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 2

Maximum Likelihood Estimation in the Presence of Missing Data

In this chapter an introduction is given to maximum likelihood estimation. In particular, the Expectation-Maximisation (EM) algorithm for maximum likelihood estimation in the presence of nonresponse will be treated extensively. The Expectation step may contain high-dimensional integrals without a closed form solution. Therefore (Markov chain) Monte Carlo methods, that have been developed to estimate high-dimensional integrals, will also be described in detail. The aim of this chapter is to present the reader with the theoretical background that is needed for a complete understanding of all topics that will be treated in the subsequent chapters of this thesis.

2.1 Maximum likelihood inference for complete data

Let $\mathbf{X}$ denote the $n \times k$ complete data matrix and let $\mathbf{X}_i$ denote the $i$th row of $\mathbf{X}$, $i = 1, \ldots, n$. Assume that the $\mathbf{X}_i$ are independently and identically distributed, $\mathbf{X}_i$ is continuous. The probability density function of $\mathbf{X}_i$ is $f(\cdot \mid \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$.

The method of maximum likelihood is looking for the values of the paramet-
ers for which the observed data are most likely, or in other words, the parameter values for which the likelihood function is maximal. The likelihood function is given by

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

The values of the parameters that maximise this likelihood function are referred to as the maximum likelihood estimates: $\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta \mid x)$. It is usually easier, however, to work with the logarithm of the likelihood function

$$\ell(\theta \mid x) = \ln L(\theta \mid x) = \sum_{i=1}^{n} \ln f(x_i \mid \theta).$$

Since the logarithm is a monotonic function, the values that maximise the likelihood are the same as the values that maximise the loglikelihood. If the parameter space $\Theta$ is compact and the loglikelihood function is continuous on $\Theta$, then the maximum likelihood estimate exists. In addition to this, if $\Theta$ is convex and the loglikelihood function is strictly concave on $\Theta$, then the maximum likelihood estimate is unique.

Assuming that

- the maximum is in the interior of the parameter space $\Theta$,
- the range of the random variable $X_i$ is independent of the parameter $\theta$,
- $f(x_i \mid \theta)$ is twice differentiable with respect to $\theta$, for all $\theta \in \Theta$,

see e.g. Fomby et al. (1984), we need to solve

$$\frac{\partial \ell(\theta \mid x)}{\partial \theta} = 0 \quad (2.1)$$

in order to find the maximum likelihood estimates. Note that in this case we can still obtain a local rather than a global maximum, or perhaps a minimum or a saddle point. In order to find a local maximum the matrix of second derivatives, also referred to as the Hessian, needs to be negative definite at the critical point. If the loglikelihood function is strictly concave, the stationary point will be a global maximum.

If equation (2.1) cannot be solved analytically, we resort to iterative schemes, such as the Newton-Raphson algorithm or Fisher scoring. The Newton-Raphson method calculates the parameters until convergence by means of a first order
Taylor series expansion of the first order derivative of $\ell(\theta \mid x)$ near $\theta^{(t)}$, where $\theta^{(t)}$ is an initial guess
\[
\frac{\partial \ell(\theta \mid x)}{\partial \theta} \simeq \left. \frac{\partial \ell(\theta \mid x)}{\partial \theta} \right|_{\theta = \theta^{(t)}} + H(\theta \mid x)_{\theta = \theta^{(t)}}(\theta - \theta^{(t)}),
\]
where $H(\theta \mid x)$ is the Hessian: $\frac{\partial^2 \ell(\theta \mid x)}{\partial \theta \partial \theta'}$. Since (2.1) holds at a maximum we solve for $\theta = \theta^{(t+1)}$
\[
\left. \frac{\partial \ell(\theta \mid x)}{\partial \theta} \right|_{\theta = \theta^{(t)}} + H(\theta \mid x)_{\theta = \theta^{(t)}}(\theta - \theta^{(t)}) = 0.
\]
This results in the Newton-Raphson algorithm
\[
\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta)_{\theta = \theta^{(t)}} \left. \frac{\partial \ell(\theta \mid x)}{\partial \theta} \right|_{\theta = \theta^{(t)}}, \quad t = 1, 2, \ldots,
\]
where $I(\theta)$ is the Fisher information matrix, which is the expectation of the observed information matrix: $I(\theta) = E[I(\theta \mid X)]$. The advantage of this method lies in the fact that (under certain regularity conditions)
\[
E \left[ \frac{\partial^2 \ell(\theta \mid X)}{\partial \theta \partial \theta'} \right] = -E \left[ \sum_{i=1}^{n} \frac{\partial \ell_i(\theta \mid \mathbf{x}_i)}{\partial \theta} \frac{\partial \ell_i(\theta \mid \mathbf{X}_i)}{\partial \theta'} \right], \quad (2.2)
\]
where $\ell_i(\theta \mid \mathbf{x}_i) = \ln f(\mathbf{x}_i \mid \theta)$. So the Hessian, which can be analytically complex, and the expectation of the Hessian, which can be difficult to determine as well, need not be calculated.

