Statistics and Computing (1996) **6**, 101–111

Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models

MARY KATHRYN COWLES

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02215, USA

Received October 1994 and accepted October 1995

The ordinal probit, univariate or multivariate, is a generalized linear model (GLM) structure that arises frequently in such disparate areas of statistical applications as medicine and econometrics. Despite the straightforwardness of its implementation using the Gibbs sampler, the ordinal probit may present challenges in obtaining satisfactory convergence.

We present a multivariate Hastings-within-Gibbs update step for generating latent data and bin boundary parameters jointly, instead of individually from their respective full conditionals. When the latent data are parameters of interest, this algorithm substantially improves Gibbs sampler convergence for large datasets. We also discuss Monte Carlo Markov chain (MCMC) implementation of cumulative logit (proportional odds) and cumulative complementary log-log (proportional hazards) models with latent data.

Keywords: Blocking, collapsing, data augmentation, Gibbs sampler, latent data

1. Introduction

Ordinal response variables are very common in biostatistical and econometric applications. For example, the outcome variable in a comparative trial of analgesics might be participants' self-report of change in pain status on a three-point scale consisting of 'improved', 'no change', and 'worse'. Generalized linear models with a cumulative link function are commonly used to analyse the relationship between an ordinal response variable and predictor variables, which may be continuous, nominal, or ordinal. Assume that for each subject i we observe a response variable W_i, which may take on any one of k ordered values labelled $1, 2, \ldots, k$. Values of a set of predictor variables $\boldsymbol{x}_i$ are also observed. A cumulative link model (Agresti, 1990) for these data would be of the form

$$\Pr(W_i \leqslant j | \boldsymbol{x}, \boldsymbol{\beta}) = G(\gamma_j - \boldsymbol{x}_i^t \boldsymbol{\beta}),$$

where G is the cumulative distribution function of a continuous random variable having positive density over the entire real line; γ_j, $j = 0, 1, \ldots k$, are ordered cutpoints dividing the real line into intervals; and $\boldsymbol{\beta}$ is a vector of coefficients of the predictors. If G is the logistic c.d.f., then the model is a *cumulative logit* or *proportional odds* model, while if G is the extreme value (minimum) c.d.f., the model is called the *cumulative complementary log-log* or *proportional hazards* model. We will first consider the model in which G is the normal c.d.f.—the *cumulative probit* or *ordinal probit* model. For each of the three link functions, maximum likelihood methods may be used to get point estimates and asymptotic standard errors of $\boldsymbol{\beta}$ and γ, although the validity of the asymptotic standard errors is questionable for small sample sizes.

Albert and Chib (1993) present Bayesian implementations of the ordinal probit model using the Gibbs sampler. They point out that the ordinal probit may be visualized in terms of an unobservable, or 'latent' continuous variable y_i^* corresponding to each observed variable w_i. The value of each y_i^* falls into one of k contiguous bins on the real line demarcated by the cutpoints $\gamma_0, \gamma_1, \ldots, \gamma_k$, and the observed values of the w$_i$'s are determined by the relationship $w_i = j$ if $y_i^* \in (\gamma_{j-1}, \gamma_j]$. Then the assumption that $y_i^* \sim \mathrm{N}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, 1)$ makes this latent variable model equivalent to the cumulative probit. The variance of the unobservable y_i^*'s is assumed to be 1 for consistency with the standard normal c.d.f. link function.

Applying the 'data augmentation' idea of Tanner and Wong (1987), Albert and Chib (1993) treat the unknown

0960-3174 © 1996 Chapman & Hall

$\boldsymbol{y}^*$ values as additional parameters to be simulated in the Gibbs sampler. Once values are obtained for $\boldsymbol{y}^*$, the problem of estimating $\boldsymbol{\beta}$ in the ordinal probit model simplifies to that of doing so in a standard normal linear model. If a flat prior is specified for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, then the full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{y}^*$, as laid out by Albert and Chib, are:

$$p(y_i^*|\boldsymbol{\beta},\boldsymbol{\gamma},w_i) = \mathrm{N}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta},1) \tag{1a}$$

truncated to $(\gamma_{w_i-1},\gamma_{w_i}]$, and

$$p(\boldsymbol{\beta}|\boldsymbol{w},\boldsymbol{y}^*) = \mathrm{N}((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}^*,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}). \tag{1b}$$

In order that the domain of the y_i^*'s may be the entire real line, the extreme bin cutpoints, γ_0 and γ_k, must be fixed at $-\infty$ and $+\infty$ respectively. Carlin and Polson (1992) present a parametric approach to the remaining cutpoints that introduces additional assumptions into the ordinal probit model (although not into a binary probit). We pursue the more general approach of Albert and Chib. They note that, in order to make the parameters of the model identifiable, one additional cutpoint must be fixed; without loss of generality they fix γ_1 at 0. (An alternative to fixing γ_1 would be to omit the intercept from the model; however, we continue under the assumption that γ_1 is fixed.) Then the full conditional distribution for each variable γ_j is uniform:

$$\begin{aligned} &p(\gamma_j|\boldsymbol{w},\boldsymbol{y}^*,\boldsymbol{\beta},\{\gamma_1, l \neq j\} \\ &\quad = U[\max(\max\{y_i^* : w_i = j\},\gamma_{j-1}), \\ &\quad\quad \min(\min\{y_i^* : w_i = j+1\},\gamma_{j+1})]. \end{aligned} \tag{1c}$$

Albert and Chib used data augmentation in their implementations of the ordinal probit solely so that the full conditionals in the Gibbs sampler would be standard densities. However, in some problems, the values of the latent variables may be of interest. For example, Albert (1992) used latent data to estimate the polychoric correlation coefficient between two ordinal variables. Similarly, Cowles *et al.* (1996) used the values of a latent continuous variable underlying an ordinal response to estimate the correlations between the ordinal response and two continuous response variables. In the latter problem, the ordinal probit was just one component of a complex random-effects model with over 6000 parameters. In order to conserve computer resources, Cowles *et al.* (1996) sought a sampling algorithm that would provide good parameter estimates based on hundreds, rather than thousands or hundreds of thousands, of iterations. The present paper gives details of that algorithm.

In Section 2 we demonstrate that, for an ordinal probit with latent data, convergence of the Gibbs sampler using univariate full conditionals may be slow when the sample size is large. In Section 3 we propose a multivariate Hastings-within-Gibbs update step that substantially accelerates convergence for the three-bin problem, and in Section 4 we extend this method to an ordinal probit problem with more than three bins and to cumulative logit and cumulative complementary log-log models. Finally, in Section 5 we suggest areas for further work.

