An Efficient Algorithm for Generating Truncated Multivariate Normal Distribution

Abstract. Sampling from truncated multivariate normal distribution (TMVND) is difficult but necessary in many statistical models. We first present an accept-reject algorithm for sampling from a truncated bivariate normal distribution and then propose a direct conditional sampling approach for a class of TMVND where the dispersion matrix possesses a random-effect covariance structure. For the general dispersion matrix, we propose a simple and efficient data augmentation (DA) algorithm to generate a TMVND by creating an incomplete-data structure. The posterior step of this DA algorithm directly generates the samples of whole vector instead of component by component, thus eliminating possible slow convergence in Markov chain Monte Carlo methods. We further show that multivariate normal distributions restricted to some commonly-encountered convex sets can be transformed into some TMVND via a simple linear mapping. The proposed method is applied to bivariate grouped model, multivariate probit models and two-way layout with ordered parameters and is further illustrated with a real example.

Keywords: Data augmentation; Gibbs sampler; MCMC; Truncated multivariate normal distribution.

1. Introduction

Sampling from a truncated multivariate normal distribution (TMVND) is an important problem occurring in many statistical models and related fields, e.g., the order restricted regression (Robertson, Wright and Dykstra, 1988), bivariate grouped models and two-way layout with ordered parameters (Gelfand, Smith and Lee, 1992), constrained multiple linear regression models (Chen and Deely, 1996), probit models for longitudinal binary data (Chib, 2000), proportional odds model for correlated ordinal data (Chen and Dey, 2000), longitu-
dinal models with informative censoring data (XXX), multivariate probit models (Chib and Greenberg, 1998; Liu, 2001), and autocorrelated ordered categorical data models (Girard and Parent, 2001). Statistical computations for the TMVND based on the numerical integration and analytic approximation are usually either infeasible or intractable. The sampling-based approaches have been the method of choice for computation in these models.

An $n$-dimensional random vector $\mu = (\mu_1, \ldots, \mu_n)^T$ is said to have a TMVND, denoted by $\mu \sim TN_n(x, \Sigma; a, b)$, if $\mu$ has a multivariate normal distribution truncated to the rectangle $[a, b] = [a_1, b_1] \times \cdots \times [a_n, b_n]$ and its density is given by

$$TN_n(\mu | x, \Sigma; a, b) = c^{-1} \exp \left\{ -\frac{(\mu - x)^T \Sigma^{-1}(\mu - x)}{2} \right\} \cdot I(a \leq \mu \leq b),$$  \hspace{1cm} (1.1)

where $c$ is the normalizing constant, $x = (x_1, \ldots, x_n)^T$ is the location parameter, $\Sigma$ is an $n \times n$ positive definite dispersion matrix, $I(\cdot)$ denotes the indicator function, $a = (a_1, \ldots, a_n)^T$ and $b = (b_1, \ldots, b_n)^T$ are known vectors with possible $-\infty$ or $+\infty$, and $a \leq x \leq b$ means that $a_i \leq x_i \leq b_i$ for all $i = 1, \ldots, n$. The primary objective is how to sample from (1.1) for given $(x, \Sigma)$ and $(a, b)$. For the univariate case, we have

$$TN(\mu_1 | x_1, \sigma_1^2; a_1, b_1) = c_1^{-1} \exp \left\{ -\frac{(\mu_1 - x_1)^2}{2\sigma_1^2} \right\} \cdot I(a_1 \leq \mu_1 \leq b_1),$$

where

$$c_1 = \sqrt{2\pi\sigma_1} \left\{ \Phi \left( \frac{b_1 - x_1}{\sigma_1} \right) - \Phi \left( \frac{a_1 - x_1}{\sigma_1} \right) \right\}$$ \hspace{1cm} (1.2)

and $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution. The inverse transformation method (Devroye, 1985, p.38) can be employed to generate random variable $\mu_1$ as $x_1 + \sigma_1 \Phi^{-1}(\Phi(\frac{a_1 - x_1}{\sigma_1}) + U \times \{\Phi(\frac{b_1 - x_1}{\sigma_1}) - \Phi(\frac{a_1 - x_1}{\sigma_1})\})$, where $U$ is a uniform $[0,1]$ random variable. This generation is easy to perform in S-PLUS since both the normal cdf $\Phi(\cdot)$ and its inverse $\Phi^{-1}(\cdot)$ are built-in functions. For the multivariate case, Robert (1995) proposed to use the Gibbs sampling to reduce the simulation problem to a sequence of one-dimensional simulations, i.e, to sample from the full conditional distributions $f(\mu_i | \mu_1, \ldots, \mu_{i-1}, \mu_{i+1}, \ldots, \mu_n)$, which is an easy-sampling truncated univariate normal distribution. Although this original Gibbs is easy to implement, it suffers from slow convergence owing to the high correlation between $\mu_i$ and $\mu_j$. Liu (1994) proposed a blocking
strategy for the Gibbs and shows that the blocked Gibbs is faster in convergence than the original Gibbs.

In Section 3, we propose a data augmentation (DA) algorithm (or two-block Gibbs sampler) for sampling from (1.1) via introducing latent data $Z$ such that both $\mu | Z$ and $Z | \mu$ are easy to simulate. By creating an incomplete-data structure, the posterior step (P-step) of this DA algorithm directly generates the samples of whole vector instead of component by component, thus eliminating possible slow convergence owing to the high correlations among $\mu_i$ in the original Gibbs.

