

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE  
CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods.

**IMPUTATION OF DATA SUBJECT TO BALANCE AND INEQUALITY RESTRICTIONS  
USING THE TRUNCATED NORMAL DISTRIBUTION**

**Supporting Paper**

Submitted by Statistics Netherlands<sup>1</sup>

**ABSTRACT:** In this paper we suggest making use of the truncated (singular) normal distribution, in order to obtain imputations that immediately satisfy both balance and inequality restrictions, while preserving the distribution of the data. We will elaborate on maximum likelihood estimation of the parameters of the truncated normal, making use of MCMC methods, for complete as well as incomplete data (using the EM algorithm).

**I. INTRODUCTION**

1. Economic data are subject to several linear restrictions, balance as well as inequality restrictions. An example of a balance edit restriction is the fact that profit must equal revenue minus costs. Examples of inequality restrictions are the fact that ratios between variables should not exceed a certain amount or that most financial variables are nonnegative. We will use a truncated multivariate (singular) normal distribution to model these types of variables.

2. The truncation of data is quite common in the field of econometrics, see for example Maddala (1983) and Amemiya (1985). The use of a truncated (normal) distribution is justified if one believes that a distribution provides a reasonable model for data inside the truncation interval while at the same time the data can never take values outside this interval. A truncated normal distribution is characterized by the same parameters as its original and an admissible region. However, unlike the nontruncated multivariate normal distribution these parameters do not correspond directly to the mean and variance of the truncated distribution.

3. First of all define the set  $G = \{\mathbf{X} \in \mathbb{R}^k : \mathbf{a} \leq \mathbf{A}\mathbf{X} \leq \mathbf{b}\}$ , in which the outcomes of the variables lie. Note that for balance restrictions it holds that  $\mathbf{A}\mathbf{X} = \mathbf{a} = \mathbf{b}$ . We will first consider the case where there are *only* linear inequality restrictions, that is where the covariance matrix  $\Sigma$  is nonsingular. Next we will extend our research to the truncated multivariate singular normal distribution.

4. In a multivariate setting a truncated density function is defined as

---

<sup>1</sup> Prepared by Caren Tempelman, DTMN@cbs.nl.

$$f(\mathbf{x} | \boldsymbol{\theta}) = \begin{cases} 0 & \mathbf{x} \notin G \\ \frac{\psi(\mathbf{x} | \boldsymbol{\theta})}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}} & \mathbf{x} \in G \end{cases}$$

where  $\psi(\mathbf{x} | \boldsymbol{\theta})$  is a multivariate nontruncated density function with parameter vector  $\boldsymbol{\theta}$ . The normalising probability  $\Pr(\mathbf{x} \in G) = \int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}$  in the denominator makes sure that the truncated probability density function integrates to one. For the truncated normal it holds that the conditional distributions are also truncated normal. The marginal distributions, however are not truncated normal in general, see Geweke (1991).

## II. MAXIMUM LIKELIHOOD ESTIMATION WHEN THE DATA ARE TRUNCATED NORMAL

### A. Derivation of the first order conditions

5. The likelihood function of a truncated density is

$$\begin{aligned} L(\boldsymbol{\theta} | \mathbf{x}) &= \prod_{i=1}^n f(\mathbf{x}_i | \boldsymbol{\theta}) \\ &= \frac{\prod_{i=1}^n \psi(\mathbf{x}_i | \boldsymbol{\theta})}{\left(\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}\right)^n} \quad \mathbf{x} \in G \end{aligned}$$

which results in the following loglikelihood

$$\ell(\boldsymbol{\theta} | \mathbf{x}) = \sum_{i=1}^n \ln \psi(\mathbf{x}_i | \boldsymbol{\theta}) - n \ln \int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x} \quad \mathbf{x} \in G$$

The probability density function of the  $k$ -variate normal distribution is

$$\psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

and therefore the loglikelihood for the truncated normal distribution is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x}) = -\frac{nk}{2} \ln 2\pi - \frac{n}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - n \ln \int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \quad \mathbf{x} \in G$$

