Imputation of restricted data
Tempelman, Caren

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2007

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 5

Imputation of Data Subject to Balance and Inequality Restrictions

In chapter 3 we discussed imputation of data that are subject to non-negativity constraints and one linear balance restriction. It was suggested to use the Dirichlet distribution for this purpose. In chapter 4 we elaborated on the imputation of data subject to multiple linear balance restrictions. As the Dirichlet distribution could not be employed in this instance, the singular normal distribution was proposed. Unfortunately, the singular normal distribution does not assume non-negativity. Besides, often other inequality restrictions occur in business surveys, such as the fact that the number of employees in fte (full-time equivalent) may not exceed the total number of employees. In this chapter we therefore suggest an extension of the method discussed in chapter 4 in order to deal with all possible types of inequality restrictions. We will discuss the use of the truncated singular normal distribution to model data subject to inequality and balance restrictions.

5.1 Introduction

Economic data have to satisfy many logical linear restrictions. In general we distinguish between two types of restrictions, i.e. balance and inequality restric-
tions. Balance restrictions refer to equalities that must hold, such as the fact that different operating expenses need to add up to the total. Inequality restrictions refer to inequalities that must hold, such as non-negativity constraints or the fact that the number of employees in fte may not exceed the total number of employees. The aim of this chapter is to develop an imputation method that can deal with several types of inequality as well as balance restrictions simultaneously. To achieve this, the singular normal distribution that was employed in chapter 4 will be truncated to the region defined by the inequality restrictions.

In section 5.2 we will discuss the different types of restrictions that may occur in business surveys. Next in section 5.3 the truncation of data is treated as well as some important properties of the truncated normal distribution. Maximum likelihood estimation is dealt with in section 5.4 and subsequently in section 5.5 the EM algorithm for maximum likelihood estimation in the presence of non-response is discussed. In section 5.6 imputation methods for truncated normal data are provided. Then in section 5.7 balance restrictions are incorporated, which means that the data are modelled with a truncated singular normal distribution. Maximum likelihood estimation for complete and incomplete data are also discussed. In section 5.8 applications of this model are treated and, finally, in section 5.9 concluding remarks are given.

5.2 Linear inequality and balance restrictions

Let $X_i$ denote the data vector for respondent $i$. Define the set $G = \{X_i \in \mathbb{R}^k : 1 \leq BX_i \leq u\}$, which defines all possible outcomes of the variables. The matrix $B$ is an $r \times k$ matrix containing $r$ inequality restrictions on the data, $B$ need not be of full row rank as it is allowed to contain non-negativity as well as other linear inequality constraints. The upper and lower bounds $u$ and $l$ may equal plus or minus infinity, which means that the variables are truncated from one side, e.g. by non-negativity constraints.

Additionally, consider a $p \times k$ restriction matrix $A$, with $p$ the number of linear balance restrictions, for which it holds that $AX_i = 0$. We assume that every variable is present in at least one balance restriction and that there are no redundant balance restrictions, which means that $A$ is of full row rank.

So, the data $X_i$, $i = 1, \ldots, n$, which are subject to $1 \leq BX_i \leq u$ and $AX_i = 0$ need to be modelled. Note that nonlinear transformations cannot be used in this instance as the balance edit structure would be lost then. These different types of restrictions lead to various issues when modelling the data, we will first assume that the data are solely subject to linear inequality restrictions,
so the covariance matrix $\Sigma$ will be nonsingular in this case. Next in section 5.7 we will extend our research to modelling both inequality and balance restrictions simultaneously.

5.3 Truncation of data

A truncated multivariate normal distribution will be used to model variables subject to inequality restrictions. Truncation of data is quite common in the field of econometrics, see for example Maddala (1983), Amemiya (1985) and Hajivassiliou and Ruud (1994). 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.

In a multivariate setting the probability density function of truncated data is defined as

$$f(x_i | \theta) = \left\{ \begin{array}{ll}
0 & x_i \not\in F \\
\frac{\psi(x_i | \theta)}{\int_G \psi(x_i | \theta) \, dx_i} & x_i \in F
\end{array} \right.,$$

where $\psi(x_i | \theta)$ is a multivariate nontruncated normal density function with parameter vector $\theta$, $\theta' = (\mu', \text{vec}(\Sigma)')$. The normalising probability $Pr(X_i \in F) = \int \cdots \int_F \psi(x_i | \theta) \, dx_i$ in the denominator makes sure that the probability density function of the truncated data integrates to one.

5.3.1 Some properties of the truncated multivariate normal distribution

A truncated normal distribution is characterised by the same parameters as its original nontruncated version plus an admissible region, which is independent of the parameters. However, unlike the nontruncated multivariate normal distribution these parameters do not correspond directly to the mean and variance of the truncated distribution. This is illustrated in figure 5.1 for the univariate case. The truncated normal density of a variable $X$, with $\mu = 2, \sigma = 3$ and $X > 0$, is plotted. Clearly the sample mean, $\bar{x}$, is positively biased for $\mu$, as $X$ is truncated from below. Moreover, the sample standard deviation, $s$, underestimates the parameter $\sigma$ as there is less variation in $X$ in a truncated sample as opposed to a nontruncated sample.

It is interesting to establish which properties of the multivariate normal distribution remain valid after truncation. Let $X_i$ be distributed according to a
truncated normal distribution

\[ X_i \sim \mathcal{N}_k^T(\mu, \Sigma), \quad X_i \in G. \]

Multiplying the vector \( X_i \) with a \( k \times k \) matrix \( D \) of full rank leads to

\[ DX_i \sim \mathcal{N}_k^T(D\mu, D\Sigma D'), \quad X_i \in G. \]

This can be established by using the change of variables formula, which is defined as follows. If \( X \) is a random vector with probability density function \( f_X(x) \) and if \( Y = g(X) \), where \( g \) is a continuously differentiable function defined on the region \( G \), then

\[ f_Y(y) = f_X(g^{-1}(y))|\text{det} J| \]

is the probability density function of \( Y \), where \( J \) is the matrix of partial derivatives of \( g^{-1}(y) \). Let \( Y_i = DX_i, \ Y_i \in T, \) with \( T := \{ DX_i, \ X_i \in G \} \). Thus
\[ g^{-1}(y_i) = D^{-1}y_i \text{ and } \frac{\partial g^{-1}(y_i)}{\partial y_i} = (D^{-1})'. \] This means that

\[ f_Y(y_i) = f_{X_i}(D^{-1}y_i) \frac{\text{abs}(|D^{-1}'|)}{\Pr(X_i \in G)} \]

\[ = \frac{1}{(2\pi)^{-k/2} \Pr(X_i \in G)} \Sigma^{-\frac{1}{2}} \exp \left( \frac{1}{2} (D^{-1}y_i - \mu)(D^{-1}y_i - \mu)' \Sigma^{-1} (D^{-1}y_i - \mu) \right) \text{abs}(|D^{-1}|). \]

Note that

\[ (D^{-1}y_i - \mu)' \Sigma^{-1} (D^{-1}y_i - \mu) = (y_i - D\mu)'(D\Sigma D')^{-1}(y_i - D\mu) \]

and

\[ |D^{-1}| = |D|^{-1} = |D|^{-\frac{k}{2}} |D'|^{-\frac{k}{2}} = |D|^{-\frac{k}{2}} |D'|^{-\frac{k}{2}}. \]

Then

\[ |\Sigma|^{-\frac{k}{2}} \text{abs}(|D^{-1}|) = |\Sigma|^{-\frac{k}{2}} \text{abs}(|D|^{-\frac{k}{2}} |D'|^{-\frac{k}{2}}) = \text{abs}(|D\Sigma D'|^{-\frac{k}{2}}), \]

as \( \Sigma \) is positive definite and therefore the determinant of \( \Sigma \) is positive. This also implies that \( D\Sigma D' \) is always positive definite and therefore \( |D\Sigma D'| > 0 \), so \( \text{abs}(|D\Sigma D'|^{-\frac{1}{2}}) = |D\Sigma D'|^{-\frac{1}{2}} \). This means that the density function for \( Y_i \) becomes

\[ f_Y(y_i) = \frac{(2\pi)^{-k/2}}{\Pr(X_i \in G)} |D\Sigma D'|^{-\frac{k}{2}} \exp \left( \frac{1}{2} (y_i - D\mu)'(D\Sigma D')^{-1}(y_i - D\mu) \right). \]

Since \( \Pr(Y_i \in T) = \Pr(X_i \in G) \), the probability density function of \( Y_i \) is truncated normal as well, with mean vector \( D\mu \) and covariance matrix \( D\Sigma D' \). Note that this will only hold if the matrix \( D \) is nonsingular.

