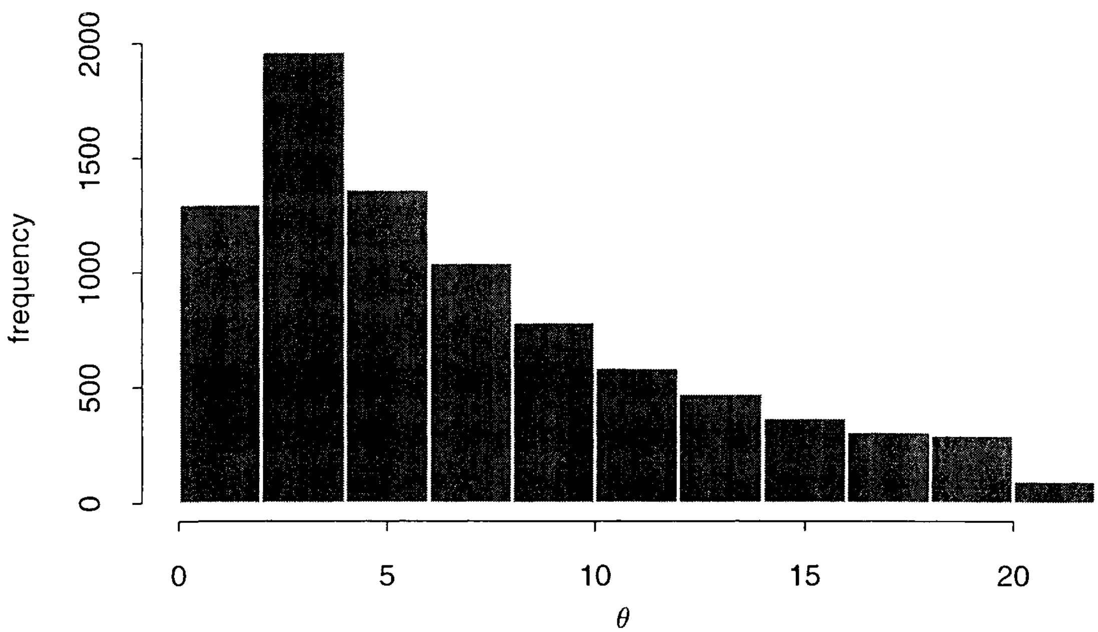

loglike
-50
-100
-150
θ
4
2
β
-15
-10
-5
0

(d)


loglike
-50
-100
θ
4
2
β
-15
-10
-5
0

(e)


frequency
0
500
1000
1500
-30
-20
-10
β


frequency
0
500
1000
1500
2000
0
5
10
15
20
θ