Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression

Gabriel Rodriguez-Yam, Richard A. Davis, and Louis L. Scharf*

March, 2004

Abstract

In this paper we propose an efficient Gibbs sampler for simulation of a multivariate normal random vector subject to inequality linear constraints. Inference in a Bayesian linear model, where the regression parameters are subject to inequality linear constraints, is the primary motivation behind this research. In the literature, implementations of the Gibbs sampler for the multivariate normal distribution subject to inequality linear constraints and for the multiple linear regression with inequality constraints often exhibit poor mixing and slow convergence. This paper overcomes these limitations

* Gabriel Rodriguez-Yam is postdoctoral fellow and Richard Davis is Professor and chair, Department of Statistics, Colorado State University, Fort Collins, CO. 80523-1877 (email: rdavis@stat.colostate.edu). Louis Scharf is Professor, Departments of Electrical and Computer Engineering and Statistics, Colorado State University, Fort Collins, CO, 80523-1877 (email: scharf@engr.colostate.edu). This work forms part of the PhD dissertation of Gabriel Rodriguez-Yam, supported in part part by Colorado Advanced Software Institute (CASI) and received a scholarship from Consejo Nacional de Ciencia y Tecnologia (CONACYT). The work of Richard Davis was supported in part by NSF grant DMS-0308109. and, in addition, allows for the number of constraints to exceed the vector size and is able to cope with equality linear constraints.

KEY WORDS: Bayesian, Markov chain Monte Carlo, inequality linear constraints.

1 Introduction

In the classical linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\tag{1}$$

where $\mathbf{Y} = [Y_1, ..., Y_n]^T$ is the data vector, \mathbf{X} is an $n \times k$ (n > k) design matrix having full rank, $\boldsymbol{\epsilon}$ is a vector of errors that are independent and $N(0, \sigma^2)$ distributed, and $\boldsymbol{\beta}$ is the vector of regression parameters. The maximum likelihood estimate of $\boldsymbol{\beta}$, which coincides with the least squares estimator, is multivariate normal. Often times, there are applications in which inequality constraints are placed on $\boldsymbol{\beta}$. For example, in hyperspectral imaging, the spectrum signature of a composite substance in a pixel can be analyzed with the model in (1), where the columns of \mathbf{X} are the spectra of the k materials in the pixel (see Manolakis and Shaw 2002). Due to physical considerations, the components of $\boldsymbol{\beta}$, the abundance parameters, are required to be non-negative, i.e., $\boldsymbol{\beta} \geq 0$. This example fits into the more general framework where the vector $\boldsymbol{\beta}$ is subject to a set of inequality linear constraints which can be written as

$$\mathbf{B}\boldsymbol{\beta} \le \mathbf{b}.\tag{2}$$

As long as the set defined in (2) has positive Lebesgue measure, there is a positive probability that the least squares estimator of β may not satisfy the constraints. When it does, it coincides with the maximum likelihood estimate as in the unconstrained case. Except in simple cases, it is very difficult to obtain sampling properties of the inequality restricted least squares estimator of β .

Judge and Takayama (1966) and Liew (1976) give the inequality constrained least-squares (ICLS) estimate of β using the Dantzig-Cottle algorithm. The ICLS estimator reduces to the ordinary least squares estimator for a sufficiently large sample. Conditioned on knowing which constraints are binding and which are not they compute an untruncated covariance matrix of the ICLS estimator. Geweke (1986) points out that this variance matrix is incorrect, since in practice it is not known ahead of time which constraints will be binding. Thus, inferences based on this matrix can be seriously misleading (Lovell and Prescott 1970).

In this paper, we consider a Bayesian approach to this constrained inference problem. Geweke (1986) uses a prior that is the product of a conventional uninformative distribution and an indicator function representing the inequality constraints. The posterior distribution and expected values of functions of interest are then computed using importance sampling. In this case, an importance function is easy to find due to the simplicity of the prior. However, this method can be extremely slow, especially when the truncation region has a small probability with respect to the unconstrained posterior distribution.

Gelfand, Smith and Lee (1992) suggest an approach to routinely analyze problems with constrained parameters using the Gibbs sampler, a Monte Carlo Markov chain (MCMC) technique. Let \mathcal{D} denote the data and $\boldsymbol{\theta}$ a parameter vector with some prior distribution. Suppose it is difficult or impossible to draw samples from the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$. The Gibbs sampler, introduced by Geman and Geman (1984) in the context of image restoration, provides a method for generating samples from $p(\boldsymbol{\theta}|\mathcal{D})$. Suppose $\boldsymbol{\theta}$ can be partitioned as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$, where the $\boldsymbol{\theta}_i$'s are either uni- or multidimensional and that we can simulate from the conditional posterior densities $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. The Gibbs sampler generates a Markov chain by cycling through $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. In each cycle, the most recent information updates the posterior conditionals. Starting from some $\boldsymbol{\theta}^{(0)}$, after t cycles we have a realization $\boldsymbol{\theta}^{(t)}$ that under regularity conditions (Gelfand and Smith 1990), approximates a drawing from $p(\boldsymbol{\theta}|\mathcal{D})$ for large t. O'Hagan (1994), Gilks and Roberts (1996), Roberts (1996) comment that the rate of convergence depends on the posterior correlation between the components of the vector $\boldsymbol{\theta}$.

Geweke (1996) implements the Gibbs sampler for the problem of multiple linear regression with at most k independent inequality linear constraints given by

$$\mathbf{c} \le \mathbf{B}\boldsymbol{\beta} \le \mathbf{d},\tag{3}$$

where **B** is a square matrix of full rank, $\mathbf{c} < \mathbf{d}$ and the elements of \mathbf{c} and \mathbf{d} are allowed to be $-\infty$ and $+\infty$, respectively. Notice that these constraints can be easily rewritten in the form given in (2). However, Geweke's implementation may suffer from poor mixing (i.e., the chain does not move rapidly through the "entire" support of the posterior distribution). In our implementation we do not impose any limitation on the number of constraints given in (2). A major difference however, is that our implementation has faster mixing, requiring substantially fewer iterations of the Markov chain than previously published Gibbs sampler implementations.

In Rodriguez-Yam, Davis and Scharf (2002), a Gibbs sampler implementation with good mixing is provided for the hyperspectral imaging problem when only the non-negativity constraints on the abundance parameters are considered. For this case, the constraints are linearly independent and the number of inequality linear constraints coincides with the number of regression coefficients.