Now partition \( X'_i = (X'_{i,1}, X'_{i,2}) \) and partition \( \mu, \Sigma \) and \( B \) accordingly. Next define the matrix \( D \) as follows in order to derive the conditional distribution

\[ D = \left( \begin{array}{c} I_m \\ 0 \end{array} \right) - \Sigma_{12} \Sigma_{22}^{-1} \]

where \( m \) is the length of data vector \( X_{i,1} \). Now calculate

\[ D(X_i - \mu) = \left( \begin{array}{c} X_{i,1} - \mu_1 - \Sigma_{12} \Sigma_{22}^{-1} (X_{i,2} - \mu_2) \\ X_{i,2} - \mu_2 \end{array} \right). \]
This will still be truncated multivariate normal with zero mean vector and covariance matrix $D\Sigma D'$. Note that

$$D\Sigma D' = \begin{pmatrix} \Sigma_{11} & -\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

and so the covariances are zero, which means that $X_{i,1} \sim -\boldsymbol{\mu}_1 \mid X_{i,2}$ remains truncated multivariate normal with mean parameter $\Sigma_{12}\Sigma_{22}^{-1}(X_{i,2} - \boldsymbol{\mu}_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. So

$X_{i,1} \mid X_{i,2} \sim \mathcal{N}_m^{T}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_{i,2} - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, $X_{i,1} \in G_1(X_{i,2})$,

with $G_1(X_{i,2}) = \{ Y_i \in \mathbb{R}^m : 1 - B_2 X_{i,2} \leq B, Y_i \leq u - B_2 X_{i,2} \}$.

The fact that the conditional densities remain truncated normal is convenient in an imputation context as the missing data are conditioned on the observed data, which consequently can still be modelled with a truncated normal distribution. Furthermore, this property is also convenient if one wants to make use of the Gibbs sampler, which will be further discussed in subsection 5.4.1.2.

The marginal distribution of $X_{i,1}$ or $X_{i,2}$ is, however, not truncated normal in general. Note that the marginal distribution of $X_{i,2}$ is defined as

$$f_{X_{i,2}}(x_{i,2}) = \frac{f_{X_{i,1},X_{i,2}}(x_{i,1},x_{i,2})}{f_{X_{i,1}\mid X_{i,2}}(x_{i,1}\mid x_{i,2})}.$$  

If the joint distribution of $X_{i,1}$ and $X_{i,2}$ is truncated normal, the conditional distribution of $X_{i,1}$ given $X_{i,2}$ is truncated normal as well. Let $\psi$ denote the nontruncated multivariate normal density. This means that

$$f_{X_{i,2}}(x_{i,2}) = \frac{\psi(x_{i,1},x_{i,2})}{\int \cdots \int \psi(x_{i,1},x_{i,2}) \, dx_{i,1} \, dx_{i,2}} \frac{\int \cdots \int \psi(x_{i,1} \mid x_{i,2}) \, dx_{i,1} \, dx_{i,2}}{\psi(x_{i,2}) \int \cdots \int \psi(x_{i,1} \mid x_{i,2}) \, dx_{i,1} \, dx_{i,2}} = \frac{\psi(x_{i,2})}{\int \cdots \int \psi(x_{i,1},x_{i,2}) \, dx_{i,1} \, dx_{i,2}} \frac{\int \cdots \int \psi(x_{i,1} \mid x_{i,2}) \, dx_{i,1} \, dx_{i,2}}{\psi(x_{i,2}) \int \cdots \int \psi(x_{i,1} \mid x_{i,2}) \, dx_{i,1} \, dx_{i,2}},$$

where

$$\psi(x_{i,1} \mid x_{i,2}) = (2\pi)^{-\frac{1}{2}}|\Sigma_{11,2}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2}(x_{i,1} - \mu_{i,2})'\Sigma_{11,2}^{-1}(x_{i,1} - \mu_{i,2}) \right),$$
with
\[
\begin{align*}
\mu_{1,2} &= \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (X_{i,2} - \mu_2) \\
\Sigma_{11,2} &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\end{align*}
\]
So the term \( \int \cdots \int \psi(x_{i,1} | x_{i,2}) \, dx_{i,1} \) is still dependent on \( X_{i,2} \) after integration, which means that the marginal distribution of \( X_{i,2} \) is not truncated multivariate normal in general. Only if \( \Sigma_{12} = 0 \), the marginal distributions of \( X_{i,1} \) and \( X_{i,2} \) are truncated normal, as in this case it holds that
\[
\int \cdots \int \psi(x_{i,1} | x_{i,2}) \, dx_{i,1} = \int \cdots \int \psi(x_{i,1}) \, dx_{i,1} = 1.
\]
For several other properties of the truncated normal distribution, see Horrace (2005).

### 5.4 Maximum likelihood estimation for truncated normal data

Due to truncation of the data, the sample mean and sample covariance matrix do not correspond directly to the distribution parameters \( \mu \) and \( \Sigma \) of the truncated multivariate normal, which was illustrated in Figure 5.1 for the univariate case. Parameter estimates of the truncated normal can be obtained using maximum likelihood estimation.

The likelihood function of a sample from a truncated density \( f(\cdot | \theta) \) is
\[
L(\theta | x) = \prod_{i=1}^{n} f(x_i | \theta) = \frac{\prod_{i=1}^{n} \psi(x_i | \theta)}{\int_{G} \psi(x_i | \theta) \, dx_i}^{n}, \quad x_i \in G,
\]
where \( \psi(\cdot | \theta) \) is the nontruncated density. This results in the following loglikelihood
\[
\ell(\theta | x) = \sum_{i=1}^{n} \ln \psi(x_i | \theta) - n \ln \int_{G} \psi(x_i | \theta) \, dx_i,
\]
and therefore the loglikelihood for the truncated normal distribution is
\[
\ell(\mu, \Sigma | x) = -\frac{nk}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)'\Sigma^{-1}(x_i - \mu) + \\
- n \ln \int \cdots \int_{G} \psi(x_i | \mu, \Sigma) \, dx_i.
\]
This loglikelihood is difficult to compute and maximise as it contains a multidiagonal integral that does not have a closed form or a rapid numerical approximation. Several simulation methods have been developed in order to find maximum likelihood estimates in the presence of high-dimensional integrals, see Hajiavassiliou and Ruud (1994) or Train (2003, chapter 10) for a detailed description of these methods. In general, there is a method that simulates the likelihood or loglikelihood function which is then maximised. This is referred to as maximum simulated likelihood (MSL). Simulation of the loglikelihood function, however, results in bias as $\ln \hat{L}$ is not unbiased for $\ln L$, even if $\hat{L}$ is an unbiased estimator of $L$. This has some consequences for the asymptotic properties of the MSL estimator. Assume that the same regularity conditions as for nonsimulated maximum likelihood (see chapter 2) hold. Let $R$ denote the number of replications used in the simulation, then if $R$ is fixed the MSL estimator will not converge to the true parameter due to the aforementioned bias. If instead $R$ rises with the sample size $n$, i.e. $R$ goes to infinity, the simulation bias will vanish and in this case the MSL estimator is consistent. Finally, if $R$ rises faster than $\sqrt{n}$, the MSL estimator is also efficient, asymptotically normal and asymptotically equivalent to the maximum likelihood estimator.

Other approaches, that can be consistent for fixed $R$, are the method of simulated moments (MSM), where the moment functions are simulated, and the method of simulated scores (MSS), within which the score functions of the (log)likelihood are simulated. Both the moment and the score functions are linear in the integrals that need to be simulated, meaning that they can be approximated through unbiased estimation. In that case both methods will be consistent for fixed $R$. Moreover, as $R$ goes to infinity the simulation noise will disappear and MSM will be asymptotically equivalent to the method of moments and MSS will be asymptotically equivalent to maximum likelihood (Train, 2003).

So, MSS with unbiased simulators of the score has better properties than MSL. With respect to MSM, an advantage of the MSS is that it can be straightforwardly applied to any model for which the score functions can be derived, whereas for the MSM (optimal) instruments need to be specified. MSS automatically uses the optimal instruments for asymptotic efficiency (Hajiavassiliou and McFadden, 1998). Besides, simulation of the score fits naturally with the EM algorithm that is used in the presence of missing data, which will be clarified in section 5.5. For these reasons we will use the method of simulated scores to obtain the maximum likelihood estimates of the parameters of the truncated normal. In Hajiavassiliou and McFadden (1998) this method is discussed extensively. The downside of this approach is that it can be difficult to find unbiased simulators of the score function with good numerical properties.
A last point that is important to realise is that the random variates underlying the simulations should be drawn only once and used at every iteration of the optimisation algorithm. If the random variates do vary, the same values of $\mu$ and $\Sigma$ may lead to (small) differences in the estimates for the integrals and consequently the algorithm may never converge.