This identity can be derived as follows. Note that
\[
\frac{\partial \ln f(\mathbf{x}_i \mid \theta)}{\partial \theta} = \frac{1}{f(\mathbf{x}_i \mid \theta)} \frac{\partial f(\mathbf{x}_i \mid \theta)}{\partial \theta}
\]
and therefore
\[
\frac{\partial f(x_i | \theta)}{\partial \theta} = \frac{\partial \ln f(x_i | \theta)}{\partial \theta} f(x_i | \theta). \tag{2.3}
\]

For a probability density function the following holds by definition
\[
\int f(x_i | \theta) \, dx_i = 1. \tag{2.4}
\]

The density \( f(x_i | \theta) \) is regular if the range of the random variable \( X_i \) is independent of the parameter vector \( \theta \) and if the first and second order derivatives of \( f(x_i | \theta) \) with respect to \( \theta \) exist and are bounded by integrable functions of \( X_i \) (see Fomby et al., 1984). Assuming that the density \( f(x_i | \theta) \) is regular, differentiating (2.4) with respect to \( \theta \) leads to
\[
\int \frac{\partial f(x_i | \theta)}{\partial \theta} \, dx_i = 0.
\]

Using equation (2.3), this results in
\[
\int \frac{\partial \ln f(x_i | \theta)}{\partial \theta} f(x_i | \theta) \, dx_i = 0, \tag{2.5}
\]

which means that
\[
E \left[ \frac{\partial \ln f(X_i | \theta)}{\partial \theta} \right] = 0. \tag{2.6}
\]

Differentiating (2.5) with respect to \( \theta' \) yields
\[
\int \left( \frac{\partial^2 \ln f(x_i | \theta)}{\partial \theta \partial \theta'} f(x_i | \theta) + \frac{\partial \ln f(x_i | \theta)}{\partial \theta} \frac{\partial f(x_i | \theta)}{\partial \theta'} \right) \, dx_i = 0.
\]

Again using the equality in (2.3), we obtain
\[
\int \left( \frac{\partial^2 \ln f(x_i | \theta)}{\partial \theta \partial \theta'} f(x_i | \theta) + \frac{\partial \ln f(x_i | \theta)}{\partial \theta} \frac{\partial \ln f(x_i | \theta)}{\partial \theta'} f(x_i | \theta) \right) \, dx_i = 0,
\]

which leads to
\[
E \left[ \frac{\partial^2 \ln f(X_i | \theta)}{\partial \theta \partial \theta'} + \frac{\partial \ln f(X_i | \theta)}{\partial \theta} \frac{\partial \ln f(X_i | \theta)}{\partial \theta'} \right] = 0.
\]
This results in the information equality given in equation (2.2) as follows

\[
E\left[ \frac{\partial^2 \ell(\theta \mid X)}{\partial \theta \partial \theta'} \right] = E\left[ \frac{\partial^2}{\partial \theta \partial \theta'} \sum_{i=1}^{n} \ln f(X_i \mid \theta) \right] \\
= E\left[ \sum_{i=1}^{n} \frac{\partial^2 \ln f(X_i \mid \theta)}{\partial \theta \partial \theta'} \right] \\
= -E\left[ \sum_{i=1}^{n} \frac{\partial \ln f(X_i \mid \theta)}{\partial \theta} \frac{\partial \ln f(X_i \mid \theta)}{\partial \theta'} \right].
\]

From this equality it follows immediately that the expected information is always positive definite, as

\[
E\left[ \sum_{i=1}^{n} \frac{\partial \ln f(X_i \mid \theta)}{\partial \theta} \frac{\partial \ln f(X_i \mid \theta)}{\partial \theta'} \right] = n \text{Var}\left( \frac{\partial \ln f(X_i \mid \theta)}{\partial \theta} \right) 
\]

because of (2.6). The expected Hessian will therefore always be negative definite, assuring that Fisher scoring is an ascent algorithm. Another major advantage of this method is that we need not calculate the Hessian explicitly, which can be analytically difficult.

Both methods described above converge to the maximum likelihood estimate \( \hat{\theta} \) if the loglikelihood function is concave and unimodal. Moreover, both methods yield an approximation to \( \text{Var}(\hat{\theta}) \) as a by-product, which we will show momentarily.

Maximum likelihood estimates are attractive because of their asymptotic behaviour. If certain regularity conditions are met, a scaling of the maximum likelihood estimator is asymptotically normally distributed: \( \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(\theta, \mathcal{I}^{-1}(\theta)) \). These regularity conditions require that the domain of the random variables does not depend on the parameter vector \( \theta \) and that the density \( f(x_i \mid \theta) \) possesses derivatives of at least third order with respect to \( \theta \), which are bounded by integrable functions of \( X_i \) (Fomby et al., 1984). Now consider the Taylor series expansion of the first order derivative of \( \ell(\theta \mid x) \) about the true parameter value \( \theta \)

\[
\left. \frac{\partial \ell(\theta \mid x)}{\partial \theta} \right|_{\theta = \hat{\theta}} \simeq \left. \frac{\partial \ell(\theta \mid x)}{\partial \theta} \right|_{\theta = \hat{\theta}} + H(\theta \mid x)(\hat{\theta} - \theta).
\]

Since (2.1) holds for the maximum likelihood estimate we find that

\[
\hat{\theta} - \theta \simeq -H^{-1}(\theta \mid x) \frac{\partial \ell(\theta \mid x)}{\partial \theta}.
\]
Multiply both sides by \( \sqrt{n} \)

\[
\sqrt{n}(\hat{\theta} - \theta) = -H^{-1}(\theta | x)\sqrt{n} \frac{\partial \ell(\theta | x)}{\partial \theta} = -\left(\frac{1}{n}H(\theta | x)\right)^{-1}\sqrt{n} \frac{1}{n} \frac{\partial \ell(\theta | x)}{\partial \theta}.
\]