The rest of the paper is organized as follows. Section 2 gives an accept-reject algorithm for sampling from a truncated bivariate normal distribution and proposes a direct conditional sampling approach for a class of TMVND where the dispersion matrix possesses a random-effect covariance structure. Section 4 shows that restricted multivariate normal distributions usually encountered in practice can be transformed into some TMVND via a linear mapping. Applications of the proposed method in bivariate grouped model, multivariate probit models and two-way layout with ordered parameters are presented in Section 5. The method is illustrated with the New Zealand apple data in Section 6 and a discussion is given finally.

2. Accept-Reject Sampling and Conditional Sampling

In this section, we give a straightforward accept-reject sampling to obtain iid samples from a truncated bivariate normal distribution. However, this kind of accept-reject sampling can not be generalized to TMVND. Therefore, we propose a direct conditional sampling approach to obtain iid samples for a class of TMVND where the dispersion matrix $\Sigma$ possesses a random-effect covariance structure.

2.1 Truncated bivariate normal distribution

Let $\mu = (\mu_1, \mu_2)^T \sim TN_2(x, \Sigma; a, b)$, where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}, \quad a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

Then the conditional density of $\mu_2$ given $\mu_1$ is a truncated univariate normal, i.e.,

$$\mu_2|\mu_1 \sim TN(x_2 + \rho \sigma_2 \sigma_1^{-1}(\mu_1 - x_1), \sigma_2^2(1 - \rho^2); a_2, b_2),$$
and \( \mu_1 \) is independent of \( \mu_2 \) if and only if \( \rho = 0 \). The marginal density of \( \mu_1 \) is given by

\[
f_1(\mu_1) = c^{-1}_1 \Psi(\mu_1) \exp \left\{ -\frac{(\mu_1 - x_1)^2}{2\sigma_1^2} \right\} \cdot I(a_1 \leq \mu_1 \leq b_1),
\]

(2.1)

where

\[
\Psi(\mu_1) = \Phi \left( \frac{b_2 - x_2 - \rho \sigma_2 \sigma_1^{-1}(\mu_1 - x_1)}{\sigma_2 \sqrt{1 - \rho^2}} \right) - \Phi \left( \frac{a_2 - x_2 - \rho \sigma_2 \sigma_1^{-1}(\mu_1 - x_1)}{\sigma_2 \sqrt{1 - \rho^2}} \right),
\]

\[c_* = \sqrt{2\pi} \sigma_1 \{ \Phi_2(b_1, b_2; x, \Sigma) - \Phi_2(b_1, a_2; x, \Sigma) - \Phi_2(a_1, b_2; x, \Sigma) + \Phi_2(a_1, a_2; x, \Sigma) \},\]

(2.2)

is the normalizing constant, \( \Phi_2(\cdot, \cdot; x, \Sigma) \) denotes the cdf of \( N_2(x, \Sigma) \) and it is a built-in function in S-PLUS.

To generate a sample from the truncated bivariate normal distribution, we only need to consider the sampling from the marginal density \( f_1(\mu_1) \) given by (2.1). Consider the accept-reject sampling with envelope function \( TN(\mu_1|x_1, \sigma_1^2; a_1, b_1) \), we have

\[
f_1(\mu_1) \leq \{ c_1 c_*^{-1} \Psi(\mu_{\text{max}}) \} \cdot TN(\mu_1|x_1, \sigma_1^2; a_1, b_1),
\]

where \( c_1 \) and \( c_* \) are given respectively by (1.2) and (2.2), and

\[
\mu_{\text{max}} = \hat{\mu}_1 \cdot I(a_1 \leq \hat{\mu}_1 \leq b_1) + a_1 \cdot I(\hat{\mu}_1 < a_1) + b_1 \cdot I(\hat{\mu}_1 > b_1),
\]

\[
\hat{\mu}_1 = x_1 + (a_2 + b_2 - 2x_2)/(2\rho \sigma_2 \sigma_1^{-1}).
\]

Remark 1. In general, the marginal density (2.1) is not truncated univariate normal. Gelfand, Smith and Lee (1992, p.529) stated that the marginal of a truncated bivariate normal distribution is still a truncated univariate normal, which is thus not correct. However, for the special case of \( a_2 = -\infty \), \( b_2 = +\infty \) or \( \rho = 0 \), the marginal density of \( \mu_1 \) becomes \( \mu_1 \sim TN(x_1, \sigma_1^2; a_1, b_1) \).

2.2 TMVND with dispersion matrix of random-effect covariance structure

Let \( \mu = (\mu_1, \ldots, \mu_n)^\top \sim TN_n(x, \Sigma; a, b) \), where all parameters \( (x, \Sigma) \) and \( (a,b) \) are known. When the dispersion matrix \( \Sigma \) is diagonal, the simulation of \( \mu \) is straightforward since each component is a truncated univariate normal variable and they are independent. When \( \Sigma \)}
is nondiagonal, the direct simulation of $\mu$ is infeasible except for some special case. For example, in probit models for longitudinal binary data (Chib, 2000), the dispersion matrix $\Sigma$ takes the form of random-effects covariance structure, i.e., $\Sigma = WDW^\top + \sigma^2 I_n$, where $W$ is the $n \times q$ covariate matrix associated with the random-effect and $D$ is a $q \times q$ positive definite symmetric matrix. In this case, we have the following result.