2. Convergence of the ordinal probit

Convergence of the Gibbs sampler implemented by simulating from the univariate full conditionals (1c), (1a), and (1b) in sequence appears to depend on how full the bins for the y_i^*'s are—that is, on the sample size. Convergence is very slow when the bins are full because the interval within which each γ_j must be generated from its full conditional (1c) is very narrow, so the cutpoint values can change very little between successive iterations. Until the cutpoints are in roughly the right places, the values of the y_i^*'s are distorted, so convergence of the β's is also retarded.

To simulate a simple example of a large-sample three-level ordinal probit, we generated $N = 2000$ data points from the model

$$y_i^* = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i = \mathrm{N}(0,1)$$

with $\beta_0 = 1$ and $\beta_1 = -2$. We then calculated the 1/3 and 2/3 quantiles of the $\boldsymbol{y}^*$ and assigned a corresponding ordinal variable w_i to each y_i^* as follows:

$$w_i = \begin{cases} 1, & y_i^* \text{ in lowest tertile} \\ 2, & y_i^* \text{ in middle tertile} \\ 3, & y_i^* \text{ in highest tertile} \end{cases}$$

Using the w's and x's from the simulated data, we ran five parallel Gibbs sampler chains for 6000 iterations. Figure 1 shows resulting convergence plots for $\boldsymbol{\beta}$ and γ_2 (the only stochastic cutpoint in a three-bin ordinal probit). To assess Gibbs sampler convergence, for each parameter we computed Gelman and Rubin's (1992) 'shrink factor'—the factor by which variance in estimation is inflated due to stopping the chain after the number of iterations run instead of continuing sampling in the limit. Gelman and Rubin suggest running Gibbs sampler chains until the estimated shrink factors are less than about 1.1 for all parameters of interest. (For a comparative review of this and other convergence diagnostics, see Cowles and Carlin, 1996). The median and 97.5th percentiles of the shrink factors are shown above the plots. The plots suggest that, after approximately 3000 iterations, all chains for the β's are traversing the same sample space, and Gelman and Rubin's diagnostic perhaps implies that the last 3000 iterations may be used for estimation.

Since we were primarily interested in algorithms that would converge rapidly, necessitating fewer than 1000 iterations, we next examined the first 800 iterations of these chains, shown in Fig. 2. Here, both the graphical impression

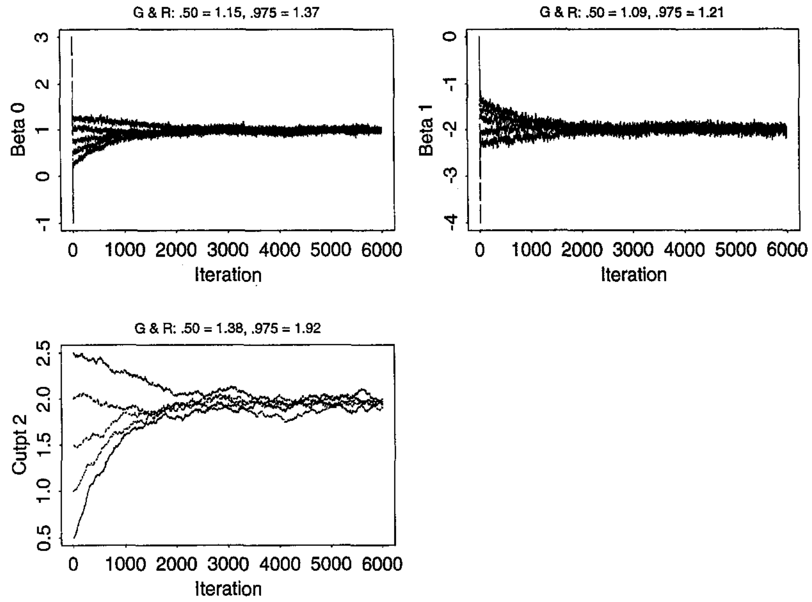

Fig. 1. *Three-bin ordinal probit, univariate full conditionals, 6000 iterations*

of the five chains and Gelman and Rubin's diagnostics indicate that the chains have not mixed well, even for the β's, let alone the cutpoint. Lag-1 autocorrelations were greater than 0.99 for the cutpoint parameter, indicating extremely slow mixing of the sample paths for this parameter.

Table 1 shows that parameter estimates based on the first 800 iterations are not good and that even a careful applied user of the Gibbs sampler might not detect this. In computing standard errors for Table 1, we used two common approaches to adjusting for correlations in the Gibbs sampler output—the batch means method (see, for example, Ripley, 1987, Section 6.2) and a method based on spectral analysis (see, for example, Geweke, 1992, Geyer 1992). As shown in Table 1, the estimate of β_1 is more than two standard errors away from the true value of -2.0 in all chains. If the truth had been that there was no linear relationship between the predictor and the ordinal response—that is, if the true value of β_1 had been zero—bias in estimation even of this small magnitude would have led to erroneous conclusions of an inverse relationship between x and w. Since with the batch means method, a rule of thumb is to use batch sizes that lead to autocorrelations between the batch means of less than 0.05, it is clear from the 'Lag-1 Autocorrelations' columns in Table 1 that too

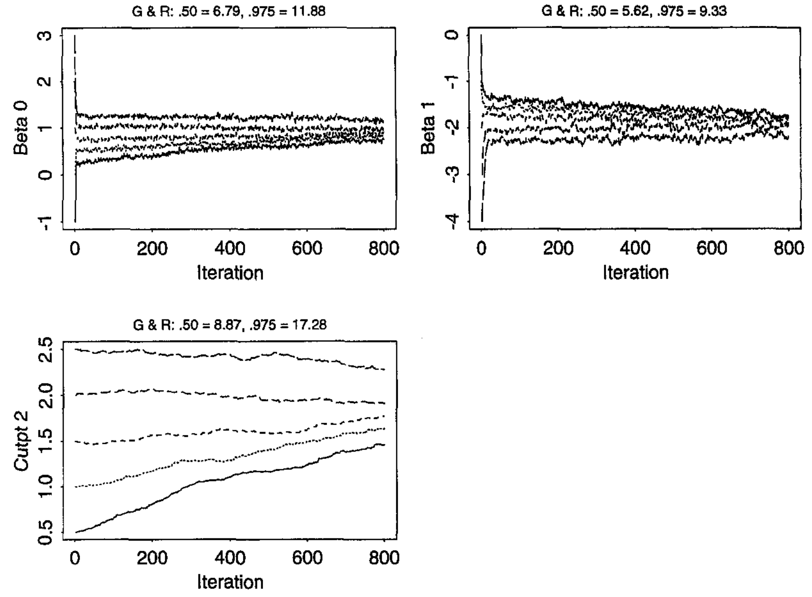

Fig. 2. *Three-bin ordinal probit, univariate full conditionals, 800 iterations*

Table 1. *Means and standard errors estimated from Gibbs samples, ordinal probit model, iterations 401-800*