The organization of this paper is as follows. In Section 2 we provide a Bayesian framework for multiple linear regression where the regression parameters are subject to the constraints in (2). In Section 3 we list standard results for the truncated multivariate normal distribution that are used in this paper and provide an efficient Gibbs sampler from this distribution. Through an example where the constraints can be written as in (3) we compare our implementation with that of Geweke's. In Section 4 we use the implementation from Section 3 to provide an implementation of the Gibbs sampler to the model in Section 2 and apply the procedure to two datasets. One is the rental data analyzed by Geweke (1986, 1996) where the regression coefficients are subject to a set of inequality linear constraints that can be written as in (3). The other is aggregate data involving smokers preferences of three leading brands of cigarettes. For this example, equality linear constraints are needed in addition to inequality linear constraints and the number of inequality linear constraints exceeds the number of regression coefficients. Section 5 contains a summary of our findings.

2 Constrained Linear Regression

In this section we construct a Bayesian model for the multiple linear regression given in (1) where the parameters satisfy the constraints in (2). Before doing this, we introduce our notation. If R is a subset of \Re^k having positive Lebesgue measure, we call the random k-vector \mathbf{Y} truncated normal and write $\mathbf{Y} \sim N_R(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its probability density function is proportional to $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})I_R(\mathbf{x})$, where $\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the k-variate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $I_R(.)$ is the indicator function for R.

Now, the inequality linear constraints in (2) define a subset of \Re^k given by

$$T := \{ \boldsymbol{\beta} \in \Re^k : \mathbf{B}\boldsymbol{\beta} \le \mathbf{b} \}.$$
(4)

Notice that the model in (1) describes the conditional distribution of \mathbf{Y} given the vector of parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma^2)$, consisting of the coefficients of regression and the common variance of the noise errors. Now assume the prior for $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\beta} \sim N_T(\boldsymbol{\mu}_0, \, \sigma_0^2(\mathbf{X}^T \mathbf{X})^{-1}),$$
 (5)

$$\sigma^2 \sim \mathrm{IG}(\nu, \lambda), \tag{6}$$

where $\boldsymbol{\beta}$ and σ^2 are independently distributed, σ_0^2 , ν and λ are known positive scalars and $\boldsymbol{\mu}_0$ is a known vector. If $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ denotes the posterior distribution of $\boldsymbol{\theta}$ given the observed vector \mathbf{y} , then,

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \propto \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) p(\boldsymbol{\beta}) p(\sigma^2)$$
 (7)

where $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ is the likelihood function based on the data \mathbf{y} from the model in (1). A sample from the posterior density $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})$ will allow us to compute posterior quantities, such as means, variances, probabilities, and so on. In Section 4 below we describe how to obtain such a sample.

3 Truncated Multivariate Normal Distribution

In order to have an efficient Gibbs sampler for the multiple linear regression problem with inequality linear constraints as given in (4), it is imperative to have an efficient sampler to the truncated multivariate normal distribution. Before pursuing this objective we begin by developing two properties of the truncated multivariate normal distribution and then propose an implementation of the Gibbs sampler for the multivariate normal distribution subject to a set of inequality linear constraints. A key feature of this implementation is the construction of variables that are independent when the constraints are ignored. Using the first example from Geweke (1991), the performance of our implementation is then compared with that of Geweke's.

For a multivariate normal random vector \mathbf{X} , all linear transformations and conditional distributions of \mathbf{X} are normal. It turns out that for a truncated normal vector, these closure properties remain valid. That is, a linear transformation of a truncated normal vector is also truncated normal and so are the one-dimensional conditional distributions. These conditional distributions play a key role in the implementation of the Gibbs sampler we propose. The specifics are as follows:

Result 1 (a) Suppose $\mathbf{X} \sim N_R(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $R \in \Re^k$ has positive Lebesgue measure, and $\boldsymbol{\Sigma}$ is positive definite. Let $\mathbf{Y} := \mathbf{A}\mathbf{X}$, where \mathbf{A} is a matrix of full rank of

dimension $r \times k$ with $r \leq k$. Then,

$$\mathbf{Y} \sim N_T(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T), \quad T := \{\mathbf{A}\mathbf{x} : \mathbf{x} \in R\}.$$
 (8)

(b) Partition \mathbf{X} , $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ X_k \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_k \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_1^T & \sigma_{kk} \end{bmatrix}.$$
(9)

Then,

$$X_k | \mathbf{X}_1 = \mathbf{x}_1 \sim N_{R_k}(\mu_k^*, \sigma_{kk}^*), \tag{10}$$

where

$$\mu_k^* = \mu_k + \Sigma_1^T \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \qquad (11)$$

$$\sigma_{kk}^* = \sigma_{kk} - \Sigma_1^T \Sigma_{11}^{-1} \Sigma_1, \qquad (12)$$

$$R_k := \{ x_k \in \Re : (\mathbf{x}_1, x_k) \in R \}.$$

$$(13)$$

The proof of (a) is immediate from the form of the density function for truncated normal random vectors. To prove (b) the expressions for the inverse of a partitioned symmetric matrix (in Hocking 1996) are used, from which the result is immediate. \Box

Gibbs Sampler

Suppose $\boldsymbol{\theta}$ is a vector of parameters with posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$, where \mathcal{D} denotes the data. Partition $\boldsymbol{\theta}$ as $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$, where the $\boldsymbol{\theta}_i$'s are either uni- or multidimensional in such a way that we can simulate from the conditional posterior densities $p(\boldsymbol{\theta}_i|\mathcal{D}, \boldsymbol{\theta}_j, j \neq i)$. The basic Gibbs sampler starts with an initial value $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_q^{(0)})$ from the support of $p(\boldsymbol{\theta}|\mathcal{D})$ and then generates $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\theta}_1^{(t)}, \ldots, \boldsymbol{\theta}_q^{(t)}), t=1, 2, \ldots$, recursively as follows:

Generate $\boldsymbol{\theta}_1^{(t)}$ from $p(\boldsymbol{\theta}_1 | \mathcal{D}, \boldsymbol{\theta}_2^{(t-1)}, \dots, \boldsymbol{\theta}_q^{(t-1)})$

Generate
$$\boldsymbol{\theta}_{2}^{(t)}$$
 from $p(\boldsymbol{\theta}_{2}|\mathcal{D}, \boldsymbol{\theta}_{1}^{(t)}, \boldsymbol{\theta}_{3}^{(t-1)}, \dots, \boldsymbol{\theta}_{q}^{(t-1)})$
:
Generate $\boldsymbol{\theta}_{a}^{(t)}$ from $p(\boldsymbol{\theta}_{q}|\mathcal{D}, \boldsymbol{\theta}_{1}^{(t)}, \boldsymbol{\theta}_{2}^{(t)}, \dots, \boldsymbol{\theta}_{q-1}^{(t)})$.