**Proposition 1.** Let $\xi \sim N_q(0, D)$ and $\mu|\xi \sim TN_n(x + W\xi, \sigma^2 I_n; a, b)$, then

$$
\mu \sim TN_n(x, WDW^\top + \sigma^2 I_n; a, b).
$$

This fact can be easily verified by integrated out $\xi$ from the joint density of $(\mu, \xi)$ to obtain the marginal density of $\mu$, where we used the following matrix identity

$$(WDW^\top + \sigma^2 I_n)^{-1} = \sigma^{-2}I_n - \sigma^{-2}W(\sigma^{-2}W^\top W + D^{-1})^{-1}W^\top \sigma^{-2}.$$

Based on Proposition 1, we can directly obtain iid samples of $\mu$ via first generating $\xi$ from $N_q(0, D)$ and then generating $\mu|\xi$ from $TN_n(x + W\xi, \sigma^2 I_n; a, b)$.

### 3. Data Augmentation Algorithm

For the general dispersion matrix, direct simulation of $\mu$ is impossible. Our idea is first to create a Bayesian incomplete-data structure where the observed data $Y_{obs}$ is augmented by latent data $Z$ such that both the complete-data posterior $f(\mu|Y_{obs}, Z)$ and the conditional predictive distribution $f(Z|Y_{obs}, \mu)$ are easy to simulate. Then we carefully chose a prior distribution such that the observed posterior distribution $f(\mu|Y_{obs})$ is exactly $TN_n(\mu|x, \Sigma; a, b)$ given by (1.1). Therefore, the data augmentation algorithm or the two-block Gibbs sampling can be applied.

The well-known Cholesky decomposition shows that there exists a unique upper triangular matrix $B_{n \times n} = (b_{ik})$ with positive diagonal elements such that $\Sigma^{-1} = B^\top B$. We create the observed data $Y_{obs} = \{y_1, \ldots, y_n\}$, where $y = (y_1, \ldots, y_n)^\top = Bx \sim N_n(B\mu, I_n)$ subject to restriction $\mu \in [a, b]$, where $x$ is treated as random vector and $\mu$ is treated as parameter. The observed-data likelihood function for $\mu$ is given by

$$
L(\mu|Y_{obs}) \propto \exp \left\{-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right\} \cdot I_{(a \leq \mu \leq b)}.
$$
Consider the truncated univariate normal density \( TN(\mu_k|\mu_{k0}, \sigma_{k0}^2; a_k, b_k) \) with known location parameter \( \mu_{k0} \) and known dispersion parameter \( \sigma_{k0}^2 \) as a candidate for prior of \( \mu_k \). If we chose the independent diffuse priors
\[
\pi(\mu) = \prod_{k=1}^{n} TN(\mu_k|\mu_{k0}, \sigma_{k0}^2; a_k, b_k), \quad \text{where} \quad \sigma_{k0}^2 \to +\infty \quad \text{for} \quad k = 1, \ldots, n, \quad (3.2)
\]
then the observed posterior \( f(\mu|Y_{\text{obs}}) = TN_n(\mu|x, \Sigma; a, b) \) is the distribution we want to simulate from.

Similar to the EM structure of Liu (2000) for discrete distributions, we augment the observed data \( Y_{\text{obs}} \) by \( n \times (n - 1) \) latent data \( Z = \{Z_{ik} : i = 1, \ldots, n, k = 1, \ldots, n - 1 \} \) to obtain a complete-data \( Y_{\text{com}} = \{Z_{ik} : i = 1, \ldots, n, k = 1, \ldots, n \} \), where
\[
Z_{ik} \overset{\text{ind}}{\sim} N(b_{ik}\mu_k, \frac{1}{n}), \quad i = 1, \ldots, n, \quad k = 1, \ldots, n, \quad (3.3)
\]
\[
y_i = \sum_{k=1}^{n} Z_{ik}, \quad i = 1, \ldots, n. \quad (3.4)
\]
The mapping (3.4) from \( Y_{\text{com}} \) to \( Y_{\text{obs}} \) preserves the observed-data likelihood function (3.1).

Using the diffuse priors (3.2), the complete-data posterior distribution of \( \mu \) is a product of \( n \) independent truncated univariate normal distributions, that is,
\[
f(\mu|Y_{\text{obs}}, Z) = \prod_{k=1}^{n} TN(\mu_k|u_k, v_k^2; a_k, b_k), \quad (3.5)
\]
where \( u_k = \sum_{i=1}^{n} b_{ik} Z_{ik} / \sum_{i=1}^{n} b_{ik}^2 \) and \( v_k^2 = \{n \sum_{i=1}^{n} b_{ik}^2\}^{-1} \) for \( k = 1, \ldots, n \). To derive the conditional predictive distribution, we first need to prove the following result.

**PROPOSITION 2.** Let random variables \( W_k \overset{\text{ind}}{\sim} N(\beta_k, \delta_k^2), k = 1, \ldots, n \), then
\[
(W_1, \ldots, W_{n-1})^\top \left| \left( \sum_{k=1}^{n} W_k = w \right) \right. \sim N_{n-1}(\beta_{n-1} + \frac{w - \sum_{k=1}^{n} \beta_k}{\sum_{k=1}^{n} \delta_k^2}, \delta_{n-1}, \text{diag}(\delta_{n-1}^2 - \frac{\sum_{k=1}^{n} \delta_k^2}{\sum_{k=1}^{n} \delta_k^2})), \quad (3.6)
\]
where \( \beta_{n-1} = (\beta_1, \ldots, \beta_{n-1})^\top \) and \( \delta_{n-1}^2 = (\delta_1^2, \ldots, \delta_{n-1}^2)^\top \). Especially, if \( \delta_1^2 = \cdots = \delta_n^2 = \delta_*^2 \), then
\[
(W_1, \ldots, W_{n-1})^\top \left| \left( \sum_{k=1}^{n} W_k = w \right) \right. \sim N_{n-1}(\beta_{n-1} + \frac{w - \sum_{k=1}^{n} \beta_k}{n}, \delta_*^2 \left( I_{n-1} - \frac{1}{n} 1_{n-1} 1_{n-1}^\top \right)), \quad (3.7)
\]