Pooled sample of 5 chains, 2000 iterates

			Batch means method			
			25 batches, size 80		10 batches, size 200	
Parameter	Mean	Naive Std Err[a]	Std Err	Lag 1 Autocorr[b]	Std Err	Lag 1 Autocorr
β_0	0.872	0.004	0.040	0.870	0.066	0.691
β_1	−1.888	0.004	0.040	0.874	0.065	0.705
γ_2	1.731	0.009	0.080	0.870	0.129	0.697

Individual chains, 400 iterates each

				25 batches, size 20		10 batches, size 50		
Chain	Parameter	Mean	Naive Std Err	Std Err	Lag 1 Autocorr	Std Err	Lag 1 Autocorr	Spectral NSE[c]
1	β_0	0.599	0.003	0.011	0.848	0.018	0.685	0.006
	β_1	−1.623	0.003	0.010	0.586	0.014	0.618	0.006
	γ_2	1.193	0.005	0.020	0.882	0.032	0.703	0.010
2	β_0	0.736	0.002	0.006	0.771	0.010	0.624	0.004
	β_1	−1.749	0.003	0.009	0.587	0.014	0.289	0.005
	γ_2	1.466	0.003	0.011	0.859	0.018	0.673	0.006
3	β_0	0.856	0.002	0.005	0.788	0.008	0.609	0.003
	β_1	−1.869	0.002	0.007	0.322	0.008	0.299	0.005
	γ_2	1.700	0.002	0.009	0.893	0.015	0.743	0.005
4	β_0	1.022	0.002	0.004	0.443	0.006	0.368	0.003
	β_1	−2.030	0.003	0.007	0.175	0.007	−0.443	0.006
	γ_2	2.020	0.001	0.005	0.834	0.008	0.561	0.003
5	β_0	1.145	0.002	0.004	0.416	0.005	0.591	0.003
	β_1	−2.166	0.002	0.006	0.166	0.008	0.011	0.005
	γ_2	2.275	0.002	0.006	0.886	0.010	0.772	0.003

[a] Assumes independent samples.
[b] Between means of batches.
[3] Numeric standard error: see for example Geweke (1992).

few iterations have been run for the batch means method of estimating standard errors to be trustworthy. However, the lag-1 autocorrelations between the batches of larger sizes are less than 0.05 for β_1 in chain 5. Thus, a user who had run only a single chain for this problem, obtaining results as in chain 5, and who had monitored convergence only of β_1 since it is the only parameter of interest, might well have been fooled into concluding satisfactory convergence and reporting poor estimates.

3. A multivariate Hastings-within-Gibbs update step for a three-level ordinal probit

Liu *et al.* (1994) show that 'grouping' or 'blocking' components usually improves the efficiency of a Gibbs sampler. Accordingly, we reasoned that generating the cutpoint γ_2 together with the $\boldsymbol{y}^*$'s might solve the problem of slow mixing caused by generating γ_2 conditional on the $\boldsymbol{y}^*$'s. The joint full conditional of $\boldsymbol{y}^*$ and γ_2 is easily determined from the identity:

$$p(\gamma_2, \boldsymbol{y}^* | \boldsymbol{\beta}, \boldsymbol{w}) = p(\gamma_2 | \boldsymbol{\beta}, \boldsymbol{w}) p(\boldsymbol{y}^* | \gamma_2, \boldsymbol{\beta}, \boldsymbol{w}), \tag{3.1a}$$

where $I(\cdot)$ is the indicator function. Now

$$p(\gamma_2 | \boldsymbol{\beta}, \boldsymbol{w}) \propto \prod_{i:w_i=1} \Phi(-x_i^{\mathrm{T}} \boldsymbol{\beta}) \prod_{i:w_i=2} [\Phi(\gamma_2 - x_i^{\mathrm{T}} \boldsymbol{\beta}) - \Phi(-x_i^{\mathrm{T}} \boldsymbol{\beta})] \times \prod_{i:w_i=3} [1 - \Phi(\gamma_2 - x_i^{\mathrm{T}} \boldsymbol{\beta})] \times I(\gamma_2 \in (0, \infty))$$

and

$$p(\boldsymbol{y}^* | \gamma_2, \boldsymbol{\beta}, \boldsymbol{w}) \propto \prod_{i:w_i=1} \left[\frac{\phi(y_i^* - x_i^{\mathrm{T}} - \boldsymbol{\beta}}{\Phi(-x_i^{\mathrm{T}} \boldsymbol{\beta})} \right] \times \prod_{i:w_i=2} \left[\frac{\phi(y_i^* - x_i^{\mathrm{T}} \boldsymbol{\beta})}{\Phi(\gamma_2 - x_i^{\mathrm{T}} \boldsymbol{\beta}) - \Phi(-x_i^{\mathrm{T}} \boldsymbol{\beta})} \right] \times \prod_{i:w_i=3} \left[\frac{\phi(y_i^* - x_i^{\mathrm{T}} \boldsymbol{\beta})}{1 - \Phi(\gamma_2 - x_i^{\mathrm{T}} \boldsymbol{\beta})} \right] \tag{3.1b}$$

where Φ denotes the standard normal cumulative distribution function and ϕ the standard normal density. The right-hand side of (3.1b) is the product of truncated normal densities. The joint full conditional of γ_2 together with the y^* 's is

$$p(\gamma_2, \boldsymbol{y}^* | \boldsymbol{\beta}, \boldsymbol{w}) \propto \prod_i \phi(y_i^* - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \times I(\gamma_2 \in (0, \infty)]. \quad (3.1c)$$

Since (3.1c) is a not a standard, normalized joint density, we chose to use a multivariate version of the Metropolis–Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970), which Tierney (1994) and Muller (1994) propose using within the Gibbs sampler for sampling from nonstandard full conditionals. To sample from target density f requires selecting a proposal density $g(v/u)$ from which it is easy to sample. At each iteration i, starting with the value $u^{(i-1)}$ from the previous iteration, a new candidate v is generated from $g(v/u^{(i-1)})$. With probability

$$\alpha = \min\left(1, \frac{f(v)g(u^{(i-1)}|v)}{f(u^{(i-1)})g(v|u^{(i-1)})}\right), \quad (3.2)$$

v is accepted, i.e. $u^{(i)}$ is set equal to v; otherwise $u^{(i)}$ is set equal to $u^{(i-1)}$. Hastings (1970) showed that this algorithm produces a sequence that converges in distribution to f.