6. The first order conditions of the truncated multivariate model are calculated as follows. First of all define  $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$ , since it is simpler to differentiate the loglikelihood with respect to the parameters of  $\boldsymbol{\Sigma}^{-1}$ , and note that the following holds true

$$\begin{aligned} \ln |\boldsymbol{\Gamma}| &= -\ln |\boldsymbol{\Sigma}| \\ \frac{\partial \ln |\boldsymbol{\Gamma}|}{\partial \boldsymbol{\Gamma}} &= (\boldsymbol{\Gamma}')^{-1} \\ \frac{\partial \mathbf{x}' \boldsymbol{\Gamma} \mathbf{x}}{\partial \boldsymbol{\Gamma}} &= \mathbf{x} \mathbf{x}' \end{aligned}$$

see Magnus and Neudecker (1988).

7. The first order derivatives, assuming that the order of integration and differentiation can be interchanged, are

$$\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x})}{\partial \boldsymbol{\mu}} = \boldsymbol{\Gamma} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}) - n \frac{\int \cdots \int_G \partial \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) / \partial \boldsymbol{\mu} d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}}$$

$$\begin{aligned}
&= \Gamma \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}) - n \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Gamma(\mathbf{x} - \boldsymbol{\mu}) d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \\
&= n\Gamma \left[ \bar{\mathbf{x}} - \boldsymbol{\mu} - \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) (\mathbf{x} - \boldsymbol{\mu}) d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \right] \\
&= n\Gamma \left[ \bar{\mathbf{x}} - \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \right] \\
\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x})}{\partial \Gamma} &= \frac{n}{2} \Gamma^{-1} - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' - n \frac{\int \cdots \int_G \partial \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) / \partial \Gamma d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \\
&= \frac{n}{2} \Gamma^{-1} - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' - \frac{n}{2} \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) [\Gamma^{-1} - (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \\
&= \frac{n}{2} \Gamma^{-1} - \frac{n}{2} \mathbf{S} - \frac{n}{2} \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) [\Gamma^{-1} - (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}} \\
&= -\frac{n}{2} \mathbf{S} + \frac{n}{2} \frac{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}}
\end{aligned}$$

where  $\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$ . We have used the fact that  $|\Gamma|^{\frac{1}{2}} = \exp\{\frac{1}{2} \ln|\Gamma|\}$  and thus the derivative with regard to  $\Gamma$  equals

$$\frac{\partial |\Gamma|^{\frac{1}{2}}}{\partial \Gamma} = \frac{\partial \exp\{\frac{1}{2} \ln|\Gamma|\}}{\partial \Gamma} = \frac{1}{2} |\Gamma|^{\frac{1}{2}} \Gamma^{-1}$$

Recall that  $\boldsymbol{\mu}$  consists of  $k$  parameters and the covariance matrix of  $k(k+1)/2$  parameters since it is symmetric.

8. In order to find the maximum likelihood estimates we need to set the first order conditions to zero. Since these are not available in closed form, we must use an iterative procedure to obtain the maximum likelihood estimates. Moreover, these first order conditions are difficult to compute because they involve multidimensional integrals that do not have closed forms or rapid numerical solutions. Therefore we will resort to Monte Carlo methods in order to approximate these integrals through simulation. We will elaborate on this in section D. First we will discuss the choice of an iterative procedure to obtain the maximum likelihood estimates.

## B. Optimisation of the loglikelihood function using Fisher scoring

9. Since the first order conditions are not available in closed form we need an iterative algorithm to solve them. In general there are several possibilities for maximising the likelihood function.

10. First we can compute the Hessian and use regular Newton methods. Define the first order derivatives of the loglikelihood as  $s(\boldsymbol{\theta} | \mathbf{x})$ , the score vector. The matrix of the second order derivatives is referred to as the Hessian,  $H(\boldsymbol{\theta} | \mathbf{x})$ . The Newton-Raphson algorithm updates the parameters as follows

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + I^{-1}(\boldsymbol{\theta}^t | \mathbf{x}) s(\boldsymbol{\theta}^t | \mathbf{x}) \quad t = 1, 2, \dots$$

where  $I(\boldsymbol{\theta}^t | \mathbf{x})$  is the observed information, which is the negative of the Hessian. This algorithm generally works well given good starting values. However, due to the complexity of the loglikelihood function the Hessian will be difficult and time consuming to derive and therefore we will not be able to use regular Newton methods. Another disadvantage is that in order to find a maximum the Hessian needs to be negative definite, which is not guaranteed.