**PROOF.** Let \( \beta = (\beta_1, \ldots, \beta_n)^\top \) and \( \delta^2 = (\delta_1^2, \ldots, \delta_n^2)^\top \). We note that \( W = (W_1, \ldots, W_n)^\top \overset{\text{ind}}{\sim} N_n(\beta, \text{diag}(\delta^2)) \), then
\[
\left( \frac{W}{\sum_{k=1}^{n} W_k} \right) = \left( \frac{1_n}{I_n^\top} \right) W \sim N_{n+1}\left( \left( \frac{\beta}{1_n^\top / n} \right), \left( \frac{\text{diag}(\delta^2)}{\delta_*^2} \frac{\delta_*^2}{1_n^\top} \right) \right).
\]
From the property of multivariate normal distributions, we have

\[(W_1, \ldots, W_n) \mid \left( \sum_{k=1}^{n} W_k = w \right) \sim N_n \left( \beta + \frac{w - \sum_{k=1}^{n} \beta_k}{I_n \delta^2}, \delta^2 \right), \]  

(3.8)

Let \( \Omega = \text{diag}(\delta^2) - \delta^2 \delta^2/\left(1_n^\top \delta^2 \right) \), then \( \Omega = D^{1/2} I_n - \Omega_1 ) D^{1/2} \), where \( D = \text{diag}(\delta) \) and \( \Omega_1 = D^{1/2} I_n (I_n - \Omega_1)^{-1} I_n D^{1/2} \). Since \( I_n - \Omega_1 \) is a projection matrix and the rank of a projection matrix is equal to its trace, then rank \((I_n - \Omega_1) = n - 1\), which means rank \((\Omega) = n - 1 < n\). That is, the distribution in (3.8) is a degenerate \(n\)-dimensional normal distribution. From (3.8), we immediately obtain (3.6) and (3.7).

Define \( Z_i = (Z_{i1}, \ldots, Z_{i,n-1})^\top \). From (3.3) and (3.7), the conditional predictive distribution is given by

\[ f(Z \mid Y_{\text{obs}}, \mu) = \prod_{i=1}^{n} N_{n-1} \left( Z_i \mid E(Z_i \mid Y_{\text{obs}}, \mu), \frac{1}{n} \left( I_{n-1} - \frac{1}{n} 1_{n-1}^\top 1_{n-1} \right) \right), \]  

(3.9)

where, for \( i = 1, \ldots, n \),

\[ E(Z_i \mid Y_{\text{obs}}, \mu) = (b_{i1} \mu_1, \ldots, b_{i,n-1} \mu_{n-1})^\top + \frac{1}{n} \left( y_i - \sum_{k=1}^{n} b_{ik} \mu_k \right). \]

Hence, the DA algorithm can be applied to simulate from (3.9) and (3.5). Let \( \{(Z^{(\ell)}, \mu^{(\ell)}) : \ell = 1, \ldots, m\} \) be the DA output, then \( \{\mu^{(\ell)} : \ell = 1, \ldots, m\} \) is a sample of size \( m \) from (1.1). We summarize the DA algorithm as follows.

**DA Algorithm:**

**Input.** A location vector \( x = (x_1, \ldots, x_n)^\top \), a positive definite dispersion matrix \( \Sigma_{n \times n} \), and two bound vectors \( a = (a_1, \ldots, a_n)^\top \) and \( b = (b_1, \ldots, b_n)^\top \).

**Output.** A sample of size \( m \) from \( T N_n(x, \Sigma; a, b) \).

**Initial.** Construct an upper triangular matrix \( B_{n \times n} = (b_{ik}) \) with positive diagonal elements via Cholesky decomposition such that \( \Sigma^{-1} = B^\top B \), and set \( y = (y_1, \ldots, y_n)^\top = Bx \).

**I-Step.** Impute missing values \( Z = \{Z_{ik} : i = 1, \ldots, n, k = 1, \ldots, n-1\} \) from (3.9);

**P-Step.** Draw \( \mu \) from (3.5).
4. Restricted Multivariate Normal Distributions

In practice, many multivariate normal distributions are restricted to some convex sets other than a rectangle \([a, b]\). In this section, we show that such restricted multivariate normal distributions can be transformed into some TMVND via a simple linear mapping for most situations. Therefore, the methods proposed in previous sections can be used to generate such restricted multivariate normal distributions.

An \(n\)-dimensional random vector \(\theta = (\theta_1, \ldots, \theta_n)^\top\) is said to have a multivariate normal distribution with location parameter \(y\) and dispersion matrix \(V\) and restricted to some convex set \(S_\theta \subset \mathbb{R}^n\), denoted by \(\theta \sim RN_n(y, V; S_\theta)\), if its density is given by

\[
f(\theta) \propto \exp\left\{ -\frac{(\theta - y)^\top V^{-1}(\theta - y)}{2} \right\} \cdot I(\theta \in S_\theta).
\]

Let \(\theta \sim RN_n(y, V; S_\theta)\) and \(A_{n \times n}\) be a non-singular matrix, then

\[
\mu = A^{-1} \theta \sim RN_n(A^{-1}y, A^{-1}VA^{-1\top}; S_\mu),
\]

where \(S_\mu = \{\mu : A\mu \in S_\theta\}\) denotes the support of \(\mu\).