Under certain regularity conditions (e.g. Gelfand and Smith 1990) the Markov chain $\{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$ has a stationary distribution which is the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$.

3.1 Gibbs sampler implementations

For comparison purposes, we first describe the implementation of the Gibbs sampler given by Geweke (1991) to a truncated normal random vector of dimension k subject to a set of at most k linearly independent inequality linear constraints. Suppose that **X** is a truncated normal random vector of dimension k, such that

$$\mathbf{X} \sim N_T(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}), \quad T := \{ \mathbf{x} \in \Re^k : \mathbf{c} \le \mathbf{B} \mathbf{x} \le \mathbf{d} \},$$
(14)

where \mathbf{c} , \mathbf{d} and \mathbf{B} are as in (3).

The Gibbs sampler in Geweke's implementation is applied to the transformed random vector $\mathbf{Y} = \mathbf{B}\mathbf{X}$. Note that

$$\mathbf{Y} \sim N_S(\mathbf{B}\boldsymbol{\mu}, \sigma^2 \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T), \quad S = \{\mathbf{y} \in \Re^k : \mathbf{c} \le \mathbf{y} \le \mathbf{d}\}.$$
 (15)

Thus, using (10)

$$Y_j|(Y_1 = y_1, \dots, Y_{j-1} = y_{j-1}, Y_{j+1} = y_{j+1}, \dots, Y_k = y_k) \sim N_{S_j}(\mu_j^*, \sigma_{jj}^*), \quad (16)$$

where $S_j = \{y_j \in \Re : c_j \leq y_j \leq d_j\}$, and μ_j^* and σ_{jj}^* must be obtained as in (11) and (12), respectively. Geweke's implementations, which we call *Sampler* TN1, is then Sampler TN1 (Geweke 1991)

Update the last component $\mathbf{y}^{(t)} = [y_1^{(t)}, y_2^{(t)}, \dots, y_k^{(t)}]^T$ of the current Gibbs path $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \dots, \mathbf{y}^{(t)}$, as follows: for $j = 1, \dots, k$

• draw $y_j^{(t+1)}$ from $p(y_j|y_1^{(t+1)}, \dots, y_{j-1}^{(t+1)}, y_{j+1}^{(t)}, \dots, y_k^{(t)}),$ (17)

where each conditional distribution is given in (16). \Box

The sampler we now propose allows for the number of constraints to exceed k. Begin with

$$\mathbf{X} \sim N_T(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}), \quad T := \{ \mathbf{x} \in \Re^k : \mathbf{B} \mathbf{x} \le \mathbf{b} \},$$
(18)

where the rows of the matrix **B** are not restricted to be linearly independent. Let **A** be a square matrix of full rank, such that $\mathbf{A}\Sigma\mathbf{A}^T = \mathbf{I}$, where **I** is the identity matrix and set

$$\mathbf{Z} := \mathbf{A}\mathbf{X}.\tag{19}$$

From (a) of Result 1, it follows that

$$\mathbf{Z} \sim N_S(\mathbf{A}\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \qquad S = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \Re^k, \mathbf{B}\mathbf{x} \le \mathbf{b}\}.$$
 (20)

The set S can be rewritten in the more suggestive way,

$$S = \{ \mathbf{z} \in \Re^k, \mathbf{D}\mathbf{z} \le \mathbf{b} \},\tag{21}$$

where

$$\mathbf{D} := \mathbf{B}\mathbf{A}^{-1}.\tag{22}$$

Thus, the transformation in (19) simplifies the functional form of the truncated multivariate distribution, but not the constraints.

If $\boldsymbol{\alpha} := \mathbf{A}\boldsymbol{\mu}$, and \mathbf{Z}_{-j} and \mathbf{z}_{-j} denote the vectors $[Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_k]^T$ and $[z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_k]^T$, respectively, then by (b) of Result 1,

$$Z_j | \mathbf{Z}_{-j} = \mathbf{z}_{-j} \sim N_{S_j}(\alpha_j, \ \sigma^2), \tag{23}$$

where from (13) and (21)

$$S_j = \{z_j \in \Re : \mathbf{z} \in S\} = \{z_j \in \Re : \mathbf{Dz} \le \mathbf{b}\}.$$

Let \mathbf{D}_{-j} be the matrix obtained from $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_k]$ by removing the *j*-th column. Then the set S_j can be easily computed from the equation

$$S_j = \{ z_j \in \Re : \mathbf{d}_j z_j \le \mathbf{b} - \mathbf{D}_{-j} \mathbf{z}_{-j} \}.$$
(24)

Since the constraints on **X** form a convex subset of \Re^k , the set S_j in (24) can be written as one of the intervals $l_j \leq z_j \leq u_j$, $-\infty < z_j \leq u_j$ or $l_j \leq \eta_j < +\infty$. The values l_j and u_j can be easily obtained from the set of one-dimensional inequalities in (24).

The idea behind the transformation in (19) is to obtain an efficient implementation of the Gibbs sampler based on the set of k conditional distributions in (23). These distributions have a simple form. That is, once the transformed mean α has been obtained, we do not need to use equations like (11) and (12) to compute the k means and variances. Also, the one-dimensional truncation intervals S_j in (24) evolve simply.

To illustrate this process, consider the following example. Let $\mathbf{X} \sim N_T(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, where

$$T = \{ \mathbf{x} \in \Re^2 : \mathbf{x} \ge \mathbf{0} \}, \text{ and } \mathbf{\Sigma} = \begin{bmatrix} 1 & 4/5 \\ 4/5 & 1 \end{bmatrix}$$

Notice that in the notation in (4), $\mathbf{B} = -\mathbf{I}$ and $\mathbf{b} = \mathbf{0}$, where \mathbf{I} is the identity matrix. For the lower-triangular Cholesky factor \mathbf{A} of $\boldsymbol{\Sigma}$ given by

$$\mathbf{A} = \begin{bmatrix} 1 & 0\\ -4/3 & 5/3 \end{bmatrix},$$

we obtain $\mathbf{A}\Sigma\mathbf{A}^T = \mathbf{I}$. The matrix $\mathbf{D} = \mathbf{B}\mathbf{A}^{-1}$ given in (22), and the submatrices \mathbf{D}_{-1} and \mathbf{D}_{-2} , are

$$\mathbf{D} = \begin{bmatrix} -1 & 0\\ -4/5 & -3/5 \end{bmatrix}, \quad \mathbf{D}_{-1} = \begin{bmatrix} 0\\ -3/5 \end{bmatrix}, \quad \mathbf{D}_{-2} = \begin{bmatrix} -1\\ -4/5 \end{bmatrix}$$