11. Another option is to use the Fisher scoring algorithm, which is a slight modification of the Newton-Raphson method. Instead of the observed information we will now use the expected information:  $\mathcal{I}(\boldsymbol{\theta}^t | \mathbf{x}) = E[I(\boldsymbol{\theta}^t | \mathbf{x})]$ . So the parameters are updated using

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \mathcal{I}^{-1}(\boldsymbol{\theta}^t | \mathbf{x})s(\boldsymbol{\theta}^t | \mathbf{x}) \quad t = 1, 2, \dots$$

The elegance of this method lies in the fact that under mild regularity conditions

$$\mathcal{I}(\boldsymbol{\theta} | \mathbf{x}) = -E\left[\frac{\partial^2 \ell(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = E\left[\sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta}} \left(\frac{\partial \ell_i(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta}}\right)'\right]$$

The latter equals  $\text{Var}\left(\frac{\partial \ell(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta}}\right)$  since  $E\left[\frac{\partial \ell(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta}}\right] = 0$ . This means that the expected information matrix will be positive definite and the expected Hessian is guaranteed to be negative definite, resulting in an ascent algorithm.

12. A third option is to use general optimisation methods such as quasi-Newton or nonlinear conjugate gradient methods, where direct calculation of the Hessian matrix is also not needed. Quasi-Newton methods typically perform better when minimising a general non quadratic function. This is partially due to the fact that quasi-Newton methods in addition to generating conjugate directions also approximate the Hessian. On the other hand compared to the conjugate gradient methods, the quasi-Newton methods require more storage space and each iteration requires more computation. One positive advantage of using an approximation instead of the actual Hessian is that the approximation can be chosen to be negative definite, ensuring that we will be attracted to a minimum.

13. Among the general purpose methods, the best is probably the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which builds upon the earlier Davidon-Fletcher-Powell (DFP) algorithm. DFP and BFGS are both robust algorithms. However, since the approximation is chosen to approximate  $H$  only in the directions needed for the Newton step, it is useless for the purpose of providing standard errors for the final estimates. If computer procedures are available, these methods are likely to take the least amount of effort. However, a line search algorithm is needed to calculate a step size, which can be time consuming and creates the need for an extra optimization algorithm. Taking all the above into account, the Fisher scoring method seems to be the best choice in our case.

### C. Monte Carlo integration

14. The original Monte Carlo approach was developed to use random number generation to compute complex integrals. Suppose we wish to compute

$$I(g) = \int_a^b g(x)dx$$

Suppose that we can write  $g(x) = f(x)p(x)$ , where  $p$  is a probability density defined on  $(a, b)$ . Let  $X$  be a random variable that has density function  $p$ , then we have

$$I(g) = \int_a^b g(x)dx = \int_a^b f(x)p(x)dx = E[f(X)]$$

So if we draw a large number of random variables  $X_i, i = 1, \dots, n$  from the density  $p$ , then

$$\int_a^b g(x)dx = E[f(X)] \approx \frac{1}{n} \sum_{i=1}^n f(X_i)$$

15. Now return to the multivariate case. The Monte Carlo method can be straightforwardly extended in order to calculate  $\int \cdots \int_G g(\mathbf{x}) d\mathbf{x}$ . Randomly generate  $\mathbf{X}_1, \dots, \mathbf{X}_n$ , which are distributed according to  $p$  in the region  $G$ , then the integral is estimated by

$$I(g) \approx \hat{I}(g) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_i)$$

If the  $\mathbf{X}_i$  are random independent variables and we take enough samples then  $\hat{I}(g)$  is a random variable with mean  $I(g)$  and variance  $\frac{1}{n} \text{Var}(f(\mathbf{X}))$ . The Central Limit theorem states that, provided the variance is finite

$$\hat{I}(g) - I(g) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{n} \text{Var}(f(\mathbf{X}))\right)$$