Robertson, Wright and Dykstra (1988) discussed the following convex sets in the setting of isotonic regressions. They are

- monotonic ordering: \(S_\theta = \{\theta : \theta_1 \leq \cdots \leq \theta_n\}\),
- umbrella ordering: \(S_\theta = \{\theta : \theta_1 \leq \cdots \leq \theta_h \geq \cdots \geq \theta_n\}, \ 1 \leq h \leq n\),
- tree ordering: \(S_\theta = \{\theta : \theta_i \leq \theta_n, \ i = 1, \ldots, n-1\}\),
- increasing convex ordering: \(S_\theta = \{\theta : 0 \leq \theta_2 - \theta_1 \leq \theta_3 - \theta_2 \leq \cdots \leq \theta_n - \theta_{n-1}\}\),
- increasing concave ordering: \(S_\theta = \{\theta : \theta_2 - \theta_1 \geq \theta_3 - \theta_2 \geq \cdots \geq \theta_n - \theta_{n-1} \geq 0\}\).

Making linear transformations \(\theta = A_s \mu\) on above five orderings corresponding to \(s = 1, \ldots, 5\), we have \(\mu = A_s^{-1} \theta\), where \(\mu \in S_\mu = \{(\mu_1, \ldots, \mu_n)^\top : -\infty < \mu_1 < +\infty, \mu_i \geq 0, i = 2, \ldots, n\}\), and \(A_1, \ldots, A_5\) are \(n \times n\) matrix given by

\[
A_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \equiv \Delta_n, \quad A_2 = \begin{pmatrix} \Delta_h \\ 1_{n-h} \end{pmatrix} \Gamma_h \equiv \Gamma_{h,n},
\]

\[
A_3 = \begin{pmatrix} 1_{n-1} & -I_{n-1} \\ 1 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & 0^\top_{n-1} \\ 1_{n-1} \end{pmatrix} \equiv \Lambda_{n-1}, \quad A_5 = \begin{pmatrix} 1_{n-1} & -\Lambda_{n-1}^\top \end{pmatrix}.
\]
and $\Lambda_{n-1}$ is defined by

$$
\Lambda_{n-1} = \begin{pmatrix}
1 & 0 & \cdots & 0 & 0 \\
2 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
n-2 & n-3 & \cdots & 1 & 0 \\
n-1 & n-2 & \cdots & 2 & 1
\end{pmatrix}.
$$

Note that the support of $\mu$ is $S_\mu = [a, b]$, where $a = (-\infty, 0, \ldots, 0)^\top$ and $b = (+\infty, \ldots, +\infty)^\top$.

If $\theta \sim RN_n(y, V; S_\theta)$, from (4.2), then we have $\mu \sim TN_n(A_s^{-1}y, A_s^{-1}VA_s^{-1}\top; a, b)$.

5. Applications

5.1 Bivariate grouped model

Gelfand, Smith and Lee (1992) consider a bivariate grouped model. Let $n_{ij}$ denote the observed number of individuals in height group $[a_i-1, a_i]$ and in weight group $[b_j-1, b_j]$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, where $\sum_{i=1}^m \sum_{j=1}^n n_{ij} = N$, $a_0 = b_0 = 0$ and $a_m = b_n = +\infty$.

Suppose that $n_{ij}$ follows a multinomial distribution with probability $p_{ij} = \Pr\{a_{i-1} \leq \xi \leq a_i, \ b_{j-1} \leq \eta \leq b_j\}$, denoted by $n_{ij} \sim \text{Multi}(N; p_{11}, \ldots, p_{mn})$, where $\sum_{i=1}^m \sum_{j=1}^n p_{ij} = 1$.

Further it is reasonable to assume that $(\xi, \eta)^\top \sim N_2(\theta, \Sigma)$. The objective is to estimate the parameters $\theta$ and $\Sigma$ based on the observed data $Y_{\text{obs}} = \{n_{ij} : i = 1, \ldots, m, \ j = 1, \ldots, n\}$.

Augment the observed data $Y_{\text{obs}}$ with the unobservable data $Z = \{Z_{ijk} : i = 1, \ldots, m, \ j = 1, \ldots, n, \ k = 1, \ldots, n_{ij}\}$, where $Z_{ijk} = (\xi_{ijk}, \eta_{ijk})^\top \overset{\text{iid}}{\sim} N_2(\theta, \Sigma)$ for all $i, j, k$. If a normal-inverse Wishart prior is used, then the joint posterior distribution of $(\theta, \Sigma)$ for the complete-data $\{Y_{\text{obs}}, Z\}$ is given by

$$
f(\theta, \Sigma|Y_{\text{obs}}, Z) = f(\theta|Y_{\text{obs}}, Z, \Sigma) \times f(\Sigma|Y_{\text{obs}}, Z),
$$

(5.1)

where $\theta|Y_{\text{obs}}, Z, \Sigma$ is an updated bivariate normal and $\Sigma|Y_{\text{obs}}, Z$ is an updated inverse Wishart. For each fixed pair $(i, j)$, $Z_{ijk}|(n_{ij}, \theta, \Sigma)$ are independent bivariate normal random vectors truncated to the region $[a_{i-1}, a_i] \times [b_{j-1}, b_j]$ for $k = 1, \ldots, n_{ij}$. Then the conditional predictive distribution is given by

$$
f(Z|Y_{\text{obs}}, \theta, \Sigma) = \prod_{i=1}^m \prod_{j=1}^n \prod_{k=1}^{n_{ij}} TN_2(Z_{ijk}|\theta, \Sigma; (a_{i-1}, b_{j-1})^\top, (a_i, b_j)^\top).
$$

(5.2)
Therefore, the accept-reject approach proposed in Section 2.1 is able to be used to sample from (5.2) so that the DA algorithm can be applied to (5.1) and (5.2).