Then,

$$Z_1 | \mathbf{Z}_{-1} = \mathbf{z}_{-1} \sim N_{S_1}(\mu_1, \sigma^2), \quad Z_2 | \mathbf{Z}_{-2} = \mathbf{z}_{-2} \sim N_{S_2}(-\frac{4}{3}\mu_1 + \frac{5}{3}\mu_2, \sigma^2),$$

where

$$S_{1} = \{z_{1} \in \Re : \begin{bmatrix} -1 \\ -4/5 \end{bmatrix} z_{1} \leq -\begin{bmatrix} 0 \\ -3/5 \end{bmatrix} z_{2}\}$$
$$= \{z_{1} \in \Re : z_{1} \geq 0; z_{1} \geq -\frac{3}{4}z_{2}\}$$
$$= [\max\{0, -\frac{3}{4}z_{2}\}, \infty)$$

$$S_2 = \{z_2 \in \Re : \begin{bmatrix} 0\\ -3/5 \end{bmatrix} z_2 \le - \begin{bmatrix} -1\\ -4/5 \end{bmatrix} z_1\}$$
$$= \{z_2 \in \Re : z_2 \ge -\frac{4}{3}z_1\}$$
$$= [-\frac{4}{3}z_1, \infty).$$

For this example, the matrix **B** needed in (14) is the identity **I**. Hence, in Sampler TN1, the Gibbs sampler is implemented on the random vector **X**. In particular, using (11)-(12) in (16), it follows that

$$Y_1|Y_2 = y_2 \sim N_{S_1}(\mu_1 + \frac{4}{5}(y_2 - \mu_2)\sigma^2, \frac{9}{25}\sigma^2),$$

$$Y_2|Y_1 = y_1 \sim N_{S_2}(\mu_2 + \frac{4}{5}(y_1 - \mu_1)\sigma^2, \frac{9}{25}\sigma^2)$$

where $S_1 = S_2 = [0, +\infty)$. \Box

Now, to obtain a sample from the distribution of \mathbf{X} we obtain first a sample from the transformed vector \mathbf{Z} in (19). A sample from \mathbf{X} is then obtained by "undoing" this transformation. For later reference this new implementation will be called *Sampler* TN2.

Sampler TN2 (This paper)

Let $\mathbf{z}_0 \in S$ be an initial value of the sampler. The last component $\mathbf{z}^{(t)} = [z_1^{(t)}, z_2^{(t)}, \dots, z_k^{(t)}]^T$ of the current Gibbs path $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \dots, \mathbf{z}^{(t)}$ is updated as follows:

- draw $z_j^{(t+1)}$ from $p(z_j | z_1^{(t+1)}, \dots, z_{j-1}^{(t+1)}, z_{j+1}^{(t)}, \dots, z_k^{(t)}), \ j = 1, \dots, k,$
- set $\mathbf{X}^{(t+1)} = \mathbf{A}^{-1}\mathbf{Z}^{(t+1)}$,

where the conditional distribution $p(z_j|z_1^{(t+1)}, \ldots, z_{j-1}^{(t+1)}, z_{j+1}^{(t)}, \ldots, z_k^{(t)})$ is given in (23). \Box

3.2 Performance comparison of Samplers TN1 and TN2

To compare the performance between Samplers TN1 and TN2, we consider an example in which $\mathbf{X} \sim N_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} 0\\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & \rho\\ \rho & 0.1 \end{bmatrix}, \quad (25)$$

and T is the region determined by the constraints,

$$c_1 \le X_1 + X_2 \le d_1, \quad c_2 \le X_1 - X_2 \le d_2.$$
 (26)

These constraints can be written in the format in (14) with $\mathbf{c} = [c_1 \ c_2]^T$, $\mathbf{d} = [d_1 \ d_2]^T$ and

$$\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

To provide some indication of the efficiency of his procedure, Geweke (1991) considered five configurations of truncation points of c_1 , c_2 , d_1 and d_2 (and $\rho = 0$). In this paper we consider three configurations of truncation points of c_1 , c_2 , d_1 and d_2 and three values of ρ . For each configuration we apply the two Gibbs sampler implementations described above and stop after 1600 iterations. As a means of comparison of the two implementations, the results of the Raftery and Lewis convergence diagnostic procedure for each chain are shown in Tables 1 and 2. In the general set up of the Gibbs sampler, this diagnostic, introduced by Raftery and

		thinning			lower	dependence				
ρ	variable	(k)	burn-in	Total	bound	factor				
$-\infty < X_1 + X_2 < \infty, -\infty < X_1 - X_2 < \infty$										
-0.7	X_1	23	115	58282	1537	37.92				
	X_2	3	12	6726	1537	4.37				
0	X_1	18	90	46206	1537	30.06				
	X_2	1	2	1551	1537	1.01				
0.7	X_1	22	154	72050	1537	23.96				
	X_2	3	12	6549	1537	5.89				
$-10 \le X_1 + X_2 \le 10, -10 \le X_1 - X_2 \le 10$										
-0.7	X_1	25	150	78650	1537	51.17				
	X_2	6	24	14160	1537	9.21				
0	X_1	10	60	30040	1537	19.54				
	X_2	1	2	1558	1537	1.01				
0.7	X_1	20	120	60380	1537	39.28				
	X_2	6	24	13272	1537	8.64				
$-1 \le X_1 + X_2 \le 1, -1 \le X_1 - X_2 \le 1$										
-0.7	X_1	3	12	6087	1537	3.96				
	X_2	1	3	1397	1537	0.91				
0	X_1	2	6	3436	1537	2.24				
	X_2	1	3	1312	1537	0.85				
0.7	X_1	3	12	5976	1537	3.89				
	X_2	1	3	1329	1537	0.86				

Table 1: Raftery and Lewis convergence diagnostics for Sampler TN1 implemented on the truncated normal random vector $[X_1, X_2]^T$ with unconstrained mean **0**, $\operatorname{var}\{X_1\} = 10$, $\operatorname{var}\{X_2\} = 0.10$ and $\operatorname{cor}\{X_1, X_2\} = \rho$, subject to the constraints $c_1 \leq X_1 + X_2 \leq d_1$, $c_2 \leq X_1 - X_2 \leq d_2$.

		lower	dependence							
ρ	variable	(k)	burn-in	Total	bound	factor				
$-\infty < X_1 + X_2 < \infty, -\infty < X_1 - X_2 < \infty$										
-0.7	X_1	1	3	1702	1537	1.11				
	X_2	1	2	1501	1537	0.98				
0	X_1	1	2	1432	1537	0.93				
	X_2	1	2	1582	1537	1.03				
0.7	X_1	1	2	1505	1537	0.98				
	X_2	1	2	1516	1537	0.99				
	$-10 \le X_1 + X_2 \le 10, -10 \le X_1 - X_2 \le 10$									
-0.7	X_1	1	2	1490	1537	0.97				
	X_2	1	2	1566	1537	1.02				
0	X_1	1	2	1509	1537	0.98				
	X_2	1	2	1490	1537	0.97				
0.7	X_1	1	2	1614	1537	1.05				
	X_2	1	3	1689	1537	1.10				
	-1	$1 \le X_1$	$-X_2 \leq$	1						
-0.7	X_1	1	2	1566	1537	1.02				
	X_2	1	2	1450	1537	0.94				
0	X_1	1	3	1390	1537	0.90				
	X_2	1	2	1520	1537	0.99				
0.7	X_1	1	2	1655	1537	1.08				
	X_2	1	2	1531	1537	1.00				

Table 2: Raftery and Lewis convergence diagnostics for Sampler TN2 implemented on the truncated bivariate normal random vector $[X_1, X_2]^T$ described in Table 1.