16. This clearly reflects the slow convergence of Monte Carlo methods; the absolute error has an average magnitude of  $O(\frac{1}{\sqrt{n}})$ . Hence to reduce the error, for example, by a factor of 10, the number of points that are sampled must be increased by a factor of 100. The slow convergence rate of  $\frac{1}{\sqrt{n}}$  means that Monte Carlo methods are usually limited to low accuracy. However, this convergence rate is independent of the number of dimensions. This makes Monte Carlo integration the preferred method for integrals in high dimensions.

#### D. Monte Carlo integration applied to the truncated multivariate normal distribution

17. In our case we want to calculate the first order conditions in order to find the maximum likelihood estimates. Recall that the first order conditions are

$$\nabla_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = n\boldsymbol{\Gamma}[\bar{\mathbf{x}} - F_1(\boldsymbol{\mu}, \boldsymbol{\Gamma})]$$

$$\nabla_{\boldsymbol{\Gamma}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = -\frac{n}{2} \mathbf{S} + \frac{n}{2} F_2(\boldsymbol{\mu}, \boldsymbol{\Gamma})$$

where

$$F_1(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) d\mathbf{x}}$$

$$F_2(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) d\mathbf{x}}$$

18. If we draw  $\mathbf{U}$  uniformly from  $G$ :  $\mathbf{U} \sim U(G)$ , then we can estimate  $F_1$  and  $F_2$  as follows

$$\hat{F}_1(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{\frac{V}{N} \sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i) \mathbf{U}_i}{\frac{V}{N} \sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i)} = \frac{\sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i) \mathbf{U}_i}{\sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i)}$$

$$\hat{F}_2(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{\frac{V}{N} \sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i) (\mathbf{U}_i - \boldsymbol{\mu})(\mathbf{U}_i - \boldsymbol{\mu})'}{\frac{V}{N} \sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i)} = \frac{\sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i) (\mathbf{U}_i - \boldsymbol{\mu})(\mathbf{U}_i - \boldsymbol{\mu})'}{\sum_{i=1}^N \boldsymbol{\psi}(\mathbf{U}_i)}$$

where  $V$  is the volume of the region  $G$ . Note that  $V$  is present in both the numerator and the denominator and therefore need not be calculated. Using uniform draws may, however, result in a large variance and therefore a less accurate estimate of the integral.

19. A distribution that puts more mass in the important regions of  $G$  would be preferable. Rewrite  $F_1$  and  $F_2$  as

$$F_1(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \int \cdots \int_G \frac{\psi(\mathbf{x})}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} \mathbf{x} d\mathbf{x} = \int \cdots \int_G g(\mathbf{x}) \mathbf{x} d\mathbf{x}$$

$$F_2(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \int \cdots \int_G \frac{\psi(\mathbf{x})}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x} = \int \cdots \int_G g(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}$$

where  $g$  is the normal distribution truncated to the region  $G$ . So if we draw  $\mathbf{V}$  from the truncated normal distribution  $g : \mathbf{V} \sim \mathcal{N}^T(\boldsymbol{\mu}, \boldsymbol{\Gamma})$ , we can estimate  $F_1$  and  $F_2$  by

$$\hat{F}_1(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{1}{N} \sum_{i=1}^N \mathbf{V}_i$$

$$\hat{F}_2(\boldsymbol{\mu}, \boldsymbol{\Gamma}) = \frac{1}{N} \sum_{i=1}^N (\mathbf{V}_i - \boldsymbol{\mu})(\mathbf{V}_i - \boldsymbol{\mu})'$$

A disadvantage of this method is that we need to generate a new Markov chain at each iteration, since it is dependent on  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ . However, since it resembles more closely the distribution of the data, the number of drawings probably need not be as large as in the uniform case.