The Central Limit Theorem states that \( \sqrt{n} \frac{1}{n} \frac{\partial \ell(\theta | x)}{\partial \theta} \) converges to a normal distribution with mean zero and variance

\[
\text{Var}\left(\sqrt{n} \frac{1}{n} \frac{\partial \ell(\theta | X)}{\partial \theta}\right) = \text{Var}\left(\frac{\partial \ln f(X_i | \theta)}{\partial \theta}\right) = \frac{1}{n}I(\theta),
\]

using equations (2.2) and (2.7). The matrix \( \frac{1}{n}H(\theta | x) \) converges to \(-\frac{1}{n}I(\theta)\). Then it follows that

\[
-\left(\frac{1}{n}H(\theta | x)\right)^{-1}\sqrt{n} \frac{1}{n} \frac{\partial \ell(\theta | x)}{\partial \theta} \overset{d}{\longrightarrow} \left(\frac{1}{n}I(\theta)\right)^{-1}Z,
\]

where \( Z \sim \mathcal{N}(0, \frac{1}{n}I(\theta)) \). Thus

\[
\left(\frac{1}{n}I(\theta)\right)^{-1}Z \sim \mathcal{N}(0, \left(\frac{1}{n}I(\theta)\right)^{-1}).
\]

This means that

\[
\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\longrightarrow} \mathcal{N}(0, \left(\frac{1}{n}I(\theta)\right)^{-1}),
\]

where \( I^{-1}(\theta) \) can be estimated consistently by \( I^{-1}(\hat{\theta}) \) or \( I^{-1}(\hat{\theta}) \), see for example Fomby et al. (1984) or Stuart and Ord (1991).

### 2.2 Maximum likelihood inference in the presence of nonresponse

#### 2.2.1 The missing data mechanism

The missing data mechanism concerns the reasons why data are missing, and in particular whether these reasons relate to the dataset. Let \( X_i \) denote the complete data vector for respondent \( i \), which can be partitioned into a missing
and an observed part, $X'_i = (X'_{i,mis}, X'_{i,obs})$.

The complete data now have a density function $f(x_{i,mis}, x_{i,obs} | \theta)$. Let $M_i = (m_{ij})$ denote the missing data indicator, where $m_{ij} = 1$ if $X_{ij}$ is observed and zero if it is missing, $j = 1, \ldots, k$. Then $\mathcal{L}(M_i | X'_{i,mis}, X'_{i,obs}, \phi)$ is the missing data mechanism, where $\phi$ is a parameter vector. The MCAR assumption states that this distribution does not depend on the data at all, so

$$f(m_i | x_{i,mis}, x_{i,obs}, \phi) = f(m_i | \phi).$$

This could for example be a Bernoulli function with density $\prod_{j=1}^{k} f(m_j | \phi_j)$, where $f(m_j | \phi_j) = \phi_j^{m_j} (1 - \phi_j)^{1-m_j}$.

The MAR assumption states that the distribution of the missing data mechanism depends only on the observed data, i.e.

$$f(m_i | x_{i,mis}, x_{i,obs}, \phi) = f(m_i | x_{i,obs}, \phi).$$

Note that the MCAR assumption is a special case of the MAR assumption.

The joint density of the data vector $X_i$ and the missing data indicator $M_i$ is

$$f(m_i, x_{i,mis}, x_{i,obs} | \theta, \phi) = f(x_{i,mis}, x_{i,obs} | \theta) f(m_i | x_{i,mis}, x_{i,obs}, \phi).$$

The joint density of the actual observed data is therefore

$$f(m_i, x_{i,obs} | \theta, \phi) = \int f(x_{i,mis}, x_{i,obs} | \theta) f(m_i | x_{i,mis}, x_{i,obs}, \phi) dx_{i,mis}.$$ 

If the data are MAR this reduces to

$$f(m_i, x_{i,obs} | \theta, \phi) = \int f(x_{i,mis}, x_{i,obs} | \theta) f(m_i | x_{i,obs}, \phi) dx_{i,mis} = f(m_i | x_{i,obs}, \phi) \int f(x_{i,mis}, x_{i,obs} | \theta) dx_{i,mis} = f(m_i | x_{i,obs}, \phi) f(x_{i,obs} | \theta).$$

The likelihood of the observed data under the MAR assumption can thus be factorised into two factors where the first factor depends on $\phi$ and the second one on $\theta$. If these parameters are distinct, in the sense that they are mutually, completely uninformative with respect to each other, then the missing data mechanism is said to be ignorable (Little and Rubin, 2002) and inference on $\theta$ can be based solely on the observed data loglikelihood, $\ell(\theta | x_{obs})$. 

2.2. Maximum likelihood inference in the presence of nonresponse


2.2.2 The EM algorithm

The Expectation-Maximisation (EM) algorithm is a technique for finding maximum likelihood estimates when the data are incomplete. The term EM was introduced by Dempster, Laird, and Rubin (1977), where proofs of general results about the behaviour of the algorithm, such as the fact that each iteration increases the observed data loglikelihood, were first given as well as a large number of applications. The EM algorithm is a popular tool in statistics and is used in several different fields of research. For more reading on the EM algorithm see, for example, Schafer (1997) or McLachlan and Krishnan (1997).

The idea behind the EM algorithm is to fill in (a function of) the missing data appearing in the loglikelihood function, based on the observed data and an estimated set of parameters \( \theta \). Next re-estimate \( \theta \) based on the observed data and the estimated (function of the) missing data and iterate this process until the estimates converge.