5.2 Multivariate probit models

Let $Y_{ij}$ denote the binary response 0 or 1 of the $j$-th variable at the $i$-th observation, and $Y_i = (Y_{i1}, \ldots, Y_{in})^T$ be the collection of responses of all $n$ variables at the $i$-th observation, where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The multivariate probit model (Chib and Greenberg, 1998) states that the conditional probability is given by

$$\Pr\{Y_i = y_i|\beta, R\} = \int_{G_{i1}} \cdots \int_{G_{in}} N_n(t|0, R) dt,$$

where $G_{ij}$ is the interval $(-\infty, x_{ij}^T \beta_j)$ if $y_{ij} = 1$ and the interval $[x_{ij}^T \beta_j, +\infty)$ if $y_{ij} = 0$, $N_n(t|0, R)$ denotes the density of an $n$-dimensional normal distribution with mean vector 0 and covariance matrix $R$, which must take the form of correlation matrix for identifiability reasons, $x_{ij} \in \mathbb{R}^{k_j}$ is known covariates, $\beta_j \in \mathbb{R}^{k_j}$ is unknown parameters and $\beta^T = (\beta_1^T, \ldots, \beta_n^T) \in \mathbb{R}^k$, $k = \sum_{j=1}^{n} k_j$.

Let $Y_{obs} = \{Y_1, \ldots, Y_m\}$ denote the observed data and $X_i = \text{diag}(x_{i1}^T, \ldots, x_{in}^T)$ be an $n \times k$ covariate matrix. Augment the observed data $Y_{obs}$ with Gaussian latent variables $Z = \{Z_i = (Z_{i1}, \ldots, Z_{in})^T : i = 1, \ldots, m\}$ via defining $Y_{ij} = I_{(Z_{ij} > 0)}$ to form the complete-data $Y_{\text{com}} = \{Y_{obs}, Z\}$, where $Z_i|(\beta, R) \sim N_n(X_i \beta, R)$. Thus (5.3) is rewritten as

$$\Pr\{Y_i = y_i|\beta, R\} = \int_{H_{i1}} \cdots \int_{H_{in}} N_n(Z_i|X_i \beta, R) dZ_i$$

where $H_{ij}$ is the interval $(0, \infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$, and let $H_i = H_{i1} \times \cdots \times H_{in}$. Liu (2001) suggests taking independent priors $\beta \sim N_k(\beta_0, V_0^{-1})$ and $\pi(R) \propto |R|^{-(n+1)/2}$. The complete-data joint posterior of $(\beta, R)$ is then given by

$$f(\beta, R|Y_{\text{obs}}, Z) = f(\beta|Y_{\text{obs}}, Z, R) \times f(R|Y_{\text{obs}}, Z),$$

$$= N_k(\hat{\beta}|\hat{V}^{-1}) \times f(R|Y_{\text{obs}}, Z),$$

where $\hat{\beta} = \hat{V}^{-1}(V_0 \beta_0 + \sum_{i=1}^{m} X_i^T R^{-1} Z_i)$ and $\hat{V} = V_0 + \sum_{i=1}^{m} X_i^T R^{-1} X_i$. Liu (2001) gives a simple algorithm to draw $R$ from $f(R|Y_{\text{obs}}, Z)$. The conditional predictive distribution of
$Z_i$ given $(Y_i, \beta, R)$ is a multivariate normal density truncated to the region specified by $H_i$, therefore we have

$$f(Z|Y_{obs}, \beta, R) = \prod_{i=1}^{m} f(Z_i|Y_i, \beta, R) = \prod_{i=1}^{m} TN_n(Z_i|X_i\beta, R; H_i). \quad (5.6)$$

Noting that $H_i$ is a rectangle, the DA algorithm proposed in Section 3 can be used to simulate from (5.6) so that another DA algorithm can be applied to (5.5) and (5.6).

### 5.3 Two-way layout with ordered parameters

Gelfand, Smith and Lee (1992) consider a model of the form

$$Y_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim iid N(0, \sigma^2), \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, \quad (5.7)$$

where the prior knowledge about the parameters constrains the $\alpha_i$ to be decreasing in $i$ and the $\beta_j$ to be increasing up to some level $h$ and then decreasing, i.e., $\alpha_1 \geq \cdots \geq \alpha_m$ and $\beta_1 \leq \cdots \leq \beta_h \geq \beta_{h+1} \geq \cdots \geq \beta_n$, where $h \in \{1, 2, \ldots, n\}$ is known or unknown integer.

The objective is to obtain the estimates of $\alpha = (\alpha_1, \ldots, \alpha_m)^\top$, $\beta = (\beta_1, \ldots, \beta_n)^\top$ and $\sigma^2$.

Such a model finds application in consumer preference studies (Green and Srinivasan, 1978): here $Y_{ij}$ might be a scoring or rating of a product, such as a candy bar, with factor $\alpha_i$ corresponding to price level and $\beta_j$ to sugar content.