### 5.4.1 Using the method of simulated scores to obtain the maximum likelihood estimates

In order to employ the method of simulated scores, the first order conditions of the truncated multivariate normal model need to be derived. First of all define $\Gamma = \Sigma^{-1}$, since it is simpler to differentiate the loglikelihood with respect to the parameters of $\Sigma^{-1}$, and note that the following holds true

$$
\frac{\ln |\Gamma|}{\partial \Gamma} = -\ln |\Sigma|
$$

$$
\frac{\partial \ln |\Gamma|}{\partial \Gamma} = (\Gamma')^{-1}
$$

$$
\frac{\partial y' \Gamma y}{\partial \Gamma} = yy',
$$

see, for instance, Magnus and Neudecker (1988). From this it follows that

$$
\frac{\partial |\Gamma|^\frac{1}{2}}{\partial \Gamma} = \frac{\partial \exp \left(\frac{1}{2} \ln |\Gamma|\right)}{\partial \Gamma} = 1/2 |\Gamma|^{\frac{1}{2}} \Gamma^{-1}.
$$

Then the first order derivatives are

$$
\frac{\partial \ell(\mu, \Gamma | x)}{\partial \mu} = \Gamma \sum_{i=1}^{n} (x_i - \mu) - \frac{n}{\int_G \psi(x_i | \mu, \Gamma) dx_i} \int_G \frac{\partial \psi(x_i | \mu, \Gamma)}{\partial \mu} dx_i
$$

$$
= \Gamma \sum_{i=1}^{n} (x_i - \mu) - \frac{n}{\int_G \psi(x_i | \mu, \Gamma) dx_i} \int_G \frac{\partial \psi(x_i | \mu, \Gamma)}{\partial \mu} dx_i
$$

$$
= \Gamma \left( \sum_{i=1}^{n} x_i - n \int_G \psi(x_i | \mu, \Gamma) dx_i \right)
$$

$$
= \Gamma \left( \sum_{i=1}^{n} x_i - n \mathbb{E}[X_i | X_i \in G, \mu, \Gamma] \right)
$$
and
\[
\frac{\partial \ell(\mu, \Gamma | x)}{\partial \Gamma} = \frac{n}{2} \Gamma^{-1} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{'} - n \int \cdots \int_{G} \psi(x_i | \mu, \Gamma) \frac{\partial \psi(x_i | \mu, \Gamma)}{\partial \Gamma} \, dx_i
\]
\[
= \frac{n}{2} \Gamma^{-1} - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{'} + \int \cdots \int_{G} \psi(x_i | \mu, \Gamma) \left( \frac{1}{2} \Gamma^{-1} - \frac{1}{2} (x_i - \mu)(x_i - \mu)^{'} \right) \, dx_i
\]
\[
= -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{'} + \frac{n}{2} \int \cdots \int_{G} \psi(x_i | \mu, \Gamma) (x_i - \mu)(x_i - \mu)^{'} \, dx_i
\]
\[
= -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^{'} + \frac{n}{2} \mathbb{E}(X_i - \mu | X_i, \mu, \Gamma) \in G, \mu, \Gamma.
\]

Recall that $\mu$ and $\Gamma$, as it is symmetric, consist of $k$ and $k(k+1)/2$ parameters respectively.

The following quantities in the first order conditions need to be estimated through simulation
\[
\mathbb{E}[X_i | X_i \in G, \mu, \Gamma] = \frac{\int \cdots \int_{G} \psi(x_i | \mu, \Gamma) x_i \, dx_i}{\int \cdots \int_{G} \psi(x_i | \mu, \Gamma) \, dx_i}
\tag{5.1}
\]
and
\[
\mathbb{E}[(X_i - \mu)(X_i - \mu)^{'} | X_i \in G, \mu, \Gamma] = \frac{\int \cdots \int_{G} \psi(x_i | \mu, \Gamma) (x_i - \mu)(x_i - \mu)^{'} \, dx_i}{\int \cdots \int_{G} \psi(x_i | \mu, \Gamma) \, dx_i}.
\tag{5.2}
\]

There are two different approaches to simulating (5.1) and (5.2). First of all draws from the multivariate truncated normal distribution can be used for unbiased estimation of these expectations. This is referred to as truncated simulation. As simulating directly from the truncated normal can, however, be rather difficult and time consuming censored simulation is often used instead.
In this case the numerator and the denominator of (5.1) and (5.2) are simulated separately using unbiased estimators. So this would result in simulating \( \int \cdots \int_G \psi(x_i \mid \mu, \Gamma) x_i \, dx_i \), \( \int \cdots \int_G \psi(x_i \mid \mu, \Gamma) (x_i - \mu)(x_i - \mu)^t \, dx_i \), and \( \int \cdots \int_G \psi(x_i \mid \mu, \Gamma) \, dx_i \). Note that the resulting estimate may be biased, as unbiased estimation of \( \int \cdots \int \psi(x_i \mid \mu, \Gamma) \, dx_i \) does not imply unbiased estimation of \( 1/\int \cdots \int \psi(x_i \mid \mu, \Gamma) \, dx_i \). This means that the MSS estimator is not consistent anymore for fixed \( R \). Hajivassiliou and McFadden (1998), however, establish that there are simulators that, although only asymptotically unbiased, possess biases that vanish at sufficiently fast rates as to guarantee consistency and asymptotic normality of the resulting MSS estimators. Methods to generate censored and truncated simulations are treated in the next sections.

### 5.4.1.1 Censored simulation

In this case the numerator and the denominator of (5.1) and (5.2) are simulated separately. This can be done accurately and fast using the well-known smooth recursive conditioning (SRC) simulator, which is also referred to as the GHK simulator named after Geweke (1991), Hajivassiliou and McFadden (1990) and Kean (1994). This simulator is convenient as it has been investigated thoroughly and its properties are well known.

If \( X_i \sim N^T(\mu, \Sigma) \) and \( 1 \leq BX_i \leq u \), then \( BX_i \sim N^T(B\mu, B\Sigma B^t) \) if we assume that \( B \) is nonsingular. Factorise \( B\Sigma B^t \) using the Cholesky decomposition into \( \Omega\Omega^t \), where \( \Omega \) is a lower-triangular matrix. This means that

\[
\int \cdots \int_{1 \leq BX_i \leq u} \psi(x_i) \, dx_i = \Pr(1 \leq BX_i \leq u), \quad X_i \sim N(\mu, \Sigma) \\
= \Pr(1^* \leq \Omega Z_i \leq u^*), \quad Z_i \sim N(0, 1),
\]

where \( 1^* = 1 - B\mu \) and \( u^* = u - B\mu \). Next draw a sequence of \( Z_j \), for \( j = 1, \ldots, k \), from the truncated standard normal distribution

\[
draw Z_1 \text{ from } N(0, 1) \text{ s.t. } l_1^* \leq \omega_{11} z_1 \leq u_1^* \\
\vdots \\
draw Z_k \text{ from } N(0, 1) \text{ s.t. } l_k^* \leq \omega_{k1} z_1 + \cdots + \omega_{kk} z_k \leq u_k^*,
\]
where $\omega_{ij}$, $i, j = 1, \ldots, k$, denotes the $ij$th element of $\Omega$.

Then

$$Q_1 = \Pr \left( \frac{l^*_1}{\omega_{11}} \leq z_1 \leq \frac{u^*_1}{\omega_{11}} \right)$$

$$\vdots$$

$$Q_k(z_1, \ldots, z_{k-1}) = \Pr \left( \frac{l^*_j - \sum_{j=1}^{k-1} \omega_{kj} z_j}{\omega_{kk}} \leq z_k \leq \frac{u^*_j - \sum_{j=1}^{k-1} \omega_{kj} z_j}{\omega_{kk}} \right).$$

The GHK simulator then becomes

$$\hat{\Pr}(X_i \in G) = \frac{1}{R} \sum_{i=1}^{R} \prod_{j=1}^{k} Q_{ij}(z_1, \ldots, z_{j-1})$$

with $R$ the number of replications. This simulation technique provides asymptotically unbiased estimators of the score. Hajivassiliou and McFadden (1998) established that if the number of simulations used grows faster than the square root of the number of observations the resulting estimator is consistent uniformly asymptotically normal.

We will, however, not be able to make use of the GHK simulator. First of all, the restriction matrix $B$ on the data needs to be square and nonsingular. This means that we can incorporate at most $k$ inequality restrictions, which is much too restrictive for economic data. Often the economic data items need to satisfy non-negativity constraints as well as other types of inequality restrictions, which leads to a number of inequalities that exceeds $k$. Furthermore, the GHK simulator can only be applied if $B$ is nonsingular. When $B$ is singular, $B\Sigma B'$ will be positive semidefinite and consequently $\Omega$ will be positive semidefinite as well, which means that some of the $\omega_{ij}$, $i, j = 1, \ldots, k$, $j \leq i$ will be equal to zero and in that instance $Q_j$ cannot be calculated. Consequently the type of inequality constraints that is allowed is restricted, whereas we are looking for a procedure that can deal with different types of inequality restrictions simultaneously.

Secondly, we would like to incorporate balance restrictions as well later on. In this case the GHK simulator cannot be applied as the upper and lower bounds will be equal, i.e. $l^*_j = u^*_j$ for balance restriction $j$. This means that $Q_j$ becomes zero and the GHK simulator will become meaningless.

Other unbiased, consistent censored simulators that have been suggested, such as importance sampling methods, cannot be used for direct unbiased simulation of the logarithmic score, unless an infinite number of simulations are
averaged (Hajivassiliou and McFadden, 1998). This means that we will not be
able to employ censored simulation.

5.4.1.2 Truncated simulation

Truncated simulation refers to the case where the terms (5.1) and (5.2) are
simulated as a whole by drawing from the truncated normal density, which
provides unbiased estimates. An advantage of this is that the resulting estimator
is consistent for a finite number of simulations. The following simulation scheme
would arise. Let \( g(\cdot | \mu, \Gamma) \) denote the truncated normal density with parameters
\( \mu \) and \( \Gamma \) \((= \Sigma^{-1})\). If we draw \( V_i \) from the truncated normal distribution:
\( V_i \sim \mathcal{N}_T(\mu, \Gamma^{-1}) \), \( V_i \in G \), then

\[
\hat{E}[X_i | X_i \in G, \mu, \Gamma] = \frac{1}{R} \sum_{i=1}^{R} V_i
\]

\[
\hat{E}[(X_i - \mu)(X_i - \mu)\prime | X_i \in G, \mu, \Gamma] = \frac{1}{R} \sum_{i=1}^{R} (V_i - \mu)(V_i - \mu)\prime.
\]

There are several methods available to generate random draws from the
truncated normal distribution. Acceptance/Rejection sampling can be used to
obtain independent draws and Markov chain Monte Carlo techniques constitute
a dependent sample.