Suppose again that we have a complete data vector \( \mathbf{X}_i \) for each respondent \( i \) with density function \( f(\cdot \mid \theta) \), where \( \theta \in \Theta \subseteq \mathbb{R}^k \). Again partition \( \mathbf{X}_i \) into a missing and an observed part, \( \mathbf{X}'_i = (\mathbf{X}'_{i,\text{mis}}, \mathbf{X}'_{i,\text{obs}}) \). If the complete data vector \( \mathbf{X}_i \) were observed, we would compute the maximum likelihood estimates of \( \theta \) based on the distribution of \( \mathbf{X}_i \). The loglikelihood function of \( \mathbf{X}_i \) is then required to be maximised. In the presence of missing data, however, only a fraction of the complete data vector \( \mathbf{X}_i \) is observed. In any incomplete data situation, the density of the complete data \( \mathbf{X}_i \) can be written as

\[
 f(x_i \mid \theta) = f(x_{i,\text{obs}}, x_{i,\text{mis}} \mid \theta) = g(x_{i,\text{obs}} \mid \theta) h(x_{i,\text{mis}} \mid x_{i,\text{obs}}, \theta).
\]

From which it follows that

\[
 \ell(\theta \mid x) = \ell(\theta \mid x_{\text{obs}}) + \ln h(x_{\text{mis}} \mid x_{\text{obs}}, \theta),
\]  

(2.8)

where \( \ell(\theta \mid x_{\text{obs}}) \) is the observed data loglikelihood and \( h(x_{\text{mis}} \mid x_{\text{obs}}, \theta) = \prod_{i=1}^n h(x_{i,\text{mis}} \mid x_{i,\text{obs}}, \theta) \). The EM algorithm is useful when maximising \( \ell(\theta \mid x_{\text{obs}}) \) can be difficult but maximising the complete data loglikelihood \( \ell(\theta \mid x) \) is simple. However, since \( \mathbf{X}_i \) is not completely observed, \( \ell(\theta \mid x) \) cannot be evaluated and hence maximised. Therefore the expectation of the complete data loglikelihood is used, which consequently leads to the maximisation of \( \ell(\theta \mid x_{\text{obs}}) \) as is shown in the next section.

In general the EM algorithm operates as follows:
2.2. Maximum likelihood inference in the presence of nonresponse

**Expectation Step**
Find the conditional expectation of \( \ell(\theta \mid X) \) given \( X_{obs} \) at \( \theta^{(t)} \):

\[
Q(\theta \mid \theta^{(t)}) = \mathbb{E}[\ell(\theta \mid X) \mid X_{obs} = x_{obs}, \theta^{(t)}].
\]

**Maximisation Step**
Maximise \( Q(\theta \mid \theta^{(t)}) \) with respect to \( \theta \) in order to obtain \( \theta^{(t+1)} \), so that

\[
Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}) \quad \text{for all } \theta \in \Theta.
\]

Repeat these two steps until the estimates converge, that is until \( \ell(\theta^{(t+1)} \mid x) - \ell(\theta^{(t)} \mid x) < \varepsilon \), with \( \varepsilon \) small.

An essential aspect of the EM algorithm, which makes it different from imputation techniques, is the fact that the missing data that are estimated and filled in do not literally have to be \( X_{mis} \); they are the results of the estimated function of \( X_{mis} \) appearing in the complete data loglikelihood function.

If the complete data loglikelihood is linear in the data \( X \), the E-step is equivalent to filling in the missing data \( X_{mis} \) by their expectation \( \mathbb{E}[X_{mis} \mid X_{obs} = x_{obs}, \theta^{(t)}] \) and the M-step is equivalent to finding \( \theta^{(t+1)} \) by maximising the resulting complete data loglikelihood. However, if the complete data loglikelihood is linear in, say, the sufficient statistics, the E-step consists of filling in the sufficient statistics with their expectations and the M-step consists of maximising the complete data loglikelihood with the sufficient statistics replaced by their expectations. We will elaborate on this in section 2.3.4 on EM and exponential families.

### 2.2.3 Theory behind the EM algorithm

Recall that

\[
\ell(\theta \mid x_{obs}) = \ell(\theta \mid x) - \ln h(x_{mis} \mid x_{obs}, \theta).
\]

Taking the expectation with respect to \( X_{mis} \) conditionally given \( X_{obs} = x_{obs} \) and \( \theta = \theta^{(t)} \) leads to

\[
\ell(\theta \mid x_{obs}) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}),
\]

where

\[
Q(\theta \mid \theta^{(t)}) = \int \ell(\theta \mid x) h(x_{mis} \mid x_{obs}, \theta^{(t)}) dx_{mis}
\]
and
\[ H(\theta | \theta^{(t)}) = \int \ln h(x_{\text{mis}} | x_{\text{obs}}, \theta) \cdot h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \, dx_{\text{mis}}. \]

Now calculate
\[
H(\theta | \theta^{(t)}) - H(\theta^{(t)} | \theta^{(t)}) \\
= \int \ln h(x_{\text{mis}} | x_{\text{obs}}, \theta) \cdot h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \, dx_{\text{mis}} \\
- \int \ln h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \cdot h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \, dx_{\text{mis}} \\
= \int \ln \left( \frac{h(x_{\text{mis}} | x_{\text{obs}}, \theta)}{h(X_{\text{mis}} | x_{\text{obs}}, \theta^{(t)})} \right) \cdot h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \, dx_{\text{mis}} \\
= \mathbb{E}[\ln \left( \frac{h(X_{\text{mis}} | x_{\text{obs}}, \theta)}{h(X_{\text{mis}} | x_{\text{obs}}, \theta^{(t)})} \right) | x_{\text{obs}} = x_{\text{obs}}, \theta^{(t)}].
\]

Note that this is the negative of the Kullback-Leibler information number defined by
\[
K(h(x | \theta^{(t)}), h(x | \theta)) = \int h(x | \theta^{(t)}) \ln \frac{h(x | \theta^{(t)})}{h(x | \theta)} \, dx,
\]
which is always non-negative and therefore \( H(\theta | \theta^{(t)}) - H(\theta^{(t)} | \theta^{(t)}) \leq 0. \) We will show this using Jensen’s inequality.