According to the discussion in Section 4, the decreasing ordering on $\alpha$ and the umbrella ordering on $\beta$ can be represented by linear transformations $\alpha = \Delta_m^\top \mu^{(1)}$ and $\beta = \Gamma_{h,n} \mu^{(2)}$, respectively, where $\Delta_m$ and $\Gamma_{h,n}$ are defined in (4.3), $\mu^{(1)} = (\mu_1, \ldots, \mu_m)^\top \in \mathbb{R}_+^{m-1} \times \mathbb{R}$, and $\mu^{(2)} = (\mu_{m+1}, \ldots, \mu_{m+n})^\top \in \mathbb{R} \times \mathbb{R}_+^{n-1}$. Therefore, model (5.7) can be represented as

$$\vec{Y} = X\mu + \vec{E}, \quad \vec{E} \sim N_{mn}(0, \sigma^2 I_{mn}), \quad (5.8)$$

subject to $\mu = (\mu_1, \ldots, \mu_{m+n})^\top \in \mathcal{S}_\mu = \mathbb{R}_+^{m-1} \times \mathbb{R}^2 \times \mathbb{R}_+^{n-1}$, where $\vec{Y}$ denotes the “vec” operator of $Y_{m \times n} = (Y_{ij}) = (Y_1, \ldots, Y_n)$, $X_{mn \times (m+n)} = (I_n \otimes \Delta_m^\top : [I_n \otimes 1_m] \Gamma_{h,n})$, and $(A : B)$ denotes the column-merged matrix of $A$ and $B$.

We consider independent priors on $\mu$ and $\sigma^2$: $\pi(\mu)$ is a flat prior on the support $\mathcal{S}_\mu$ and $\sigma^2 \sim IG(\frac{q_0}{2}, \lambda_0)$ with inverse gamma density

$$IG\left(u | \frac{q_0}{2}, \frac{\lambda_0}{2}\right) = \frac{(\lambda_0/2)^{q_0/2}}{\Gamma(q_0/2)} u^{-1-q_0/2} \exp\left\{-\frac{\lambda_0}{2u}\right\}.$$
where $q_0$ and $\lambda_0$ are known constants. From (5.8), we have

$$\sigma^2 | (Y_{\text{obs}}, \mu) \sim \text{IG}\left(\frac{q_0 + mn}{2}, \frac{\lambda_0 + (\bar{Y} - X\mu)^\top(\bar{Y} - X\mu)}{2}\right),$$

(5.9)

$$\mu | (Y_{\text{obs}}, \sigma^2) \sim \text{TN}_{mn}\left((X^\top X)^{-1}X^\top \bar{Y}, \sigma^2(X^\top X)^{-1}; S_{\mu}\right).$$

(5.10)

Noting that $S_{\mu} = IR_{m-1}^n \times IR_{n-1}^n$ is a rectangle, the DA algorithm proposed in Section 3 can be used to simulate from (5.10) so that a two-block Gibbs sampler can be applied to (5.9) and (5.10). Once a sample of $\mu$ is obtained, a sample of $(\alpha, \beta)$ can be obtained by using linear mappings $\alpha = \Delta_m^\top\mu^{(1)}$ and $\beta = \Gamma_{h,n}\mu^{(2)}$.

6. An Illustrated Example

Chen and Deely (1996) use a constrained linear regression to model the yield of New Zealand apples (in cartons) for $m = 207$ producers. The covariate is the number of trees at ages 1, 2, ..., 10, where age “10” tree means ten or older and is considered to be a mature tree. Let $Y_i$ denote the apple yield of producer $i$ ($i = 1, \ldots, m$) and $x_{ij}$ is the number of trees at age $j$ ($j = 1, \ldots, n$ and $n = 10$) for producer $i$. The model can be written as

$$Y_i = \sum_{j=1}^{n} x_{ij}\theta_j + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, m,$$

(6.1)

where $\theta_j$ represents the average over all trees of age $i$. Since older trees have an higher yield or producers do not usually allow poor trees to persist on average, it is reasonable to constrain the regression coefficients by a monotonic ordering $0 \leq \theta_1 \leq \cdots \leq \theta_n$. The objective is to estimate $\theta = (\theta_1, \ldots, \theta_n)^\top$ and $\sigma^2$.

According to the discussion in Section 4, the monotonic ordering on $\theta$ can be represented by a linear transformation $\theta = \Delta_n\mu$, where $\Delta_n$ is defined in (4.3) and $\mu \in IR_{n}^n$. Therefore, the constrained linear regression model (6.1) can be rewritten as

$$Y = W\mu + \varepsilon, \quad \varepsilon \sim N_m(0, \sigma^2I_m),$$

(6.2)

subject to restriction $\mu = (\mu_1, \ldots, \mu_n)^\top \in IR_{n}^n$, where $Y_{m\times 1} = (Y_1, \ldots, Y_m)^\top$, $W_{m\times n} = X\Delta_n$ and $X_{m\times n} = (x_{ij})$. We consider independent priors on $\mu$ and $\sigma^2$: $\pi(\mu)$ is a flat prior on the
support \( \mathcal{R}_+^a \) and \( \sigma^2 \sim \text{IG}(\frac{q_0}{2}, \frac{\lambda_0}{2}) \) with \( q_0 = \lambda_0 = 0 \). From (6.2), we have

\[
\begin{align*}
\sigma^2 | (Y_{\text{obs}}, \mu) & \sim \text{IG} \left( \frac{m}{2}, \frac{(Y - W\mu)^\top(Y - W\mu)}{2} \right), \\
\mu | (Y_{\text{obs}}, \sigma^2) & \sim \text{TN}_n \left( \frac{(W^\top W)^{-1} W^\top Y}{\sigma^2 (W^\top W)^{-1}}, \mathcal{R}_+^a \right). 
\end{align*}
\] (6.3) 

(6.4)