In constructing a suitable joint proposal density for γ_2 and the y^*'s we factored it in the same way as we had factored the joint conditional density:

$$\begin{aligned} & g(\gamma_{2,\mathrm{new}}, \boldsymbol{y}^*_{\mathrm{new}} | \gamma_{2,\mathrm{old}}, \boldsymbol{y}^*_{\mathrm{old}}, \boldsymbol{\beta}, \boldsymbol{w}) \\ & = g(\gamma_{2,\mathrm{new}} | \gamma_{2,\mathrm{old}}, \boldsymbol{y}^*_{\mathrm{old}}, \boldsymbol{\beta}, \boldsymbol{w}) g(\boldsymbol{y}^*_{\mathrm{new}} | \gamma_{2,\mathrm{new}}, \gamma_{2,\mathrm{old}}, \boldsymbol{y}^*_{\mathrm{old}}, \boldsymbol{\beta}, \boldsymbol{w}) \\ & \propto (1/\sigma_\gamma)\phi \frac{((\gamma_{2\mathrm{new}} - \gamma_{2\mathrm{old}})/\sigma_\gamma)}{\Phi(\gamma_{2\mathrm{old}}/\sigma_\gamma)} \times \prod_{i:w_i=1} \frac{\phi(\boldsymbol{y}^*_{i\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\Phi(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \\ & \times \prod_{i:w_i=2} \frac{\phi(\boldsymbol{y}^*_{i\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\Phi(\gamma_{2\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \Phi(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \\ & \times \prod_{i:w_i=3} \frac{\phi(\boldsymbol{y}^*_{i\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 - \Phi(\gamma_{2\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \end{aligned}$$

The first term in the above product is a Normal $(\gamma_{2\,\mathrm{old}}, \sigma_\gamma^2)$ density, truncated to $(0, \infty)$ to keep the cutpoints in the correct order. An appropriate value for σ_γ^2 can be chosen to obtain an acceptance rate of approximately 0.44, which Gelman *et al.* (1994) found to be optimal for univariate Metropolis–Hastings chains of certain types. With this candidate-generating density, the acceptance probability α is equal to min(1, R), where

$$\begin{aligned} R = & \frac{\Phi(\gamma_{2\,\mathrm{old}}/\sigma_\gamma)}{\Phi(\gamma_{2\,\mathrm{new}}/\sigma_\gamma)} \prod_{i \ni w_i=2} \frac{\Phi(\gamma_{2\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \Phi(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{\Phi(\gamma_{2\,\mathrm{old}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \Phi(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})} \\ & \times \prod_{i \ni w_i=3} \frac{1 - \Phi(\gamma_{2\,\mathrm{new}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1 - \Phi(\gamma_{2\,\mathrm{old}} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}. \end{aligned} \quad (3.3)$$

The term in (3.3) corresponding to $i{:}w_i = 1$ is omitted because it is equal to 1. Accordingly an efficient multivariate Metropolis–Hastings-within-Gibbs algorithm can be implemented as follows to generate new values $\gamma_2^{(k)}$ and $\boldsymbol{y}^{*(k)}$ at iteration k of the Gibbs sampler:

1. Generate a candidate value $\gamma_{2\,\mathrm{new}}$ from a Normal $(\gamma_2^{(k-1)}, \sigma_\gamma^2)$ density, truncated to $(0, \infty)$.
2. Evaluate the quantity R in (3.3) to get the acceptance probability α.
3. With probability α, set $\gamma_2^{(k)} = \gamma_{2\,\mathrm{new}}$ and generate new $\boldsymbol{y}^{*(k)}$'s from their usual full conditionals (1a), which will depend on the new $\gamma_2^{(k)}$. Otherwise, set $\gamma_2^{(k)} = \gamma_2^{(k-1)}$ and $y_i^{*(k)} = y_i^{*(k-1)}$ for all observations i.

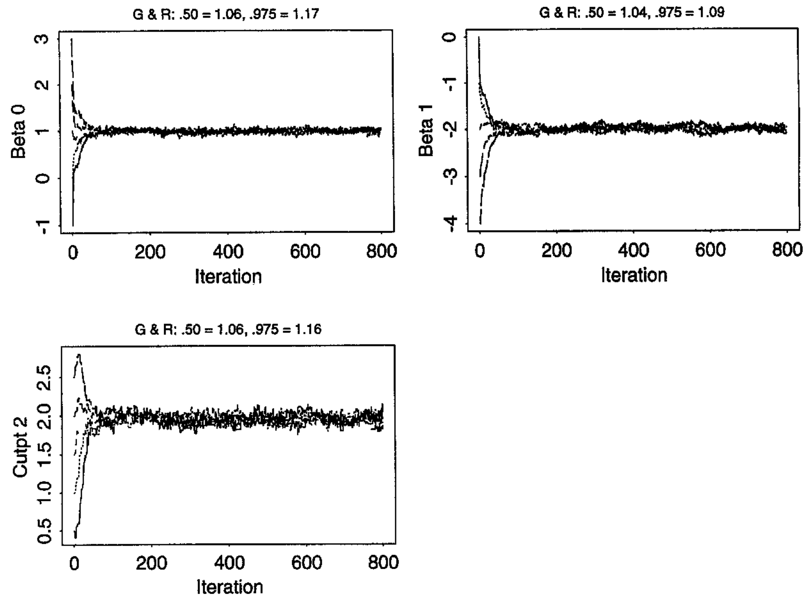

Fig. 3. *Three-bin ordinal probit, multivariate Hastings, 800 iterations*

Since the acceptance probability depends only on the old and new values of γ_2 and not on the $\boldsymbol{y}^*$'s, in step 3 new $\boldsymbol{y}^*$'s need not be generated in any iteration for which the new value of γ_2 is not accepted.

To complete the Gibbs sampler iteration, $\boldsymbol{\beta}^{(k)}$ is generated from its standard full conditional (1b). Note that, despite the fact that new $\boldsymbol{y}^*$'s are not generated at every iteration, the resulting Gibbs sampler differs from the 'collapsed' Gibbs sampler of Liu *et al.* (1994) in that, since the full conditional for $\boldsymbol{\beta}$ depends on the $\boldsymbol{y}^*$'s, generation of the $\boldsymbol{y}^*$'s cannot be omitted altogether. Since we have assumed a need for samples from the distributions of the latent data, that is just as well.

This algorithm was applied to the same simulated dataset described in Section 2. A value of 0.01 was used for σ^2_γ in the truncated normal proposal density for γ_2, producing an acceptance rate of 0.43—only 0.01 different from the rate recommended by Gelman *et al.* (1994). Figure 3 shows convergence plots and statistics for 800 iterations, which may be compared to those in Fig. 2. Despite the fact that lag-1 autocorrelations within chains still are high, Gelman and Rubin's shrink factors and the graphical impression of all five chains suggest very rapid convergence. Indeed Table 2 shows that estimation of β_1 based on either the five chains pooled or on any single chain is very good.