## E. Drawing random independent values from the integration region $G$

### 20. Uniform draws

Drawing uniformly from the region  $G$  can be a serious problem if  $G$  is a complex region. A solution could be to insert  $G$  into another region, say  $H$ , that has simple boundaries. Next generate a random uniformly distributed vector from the region  $H$ , accept this vector if it is in  $G$  and reject it otherwise. In order for this method to be efficient the regions  $G$  and  $H$  need to be close. However, since the integration region  $G$  is a rather complex region, finding an area with simple boundaries close to  $G$  may be quite difficult.

21. Smith (1984) describes a class of procedures that are more efficient due to the fact that they generate a sequence of points within the region  $G$ . Generally speaking the described algorithms work as follows. Find a starting point  $\mathbf{x}^0$  in  $G$  and generate a random direction  $\mathbf{d}$ . Then find the line set  $L = G \cap \{\mathbf{x} : \mathbf{x} = \mathbf{x}^0 + \lambda \mathbf{d}, \lambda \in \mathbb{R}\}$  and generate a random point  $\mathbf{x}$  uniformly distributed over  $L$ . Furthermore Smith shows that if a long sequence of points is generated this way and then randomly mixed, the statistical properties of the resulting sequence for large  $n$  will approximate those of  $n$  independently identically distributed uniform draws.

22. Another approach is to use Markov chain Monte Carlo (MCMC) methods. In this case an ergodic Markov chain is obtained that is in the region  $G$  with the distribution of interest as a stationary distribution. We can, for example, use Metropolis-Hastings to generate an ergodic Markov chain with the uniform distribution as stationary distribution. The Metropolis-Hastings algorithm was developed by Metropolis et al. (1953) and generalised by Hastings (1970). Again let  $V$  be the volume of  $G$ , then the stationary distribution  $\pi$  is

$$\pi(\mathbf{x}) = \begin{cases} \frac{1}{V} & \mathbf{x} \in G \\ 0 & \mathbf{x} \notin G \end{cases}$$

Set  $j = 0$  and choose a starting value  $\mathbf{x}^0 \in G$ , for example the origin. Since we are dealing with economic data the origin is always a valid option. Other starting values could be the item values of one respondent or the mean calculated based on the truncated data. The latter is a valid option since all

records satisfy the edit restrictions. In fact, starting with the mean is probably a good idea since this is near the mass of the density. At any case this means that even for very complex regions, we will not have any difficulty finding a starting value in  $G$ .

23. Now choose a candidate generating density  $q(\cdot, \mathbf{x}^j)$ , from which we draw a candidate  $\mathbf{y}$ . Next calculate the acceptance probability  $\alpha$

$$\alpha(\mathbf{x}^j, \mathbf{y}) = \min \left\{ \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x}^j)}{\pi(\mathbf{x})q(\mathbf{x}^j, \mathbf{y})}, 1 \right\}$$

If  $\mathbf{y} \notin G$ , then the acceptance probability becomes zero, since  $\pi(\mathbf{y}) = 0$ . This means that the new value is rejected and the old value retained:  $\mathbf{x}^{j+1} = \mathbf{x}^j$ . If  $\mathbf{y} \in G$ , then  $\alpha$  becomes

$$\alpha(\mathbf{x}^j, \mathbf{y}) = \min \left\{ \frac{q(\mathbf{y}, \mathbf{x}^j)}{q(\mathbf{x}^j, \mathbf{y})}, 1 \right\}$$

If we choose the candidate generating density to be symmetric ( $q(\mathbf{x}, \mathbf{y}) = q(\mathbf{y}, \mathbf{x})$ ), for example a normal density, the acceptance probability will be one and the candidate  $\mathbf{y}$  will be accepted:  $\mathbf{x}^{j+1} = \mathbf{y}$ . If  $q$  is not symmetric, we draw  $u \sim U(0,1)$  and accept  $\mathbf{y}$  if  $\alpha(\mathbf{x}^j, \mathbf{y}) \geq u$ . Otherwise we retain the previous value. Set  $j = j+1$  and generate a new candidate.

24. This immediately illustrates the major advantage of Markov chains as opposed to Acceptance/Rejection sampling. If the new draw is rejected, the old value is retained, which makes it less difficult to find a sufficiently large sample. Note, however, that due to the dependence in Markov chains we need to obtain a much larger sample to get a certain amount of accuracy as opposed to drawing independently. But the statistical properties of Markov chains are much better known than those of the method developed by Smith (1984) that was mentioned above.