Acceptance/Rejection (AR) sampling

In this instance an auxiliary density is determined which can be easily sampled
from. Each draw is either accepted or rejected based on some acceptance rule.
The accepted draws constitute the AR sample.

For truncated distributions AR sampling is straightforwardly applied by
drawing from the nontruncated distributions and rejecting values that fall out-
side the admissible region. For the truncated normal this algorithm will work
well if \( \mu \) lies near the centre of the region and the region is sufficiently large.
However, if \( \mu \) lies near the boundaries of the region or if the region is relatively
small, the number of rejections can become reasonably large and consequently
the algorithm will be inefficient.

As we mentioned before, the same underlying random variates need to be
used in each iteration of the optimisation algorithm in order to avoid conver-
gence issues. Obtaining the same underlying random variates in the case of AR
sampling can be achieved as follows. Let \( \Phi \Phi' \) be the Cholesky decomposition
of $\Sigma$. Then draws from $X_i \sim \mathcal{N}(\mu, \Sigma)$ can be generated using $X_i = \Phi Z_i + \mu$, where $Z_i \sim \mathcal{N}(0, I)$. Afterwards the draws of $X_i$ that fall outside $G$ are discarded.

Fortunately, this method can still be applied when we are confronted with balance restrictions. Instead of generating draws from the multivariate normal distribution, the multivariate singular normal distribution will be used as an auxiliary density. This means that $\Phi$ will become positive semidefinite, which poses no problems.

**Gibbs sampling**

Another popular approach to generate truncated draws is to use MCMC methods. This means that instead of drawing independently we will generate dependent draws and consequently the number of draws strongly increases in order to obtain similar accuracy.

An interesting MCMC method is Gibbs sampling, which was discussed in chapter 2. The Gibbs sampler uses draws from the univariate conditional truncated normals to construct draws from the multivariate truncated normal. The Gibbs sampler is composed by

\[
\begin{align*}
\text{draw} & \quad X_1^{(t)} \sim g(x_1 | x_2^{(t-1)}, \ldots, x_k^{(t-1)}, \theta_1) \\
\text{draw} & \quad X_2^{(t)} \sim g(x_2 | x_1^{(t)}, x_3^{(t-1)}, \ldots, x_k^{(t-1)}, \theta_2) \\
& \quad \vdots \\
\text{draw} & \quad X_k^{(t)} \sim g(x_k | x_1^{(t)}, \ldots, x_{k-1}^{(t)}, \theta_k),
\end{align*}
\]

where $g$ denotes the univariate conditional truncated normal density and $\theta_j = (\mu_{j,-j}, \Sigma_{jj,-j})$. The term $-j$ refers to the set of $k$ variables in the model with the $j$th element removed. After $t_G$ such iterations, we have a draw $X_i^{(t_G)} = (X_1^{(t_G)}, \ldots, X_k^{(t_G)})$, which is referred to as a terminal simulation. Then a sample is created of size $R$, which is the number of terminal simulations drawn.

To make sure the same underlying random variates are used draw $U$ from $U(0, 1)$, then $Z = \Phi^{-1}((\Phi(u^*) - \Phi(l^*))U + \Phi(l^*)) \sim \mathcal{N}(0, 1)$, with $l^* \leq Z \leq u^*$, where $l^* = 1/\sigma(l - \mu)$ and $u^* = 1/\sigma(u - \mu)$. Consequently, $Y = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$, with $l \leq Y \leq u$.

The Gibbs sampler is asymptotically unbiased and the resulting MSS estimator will be consistent uniformly asymptotically normal if the number of resamplings used grows with the rate $\ln n$, where $n$ is the sample size (Hajivassiliou and McFadden, 1998).
In the presence of balance restrictions this method needs to be applied with a slight modification. As in this instance some variables can be derived with certainty from the other variables, the Gibbs sampler will remain stuck in the initial values if all variables are present in the sampler. This can be solved by leaving out one variable in each balance restriction, such that the remaining data does not contain singularities. This will be explained more thoroughly in section 5.7.3.

Example 5.4.1.2. Acceptance/Rejection versus Gibbs sampling.

In summary, we can only make use of truncated sampling for the simulation of the integrals given in equation (5.1) and (5.2) that will be used for maximum likelihood estimation. A truncated sample can be obtained using either Acceptance/Rejection or Gibbs sampling. In this example both methods will be employed in order to get some insight in the accuracy of the estimates for a given sample size.

Suppose that \( \mathbf{X}_i \) follows a truncated normal distribution with parameters

\[
\mu = \begin{pmatrix} 2 \\ 3 \\ 6 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 4 & 3 \\ 2 & 3 & 5 \end{pmatrix},
\]

where \( \mathbf{X}_i \) is truncated to the positive orthant, that is \( X_{i,1}, X_{i,2}, X_{i,3} \geq 0 \). Note that in this example \( \Sigma \) is nonsingular. Let \( t = \mathbb{E}[\mathbf{X}_i \mid \mathbf{X}_i \in \mathcal{G}, \mu, \Gamma] \) and \( T = \mathbb{E}[(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^T \mid \mathbf{X}_i \in \mathcal{G}, \mu, \Gamma] \) denote the quantities that will be estimated using samples obtained by AR sampling and Gibbs sampling, while varying the sample size. Due to the stochastic nature of both methods we generate 25 samples for each sample size.

The AR sample is generated using a 3-dimensional nontruncated normal distribution and rejecting those values that fall outside the positive orthant and the Gibbs sample is based on draws from univariate truncated normals, with the number of resamplings \( t_G = 5 \). The true values and the estimates of equations (5.1) and (5.2) for both methods are presented in Table 5.1, the standard errors of the estimates are given between brackets.

In this instance, the AR sample can be generated quite fast as a relatively small part of the random draws is rejected and therefore the AR algorithm will be efficient. Both the AR and the Gibbs sample provide accurate estimates of \( t \) and \( T \), which means that the sampler that is fastest should be used.

Note that the standard errors decrease with approximately the factor \( 1/\sqrt{10} \) as the sample sizes increase with a factor 10, which was discussed in chapter 2.
Table 5.1: *Estimation of equations (5.1) and (5.2) using AR and Gibbs sampling.*

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>AR (1)</th>
<th>AR (10)</th>
<th>AR (100)</th>
<th>Gibbs (1)</th>
<th>Gibbs (10)</th>
<th>Gibbs (100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>2.625</td>
<td>2.605</td>
<td>2.631</td>
<td>2.628</td>
<td>2.639</td>
<td>2.621</td>
<td>2.627</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.159)</td>
<td>(0.049)</td>
<td>(0.013)</td>
<td>(0.152)</td>
<td>(0.046)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>$t_2$</td>
<td>3.463</td>
<td>3.453</td>
<td>3.466</td>
<td>3.464</td>
<td>3.471</td>
<td>3.441</td>
<td>3.461</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.172)</td>
<td>(0.054)</td>
<td>(0.014)</td>
<td>(0.167)</td>
<td>(0.052)</td>
<td>(0.015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.243)</td>
<td>(0.062)</td>
<td>(0.021)</td>
<td>(0.234)</td>
<td>(0.055)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>$T_{11}$</td>
<td>2.905</td>
<td>2.928</td>
<td>2.920</td>
<td>2.914</td>
<td>2.912</td>
<td>2.903</td>
<td>2.919</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.423)</td>
<td>(0.158)</td>
<td>(0.048)</td>
<td>(0.522)</td>
<td>(0.149)</td>
<td>(0.044)</td>
</tr>
<tr>
<td>$T_{12}$</td>
<td>1.390</td>
<td>1.359</td>
<td>1.411</td>
<td>1.390</td>
<td>1.422</td>
<td>1.353</td>
<td>1.400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.355)</td>
<td>(0.088)</td>
<td>(0.039)</td>
<td>(0.354)</td>
<td>(0.113)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>$T_{13}$</td>
<td>1.409</td>
<td>1.376</td>
<td>1.404</td>
<td>1.411</td>
<td>1.454</td>
<td>1.360</td>
<td>1.413</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.459)</td>
<td>(0.123)</td>
<td>(0.043)</td>
<td>(0.362)</td>
<td>(0.086)</td>
<td>(0.039)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.443)</td>
<td>(0.132)</td>
<td>(0.054)</td>
<td>(0.466)</td>
<td>(0.125)</td>
<td>(0.049)</td>
</tr>
<tr>
<td>$T_{22}$</td>
<td>2.401</td>
<td>2.367</td>
<td>2.394</td>
<td>2.401</td>
<td>2.462</td>
<td>2.354</td>
<td>2.404</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.446)</td>
<td>(0.107)</td>
<td>(0.043)</td>
<td>(0.456)</td>
<td>(0.147)</td>
<td>(0.038)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.619)</td>
<td>(0.159)</td>
<td>(0.054)</td>
<td>(0.752)</td>
<td>(0.223)</td>
<td>(0.064)</td>
</tr>
</tbody>
</table>

using the Central Limit Theorem.