We also considered other MCMC algorithms for this problem. In extremely high-dimensional ordinal probit models like the one of interest to us (with very large sample size and requiring generation of latent data), it is not practical to apply a single multivariate generation algorithm such as Hastings (1970), Hit and Run (Belisle *et al.*, 1993), or Adaptive Direction Sampling (Gilks *et al.* 1994), to the unpartitioned joint posterior distribution of the latent data, regression coefficients, and cutpoints. A feasible alternative is a 'collapsed' Gibbs sampler (Liu *et al.*, 1994), in which the parameters are generated in two groups as follows:

Table 2. *Means and standard errors estimated from Gibbs samples ordinal probit model, multivariate Hastings update step. Iterations 401–800*

Pooled sample of 5 chains, 2000 iterates

			Batch means method			
			25 batches, size 80		10 batches, size 200	
Parameter	Mean	Naive Std Err[a]	Std Err	Lag 1 Autocorr[b]	Std Err	Lag 1 Autocorr
β_0	0.003	0.001	0.005	0.058	0.0004	−0.317
β_1	−2.000	0.001	0.006	0.010	0.005	−0.144
γ_2	1.967	0.002	0.007	0.076	0.006	−0.259

Individual chains, 400 iterates each

				25 batches, size 20		10 batches, size 50		
Chain	Parameter	Mean	Naive Std Err	Std Err	Lag 1 Autocorr	Std Err	Lag 1 Autocorr	Spectral NSE[c]
1	β_0	0.986	0.002	0.007	0.542	0.010	0.383	0.005
	β_1	−1.992	0.002	0.009	0.484	0.013	−0.025	0.006
	γ_2	1.961	0.004	0.012	0.532	0.016	0.445	0.008
2	β_0	1.002	0.002	0.005	0.362	0.006	0.172	0.004
	β_1	−1.998	0.002	0.007	0.411	0.009	−0.140	0.005
	γ_2	1.979	0.003	0.009	0.509	0.012	0.071	0.006
3	β_0	0.986	0.002	0.008	0.565	0.010	0.123	0.005
	β_1	−1.996	0.003	0.010	0.376	0.013	0.363	0.007
	γ_2	1.953	0.003	0.012	0.455	0.015	0.204	0.008
4	β_0	0.995	0.002	0.007	0.555	0.010	−0.095	0.005
	β_1	−2.005	0.003	0.010	0.484	0.012	0.206	0.007
	γ_2	1.974	0.004	0.013	0.633	0.018	0.102	0.008
5	β_0	0.995	0.002	0.006	0.430	0.009	0.240	0.004
	β_1	−2.009	0.003	0.008	0.415	0.010	0.104	0.006
	γ_2	1.969	0.003	0.010	0.466	0.015	0.228	0.006

[a] Assumes independent samples.
[b] Between means of batches.
[c] Numeric standard error. See for example Geweke (1992).