Lewis (1992), determines the total number of iterations required to compute quantiles of functionals of $\boldsymbol{\theta}$. Also, the method gives the number of initial iterations that must be discarded to allow for "burn-in". Some specifics of the method are as follows. Let $\boldsymbol{\xi}$ be a function of the parameter vector $\boldsymbol{\theta}$. For a fixed probability s, a known q and accuracy r, suppose that we want to estimate the value of the quantile u, given by $P(\xi \leq u | \mathcal{D}) = q$ in such a way that $P(|\hat{q} - q| \leq r | \mathcal{D}) = s$, where \hat{q} is an estimator of q based on the sample path of the chain.

In Tables 1 and 2, the columns labeled "bound" would be the total length needed if the components of the chain were in fact an iid sample. The column labeled as "thinning" means that after the burn-in, every k-th observation is used. In both tables we set q = 0.5, r = 0.025 and s = 0.95. Based on the results from these tables, we note that the convergence of Sampler TN1 is much slower than that for Sampler TN2. Also, ρ (which is the correlation between X_1 and X_2 when no truncation is considered) affects the performance of Sampler TN1. In general, as the region of truncation gets small, the speed of convergence of Sampler TN1 improves. One possible explanation for this is that the chain must cover a "small" region faster than a "large" region. On the other hand, Sampler TN2 has the advantage of providing samples that are "close" to iid samples, regardless of the size of the region of truncation or the correlation of X_1 and X_2 . In fact, for the configuration $-\infty < X_1 + X_2 < \infty$, $-\infty < X_1 - X_2 < \infty$, this sampler provides an iid sample, since the conditional distribution in (23) does not depend on the fixed values \mathbf{z}_{-i} (e.g., see O'Hagan 1994, p. 233).

The autocorrelations of the output of a Gibbs sampler can be used to measure the performance of a simulation. Chen, Shao and Ibrahim (2000) observe that slow decay in the autocorrelations suggests slow mixing within a chain and usually slow convergence to the posterior distribution. For this example, the autocorrelations of X_1 , X_2 and $X_1 + X_2$ for two configurations of values of ρ , **c** and **d** using the output of the Sampler TN1 are shown in Figure 1. The first row of graphs contains the autocorrelations for $\rho = 0$, $-\infty < X_1 + X_2 < \infty$, and $-\infty < X_1 - X_2 < \infty$ and the second row of graphs contains the autocorrelations for $\rho = -0.7$, $-1 \le X_1 + X_2 \le 1$ and $-1 \le X_1 - X_2 \le 1$. Figure 2 contains the analogous autocorrelations for the output of Sampler TN2. For the two configurations considered in Figures 1 and 2,



Figure 1: Autocorrelation plots of X_1 , X_2 and $X_1 + X_2$ for two configurations of values of ρ , **c** and **d** obtained with Sampler TN1.



Figure 2: Autocorrelation plots of X_1 , X_2 and $X_1 + X_2$ for two configurations of values of ρ , **c** and **d** obtained with Sampler TN2.

we conclude that the mixing of the Sampler TN2 is better than that of the Sampler TN1. Notice that the column labeled "dependence factor" in Tables 1 and 2 is

related to the mixing provided by the autocorrelation plots in Figures 1 and 2. That is, the slower the mixing, the higher the dependence factor and vice versa.

A useful graphical tool to assess performance of a Gibbs sampler consists in monitoring some statistics of the output against iteration. In order to achieve stationarity, the monitored statistics must stabilize at some iteration. Thus, a monitored statistic which has not yet stabilized provides evidence for non convergence to stationarity. In Figure 3 we monitor the means of X_1 , X_2 and $X_1 + X_2$ against the iteration number using the output of the two implementations with the configurations used in the autocorrelation plots given in Figures 1 and 2. In this figure,



Figure 3: Running mean plots of X_1 , X_2 and $X_1 + X_2$ for two configurations of values of ρ , **c** and **d**. The solid lines are the running means obtained with the output of Sampler TN1. The dotted lines are the running means computed with the output of Gibbs Sampler TN2. In the first row, the horizontal lines show the means of the monitored statistics and the dotted lines are the lower and upper limits of the true 95% confidence intervals for these statistics.

the solid lines are the means obtained using the output of Sampler TN1, while the dotted lines are the means obtained with the output of Sampler TN2. Recall that Sampler TN2 provides an iid sample $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(t)}$ from the distribution of \mathbf{X} when $-\infty < X_1 + X_2 < \infty$, $-\infty < X_1 - X_2 < \infty$. Thus, the means and variances of \bar{X}_1, \bar{X}_2 and $\bar{X}_1 + \bar{X}_2$ (which are estimators of the means of X_1, X_2 and $X_1 + X_2$, respectively) are known. For example, $E(\bar{X}_1) = 0$ and $\operatorname{var}(\bar{X}_1) = 10/t$. In the first row of Figure 3, the horizontal solid lines show the expected means of the monitored statistics, while the dotted lines show the upper and lower 95% confidence limits $\mp 1.96\sigma$ of the monitored statistics. We note in this figure that the monitored means stabilize earlier for the Sampler TN2. In particular, in the upper left panel, with Sampler TN2, the monitored means stabilize after 500 iterations, while for Sampler TN1 they have not yet stabilized even after 1500 iterations.

4 Gibbs Sampler Implementations to the Constrained Linear Regression

In this section we implement the Gibbs sampler for a Bayesian linear regression model in which the regression coefficients satisfy inequality linear constraints. When the number of constraints does not exceed the number of regression coefficients we compare our procedure with the implementation described in Geweke (1996). In addition, we show through an example how the case of equality linear constraints can be handled.