Recall that Jensen’s inequality states that \( \mathbb{E}[f(X)] \leq f(\mathbb{E}[X]) \) if \( f \) is concave. Take \( f(X) = \ln X. \) Then
\[
\mathbb{E}[\ln \left( \frac{h(X_{\text{mis}} | x_{\text{obs}}, \theta)}{h(X_{\text{mis}} | x_{\text{obs}}, \theta^{(t)})} \right) | x_{\text{obs}} = x_{\text{obs}}, \theta^{(t)}] \\
\leq \ln \mathbb{E}[\left( \frac{h(X_{\text{mis}} | x_{\text{obs}}, \theta)}{h(X_{\text{mis}} | x_{\text{obs}}, \theta^{(t)})} \right) | x_{\text{obs}} = x_{\text{obs}}, \theta^{(t)}] \\
= \ln \int \frac{h(x_{\text{mis}} | x_{\text{obs}}, \theta)}{h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)})} \cdot h(x_{\text{mis}} | x_{\text{obs}}, \theta^{(t)}) \, dx_{\text{mis}} \\
\leq \ln \int h(x_{\text{mis}} | x_{\text{obs}}, \theta) \, dx_{\text{mis}} \\
= \ln 1 = 0.
\]

This means that
\[
H(\theta | \theta^{(t)}) \leq H(\theta^{(t)} | \theta^{(t)}) \text{ for any } \theta \in \Theta.
\]
The difference in loglikelihood is
\[ \ell(\theta^{(t+1)} | x_{\text{obs}}) - \ell(\theta^{(t)} | x_{\text{obs}}) = Q(\theta^{(t+1)} | \theta^{(t)}) - H(\theta^{(t+1)} | \theta^{(t)}) + Q(\theta^{(t)} | \theta^{(t)}) + H(\theta^{(t)} | \theta^{(t)}), \]
which is non-negative, as for any \( \theta \in \Theta \)
\[ Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)}) \]
and
\[ H(\theta^{(t+1)} | \theta^{(t)}) \leq H(\theta^{(t)} | \theta^{(t)}). \]
This means that at each iteration the observed data loglikelihood increases.

2.2.4 Starting values

In order to start the EM algorithm we need an initial estimate of \( \theta \). This initial estimate can be obtained in several ways, for example by using the complete cases or the available cases, but also by using a simple imputation method to impute all missing items and estimating \( \theta \) based on this completed data file.

Unless the fraction of missing data is very high, the choice of starting values usually is not crucial (Schafer, 1997). If the observed data loglikelihood function is concave and unimodal over the entire parameter space, the EM algorithm will converge to the unique maximum likelihood estimate from any starting value. In less well-behaved situations it would be wise to apply the algorithm several times with different starting values to see if it is stable.

Besides, as we already mentioned in chapter 1, if we are dealing with data that have to satisfy certain linear balance restrictions the starting values do matter. This will be treated extensively in chapter 4.

2.2.5 The rate of convergence of the EM algorithm

The rate of convergence of the EM algorithm clearly is of practical interest. Dempster et al. (1977) have shown that the rate of convergence is linear, which we will show momentarily. This means that the EM algorithm converges much slower than Newton-type methods, which typically display quadratic convergence. Furthermore, they also showed that the convergence rate depends on the proportion of information in the observed data, meaning that the algorithm may converge quite slowly if a large amount of the data is missing. This will be treated in the next section.
Any iterative algorithm defines a mapping $\mathbf{M}$, such that $\mathbf{M}(\theta^{(t)}) = \theta^{(t+1)}$, $t = 0, 1, \ldots$. Suppose that $\mathbf{M}$ is differentiable. The parameter $\theta^{(t)}$ converges to $\theta^*$, so $\theta^* = \mathbf{M}(\theta^*)$, which means that $\theta^*$ is a fixed point of $\mathbf{M}$. Consider the first order Taylor series expansion of $\theta^{(t+1)} = \mathbf{M}(\theta^{(t)})$ about $\theta^*$

$$
\mathbf{M}(\theta^{(t)}) \simeq \mathbf{M}(\theta^*) + (\theta^{(t)} - \theta^*) \frac{\partial \mathbf{M}(\theta)}{\partial \theta} \bigg|_{\theta=\theta^*}.
$$

This leads to

$$
\theta^{(t+1)} - \theta^* \simeq \frac{\partial \mathbf{M}(\theta)}{\partial \theta} \bigg|_{\theta=\theta^*} (\theta^{(t)} - \theta^*).
$$

The matrix $\frac{\partial \mathbf{M}(\theta)}{\partial \theta} \bigg|_{\theta=\theta^*}$ is the $k \times k$ matrix of first derivatives of $\mathbf{M}(\theta^*)$. Thus near $\theta^*$ the convergence of the EM algorithm is said to be linear, since $\theta^{(t+1)} - \theta^*$ is approximately a linear transformation of $\theta^{(t)} - \theta^*$, with rate matrix $\frac{\partial \mathbf{M}(\theta)}{\partial \theta} \bigg|_{\theta=\theta^*}$.