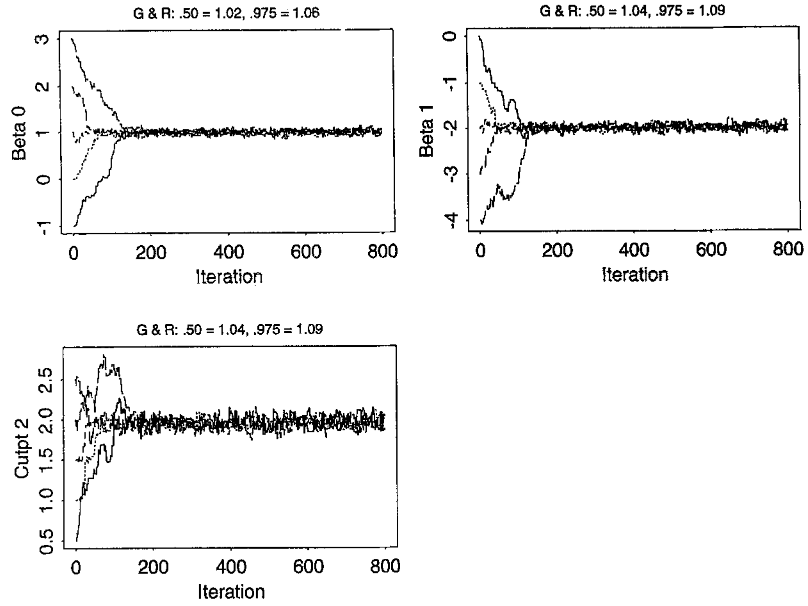

Fig. 4. *Three-bin ordinal probit, collapsed algorithm, 800 iterations*

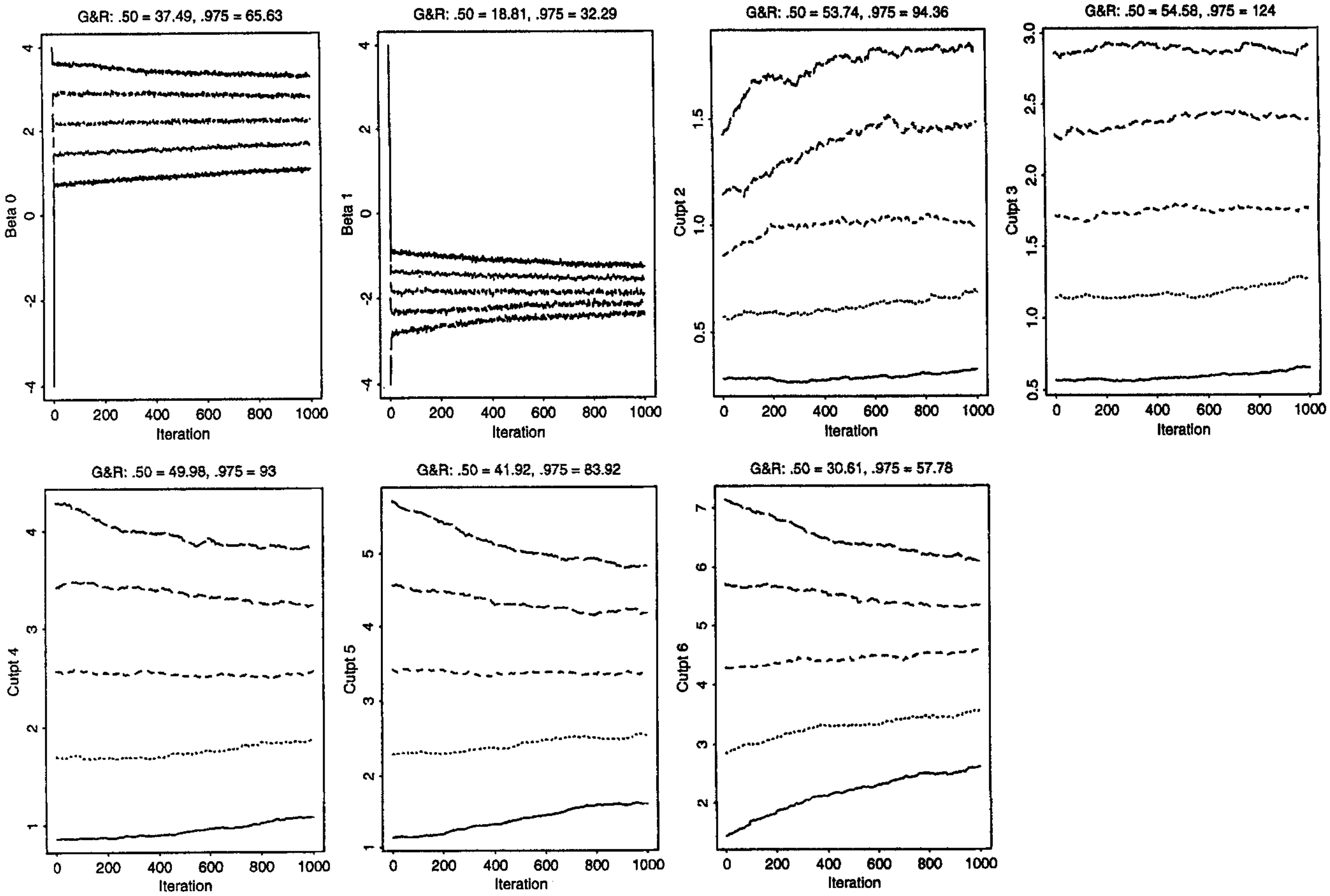

Fig. 5. *Seven-bin ordinal probit, univariate full conditionals, 1000 iterations*

1. Use one of the multivariate methods mentioned above to generate β and γ_2 together from their joint posterior distribution,

$$\begin{aligned} p(\beta, \gamma_2 | \boldsymbol{x}, \boldsymbol{w}) &\propto \prod_{i:w_i=1} \Phi(-\boldsymbol{x}_i^{\mathrm{T}}\beta) \\ &\times \prod_{i:w_i=2} [\Phi(\gamma_2 - \boldsymbol{x}_i^{\mathrm{T}}\beta) - \Phi(-\boldsymbol{x}_i^{\mathrm{T}}\beta)] \\ &\times \prod_{i:w_i=3} [1 - \Phi(\gamma_2 - \boldsymbol{x}_i^{\mathrm{T}}\beta)] \times I(\gamma_2 \in (0, \infty)). \end{aligned} \tag{3.4}$$

2. Generate the latent data points from their standard full conditionals (1a). (This step could be omitted if values of the latent data were not needed.)

We applied this algorithm to our simulated data, using the Hastings algorithm for step 1. The proposal density was multivariate normal with the first component, corresponding to the cutpoint, truncated to the positive line and with covariance matrix proportional to the sample covariance matrix obtained among the Gibbs iterates for γ_2, β_0, and β_1 in the chains shown in Fig. 3. The multiplicative constant for the covariance matrix in the proposal density was chosen to optimize the acceptance rate for a three-dimensional problem according to Gelman *et al.* (1994). Convergence plots shown in Fig. 4 for five chains run for 800 iterations using the collapsed algorithm indicate that the sample paths for each parameter take a few iterations longer to mix with this sampler than with the multivariate Hastings-within-Gibbs but that the Gelman and Rubin shrink factors are slightly smaller with the collapsed sampler. We prefer the multivariate Hastings-within-Gibbs approach because it takes slightly less computer time for the same number of iterations and does not require higher-dimensional proposal densities when the number of predictor variables increases.

4. Extension to other ordinal models

4.1. *Ordinal probit problems with more than three levels of the response variable*

To determine whether the same algorithm could be extended for use in ordinal probit problems with more than three bins, we simulated a dataset in exactly the same manner as in Section 2 except that the y^*'s were divided into septiles and the corresponding w_i's took on values from 1 up to 7. To analyse this seven-bin problem, five cutpoint parameters must be estimated. For each parameter, Fig. 5 shows Gelman and Rubin's shrink factors above plots of traces of five parallel Gibbs sampler chains run for 1000 iterations using all univariate full conditionals.

To apply our multivariate Hastings update algorithm in this setting, at iteration k of the Gibbs sampler we generated a vector of candidate cutpoint values, $\gamma_{j\,\mathrm{new}}, j = 2, \ldots, 6$, each from the truncated normal density

$$\begin{aligned} &g(\gamma_{j\,\mathrm{new}} | \gamma_j^{(k-1)}, \gamma_{j+1}^{(k-1)}, \gamma_{j-1\,\mathrm{new}}) = \mathrm{N}(\gamma_j^{(k-1)}, \sigma_\gamma^2) \\ &\text{truncated to the interval } (\gamma_{j-1\,\mathrm{new}}, \gamma_{j+1}^{(k-1)}). \end{aligned} \tag{4.1}$$