Combining the prior distribution of $(\boldsymbol{\beta}, \sigma^2)$ given in (5)-(6) with the likelihood of the model in (1) we have

$$\boldsymbol{\beta}|(\sigma^2, \mathbf{y}) \sim N_T(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$
 (27)

$$\sigma^{-2}|(\boldsymbol{\beta}, \mathbf{y}) \sim (SS(\boldsymbol{\beta}) + 2\lambda)^{-1}\chi^2_{n+2\nu}, \qquad (28)$$

where $\chi^2_{n+2\nu}$ denotes a chi-squared distribution with $n + 2\nu$ degrees of freedom, T

is defined in (4), and

$$\mu_{1} = \gamma \boldsymbol{\beta} + (1 - \gamma) \mu_{0}$$

$$\Sigma_{1} = \sigma^{2} \gamma (\mathbf{X}^{T} \mathbf{X})^{-1}$$

$$SS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})^{T} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})$$

$$\gamma = \sigma_{0}^{2} / (\sigma_{0}^{2} + \sigma^{2})$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{T} \mathbf{X})^{-1} \mathbf{X}^{T} \mathbf{y}.$$

To show (27) and (28), an analogous procedure for the unconstrained linear model (e.g., Tanner 1996, pp. 17-18) can be followed.

For comparison purposes, we describe now the implementation of the Gibbs sampler given by Geweke (1996) for a multiple linear regression model where the regression coefficients are subject to a set of at most k linearly independent inequality linear constraints. That is,

$$T := \{ \boldsymbol{\beta} \in \Re^k : \mathbf{c} \le \mathbf{B} \boldsymbol{\beta} \le \mathbf{d} \},\$$

where \mathbf{c} , \mathbf{d} and \mathbf{B} are as in (3).

As in Sampler TN1, the vector of regression coefficients β is transformed to $\eta = \mathbf{B}\beta$. Then,

$$\boldsymbol{\eta}|(\sigma^2, \mathbf{y}) \sim N_S(\mathbf{B}\boldsymbol{\mu}_1, \mathbf{B}\boldsymbol{\Sigma}_1\mathbf{B}^T), \quad S = \{\boldsymbol{\beta} \in \Re^k : \mathbf{c} \le \boldsymbol{\beta} \le \mathbf{d}\}.$$
 (29)

The full implementation of the Gibbs sampler to the vector $\boldsymbol{\theta} := (\boldsymbol{\eta}, \sigma^2)$ proposed by Geweke is summarized in the following sampler.

Sampler CLR1

Let $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(t)}$ be the current path of the Gibbs sampler. The last component $\boldsymbol{\theta}^{(t)} = (\eta_1^{(t)}, \dots, \eta_k^{(t)}, \sigma^{2(t)})$ is updated as follows:

• Generate $\eta_1^{(t+1)}$ from $p(\eta_1|\eta_2^{(t)},\ldots,\eta_k^{(t)},\sigma^{2(t)},\mathbf{y})$

- Generate $\eta_k^{(t+1)}$ from $p(\eta_k | \eta_1^{(t+1)}, \eta_2^{(t+1)} \dots, \eta_{k-1}^{(t+1)}, \sigma^{2(t)}, \mathbf{y})$
- Generate $\sigma^{2(t+1)}$ from $p(\sigma^2 | \eta_1^{(t+1)}, \eta_2^{(t+1)} \dots, \eta_k^{(t+1)}, \mathbf{y}),$

where, due to (29), for j = 1, ..., k, the distribution

$$p(\eta_j|\eta_1^{(t+1)},\ldots,\eta_{j-1}^{(t+1)},\eta_{j+1}^{(t)},\ldots,\eta_k^{(t)},\sigma^{2(t)},\mathbf{y}),$$

is univariate normal truncated below by c_i , truncated above by d_i , and its mean and variance found using (29) along with the expressions in (11) and (12). Also, $p(\sigma^2|\eta_1^{(t+1)}, \eta_2^{(t+1)}, \dots, \eta_k^{(t+1)}, \mathbf{y})$ can be obtained from (28). \Box

Now, we give a new implementation, similar to Sampler TN2, for the case where the number of inequality linear constraints can exceed the number of regression parameters. For this case, T is given in (4). Let \mathbf{A} be a non-singular matrix for which $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A} = \mathbf{I}$, and set

$$\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\beta}.\tag{30}$$

Then, from (8) and (27)

$$\boldsymbol{\eta}|(\sigma^2, \mathbf{y}) \sim N_S(\mathbf{A}\boldsymbol{\mu}_1, \sigma^2 \gamma \mathbf{I}), \qquad S = \{\boldsymbol{\eta} \in \Re^k : \mathbf{D}\boldsymbol{\eta} \le \mathbf{b}\},$$
 (31)

where $\mathbf{D} = \mathbf{B}\mathbf{A}^{-1}$ and \mathbf{B} and \mathbf{b} are defined as in (2). We implement the Gibbs sampler for the transformed vector $\boldsymbol{\theta} = (\boldsymbol{\eta}, \sigma^2)$. The details are given in Sampler CLR2.

Sampler CLR2

Update the last component $\boldsymbol{\theta}^{(t)} = (\eta_1^{(t)}, \dots, \eta_k^{(t)}, \sigma^{2(t)})$ of the current path $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(t)}$ of the Gibbs sampler as follows

- Generate η₁^(t+1) from p(η₁|η₂^(t),...,η_k^(t), σ^{2(t)}, y)
 Generate η₂^(t+1) from p(η₂|η₁^(t+1), η₃^(t)...,η_k^(t), σ^{2(t)}, y)
 Generate η_k^(t+1) from p(η_k|η₁^(t+1), η₂^(t+1)...,η_{k-1}^(t+1), σ^{2(t)}, y)
- Generate $\sigma^{2(t+1)}$ from $p(\sigma^2 | \eta_1^{(t+1)}, \eta_2^{(t+1)}, \dots, \eta_k^{(t+1)}, \mathbf{y}),$

where due to (31), for j = 1, ..., k, the distribution

$$p(\eta_j | \eta_1^{(t+1)}, \dots, \eta_{j-1}^{(t+1)}, \eta_{j+1}^{(t)}, \dots, \eta_k^{(t)}, \sigma^{2(t)}, \mathbf{y}),$$

can be obtained similarly as in Sampler TN2, and $p(\sigma^2|\eta_1^{(t+1)}, \eta_2^{(t+1)}, \dots, \eta_k^{(t+1)}, \mathbf{y})$ can be obtained from (28). \Box