In practice, the convergence rate can be defined as

$$
r = \lim_{t \to \infty} \frac{||\theta^{(t+1)} - \theta^*||}{||\theta^{(t)} - \theta^*||}.
$$

Under regularity conditions it can be shown that $r$ is the largest eigenvalue of the rate matrix $\frac{\partial \mathbf{M}(\theta)}{\partial \theta} \bigg|_{\theta=\theta^*}$. (see Schafer, 1997).

### 2.2.6 The missing information principle

Recall that we can write the observed data loglikelihood as follows

$$
-\ell(\theta \mid x_{ob}) = -\ell(\theta \mid x) + \ln h(x_{mis} \mid x_{ob}, \theta).
$$

Differentiating this equation twice with respect to $\theta$ leads to

$$
-\frac{\partial^2 \ell(\theta \mid x_{ob})}{\partial \theta \partial \theta'} = -\frac{\partial^2 \ell(\theta \mid x)}{\partial \theta \partial \theta'} + \frac{\partial^2 \ln h(x_{mis} \mid x_{obs}, \theta)}{\partial \theta \partial \theta'},
$$

which results in

$$
\mathbf{I}_{obs}(\theta \mid x_{ob}) = \mathbf{I}_{com}(\theta \mid x) + \frac{\partial^2 \ln h(x_{mis} \mid x_{obs}, \theta)}{\partial \theta \partial \theta'},
$$

where $\mathbf{I}_{obs}(\theta \mid x_{ob})$ is the observed information matrix and $\mathbf{I}_{com}(\theta \mid x)$ the complete information matrix. Taking the expectation over $X_{mis}$ conditionally
given $X_{\text{obs}} = x_{\text{obs}}$ and $\theta = \theta^{(i)}$ of the left-hand side results in the observed information matrix since it does not depend on $X_{\text{mis}}$. Taking the expectation over $X_{\text{mis}}$ conditionally given $X_{\text{obs}} = x_{\text{obs}}$ and $\theta = \theta^{(i)}$ of the right-hand side and assuming that the order of integration and differentiation can be interchanged we obtain

$$I_{\text{obs}}(\theta \mid x_{\text{obs}}) = -\frac{\partial^2}{\partial \theta \partial \theta^\ast} Q(\theta \mid \theta^{(i)}) + \frac{\partial^2}{\partial \theta \partial \theta^\ast} I(\theta \mid \theta^{(i)}),$$

which results in the missing information principle (Orchard and Woodbury, 1972):

$$I_{\text{obs}}(\theta \mid x_{\text{obs}}) = I_{\text{com}}(\theta \mid x_{\text{obs}}) - I_{\text{mis}}(\theta \mid x_{\text{obs}}),$$

where $I_{\text{com}}(\theta \mid x_{\text{obs}})$ is the expected complete information matrix and $I_{\text{mis}}(\theta \mid x_{\text{obs}})$ is the expected missing information matrix. In other words the missing data principle states that the observed information equals the complete information minus the missing information.

In maximum likelihood estimation, the large-sample covariance matrix of the maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$ is usually estimated by the inverse of the observed information matrix, $I_{\text{obs}}^{-1}(\hat{\theta}_{\text{MLE}} \mid x_{\text{obs}})$. However, this involves a calculation which could be rather difficult; if it were not we would not have used the EM algorithm to obtain the maximum likelihood estimates in the first place. Then we probably would have applied Newton-Raphson which uses the observed information matrix to find the maximum likelihood estimates. An alternative way to calculate the covariance matrix would be to use the missing information principle stated above. For an application of this, see Louis (1982) and Ibrahim (1990).

Another important aspect of these information quantities was given by Dempster, Laird and Rubin (1977) who showed that for the EM algorithm the rate matrix is related to these information matrices as follows

$$\frac{\partial M(\theta)}{\partial \theta} \bigg|_{\theta = \theta^\ast} = I_{\text{mis}}^{-1}(\theta^\ast \mid x_{\text{obs}}) I_{\text{mis}}(\theta^\ast \mid x_{\text{obs}}).$$

### 2.2.7 Advantages and disadvantages of the EM algorithm

The EM algorithm has some very appealing properties. As we have shown, the observed data loglikelihood increases at each iteration. This means that the EM algorithm is guaranteed to converge to a (local) maximum. Another advantage
is that compared to maximising the observed data loglikelihood directly the E-step and M-step may be much easier to calculate.

However, the EM algorithm does also have some disadvantages. First of all, the convergence rate may be extremely slow since the convergence rate of the EM algorithm is linear, whereas the convergence rate of Newton-Raphson is quadratic. Another disadvantage is that the EM algorithm does not automatically provide an estimate of variability for the parameter of interest.

### 2.2.8 Generalisations of the EM algorithm

Recall that in the M-step we maximise $Q(\theta \mid \theta^{(t)})$, in other words we look for a $\theta^{(t+1)}$ such that

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)})$$

holds for all $\theta \in \Theta$. In a generalised version of the EM Algorithm we will require only that $\theta^{(t+1)}$ be chosen such that

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta^{(t)} \mid \theta^{(t)})$$

holds. The parameter $\theta^{(t+1)}$ is chosen to increase $Q(\theta \mid \theta^{(t)})$ over its value at $\theta^{(t)}$ at each iteration $t$, rather than to maximise it. This is sufficient to ensure that

$$\ell(\theta^{(t+1)} \mid x) \geq \ell(\theta^{(t)} \mid x)$$

at each iteration, so the sequence of iterates generated by a generalised EM algorithm also converges to a local maximum.