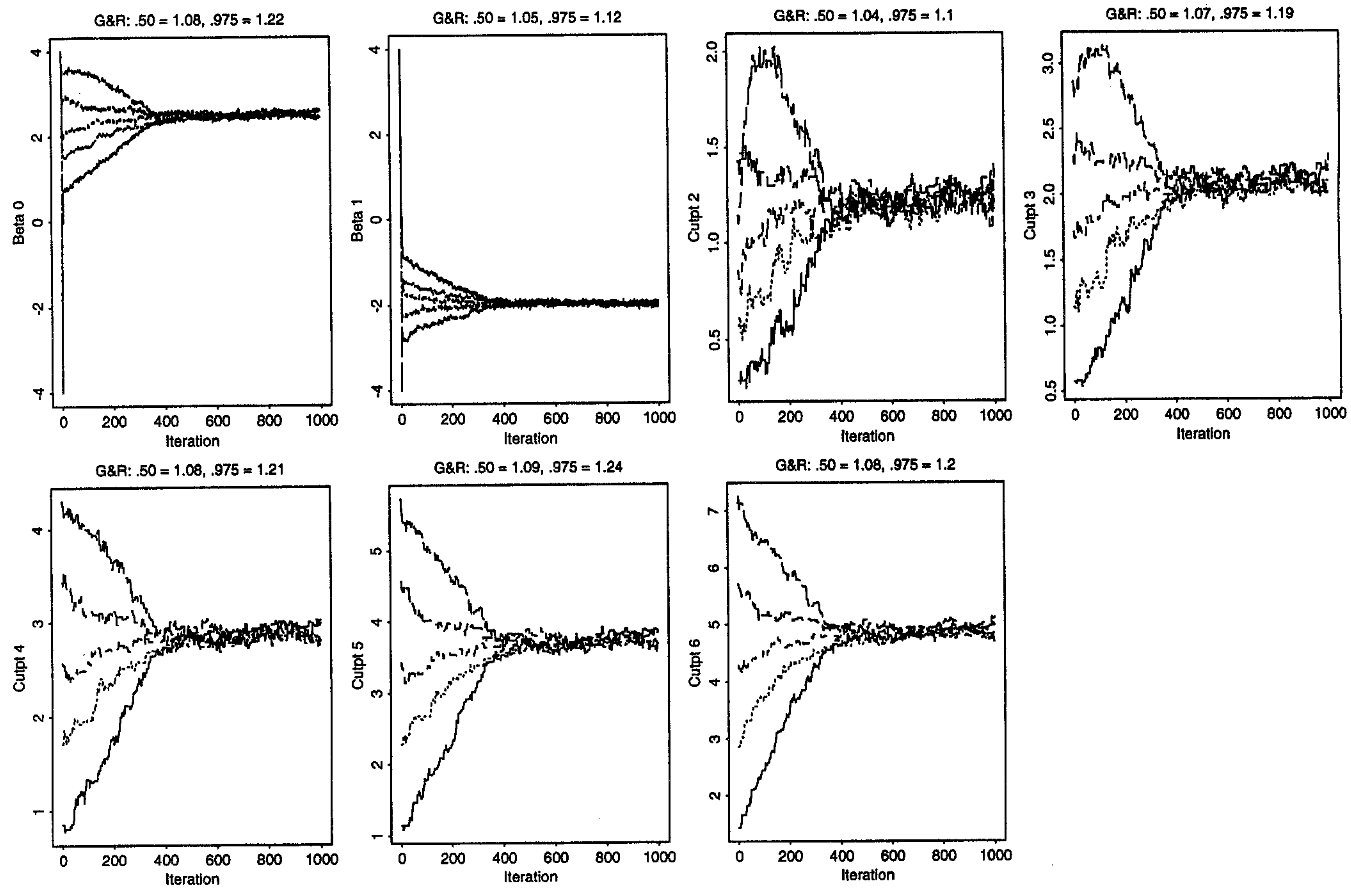

Fig. 6. *Seven-bin ordinal probit, multivariate Hastings, 1000 iterations*

The acceptance probability for the vector of new cutpoints was min(1, R) where

$$R = \prod_{j=2}^{6} \frac{\Phi((\gamma_{j+1}^{(k-1)} - \gamma_j^{(k-1)})/\sigma_\gamma) - \Phi((\gamma_{j-1\,\mathrm{new}} - \gamma_j^{(k-1)})/\sigma_\gamma)}{\Phi((\gamma_{j+1\,\mathrm{new}} - \gamma_{j\,\mathrm{new}})/\sigma_\gamma) - \Phi((\gamma_{j-1}^{(k-1)} - \gamma_{j\,\mathrm{new}})/\sigma_\gamma)}$$
$$\times \prod_i \frac{\Phi(\gamma_{w_i\,\mathrm{new}} - x_i^{\mathrm{T}}\beta) - \Phi(\gamma_{w_i-1\,\mathrm{new}} - x_i^{\mathrm{T}}\beta)}{\Phi(\gamma_{w_i}^{(k-1)} - x_i^{\mathrm{T}}\beta) - \Phi(\gamma_{w_i-1}^{(k-1)} - x_i^{\mathrm{T}}\beta)},$$

where γ_0, γ_1, and γ_7 are fixed at $-\infty$, 0, and $+\infty$ respectively. Figure 6 shows plots and Gelman and Rubin shrink factors for five parallel chains run using this algorithm on the simulated dataset for the seven-level ordinal probit. Within-chain autocorrelations (not shown) are much smaller with the multivariate Hastings algorithm than when the univariate full conditionals are used

4.2. *Cumulative logit and cumulative complementary log-log models*

It is possible to implement cumulative logit and cumulative complementary log-log models with the Gibbs sampler using latent variables. Random variates from truncated logistic and extreme value distributions (needed for the errors of the latent variables in the two respective models) are easily generated by the inversion method. As in the ordinal probit, latent data enable generation of the cutpoint parameters from the same simple full conditional shown in (1c). However, in cumulative logit and complementary log-log models, the full conditional for β is not a standard form even when latent data are used. Adaptive rejection sampling (see Gilks and Wild, 1992, and, for an application to generalized linear models, Dellaportas and Smith, 1993) or the Metropolis algorithm may be used within the Gibbs sampler to generate β from the univariate full conditional $p(\beta|\gamma, y^*) \propto \prod_i f(y_i^* - x_i^{\mathrm{T}}\beta)$, where f is the standard logistic or extreme value density.

To evaluate the performance of the Gibbs sampler in these models, we simulated two new datasets of 2000 data points each, exactly as we had for the three-bin ordinal probit in Section 2 except that the errors were generated from the standard logistic distribution and the standard extreme value (minimum) distribution respectively. We then ran two Gibbs samplers for each model, the first using univariate full conditionals and the second using our multivariate Hastings-within-Gibbs algorithm to generate the cutpoint and the latent data from their joint full conditional. As shown in Fig. 7, the results are very similar to those for the ordinal probit, with much faster convergence with the