4.1 Example: Rental Data

We consider the 32 observations provided by Pindyck and Rubinfeld (1981, p. 44) on rent paid, number of rooms rented, number of occupants, sex and distance from campus in blocks for undergraduates at the University of Michigan. Geweke (1986, 1996) considers the model

$$y_i = \beta_1 + \beta_2 s_i r_i + \beta_3 (1 - s_i) r_i + \beta_4 s_i d_i + \beta_5 (1 - s_i) d_i + \epsilon_i,$$

where y_i denotes rent paid per person, r_i number of rooms per person, d_i distance from campus in blocks, s_i is a dummy variable representing gender (one for male and zero for female), ϵ_i is normally distributed error with mean 0 and variance σ^2 , and the β 's are subject to the constraints

$$\beta_2 \ge 0, \ \beta_3 \ge 0, \ \beta_4 \le 0, \ \beta_5 \le 0.$$
 (32)

Since the number of constraints does not exceed the number of regression coefficients, Sampler CLR1 can be used to draw a sample from the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$. For this case, the matrix **B** is the identity of size 5, **c** := $[-\infty, 0, 0, -\infty, -\infty]^T$ and $\mathbf{d} := [\infty, \infty, \infty, 0, 0]^T$. For Sampler CLR2, $\mathbf{b} = \mathbf{0}$ and \mathbf{B} is given by

$$\mathbf{B} = \left| \begin{array}{ccccccc} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right|$$

Taking μ_0 to be the constrained MLE of β , $\sigma_0^2 = 1000$, and $\nu = \lambda = 0.001$ as the values needed for the prior distribution of (β, σ^2) , we obtained Gibbs paths of length 1600 for the posterior distribution of (β, σ^2) using Samplers CLR1 and CLR2. In Figure 4 we show the autocorrelation function plots of β_2 , β_3 and β_4 using these outputs. From this figure, we note that the mixing of the Sampler CLR2 is faster than that from Sampler CLR1.



Figure 4: Autocorrelation plots of β_2 , β_3 , and β_4 obtained with Gibbs paths of length 1600. The first row was obtained with the output of Sampler CLR1. The second was obtained with the output of Sampler CLR2.

To compare the speed of convergence of both samplers, we increased the length of both paths to a total length of 20000 (each path). In Figure 5 we show two sections of the running mean plots of σ^2 and all the components of β . In each panel, the thick lines correspond to the section 9401 $\leq t \leq$ 10000 while the thin lines correspond to the section 19401 $\leq t \leq$ 20000. The solid lines were obtained using the output of Sampler CLR1 and the dotted lines using the output of Sampler CLR2. While it appears that for Sampler CLR2, 20000 iterations are enough to stabilize the mean of σ^2 and β , this is not the case for Sampler CLR1.



Figure 5: Two sections of the running means of β_1, \ldots, β_5 and σ^2 . The thick lines show the first sections (9401 $\leq t \leq$ 10000) and the thin lines the second sections (19401 $\leq t \leq$ 20000) of these running means. The values obtained with the output of Sampler CLR1 are shown with solid lines and those obtained with the output of Sampler CLR2 with dotted lines.

4.2 Example: Application to the Cigarette-brand Preference Data

This example considers the estimation of the transition probability matrix of a finite Markov process when only the time series of the proportion of visits to each state is known. To estimate the transition probability matrix, Telser (1963), proposes least-squares estimation based on a set of regression models subject to constraints that the coefficients are non-negative and each row sums to 1. This problem is analyzed again by Judge and Takayama (1966), who take these constraints into account explicitly. The numerical example given by Telser (1963) and Jugde and Takayama (1966), consists of the annual sales in billions of cigarettes for the three leading brands from 1925 to 1943. Given the time ordered market shares of these brands and assuming that the probability of a transition, p_{ij} , from brand *i* to brand *j* is constant over time, Telser gives the regression models

$$y_{jt} = \sum_{i=1}^{3} y_{i,t-1} p_{ij} + u_{jt}, \quad j = 1, 2, 3,$$
(33)

where y_{jt} is the proportion of individuals in state j at time t and u_{jt} , t = 1, ..., Tare independent errors. The probabilities p_{ij} are subject to the constraints

$$\sum_{j=1}^{3} p_{ij} = 1, \text{ for all } i, \tag{34}$$

$$p_{ij} \geq 0$$
, for all *i*, and *j*. (35)

For the cigarettes data, the three models in (33) can be combined as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \mathbf{p}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix}, \quad (36)$$

where $\mathbf{y}_j := [y_{j2}, \dots, y_{jT}]^T$, \mathbf{W} is the common design matrix of dimension $3 \times T - 1$ from the models in (33), \mathbf{p}_j is the *j*-th column of the probability transition matrix **P** of the finite Markov process, and \mathbf{u}_j is the vector of errors from the model in (33).

We propose to treat the equality constraints in (34) as in the frequentist approach. Hocking (1996, p. 70) incorporates equality linear constraints into the so called *full model* to obtain a *reduced model*. For this Bayesian approach, a new feature appears. The equality constraints need to be "incorporated" in the support of the full model. Denote by **y** the response vector of the full model in (36), by \mathbf{W}_1 , \mathbf{W}_2 and \mathbf{W}_3 the matrices having the columns 1 through 3, 4 through 6 and 7 through 9, respectively of the design matrix in (36). Substituting $p_{i3} = 1 - p_{i1} - p_{i2}$, i = 1, 2, 3, in this model, we obtain

$$\mathbf{y} - \mathbf{W}_3 \begin{bmatrix} 1\\1\\1 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 - \mathbf{W}_3 & \mathbf{W}_2 - \mathbf{W}_3 \end{bmatrix} \begin{bmatrix} \mathbf{p}_1\\\mathbf{p}_2 \end{bmatrix} + \mathbf{u}, \quad (37)$$

subject to the constraints

$$p_{i1} + p_{i2} \leq 1, \quad i = 1, 2, 3,$$
(38)

$$p_{ij} \geq 0, \quad i = 1, 2, 3, \ j = 1, 2,$$
(39)

where **u** is the vector of errors from the model in (36). In their method, Judge and Takayama (1966) assume that $\operatorname{var}(\mathbf{u}) = \sigma^2 \mathbf{I}$. For simplicity we also assume that $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. However, a more general matrix of variance-covariance for **u** can be used, e.g., $\operatorname{var}\{\mathbf{u}_j\} = \sigma_j^2 \mathbf{I}$ and $\operatorname{cov}\{\mathbf{u}_{j_1}, \mathbf{u}_{j_2}\} = \mathbf{0}, j_1 \neq j_2$. For this case, a prior distribution for the vector $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ needs to be specified.