25. *Truncated normal draws*

As we mentioned in section D we could also estimate the first order derivatives using draws from the truncated normal distribution. Again a Markov chain can be generated using Metropolis-Hastings. In this case the stationary distribution  $\pi$  is

$$\pi(\mathbf{x}) = \begin{cases} \frac{\psi(\mathbf{x})}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} & \mathbf{x} \in G \\ 0 & \mathbf{x} \notin G \end{cases}$$

with  $\psi$  the normal probability density function. Again set  $j=0$  and choose a starting value  $\mathbf{x}^0$ . Generate a candidate  $\mathbf{y}$  from  $q(\cdot, \mathbf{x}^j)$ . If  $\mathbf{y} \notin G$ , we reject  $\mathbf{y}$  with probability one and retain  $\mathbf{x}^j$ . If  $\mathbf{y} \in G$ , the acceptance probability becomes

$$\alpha(\mathbf{x}^j, \mathbf{y}) = \min \left\{ \frac{\psi(\mathbf{y})q(\mathbf{y}, \mathbf{x}^j)}{\psi(\mathbf{x})q(\mathbf{x}^j, \mathbf{y})}, 1 \right\}$$

Note that the normalising probability need not be calculated since it is present in both numerator and denominator. Again draw  $u \sim U(0,1)$  and accept  $\mathbf{y}$  if  $\alpha(\mathbf{x}^j, \mathbf{y}) \geq u$ .

## F. THE ALGORITHM

26. The algorithm that results for maximum likelihood estimation in the presence of truncated normal data is

$$\max_{\boldsymbol{\mu}, \boldsymbol{\Gamma}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = -\frac{nk}{2} \ln 2\pi + \frac{n}{2} \ln |\boldsymbol{\Gamma}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Gamma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) - n \ln \int \cdots \int_G \psi(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) d\mathbf{x}$$

with

$$\nabla_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = n\boldsymbol{\Gamma} \left[ \bar{\mathbf{x}} - \frac{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) d\mathbf{x}} \right]$$

$$\nabla_{\boldsymbol{\Gamma}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = -\frac{n}{2} \mathbf{S} + \frac{n}{2} \frac{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \boldsymbol{\psi}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Gamma}) d\mathbf{x}}$$

Define

$$\nabla \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) = \begin{pmatrix} \nabla_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x}) \\ \text{vech}(\nabla_{\boldsymbol{\Gamma}} \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} | \mathbf{x})) \end{pmatrix}$$

For a symmetric matrix the vech of that matrix is the  $\frac{1}{2}k(k+1) \times 1$  vector obtained by stacking the unique elements of the matrix into a single vector.

- Step 0.

Choose initial values  $\boldsymbol{\mu}^0$ ,  $\boldsymbol{\Gamma}^0$  and generate a Markov chain  $\{\mathbf{U}^t\}$ ,  $t=1, \dots, N$  uniformly distributed in  $G$  using the Metropolis Hastings algorithm described in section E and set  $k=0$ .

- Step 1.

Calculate  $\hat{F}_1(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k)$  and  $\hat{F}_2(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k)$  based on the estimates  $\boldsymbol{\mu}^k$  and  $\boldsymbol{\Gamma}^k$  and the Markov chain  $\{\mathbf{U}^t\}$ . Or generate a Markov chain  $\{\mathbf{V}^t\}$ ,  $t=1, \dots, N$  representing truncated normal draws in  $G$  based on the estimates  $\boldsymbol{\mu}^k$  and  $\boldsymbol{\Gamma}^k$  and estimate  $F_1(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k)$  and  $F_2(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k)$ .

Next calculate  $\nabla \ell(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x})$ . If all of the elements of the gradient are less than  $\varepsilon$ , with  $\varepsilon$  sufficiently small, then stop. The estimates  $\boldsymbol{\mu}^k$  and  $\boldsymbol{\Gamma}^k$  are local optima of  $\ell$ .

- Step 2.