Noting that \( \mathcal{R}_+^a \) is a rectangle, the DA algorithm proposed in Section 3 can be used to simulate from (6.4) so that a two-block Gibbs sampler can be applied to (6.3) and (6.4). Once a sample of \( \mu \) is obtained, a sample of \( \theta \) can be obtained by \( \theta = \Delta_n \mu \).

\[ \text{Table 1. Bayesian estimates of } \theta \text{ and } \sigma^2 \]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Posterior mean</th>
<th>Posterior sd</th>
<th>95% Posterior interval estimates</th>
<th>PSR at 500 iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta_1 )</td>
<td>0.0876</td>
<td>0.0383</td>
<td>[0.0176, 0.1624]</td>
<td>1.021</td>
</tr>
<tr>
<td>( \theta_2 )</td>
<td>0.1382</td>
<td>0.0400</td>
<td>[0.0622, 0.2190]</td>
<td>1.041</td>
</tr>
<tr>
<td>( \theta_3 )</td>
<td>0.1912</td>
<td>0.0452</td>
<td>[0.1033, 0.2836]</td>
<td>1.126</td>
</tr>
<tr>
<td>( \theta_4 )</td>
<td>0.2642</td>
<td>0.0528</td>
<td>[0.1651, 0.3726]</td>
<td>1.024</td>
</tr>
<tr>
<td>( \theta_5 )</td>
<td>0.4212</td>
<td>0.1023</td>
<td>[0.2498, 0.6443]</td>
<td>1.051</td>
</tr>
<tr>
<td>( \theta_6 )</td>
<td>0.6050</td>
<td>0.1719</td>
<td>[0.3328, 0.9927]</td>
<td>1.054</td>
</tr>
<tr>
<td>( \theta_7 )</td>
<td>0.7702</td>
<td>0.2035</td>
<td>[0.4362, 1.2288]</td>
<td>1.038</td>
</tr>
<tr>
<td>( \theta_8 )</td>
<td>0.8790</td>
<td>0.2161</td>
<td>[0.5108, 1.3460]</td>
<td>1.163</td>
</tr>
<tr>
<td>( \theta_9 )</td>
<td>0.9955</td>
<td>0.2305</td>
<td>[0.5983, 1.4997]</td>
<td>1.054</td>
</tr>
<tr>
<td>( \theta_{10} )</td>
<td>1.4686</td>
<td>0.4276</td>
<td>[0.8030, 2.4303]</td>
<td>1.071</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>15551.7</td>
<td>4156.326</td>
<td>[9348.598, 25498.853]</td>
<td>1.096</td>
</tr>
</tbody>
</table>

Table 1.1 of Chen, Shao and Ibrahim (2000) gives the first 40 observations. Using this subset of New Zealand apple data, we implement the two-block Gibbs sampler based on (6.3) and (6.4). The criterion of potential scale reduction (PSR) proposed by Gelman and Rubin (1992) is utilized to check the convergence of the Gibbs sampler. PSR values close to 1 are indicative of convergence of the Markov chain to the target distribution. We run 8 multiple chains with length 3000 by using dispersed initial values. The PSR’s at 500 iterations are listed in Table 1. The 8 \( \times \) 1500 samples from the second half of each sequence is collected and the corresponding Bayesian estimates of \( \theta \) and \( \sigma^2 \) are given in Table 1. The marginal posterior densities of \( \theta_1, \theta_2, \theta_5, \) and \( \theta_{10} \) are plotted in Figure 1. Figure 2 gives a comparison of Bayesian estimates of \( \theta \) computed in this paper and those given in Chen and Deely (1996). There is a slight difference between our results and Chen and Deely (1996)’s, partial reason
may be that we only use 40 observations and a different prior while Chen and Deely (1996) used all 207 observations.

![Marginal posterior densities of various parameters](image1.png)

**Figure 1.** The marginal posterior densities of $\theta_1$, $\theta_2$, $\theta_5$, and $\theta_{10}$.

![Comparison of Bayesian estimates](image2.png)

**Figure 2.** Comparison of Bayesian estimates of $\theta$ computed in this paper and those given in Chen and Deely (1996).
7. Discussion

By creating a DA structure, we develop a simple and efficient algorithm to generate samples from TMVND. The posterior step of this DA algorithm directly generates samples of the whole vector instead of component by component, thus eliminating possible slow convergence owing to the high correlations among components in the original Gibbs. For two special cases, we provide an accept-reject algorithm to obtain iid samples from a truncated bivariate normal distribution and propose a direct conditional sampling approach to generate iid samples from a class of TMVND where the dispersion matrix possesses a random-effect covariance structure. We further show that multivariate normal distributions restricted to some commonly-encountered convex sets can be transformed into some TMVND via a simple linear mapping.

Noting that in Section 3 the Cholesky decomposition of \( \Sigma^{-1} = B^\top B \) yields an upper triangular matrix \( B_{n \times n} = (b_{ik}) \). The advantage of the Cholesky decomposition of \( B^\top B \) over the general decomposition \( (\Sigma^{-1/2})^\top \Sigma^{-1/2} \) is that the former can facilitate the computation in the DA algorithm based on (3.5) and (3.9) since there are \( n(n - 1)/2 \) zero elements in \( B \), i.e., \( b_{ik} = 0 \) for \( k < i \) and \( i = 2, \ldots, n \). This Cholesky decomposition reduces the number of latent variables introduced in (3.3) from \( n(n - 1) \) to \( n(n - 1)/2 \), which further accelerates the DA algorithm.

Furthermore, it would be of interest to extend the method to cases where the restricted multivariate normal distributions can not be transformed into a TMVND.

References