We use generalised EM-type algorithms when a maximiser of $Q(\theta \mid \theta^{(t)})$ does not exist in closed form. In this case, possibly an iterative method is required to accomplish the M-step, which might prove to be a computationally infeasible procedure. Moreover, even when it is not computationally infeasible it can seriously slow down the convergence of the EM algorithm. Since it is not essential to actually maximise $Q(\theta \mid \theta^{(t)})$ in a generalised EM algorithm, but only to increase the likelihood, we may replace the M-step with a step that achieves that. One possibility of such a step is a single iteration of the Newton-Raphson algorithm:

$$\theta^{(t+1)} = \theta^{(t)} - \left( \frac{\partial^2 Q(\theta \mid \theta^{(t)})}{\partial \theta \partial \theta'} \bigg|_{\theta = \theta^{(t)}} \right)^{-1} \frac{\partial Q(\theta \mid \theta^{(t)})}{\partial \theta} \bigg|_{\theta = \theta^{(t)}}.$$

The matrix $\frac{\partial^2 Q(\theta \mid \theta^{(t)})}{\partial \theta \partial \theta'} \bigg|_{\theta = \theta^{(t)}}$ may, however, not be negative definite and consequently we could find an iteration that does not increase $Q(\theta \mid \theta^{(t)})$. 
2.2.9 Simulated EM algorithms

Even though the EM algorithm avoids the integration involved in finding the observed data loglikelihood, some integration is still needed in the E-step in order to find the conditional expectation $Q(\theta | \theta^{(t)})$. In some cases this may not be feasible, for instance when the expectation is a high-dimensional integral without a closed form expression. This problem can be solved by estimating $Q(\theta | \theta^{(t)})$ by simulation, which results in the following general simulated EM algorithm.

**S-step**
Simulate values $x_{mis}$ from the conditional distribution of $X_{mis}$ given $X_{obs} = x_{obs}$ and $\theta = \theta^{(t)}$. Calculate the complete data loglikelihood based on the simulated data and estimate $Q(\theta | \theta^{(t)})$ by $\hat{Q}(\theta | \theta^{(t)})$.

**M-step**
Maximise the estimated $\hat{Q}(\theta | \theta^{(t)})$ with respect to $\theta$.

Some simulated EM algorithms, that have been suggested, are

1. **Monte Carlo EM (MCEM)**
   Wei and Tanner (1990) suggested using Monte Carlo integration in order to calculate the E-step. Simulate $x_{mis}^{(1)}, \ldots, x_{mis}^{(M)}$ from the conditional distribution of the missing data. Then estimate the complete data loglikelihood by calculating

   \[ \hat{Q}(\theta | \theta^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} \ell(\theta | x_{obs}, x_{mis}^{(m)}). \]

If $M$ goes to infinity this quantity converges to $Q(\theta | \theta^{(t)})$ and therefore the limiting form of MCEM is the regular EM algorithm. Monte Carlo integration will be discussed in detail in the section 2.4.

It is important to note that MCEM, unlike EM, does not deterministically increase the likelihood at each iteration, so the monotonicity property of the EM algorithm is lost, which means that convergence is not guaranteed anymore. Secondly, $M$ needs to be specified. Wei and Tanner (1990) recommend small values of $M$ be used in the initial stages of the algorithm, which should be increased if the algorithm is closer to convergence. With regard to monitoring convergence, they suggest plotting $\theta^{(t)}$ against the
iteration number. When convergence is indicated by the stabilisation of the process with random fluctuations about \( \theta \), the process can either be terminated or continued with a larger value of \( M \) in order to decrease system variability.

2. Stochastic EM (StEM)

The Stochastic EM algorithm was first introduced by Celeux and Diebolt (1985). In this case the missing data are drawn from the conditional distribution once, and the complete data loglikelihood is estimated based on these generated missing data. Next this estimated complete data loglikelihood is maximised with respect to \( \theta \). The sequence \( \hat{\theta}^{(j)} \), \( j = 1, \ldots, t \) forms a Markov chain and the estimator \( \hat{\theta}^{(t)} \) is a random variable drawn from the stationary distribution of the chain.

Note that the StEM algorithm is equal to the MCEM algorithm with \( M = 1 \).

2.3 The exponential family

2.3.1 Introduction

It is well known that the normal distribution has several convenient properties. Most of these properties are shared by a wider class of distributions called an exponential family.

The distribution of a random vector belongs to an exponential family if it can be written in the form:

\[
f(x \mid \theta) = a(\theta)b(x) \exp \left( c(\theta)’t(x) \right),
\]

where \( a, b \) and \( t \) are known functions and \( c(\theta) \) is often called the natural parameter. If \( t(x) = x \), the parameterisation is said to be canonical. It is evident from the form of the density of the exponential family that the \( k \)-tuple of statistics \( t(x) = (t_1(x), \ldots, t_k(x)) \) are sufficient for \( \theta \).

This parameterisation is not unique, in fact many other parameterisations are possible. A parameterisation that is often useful for theory is the so-called natural parameterisation

\[
f(x \mid \eta) = \exp \left( \eta’t(x) + d(x) + g(\eta) \right), \tag{2.9}
\]

where \( \eta = c(\theta) \) is the natural parameter. If \( \eta \) and \( t(x) \) are linearly independent and \( \{ \eta : \eta = (c_1(\theta), \ldots, c_k(\theta)) \mid \theta \in \Theta \} \) is an open set, then the family is
a regular exponential family, which means that the sufficient statistic \( t(x) \) is complete. For a detailed description of exponential families we refer to Brown (1986).