Since $X_i$ is truncated from below its expected value will be larger than $\mu$ and its expected variance will be smaller than $\Sigma$, which we already observed for the univariate case in figure 5.1. These differences will increase as $\mu$ approaches the boundaries of $G$.  

5.5 The EM algorithm applied to truncated normal data

For maximum likelihood estimation in the presence of nonresponse we will use the EM algorithm. Recall that the E-step of the EM algorithm consists of es-
5.5. The EM algorithm applied to truncated normal data

timating the expected complete data loglikelihood \( Q(\theta \mid \theta^{(t)}) \) defined as
\[
Q(\theta \mid \theta^{(t)}) = E[\ell(\theta \mid \mathbf{X}_{\text{mis}}, \mathbf{X}_{\text{obs}}) \mid \mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}}, \theta^{(t)}] = \int \cdots \int_{G(\mathbf{x}_{\text{obs}})} \ell(\theta \mid \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}) h(\mathbf{x}_{\text{mis}} \mid \mathbf{x}_{\text{obs}}, \theta^{(t)}) d\mathbf{x}_{\text{mis}}.
\]
The M-step then consists of maximising \( Q(\theta \mid \theta^{(t)}) \) with respect to \( \theta \) to obtain an update \( \theta^{(t+1)} \). If an analytic solution to \( Q(\theta \mid \theta^{(t)}) \) cannot be found, this quantity can be estimated unbiasedly by
\[
\hat{Q}(\theta \mid \theta^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} \ell(\theta \mid \mathbf{X}_{\text{mis}}^{(m)}, \mathbf{x}_{\text{obs}}),
\]
where \( \mathbf{X}_{\text{mis}}^{(m)}, m = 1, \ldots, M \) are draws from the conditional distribution of the missing data given the observed data and the current value of the parameter vector; \( h(\cdot \mid \mathbf{X}_{\text{obs}} = \mathbf{x}_{\text{obs}}, \theta^{(t)}) \). This is known as the Monte Carlo EM (MCEM) algorithm and was developed by Wei and Tanner (1990).

For the truncated normal the EM algorithm would result in iteratively calculating the expected score functions and finding \( \mu^{(t+1)} \) and \( \Gamma^{(t+1)} = (\Sigma^{(t+1)})^{-1} \) for which \( \nabla_{\mu} Q(\mu, \Gamma \mid \mu^{(t)}, \Gamma^{(t)}) = 0 \) and \( \nabla_{\Gamma} Q(\mu, \Gamma \mid \mu^{(t)}, \Gamma^{(t)}) = 0 \), until the parameter estimates converge. As the area \( G \) is independent of the parameters \( \mu \) and \( \Gamma \), this means that the updated \( \mu^{(t+1)} \) and \( \Gamma^{(t+1)} \) are the parameters for which
\[
\nabla_{\mu} Q(\mu, \Gamma \mid \mu^{(t)}, \Gamma^{(t)}) = \Gamma \sum_{i=1}^{n} E[\mathbf{X}_i \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] + \\
- \Gamma E[\mathbf{X}_i \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] \\
\nabla_{\Gamma} Q(\mu, \Gamma \mid \mu^{(t)}, \Gamma^{(t)}) = - \frac{n}{2} E[\mathbf{S} \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] + \\
+ \frac{n}{2} E[(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)'] \mid \mathbf{X}_i \in G, \mu^{(t)}, \Gamma^{(t)}]
\]
are equal to zero. So, similar to the EM algorithm for nontruncated normal data, we need to calculate \( E[\mathbf{X}_i \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] \) and \( E[\mathbf{X}_i \mathbf{X}_i' \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] \). Note that for each record \( i \)
\[
E[\mathbf{X}_i \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] = E[(\mathbf{X}_{i, \text{mis}}, \mathbf{X}_{i, \text{obs}}) \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}] \\
= (E[\mathbf{X}_{i, \text{mis}} \mid \mathbf{X}_{i, \text{obs}} = \mathbf{x}_{i, \text{obs}}, \mu^{(t)}, \Gamma^{(t)}], \mathbf{x}_{i, \text{obs}}')
\]
and
\[
\mathbb{E}[X_i, X'_i | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}] = \mathbb{E}\left[\begin{pmatrix} X_{i, mis}^t \\ X_{i, obs}^t \end{pmatrix} | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}\right] \\
= \mathbb{E}\left[\begin{pmatrix} X_{i, mis}^t \\ x_{i, obs} \end{pmatrix} | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}\right]
\]

In this case
\[
\mathbb{E}[X_{i, mis} | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}] = \int \cdots \int_{G(x_{i, obs})} x_{i, mis} h(x_{i, mis} | x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}) \, dx_{i, mis}
\]

and
\[
\mathbb{E}[X_{i, mis} X_{i, mis}^t | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}] = \int \cdots \int_{G(x_{i, obs})} x_{i, mis} x_{i, mis}^t h(x_{i, mis} | x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}) \, dx_{i, mis}.
\]

These quantities can be estimated through Monte Carlo integration, that is
\[
\hat{\mathbb{E}}[X_{i, mis} | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}] = \frac{1}{M} \sum_{m=1}^{M} V^{(m)}_{i, mis} (5.3)
\]
\[
\hat{\mathbb{E}}[X_{i, mis} X_{i, mis}^t | X_{i, obs} = x_{i, obs}, \mu^{(t)}, \Gamma^{(t)}] = \frac{1}{M} \sum_{m=1}^{M} V^{(m)}_{i, mis} (V^{(m)}_{i, mis})^t (5.4)
\]

for \( V_{i, mis} | X_{i, obs} = x_{i, obs} \sim N_T(\mu^{(t)}_{mis, obs}, \Sigma^{(t)}_{mis, mis, obs}) \) with \( V_{i, mis} \in G(x_{i, obs}) \) and
\[
\mu^{(t)}_{mis, obs} = \mu^{(t)}_{mis} + \Sigma^{(t)}_{mis, obs} (\Sigma^{(t)}_{obs, obs})^{-1} (x_{i, obs} - \mu^{(t)}_{obs}) \\\n\Sigma^{(t)}_{mis, mis, obs} = \Sigma^{(t)}_{mis, mis} - \Sigma^{(t)}_{mis, obs} (\Sigma^{(t)}_{obs, obs})^{-1} \Sigma^{(t)}_{obs, mis}.
\]

Random draws for \( V_{i, mis} \) can be obtained through either AR or Gibbs sampling methods, which were treated in section 5.4.1.2.

Note that as the missingness varies across observations, the equations (5.3) and (5.4) need to be calculated for each record separately. The process will therefore be quite time consuming. Fortunately, as Wei and Tanner (1990) state that
in the initial stages of the algorithm the number of drawings $M$ may be small, increasing $M$ once the process seems to converge, which will reduce computer time.

### 5.6 Imputation of missing data items

Once the parameters of the truncated normal are estimated, these quantities can be used for imputation purposes. Let $\hat{\mu}$ and $\hat{\Sigma}$ denote the maximum likelihood estimates obtained through the EM algorithm. The missing data items $X_{i,mis}$ can now be imputed either stochastically by drawing from: $\hat{X}_{i,imp} | \hat{X}_{i,obs} = x_{i,obs} \sim \mathcal{N}^T(\hat{\mu}_{mis,obs}; \hat{\Sigma}_{mis,obs})$, $\hat{X}_{i,imp} \in G(x_{i,obs})$ or deterministically by retaining the estimated values for (5.3) obtained in the E-step of the EM algorithm. So $\hat{X}_{i,imp} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{V}^{(m)}_{i,mis}$.

### 5.7 Handling balance and inequality restrictions simultaneously

Now that we have developed an imputation method that can handle any type of linear inequality restriction we would like to incorporate linear balance restrictions as well, as economic data usually consist of a combination of both types of restrictions.

Let

$$ R = \begin{pmatrix} B \\ A \end{pmatrix} $$

denote an $(r + p) \times k$ restriction matrix that contains both inequality as well as balance restrictions. This means that the data now need to lie in $H := \{ X_i \in \mathbb{R}^k : l \leq RX_i \leq u \}$. For balance restriction $j$, $j = r + 1, \ldots, r + p$ it holds that $l_j = u_j$. This type of data can be modelled with a truncated singular normal distribution.

#### 5.7.1 The truncated singular normal distribution

Assume that $X_i$ is distributed according to a truncated singular normal distribution, that is $X_i \sim \mathcal{N}^T(\mu, \Sigma)$, with $l \leq BX_i \leq u$ and where $\Sigma$ is singular with rank $q$ ($= k - p$). This density is defined on a subspace of $\mathbb{R}^k$, as was the case for the nontruncated singular normal. The situation is illustrated graphically
in a 3-dimensional setting in Figure 5.2, where the contour lines of the normal density are plotted in the region $H$.

The singular covariance matrix $\Sigma$ can be decomposed by means of an eigenvalue decomposition into $\mathbf{C} \Lambda \mathbf{C}^t$, where $\mathbf{C}$ is the orthogonal matrix of eigenvectors of $\Sigma$ and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_q, 0, \ldots, 0\}$ is the diagonal matrix of eigenvalues. Again let $\Lambda_1 = \text{diag}\{\lambda_1, \ldots, \hat{\lambda}_j, \ldots\}$ be the matrix of nonzero eigenvalues. Partition $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2)$, where $\mathbf{C}_2$ is the matrix of eigenvectors that corresponds to the zero eigenvalues of $\Sigma; \Sigma \mathbf{C}_2 = \mathbf{0}$.