Now calculate  $\mathbf{d}^k = \mathcal{I}^{-1}(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x}) \nabla \ell(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x})$ , where

$$\mathcal{I}(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x}) = E \left[ \sum_{i=1}^n \nabla \ell_i(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x}) \nabla \ell_i(\boldsymbol{\mu}^k, \boldsymbol{\Gamma}^k | \mathbf{x})' \right]$$

- Step 3.

Update the estimates

$$\begin{pmatrix} \boldsymbol{\mu}^{k+1} \\ \text{vech}(\boldsymbol{\Gamma}^{k+1}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}^k \\ \text{vech}(\boldsymbol{\Gamma}^k) \end{pmatrix} + \mathbf{d}^k$$

Set  $k := k+1$  and go back to step 1.

### III. MAXIMUM LIKELIHOOD ESTIMATION WHEN THE DATA ARE TRUNCATED SINGULAR NORMAL

27. Regular economic data consist of linear balance restrictions as well as linear inequality restrictions, and we want to impute the missing data items satisfying both types of restrictions. This means we need to incorporate them simultaneously when we are estimating the parameters using the maximum likelihood principle. Let  $\mathbf{X}$  represent the data matrix. The balance and inequality restrictions on  $\mathbf{X}$  can be written as:  $\mathbf{l} \leq \mathbf{A}\mathbf{X} \leq \mathbf{u}$ . Let  $r$  be the total number of restrictions on the data  $\mathbf{X}$ . If restriction  $j$ ,  $j=1, \dots, r$  is a balance restriction it holds that  $\mathbf{l}_j = \mathbf{u}_j$ . Let  $p$  denote the number of balance restrictions, and  $q$  the number of inequality restrictions. Assume that  $\mathbf{X}$  is distributed according to a truncated singular normal distribution, that is  $\mathbf{X} \sim \mathcal{N}_k^T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\Sigma}$  singular.

28. The singular covariance matrix  $\boldsymbol{\Sigma}$  can be decomposed by means of an eigenvalue decomposition into  $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$ , where  $\mathbf{C}$  is the orthogonal matrix of eigenvectors and  $\boldsymbol{\Lambda}$  is the diagonal matrix of

eigenvalues of  $\Sigma$ ,  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_{k-p}, 0, \dots, 0\}$ . Let  $\tilde{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_{k-p}\}$ , the matrix of nonzero eigenvalues. The density function of the singular normal is (see Khatri, 1968)

$$\varphi(\mathbf{x}) = (2\pi)^{-\frac{q}{2}} \left( \prod_{j=1}^q \lambda_j \right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^+ (\mathbf{x} - \boldsymbol{\mu})\right\} \text{ for } \mathbf{x} \in \boldsymbol{\mu} + L^\perp$$

where  $\prod_{j=1}^q \lambda_j = \det(\tilde{\Lambda})$ ,  $\Sigma^+ = \mathbf{C}_1 \Lambda^{-1} \mathbf{C}_1'$ , and  $L^\perp$  is the orthogonal complement of  $L$ , which is the kernel of  $\Sigma$ .

29. Maximum likelihood estimation of the parameters of the truncated singular normal distribution is similar to maximum likelihood estimation of parameters of the truncated normal. Khatri (1968) showed that for the nontruncated singular normal the maximum likelihood estimates of the parameters are the same as in the nonsingular case. His proof can be straightforwardly extended to the truncated case. This results in the following first order conditions

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} &= \bar{\mathbf{x}} - \frac{\int \cdots \int_G \varphi(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \varphi(\mathbf{x}) d\mathbf{x}} \\ \nabla_{\Gamma} &= -\mathbf{S} + \frac{\int \cdots \int_G \varphi(\mathbf{x}) (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \varphi(\mathbf{x}) d\mathbf{x}} \end{aligned}$$

In order to obtain the maximum likelihood estimates we need to set these first order conditions to zero. Note that the only difference with the nonsingular case is the fact that we now use the singular normal density  $\varphi$ . This means that the algorithm derived in the previous section can be used to obtain the maximum likelihood estimates of data subject to both inequality and balance restrictions.