Notice that the number of constraints in (38) and (39) to the regression model in (37) exceeds the number of regression coefficients. This time, Sampler CLR1 can not be carried out. To implement the Sampler CLR2, set $\boldsymbol{\mu}_0$ equal to the constrained MLE of $\boldsymbol{\beta} := [\mathbf{p}_1^T \quad \mathbf{p}_2^T]$, σ_0^2 large (100), and $\nu = \lambda = 0.001$ as the values needed for the prior distribution of $(\boldsymbol{\beta}, \sigma^2)$. A path of length 5000 for the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ is then generated. Based on the last 2500 iterates of this sample, the estimate $\hat{\mathbf{P}}$ of the probability transition matrix and the matrix $\hat{\sigma}_{\hat{\mathbf{P}}}$ having in its entries the estimated standard error of each component of $\hat{\mathbf{P}}$ are

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.690 & 0.118 & 0.192 \\ 0.035 & 0.844 & 0.121 \\ 0.334 & 0.060 & 0.606 \end{bmatrix},$$
(40)
$$\hat{\sigma}_{\hat{\mathbf{P}}} = \begin{bmatrix} 0.0016 & 0.0009 & 0.0016 \\ 0.0006 & 0.0008 & 0.0009 \\ 0.0023 & 0.0010 & 0.0024 \end{bmatrix}.$$

The restricted least-squares estimates obtained by Judge and Takayama (1966) are given by

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.6686 & 0.1423 & 0.1891 \\ 0 & 0.8683 & 0.1317 \\ 0.4019 & 0 & 0.5981 \end{bmatrix}.$$
(41)

The estimates in (40) differ slightly from the restricted least-squares in (41). Perhaps the most important difference is the fact that the estimates of p_{21} and p_{32} are non zero. The zero estimates of the elements of **P** can induce misleading interpretations. For example, because $\hat{p}_{21} = 0$, a smoker of the second brand never tries cigarettes of the first brand, unless he tries cigarettes of the third brand. This unlikely behavior does not show up with the estimates in (40).

To get an indication of how the sampler performs, the autocorrelation plots of the components of the matrix \mathbf{P} and the running means of these components are shown in Figures 6 and 7, respectively. In Figure 6 we observe a fast decay on the autocorrelations. Following Chen, et al. (2000), we expect a good mixing and fast convergence. This is in fact corroborated by the results in Figure 7, where the monitored statistics seem to be stabilized after a relatively small number of iterations.



Figure 6: Autocorrelation plots of the components of the transition probability matrix **P** of the cigarattes data obtained with a Gibbs path of length 5000.



Figure 7: Running mean plots of the components of the transition probability matrix \mathbf{P} of the cigarattes data obtained with a Gibbs path of length 5000.

5 Conclusions

In this paper, Bayesian analysis of a linear regression model where the parameters are subject to inequality linear constraints has been considered. Our method is based on an efficient Gibbs sampler for the truncated multivariate normal distribution. This sampler mixes fast, a property that is not always enjoyed by other implementations and can cope with non-standard situations such as when the constraints are linearly dependent and when the number of constraints exceed the number of regression coefficients. We have shown with an example how to manage equality linear constraints; a case in which other implementations do not apply.

References

- Chen, M-H., Shao, Q-M., and Ibrahim, J. G. (2000), Monte Carlo Methods in Bayesian Computation, New York: Springer.
- [2] 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.
- [3] Gelfand, A. E., Smith, A. F. M., and Lee, T. M. (1992), "Bayesian Analysis of Constrained Parameters and Truncated Data Problems," *Journal of the American Statistical Association*, 87, 523-532.
- [4] 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, 721-741.
- [5] Geweke, J. (1986), "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics*, 1, 127-141.

- [6] Geweke, J. (1991), "Efficient Simulation From the Multivariate Normal and Student t-Distributions Subject to Linear Constraints," in Computer Sciences and Statistics Proceedings of the 23d Symposium on the Interface, pp. 571-578.
- [7] Geweke, J. (1996), "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints," in *Modeling and Prediction: Honouring Seymour Geisser*, eds. W. O. Johnson, J. C. Lee, and A. Zellner, New York, Springer, pp. 248-263.
- [8] Gilks, W. R., and Roberts, G. O. (1996), "Strategies for Improving MCMC," in Markov Chain Monte Carlo in Practice, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman & Hall/CRC, pp. 89-114.
- Hocking, R. R. (1996), Methods and Applications of Linear Models: Regression and the Analysis of Variance, New York: Wiley.
- [10] Judge, G. C., and Takayama, T. (1966), "Inequality Restrictions In Regression Analysis," *Journal of the American Statistical Association*, 61, 166-181.
- [11] Liew, C. K. (1976), "Inequality Constrained Least-Squares Estimation," Journal of the American Statistical Association, 71, 746-751.
- [12] Lovell, M. C., and Prescott, E. (1970), "Multiple Regression with inequality constraints: Pretesting Bias, Hypothesis Testing, and Efficiency," *Journal of* the American Statistical Association, 65, 913-925.
- [13] Manolakis, D., and Shaw, G. (2002), "Detection Algorithms for Hyperspectral Imaging Applications," *IEEE Signal Processing Magazine*, 19, 29-43.
- [14] O'Hagan, A. (1994), Kendall's Advanced Theory of Statistics 2B: Bayesian Inference, New York: Oxford University Press Inc.
- [15] Pindyck, R. S., and Rubinfeld, D. L. (1981), Econometric Models and Economic Forecasts (2nd ed.), New York: McGraw-Hill.

- [16] Raftery, A. L., and Lewis, S. (1992), "How Many Iterations in the Gibbs Sampler?" in *Bayesian Statistics* 4, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press.
- [17] Roberts, G. O. (1996), "Markov Chain Concepts Related to Sampling Algorithms," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 45-57, London: Chapman & Hall/CRC, pp. 45-57.
- [18] Rodriguez-Yam, G. A., Davis, R. A., and Scharf, L. L. (2002), "A Bayesian Model and Gibbs Sampler for Hyperspectral Imaging," in *Proceedings of the* 2002 IEEE Sensor Array and Multichannel Signal Processing Workshop, Washington, D.C., pp. 105-109.
- [19] Tanner, M. A. (1996), Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions (3rd ed.), New York: Springer-Verlag Inc.
- [20] Telser, L. G. (1963), "Least Squares Estimates of Transition Probabilities," in Measurement in Economics: Studies in Mathematical Economics and Econometrics: In memory of Yehuda Grunfeld, eds. C. F. Christ, M. Friedman, L. A. Goodman, Z. Griliches, A. C. Harberger, N. Liviatan, J. Mincer, Y. Mundlak, M. Nerlove, D. Patinkin, L. G. Telser, and H. Theil, Stanford: Stanford University Press, pp. 270-292.