Recall that the density function of the singular normal is

$$\varphi(x_i | \mu, \Sigma) = (2\pi)^{-n/2} (\prod_{j=1}^q \lambda_j)^{-1/2} \exp \left( -\frac{1}{2} (x_i - \mu)' \Sigma^+ (x_i - \mu) \right), \text{ for } x_i \in \mu + L^\perp,$$

where $\prod_{j=1}^q \lambda_j = \det(\Lambda_1)$ and $\Sigma^+ = \Lambda_1 \Lambda_1^{-1} \Lambda_1^t$. This means that the truncated singular normal density is defined as

$$g(x_i | \mu, \Sigma) = \begin{cases} 0 & x_i \notin H \\ \frac{\varphi(x_i | \mu, \Sigma)}{\int_0 \varphi(x_i | \mu, \Sigma) \, dx} & x_i \in H \end{cases}.$$
5.7.2 Maximum likelihood estimation for truncated singular normal data

Similar to the results found for maximum likelihood estimation of the singular normal, we will show in this section that the singular part of the data does not influence maximum likelihood estimation of the truncated singular normal. For this purpose one variable in each balance restriction will be left out, such that the remaining data does not contain singularities.

Create
\[
A_c = \begin{pmatrix}
    A \\
    A^*
\end{pmatrix},
\]
where \(A^*\) is a \(q \times k\) matrix chosen such that \(A_c\) is nonsingular. Partition the restriction matrix \(A\) into \(A = (A_p \ A_q)\), where \(A_p\) is a \(p \times p\) nonsingular matrix, which always exists as \(\text{rank}(A) = p\). Now partition \(A^*\) accordingly: \(A^* = (A_p^* \ A_q^*)\), where \(A_p^*\) is a \(q \times p\) matrix. Taking \(A_p^* = 0\) and \(A_q^* = I_q\) results in a nonsingular \(A^c\) as

\[
det(A^c) = det\left( \begin{pmatrix}
    A_p & A_q \\
    0 & I_q
\end{pmatrix}\right) \\
= det(A_p) \det(I_q - 0A_p^{-1}A_q) \\
= det(A_p) \neq 0.
\]

Now partition \(B, X_i, \mu\) and \(\Sigma\) accordingly. Using the fact that multiplying a truncated normal variable with a nonsingular matrix results in a truncated normal variable as well, we find that

\[
A_cX_i = \begin{pmatrix}
    AX_i \\
    A^*X_i
\end{pmatrix} \sim \mathcal{N}^T\left( \begin{pmatrix}
    A\mu \\
    A^*\mu
\end{pmatrix}, \begin{pmatrix}
    A\Sigma A' & A\Sigma(A^*)' \\
    A^*\Sigma A' & A^*\Sigma(A^*)'
\end{pmatrix}\right),
\]
where \(1 \leq BX_i \leq u\). As \(A\mu = 0\) and \(\Sigma A' = 0\) (see chapter 4), it follows that

\[
\begin{pmatrix}
    AX_i \\
    A^*X_i
\end{pmatrix} \sim \mathcal{N}^T\left( \begin{pmatrix}
    0 \\
    0
\end{pmatrix}, \begin{pmatrix}
    0 & 0 \\
    0 & \Sigma(A^*)'
\end{pmatrix}\right), \text{ with } 1 \leq BX_i \leq u.
\]

So \(AX_i = 0\) with probability one and therefore \(X_{i,p} = -A_p^{-1}A_qX_{i,q}\). Furthermore \(A^*X_i \sim \mathcal{N}^T(A^*\mu, A^*\Sigma(A^*)')\). Using the fact that \(A^* = (0 \ I_q)\) we find that \(X_{i,q} \sim \mathcal{N}^T(\mu_q, \Sigma_{qq})\), with \(1 \leq (B_q - B_pA_p^{-1}A_q)X_{i,q} \leq u\). This means that the maximum likelihood estimates for \(\mu_q\) and \(\Gamma_{qq} (= \Sigma_{qq}^{-1})\) can be obtained by
setting

\[ \nabla_{\mu_q} \ell(\mu_q, \Gamma_{qq} | x_q) = \Gamma_{qq} \left( \sum_{i=1}^{n} x_{i,q} - n E[X_{i,q} | X_{i,q} \in H_q, \mu_q, \Gamma_{qq}] \right) \]

\[ \nabla_{\Gamma_{qq}} \ell(\mu_q, \Gamma_{qq} | x_q) = -\frac{1}{2} \sum_{i=1}^{n} (x_{i,q} - \mu_q)(x_{i,q} - \mu_q)' + \frac{n}{2} E[(x_{i,q} - \mu_q)(x_{i,q} - \mu_q)'] | X_{i,q} \in H_q, \mu_q, \Gamma_{qq}] \]

equal to zero, where \( H_q := \{ Y \in \mathbb{R}^q : 1 \leq (B_q - B_p A_p^{-1} A_q) Y \leq u \} \). Once \( \hat{\mu}_q \) and \( \hat{\Sigma}_{qq} (= \hat{\Sigma}_{qq}^{-1}) \) are found, \( \hat{\mu} \) and \( \hat{\Sigma} \) can be derived by using the fact that \( A \hat{\mu} = A_p \hat{\mu}_p + A_q \hat{\mu}_q = 0 \) and

\[ \hat{\Sigma} A' = \left( \begin{array}{cc} \hat{\Sigma}_{pp} & \hat{\Sigma}_{pq} \\ \hat{\Sigma}_{qp} & \hat{\Sigma}_{qq} \end{array} \right) \left( \begin{array}{c} A_p' \\ A_q' \end{array} \right) = 0. \]

So

\[ \hat{\mu}_p = -A_p^{-1} A_q \hat{\mu}_q \]

\[ \hat{\Sigma}_{qp} = -\hat{\Sigma}_{qq} A_q' (A_p')^{-1} \]

\[ \hat{\Sigma}_{pq} = -A_p^{-1} A_q \hat{\Sigma}_{qq} \]

\[ \hat{\Sigma}_{pp} = -A_p^{-1} A_q \hat{\Sigma}_{qq} A_q' (A_p')^{-1}. \]

This means that the estimates for \( \mu_p, \Sigma_{qp}, \Sigma_{pq} \) and \( \Sigma_{pp} \) can be based completely on the nonsingular part of the data and consequently the singular data will yield the same maximum likelihood estimates as the nonsingular data.

### 5.7.3 The EM algorithm applied to truncated singular normal data

If the data contain missing items the situation alters, as the vector containing the redundant variables, \( X_{i,p} \), may now be partly or completely observed and \( X_{i,q} \) may contain missing items. In that case the vector \( X_{i,p} \) contains information about \( X_{i,q} \), which is not present in \( X_{i,q} \) and consequently \( X_{i,p} \) is not completely redundant anymore. Clearly, we need to use this information as well for the maximum likelihood estimation of \( \mu \) and \( \Sigma \). This can be done as follows.
The E-step consists of calculating $Q(\mu, \Sigma \mid \mu^{(t)}, \Sigma^{(t)})$

$$Q(\mu, \Sigma \mid \mu^{(t)}, \Sigma^{(t)})$$

$$= E[\ell(\mu, \Sigma \mid X_{\text{mis}}, x_{\text{obs}}) \mid X_{\text{obs}} = x_{\text{obs}}, \mu^{(t)}, \Sigma^{(t)}]$$

$$= -\frac{nq}{2} \ln 2\pi - \frac{n}{2} \ln |\Lambda_1| - \frac{1}{2} \sum_{i=1}^{n} E[(X_i - \mu)'\Sigma^+(X_i - \mu)] +$$

$$-n \ln \int \cdots \int_H \varphi(x_i \mid \mu, \Sigma) dx_i. \quad (5.5)$$

As $\Sigma^+ = C_1 \Lambda^{-1}_k C_1'$, this expected loglikelihood actually consists of three parameters; $\mu, \Lambda_1$ and $C_1$. So in order to obtain updated parameter values for $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$, we need to differentiate (5.5) with respect to $\mu, \Lambda_1^{-1}$ and $C_1$ and set these first order conditions equal to zero. As the integration region $H$ is independent of the parameters, these first order conditions become

$$\nabla_\mu Q(\mu, \Lambda_1, C_1 \mid \mu^{(t)}, \Lambda_1^{(t)}, C_1^{(t)})$$

$$= \Sigma^+ \sum_{i=1}^{n} E[X_i \mid X_{i,\text{obs}} = x_{i,\text{obs}}, \mu^{(t)}, \Sigma^{(t)}] +$$

$$-n \Sigma^+ E[X_i \mid X_i \in H, \mu^{(t)}, \Sigma^{(t)}]$$

$$\nabla_{\Lambda_1} Q(\mu, \Lambda_1, C_1 \mid \mu^{(t)}, \Lambda_1^{(t)}, C_1^{(t)})$$

$$= -\frac{1}{2} \sum_{i=1}^{n} E[C_1'(X_i - \mu)(X_i - \mu)'C_1 \mid X_{i,\text{obs}} = x_{i,\text{obs}}, \mu^{(t)}, \Sigma^{(t)}] +$$

$$+\frac{n}{2} E[C_1'(X_i - \mu)(X_i - \mu)'C_1 \mid X_i \in H, \mu^{(t)}, \Sigma^{(t)}]$$

$$\nabla_{C_1} Q(\mu, \Lambda_1, C_1 \mid \mu^{(t)}, \Lambda_1^{(t)}, C_1^{(t)})$$

$$= -\sum_{i=1}^{n} E[(X_i - \mu)(X_i - \mu)'C_1 \Lambda_1^{-1} \mid X_{i,\text{obs}} = x_{i,\text{obs}}, \mu^{(t)}, \Sigma^{(t)}] +$$

$$+n E[(X_i - \mu)(X_i - \mu)'C_1 \Lambda_1^{-1} \mid X_i \in H, \mu^{(t)}, \Sigma^{(t)}].$$