#### IV. INCOMPLETE DATA

30. In the previous sections we have derived maximum likelihood estimates for the multivariate truncated normal distribution assuming there was no missing data. However, as we mentioned before, incomplete data is a prevalent problem in survey research. Therefore we will now extend our method in order to handle incomplete data. We will make use of the EM algorithm (Dempster, Laird and Rubin, 1988), which was developed for maximum likelihood estimation in the presence of missing data.

31. Partition the data matrix  $\mathbf{X}$  into an observed and a missing part:  $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$ . Partition the data vector corresponding to each record  $i$  accordingly. The EM algorithm consists of an E-step, which calculates the expected observed data loglikelihood, and an M-step, which maximises this loglikelihood. For each iteration  $t$  the E-step consists of finding  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t)$  which is defined as

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t) = E[\ell(\boldsymbol{\theta} | \mathbf{X}) | \mathbf{X}_{obs}, \boldsymbol{\theta}^t]$$

where  $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$ . The M-Step maximizes  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t)$  with respect to  $\boldsymbol{\theta}$  and this results in  $\boldsymbol{\theta}^{t+1}$ .

32. We will start by differentiating  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t)$  with respect to  $\boldsymbol{\theta}$ .

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t)}{\partial \boldsymbol{\mu}} &= E \left[ \Sigma^{-1} \left( \bar{\mathbf{x}} - \frac{\int \cdots \int_G \psi(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} \right) \right] \\ &= \Sigma^{-1} \left( E[\mathbf{X}] - \frac{\int \cdots \int_G \psi(\mathbf{x}) \mathbf{x} d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} \right) \end{aligned}$$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta} | \boldsymbol{\theta}')}{\partial \boldsymbol{\Sigma}^{-1}} &= \mathbb{E} \left[ -\mathbf{S} + \frac{\int \cdots \int_G \psi(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}} \right] \\ &= -\mathbb{E}[\mathbf{S}] + \frac{\int \cdots \int_G \psi(\mathbf{x})(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' d\mathbf{x}}{\int \cdots \int_G \psi(\mathbf{x}) d\mathbf{x}}\end{aligned}$$

These conditions need to be set equal to zero in order to find the maximum likelihood estimates. This means that we need to use  $\mathbb{E}[\mathbf{X}_{i,mis} | \mathbf{X}_{i,obs}, \boldsymbol{\theta}']$ , and  $\mathbb{E}[\mathbf{X}_{i,mis} \mathbf{X}_{i,mis}' | \mathbf{X}_{i,obs}, \boldsymbol{\theta}']$ , to calculate the above mentioned conditions.

33. Since  $\mathbf{X}_{i,mis}$  conditional on  $\mathbf{X}_{i,obs}$  is truncated normal, we can derive these expectations using Monte Carlo integration with draws from the truncated normal distribution, so that the algorithm results in the so-called MCEM algorithm (Wei and Tanner, 1990).

#### IV. DISCUSSION

34. We have derived an algorithm for the maximum likelihood estimation of the parameters of a truncated multivariate normal distribution, for the non-singular as well as the singular case, which means that the algorithm can handle data that need to satisfy both linear balance and inequality restrictions. In the presence of missing data the MCEM algorithm will be used. Future research will concentrate on applying the algorithm to synthetic and realistic data. Moreover the MCEM procedure will be investigated further since it is computationally expensive.

#### References

- Amemiya, T. (1985), *Advanced Econometrics*, Oxford: Basil Blackwell.
- Dempster, A.P., Laird, N.M. and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society B*, 39, 1-38.
- Geweke, J. (1991), "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 571-578.
- Hastings, W.K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97-109.
- Khatri, C.G. (1968), "Some results for the singular normal multivariate regression models," *Sankhyā, Series A*, 267-280.
- Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.
- Magnus, J.R. and Neudecker, H. (1988), *Matrix differential Calculus*, New York: Wiley.
- 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, 1087-1092.
- Smith, R.L. (1984), "Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions," *Operations research*, 32, 1296-1308.
- Wei, G.C.G. and Tanner, M.A. (1990), "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm," *Journal of the American Statistical Association*, 85, 699-704.