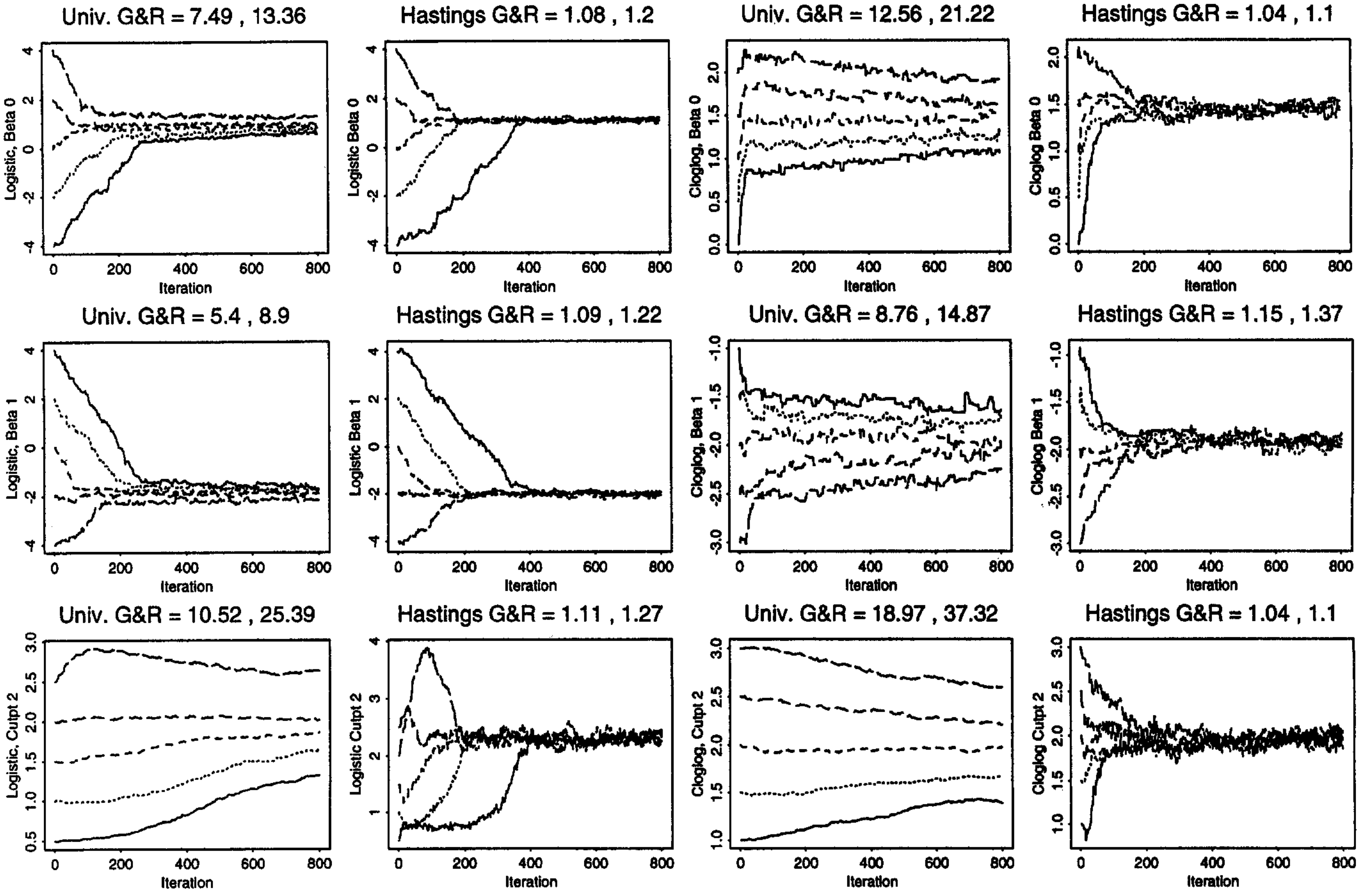

Fig. 7. *Three-bin cumulative logit and complementary log-log models, 800 iterations*

new algorithm. However, in cumulative logit and complementary log-log models, it is more efficient (and no more difficult) to use a collapsed Gibbs sampler as described at the end of Section 3, with the cumulative normal distribution in (3.4) replaced by the cumulative logistic or cumulative extreme value distribution.

5. Discussion and areas for further research

It is clear that our simple method of generating from the joint full conditional of the cutpoint parameters and the latent data accelerates Gibbs sampler convergence for cumulative-link GLMs applied to large datasets. The algorithm is very general. It is always possible to factor a joint full conditional as in (3.1a):

$$p(a, b|\text{rest}) = p(a|\text{rest})p(b|a, \text{rest}),$$

where a and b may be either vector or scalar parameters and 'rest' refers to all the other parameters in the model. If the joint proposal density is then constructed as the product of a suitable proposal density for a times the full conditional of b:

$$g(a_{\text{new}}, b_{\text{new}}|a_{\text{old}}, b_{\text{old}}, \text{rest}) \\ = g(a_{\text{new}}|a_{\text{old}}, \text{rest})p(b_{\text{new}}|a_{\text{new}}, \text{rest}),$$

then the full conditional of b will cancel out of the acceptance probability α. Thus this algorithm would be easy to implement when the marginal conditional of a is easy to derive and the full conditional of b (which is sampled in exactly the same way with this algorithm as in the usual Gibbs sampler) is easy to sample. It would be efficient when the full conditional of a is slower-mixing than the joint full conditional of a and b, particularly if, in addition, the full conditional of b is slow to sample so that it is advantageous not to have to sample from it when a new candidate for a is rejected. All of these factors hold in the ordinal probit. Further work is needed to identify other types of problems in which the same simple algorithm would be useful.

The Gibbs sampler probably is the most popular MCMC method at present. However, as illustrated in this paper, there are problems for which the pure Gibbs sampler with exclusively univariate full conditionals is poorly suited. For such problems, other MCMC algorithms, or hybrids of algorithms, must be employed. Even with faster algorithms, there is no way of determining with certainty by examining or analysing output from a Monte Carlo Markov chain that the samples are representative of the true underlying stationary distribution. It is to be hoped that progress will be made in the area of determining theoretical convergence bounds for generalized linear models.

Acknowledgements

The author thanks Dr Bradley P. Carlin, whose National Institute of Allergy and Infectious Diseases FIRST Award 1-R29-AI33466 supported this work in part. In addition, Dr Ming-Hui Chen and two reviewers provided helpful suggestions which greatly improved the manuscript.

References

Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York.

Albert, J. H. (1992) Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation*, **44**, 47–61.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–79.

Belisle, C. J. P., Romeijn, H. E. and Smith, R. O. (1993) Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, **18**, 255–66.

Carlin, B. P. and Polson, N. G. (1992) Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics* **4** (eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 577–86. Oxford University Press.

Casella, G. and George, E. I. (1992) Explaining the Gibbs sampler. *American Statistician*, **46**, 167–74.

Cowles, M. K., Carlin, B. P. and Connett, J. E. (1996) Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable nonresponse. *Journal of the American Statistical Association.*

Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Assocation.*

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalised linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443–59.

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.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Gelman, A., Roberts, G., and Gilks, W. (1996) Efficient Metropolis jumping rules. In *Bayesian Statistics* **5** (Eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), Oxford University Press..

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–41.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics* 4 (Eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 169–193. Oxford University Press.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo. *Statistical Science*, **7**, 473–83.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–48.

Gilks, W. R., Roberts, G. O., and George, E. (1994) Adaptive direction sampling. *The Statistician*, **43**, 179–88.

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

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–91.

Muller, P. (1994) A generic approach to posterior integration and Gibbs sampling. *Journal of the American Statistical Association.*

Raftery, A. E. and Lewis S. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics* **4** (Eds. J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 763–73. Oxford University Press.

Ripley, B. D. (1987) *Stochastic Simulation.* Wiley, New York.

Tanner, T. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–50.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–86.