This means that for each record we need to estimate the following quantities

$$E[X_{i,\text{mis}} \mid X_{i,\text{obs}} = x_{i,\text{obs}}, \mu^{(t)}, \Sigma^{(t)}]$$

$$= \int \cdots \int_H x_{i,\text{mis}} g(x_{i,\text{mis}} \mid x_{i,\text{obs}}, \mu^{(t)}, \Sigma^{(t)}) dx_{i,\text{mis}}$$
Gibbs sampling

In the case of AR sampling (AR sampling can be generated using the truncated singular
normal distribution, where the singular matrix is decomposed by means of an eigenvalue decomposition into CAC. Draws of are accepted
which can be done by generating a draw from the truncated singular normal

where

is singular and

is singular.

Then draw

$$V_{\text{max}} \sim \mathcal{N}(\mu_{\|\mu\|}, \Sigma_{\|\mu\|})$$

for

and

where

$$\Sigma_{\|\mu\|} = \Sigma_{\|\mu\|}$$

with

$$V_{\text{max}} \in \mathcal{N}(\mu_{\|\mu\|}, \Sigma_{\|\mu\|})$$

and

$$\Sigma_{\|\mu\|}$$

The

is not contained in the data.

First determine

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and remove any redundant rows from

$$\Sigma_{\|\mu\|}$$

in order to obtain

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and

$$\Sigma_{\|\mu\|}$$

where

$$\Sigma_{\|\mu\|}$$

and

$$\Sigma_{\|\mu\|}$$

are not contained in the data.

Then determine

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and remove any redundant rows from

$$\Sigma_{\|\mu\|}$$

in order to obtain

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and

$$\Sigma_{\|\mu\|}$$

where

$$\Sigma_{\|\mu\|}$$

and

$$\Sigma_{\|\mu\|}$$

are not contained in the data.

Then determine

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and remove any redundant rows from

$$\Sigma_{\|\mu\|}$$

in order to obtain

$$A = \text{rank}(\Sigma_{\|\mu\|})$$

and

$$\Sigma_{\|\mu\|}$$

where

$$\Sigma_{\|\mu\|}$$

and

$$\Sigma_{\|\mu\|}$$

are not contained in the data.
where

\[ 1^* \leq (B_{i, \text{mis}, q_i} - B_{i, \text{mis}, p_i} (A_{i, \text{mis}, p_i}^n)^{-1} A_{i, \text{mis}, q_i}^n) V_{i, \text{mis}, q_i} \leq u^* , \]

with

\[ 1^* = 1 - B_{i, \text{obs}} x_{i, \text{obs}} + B_{i, \text{mis}, p_i} (A_{i, \text{mis}, p_i}^n)^{-1} A_{i, \text{obs}} x_{i, \text{obs}} \]

and

\[ u^* = u - B_{i, \text{obs}} x_{i, \text{obs}} B_{i, \text{mis}, p_i} (A_{i, \text{mis}, p_i}^n)^{-1} A_{i, \text{obs}} x_{i, \text{obs}} , \]

making use of the Gibbs sampler. Next calculate \( V_{i, \text{mis}, p_i} \) based on the generated sample and using the fact that \( V_{i, \text{mis}, p_i} = -(A_{i, \text{mis}, p_i}^n)^{-1} A_{i, \text{mis}, q_i} V_{i, \text{mis}, q_i} \).

Note that this method can be time-consuming as \( A_{i, \text{mis}, p_i}^n \) needs to be determined for each missing data pattern separately.

Some additional issues that arise in practice for maximum likelihood estimation, such as the fact that \( \Sigma \) needs to be positive definite and the use of a step size in the Fisher scoring algorithm, are discussed in the appendices.

### 5.7.4 Imputation of missing data items

Once the parameters of the truncated singular normal are estimated with the EM algorithm, the missing data items can be imputed. Let \( \hat{\mu} \) and \( \hat{\Sigma} \) denote the maximum likelihood estimates. The missing data items \( X_{i, \text{mis}} \) can be imputed either stochastically by drawing from the truncated singular normal \( \hat{X}_{i, \text{imp}} | X_{i, \text{obs}} = x_{i, \text{obs}} \sim N^T(\hat{\mu}_{\text{mis}, \text{obs}}, \hat{\Sigma}_{\text{mis}, \text{mis}, \text{obs}}) \), \( \hat{X}_{i, \text{imp}} \in H(x_{i, \text{obs}}) \) or deterministically by retaining the estimated values that were obtained in the E-step of the EM algorithm. So \( \hat{X}_{i, \text{imp}} = \frac{1}{M} \sum_{m=1}^{M} V_{i, \text{mis}}^{(m)} \).

### 5.8 Imputation performance

Similarly to chapters 3 and 4 we would now like to assess the performance of this imputation method by applying it to empirical data. Unfortunately, the data we have at our disposal displays highly non-normal behaviour and consequently the algorithm does not converge. Histogram plots of the data reveal that the peak of the data may lie at the origin (which is usually the truncation point) or even to the left of the origin. Clearly, this means that the location of the mean parameter cannot be derived from the observed data. In this instance it is therefore impossible to obtain parameter estimates for mean and dispersion. So, a major drawback of this imputation method is that it is highly dependent
on the model assumptions and that the point of truncation needs to lie in the
tail of the distribution.

In addition to this, the dispersion parameters can be much larger than the
mean parameters in economic data, for example it is not uncommon that these
parameters differ with a factor of $10^5$. In this case the expected information
matrix becomes nearly singular, leading to difficulties in the optimisation
algorithm.

In conclusion, the multivariate truncated singular model cannot be applied
to the present economic dataset. In order to get an impression of the imputation
performance of the model, we will investigate the effects of imputation on
parameter estimation for a synthetic dataset.

### 5.8.1 Generation of data and missing values

A sample, $\mathbf{X}$, will be generated from a three-dimensional truncated singular
normal distribution: $\mathbf{X} \sim \mathcal{N}^T(\mathbf{\mu}, \mathbf{\Sigma})$, where $X_1, X_2, X_3 \geq 0$ and $X_1 + X_2 = X_3$.
The data will be generated with the following parameter values

$$\mathbf{\mu} = \begin{pmatrix} 50 \\ 100 \\ 150 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} 1000 & 750 & 1750 \\ 750 & 1500 & 2250 \\ 1750 & 2250 & 4000 \end{pmatrix}.$$ 

The size of the synthetic dataset is chosen close to the sample sizes of the
empirical datasets that were used in the previous chapters, i.e. $n = 600$.

Next, 100 random samples of missing data are drawn from the synthetic
data using Bernoulli draws, where the amount of missingness is approximately
20%. The number of missing values introduced is somewhat smaller than the
number of missing values that was observed in the business surveys. This is
done to reduce computing time, as the EM algorithm consists of calculating
expectations for each record separately, a high amount of missingness will slow
down the EM algorithm. The remaining datasets, containing missing data items,
are used for parameter estimation.

### 5.8.2 The effects of imputation on parameter estimation

The effects of imputation will be investigated with respect to estimation of the
population parameters $\mathbf{\mu}$ and $\mathbf{\Sigma}$. The EM algorithm for truncated singular normal
data, described in section 5.7.3, will be used to obtain parameter estimates
for the datasets with missing values. The complete cases mean and covariance
matrix are used as starting values. The results are shown in Table 5.2. Note
that the estimates for $\mu_3$, $\Sigma_{13}$, $\Sigma_{23}$ and $\Sigma_{33}$ are not given as they can be derived from the other estimates. In Table 5.2 the true values of the parameters are given as well as the estimates that are obtained through a nontruncated estimation procedure, i.e. the sample mean and covariance matrix. The estimates that are obtained if we use maximum likelihood estimation for the truncated singular normal are presented for completely observed data and data with missings, where the standard errors are given between brackets.

As was expected, the (nontruncated) sample variances underestimate the true variances. The truncated estimation procedure for the complete data corrects for this, but as a consequence $\Sigma_{11}$ and $\Sigma_{12}$ are somewhat overestimated. This might be due to the fact that the integrals in the score functions are estimated and therefore are subject to estimation error, which can be solved by increasing the sample on which the estimation is based. This will slow down the algorithm, however.

Note that this effect is reduced for maximum likelihood estimation in the presence of missing data. This is probably caused by the fact that the starting values of the EM algorithm are based on the completely observed cases in the dataset with missings. These variances are likely to be smaller than the sample variances for the complete dataset as they are based on a smaller sample. The correction made by the truncated estimation procedure will consequently result in variance estimates that are closer to the true values.

The estimates obtained by the EM algorithm for truncated singular normal data therefore need to be assessed at their ability to preserve the truncated complete data estimates. Unfortunately, even in the case of synthetic data, that follow a truncated normal distribution, the imputation model appears to have difficulties in providing accurate point estimates for $\Sigma$.

<table>
<thead>
<tr>
<th></th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\Sigma_{11}$</th>
<th>$\Sigma_{12}$</th>
<th>$\Sigma_{22}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>50</td>
<td>100</td>
<td>1000</td>
<td>750</td>
<td>1500</td>
</tr>
<tr>
<td>Nontruncated</td>
<td>56</td>
<td>104</td>
<td>873</td>
<td>661</td>
<td>1350</td>
</tr>
<tr>
<td>Truncated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete data</td>
<td>51</td>
<td>101</td>
<td>1109</td>
<td>843</td>
<td>1503</td>
</tr>
<tr>
<td>Missing data</td>
<td>53</td>
<td>102</td>
<td>978</td>
<td>739</td>
<td>1408</td>
</tr>
<tr>
<td></td>
<td>(1.1)</td>
<td>(1.2)</td>
<td>(61.8)</td>
<td>(55.5)</td>
<td>(53.4)</td>
</tr>
</tbody>
</table>
5.9 Concluding remarks

In this chapter we have discussed the use of the multivariate truncated singular normal distribution to impute data that are subject to both balance and inequality restrictions. First, a procedure to obtain the maximum likelihood estimates for the truncated nonsingular normal distribution, using Monte Carlo methods to estimate high-dimensional integrals, is described. We have shown that the maximum likelihood estimates remain the same for the corresponding singular case. Furthermore, we have developed an application of the (Monte Carlo) EM algorithm for the truncated singular normal distribution.

Although the model is theoretically sound and can incorporate multiple balance restrictions and different types of inequality restrictions, it can not be straightforwardly applied in practice as strong model assumptions need to be met, which are likely not to hold in reality, such as multivariate normality.

This means that it is hard to find a multivariate model that can incorporate all sorts of restrictions on the data with model assumptions that hold in practice. Therefore, in the next chapter, we will investigate the use of univariate conditional distributions to sequentially model and impute the data in order to approximate a joint model.

5.A Positive (semi)definiteness of $\Sigma$

The covariance matrix $\Sigma$ needs to be positive (semi)definite. In practice this may not always be the case in the Fisher scoring algorithm. For example, when one of the variances is near zero, the step size may be such that this variance becomes negative. To prevent this from happening we can use either constrained optimisation or a parameterisation that ensures positive definiteness. Constrained optimisation is, however, rather difficult to implement as the constraints of a positive (semi)definite matrix are difficult to impose during the optimisation process. Therefore we suggest using the parameterisation that $\Sigma = \Phi \Phi^\prime$. It is clear that any $\Sigma$ defined by $\Phi$ is positive (semi)definite. There are several choices available for $\Phi$ that lead to different parameterisations of $\Sigma$. We propose using the Cholesky factorization, which defines $\Phi$ as an lower triangular matrix. Note that, if $\Upsilon = \Phi^{-1}$

$$\Gamma = \Sigma^{-1} = (\Phi \Phi')^{-1} = (\Phi')^{-1} \Phi^{-1} = (\Phi^{-1})' \Phi^{-1} = \Upsilon^\prime \Upsilon.$$ 

In order to determine the first order conditions we need the following identities

$$\ln |\Gamma| = \ln |\Upsilon^\prime \Upsilon| = \ln |\Upsilon^\prime||\Upsilon| = \ln |\Upsilon|^2 = 2 \ln |\Upsilon|$$
and  
\[ \frac{\partial |Y|}{\partial Y} = \frac{\partial \exp(\ln |Y|)}{\partial Y} = \exp(\ln |Y|) \frac{\partial \ln |Y|}{\partial Y} = |Y|(Y')^{-1}. \]

Also note that  
\[ \frac{\partial y'Y'Yy}{\partial Y} = 2Y'y'. \]

The loglikelihood can now be rewritten as  
\[ \ell(\mu, Y | x) = -\frac{nk}{2} \ln 2\pi + n \ln |Y| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)'Y'Y(x_i - \mu) + \]
\[ -n \ln \int \cdots \int \Psi(x_i | \mu, Y) \, dx_i, \]

From this it follows that the first order derivatives become  
\[ \nabla_{\mu} \ell(\mu, Y | x) = Y'Y \left( \sum_{i=1}^{n} x_i - nE[X_i | X_i \in G, \mu, Y] \right) \]
\[ \nabla_{Y} \ell(\mu, Y | x) = -Y' \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)' \]
\[ + nY'E[(X_i - \mu)(X_i - \mu)' | X_i \in G, \mu, Y]. \]

So instead of the unique elements of \( \Gamma \), the lower triangular elements of \( Y \) are estimated, from which \( \Gamma \) can be derived by \( Y'Y \).

5.B The step size of the Fisher scoring algorithm

Although the Fisher scoring method makes sure we are going in the direction of the maximum, the maximal step size that should be taken in that direction is not calculated and therefore the danger arises that we overshoot the maximum and consequently the algorithm may not converge. If this occurs a step size can be introduced that is chosen such that the loglikelihood increases.

In Fisher scoring the parameter estimates are updated as follows  
\[ \left( \begin{array}{c} \mu^{(t+1)} \\ \text{vch}(Y^{(t+1)}) \end{array} \right) = \left( \begin{array}{c} \mu^{(t)} \\ \text{vch}(Y^{(t)}) \end{array} \right) + \]
\[ + a\hat{I}^{-1}(\mu, Y)|_{\mu=\mu^{(t)}, Y=Y^{(t)}} \nabla \ell(\mu, Y | x)|_{\mu=\mu^{(t)}, Y=Y^{(t)}}, \]
where \(a\) denotes the step size, and 
\[
\hat{I}(\mu, \gamma) = \sum_{i=1}^n \nabla \ell_i(\mu, \gamma | x) \nabla \ell_i(\mu, \gamma | x)',
\]
with 
\[
\nabla \ell_i(\mu, \gamma | x) = (\nabla_{\mu} \ell_i(\mu, \gamma | x'), \text{vech}(\nabla_{\gamma} \ell_i(\mu, \gamma | x'))' \text{ and } \nabla \ell(\mu, \gamma | x) = \sum_{i=1}^n \nabla \ell_i(\mu, \gamma | x).
\]
The vech-operator stacks the elements of a matrix that lie above or on the main diagonal.

Let \(\ell^{(t)}\) denote the loglikelihood at iteration \(t\): 
\[
\ell^{(t)} = \ell(\mu^{(t)}, \gamma^{(t)} | x).
\]
Then the difference between loglikelihoods in subsequent iterations is
\[
\Delta(\ell^{(t+1)}, \ell^{(t)}) = \ell^{(t+1)} - \ell^{(t)} - n \ln \frac{\int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t+1)}, \gamma^{(t+1)}) dx_i}{\int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t)}, \gamma^{(t)}) dx_i},
\]
where \(\ell_{nt}\) denotes the loglikelihood function of the nontruncated normal distribution. Now define
\[
R = \frac{\int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t+1)}, \gamma^{(t+1)}) dx_i}{\int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t)}, \gamma^{(t)}) dx_i}.
\]
The term \(\ln R\) cannot be estimated without bias by \(\ln \hat{R}\), as \(\text{E}[\ln R] \neq \ln \text{E}[R]\). However, recall that Jensen’s inequality states that \(\text{E}[g(X)] \leq g(\text{E}[X])\) for a concave function \(g\). This means that
\[
\hat{\Delta}(\ell^{(t+1)}, \ell^{(t)}) = \ell^{(t+1)} - \ell^{(t)} - n \text{E}[\ln R] \geq \ell^{(t+1)} - \ell^{(t)} - n \ln \text{E}[R].
\]
So if the latter equation is non-negative, the first will be non-negative as well and we can therefore use the latter one to calculate the step size. The quantity \(R\) can be estimated unbiasedly using importance sampling techniques
\[
R = \int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t+1)}, \gamma^{(t+1)}) \psi(\mathbf{x}_i | \mu^{(t)}, \gamma^{(t)}) dx_i = \int \cdots \int_G \psi(\mathbf{x}_i | \mu^{(t+1)}, \gamma^{(t+1)}) \psi(\mathbf{x}_i | \mu^{(t)}, \gamma^{(t)}) dx_i = \text{E} \left[ \frac{\psi(\mathbf{x}_i | \mu^{(t+1)}, \gamma^{(t+1)})}{\psi(\mathbf{x}_i | \mu^{(t)}, \gamma^{(t)})} \right].
\]
So if \(V_i\) represents draws from the truncated normal with parameters \(\mu^{(t)}\) and \(\Sigma^{(t)} = ((\gamma^{(t)})'\gamma^{(t)})^{-1}\), then \(R\) can be estimated by
\[
\hat{R} = \frac{1}{N} \sum_{i=1}^N \frac{\psi(V_i | \mu^{(t+1)}, \gamma^{(t+1)})}{\psi(V_i | \mu^{(t)}, \gamma^{(t)})}.
\]
An appropriate step size can now be easily found by choosing an initial step size, for example $a = 1$, and calculating the corresponding $\Delta$. If $\Delta$ is negative the step size needs to be reduced, for example by taking $a := a/2$ until $\Delta$ is non-negative.