Several well-known distributions are members of exponential families, such as the normal, Bernoulli, binomial, Poisson, exponential, gamma, and Dirichlet distributions.

### 2.3.2 Mean and variance of exponential families

We will start with the general property

\[
\int f(x \mid \eta) dx = 1. \tag{2.10}
\]

Under suitable conditions we are allowed to reverse the order of differentiation and integration, then if we differentiate (2.10) with respect to \( \eta_j \), \( j = 1, \ldots, k \), we obtain

\[
\int \frac{\partial f(x \mid \eta)}{\partial \eta_j} dx = 0. \tag{2.11}
\]

Similarly if we differentiate (2.10) twice and reverse the order of integration and differentiation, we obtain

\[
\int \frac{\partial^2 f(x \mid \eta)}{\partial \eta_j \partial \eta_l} dx = 0, \quad j, l = 1, \ldots, k. \tag{2.12}
\]

Applying the property in equation (2.11) to the natural parameterisation for an exponential family given in (2.9) yields

\[
\int \frac{\partial f(x \mid \eta)}{\partial \eta_j} dx = \int f(x \mid \eta) \left( t_j(x) + \frac{\partial g(\eta)}{\partial \eta_j} \right) dx = 0, \quad j = 1, \ldots, k,
\]

which results in

\[
\text{E}[t_j(X)] = -\frac{\partial g(\eta)}{\partial \eta_j}, \quad j = 1, \ldots, k. \tag{2.13}
\]

For the variance we use equation (2.12) which, again using the natural parameterisation for an exponential family, becomes

\[
\int f(x \mid \eta) \left( t_j(x) + \frac{\partial g(\eta)}{\partial \eta_j} \right) \left( t_l(x) + \frac{\partial g(\eta)}{\partial \eta_l} \right) + \frac{\partial^2 g(\eta)}{\partial \eta_j \partial \eta_l} dx = 0.
\]
Using the result from (2.13) this leads to

\[ E[t_j(X)t_l(X)] = \frac{\partial g(\eta)}{\partial \eta_j} \frac{\partial g(\eta)}{\partial \eta_l} - \frac{\partial^2 g(\eta)}{\partial \eta_j \partial \eta_l}, \quad j, l = 1, \ldots, k, \]

and

\[
\text{Cov}(t_j(X), t_l(X)) = E[t_j(X)t_l(X)] - E[t_j(X)]E[t_l(X)] = -\frac{\partial^2 g(\eta)}{\partial \eta_j \partial \eta_l}, \quad j, l = 1, \ldots, k.
\]

So the covariance matrix of \( t(X) \) will be

\[
\text{Var}(t(X)) = \begin{pmatrix}
\frac{\partial^2 g(\eta)}{\partial \eta_1 \partial \eta_1} & \cdots & -\frac{\partial^2 g(\eta)}{\partial \eta_1 \partial \eta_k} \\
\vdots & \ddots & \vdots \\
-\frac{\partial^2 g(\eta)}{\partial \eta_k \partial \eta_1} & \cdots & -\frac{\partial^2 g(\eta)}{\partial \eta_k \partial \eta_k}
\end{pmatrix}.
\]

### 2.3.3 Maximum likelihood estimation for exponential families

The loglikelihood function of the natural representation for an exponential family is

\[
\ell(\eta \mid x) = \eta' \sum_{i=1}^{n} t(x_i) + \sum_{i=1}^{n} d(x_i) + ng(\eta). \tag{2.14}
\]

The first order derivative is

\[
\frac{\partial \ell(\eta \mid x)}{\partial \eta_j} = \sum_{i=1}^{n} t_j(x_i) + n \frac{\partial g(\eta)}{\partial \eta_j}, \quad j = 1, \ldots, k,
\]

which is set to zero to calculate the stationary points of \( \ell(\eta \mid x) \). The second order derivatives provide information about the nature of the stationary point. These are

\[
\frac{\partial^2 \ell(\eta \mid x)}{\partial \eta_j \partial \eta_l} = n \frac{\partial^2 g(\eta)}{\partial \eta_j \partial \eta_l}, \quad j, l = 1, \ldots, k.
\]
This means that the Hessian is

$$
H(\eta) = n \begin{pmatrix}
\frac{\partial^2 g(\eta)}{\partial \eta_1 \partial \eta_1} & \cdots & \frac{\partial^2 g(\eta)}{\partial \eta_1 \partial \eta_k} \\
\cdots & \cdots & \cdots \\
\frac{\partial^2 g(\eta)}{\partial \eta_k \partial \eta_1} & \cdots & \frac{\partial^2 g(\eta)}{\partial \eta_k \partial \eta_k}
\end{pmatrix} = -n \text{Var}(t(X)).
$$

Since \( \text{Var}(t(X)) \) is positive definite, the Hessian will be negative definite. This means that the loglikelihood function for the natural representation for an exponential family is strictly concave and the stationary point will be a global maximum.

### 2.3.4 EM and exponential families

If we are confronted with missing data and need to use the EM algorithm to obtain maximum likelihood estimates we need to calculate \( Q(\eta \mid \eta^{(t)}) \) in the E-step

$$
Q(\eta \mid \eta^{(t)}) = E[\ell(\eta \mid X) \mid X_{obs} = x_{obs}, \eta^{(t)}] = \frac{n\eta' E[t(X) \mid X_{obs} = x_{obs}, \eta^{(t)}]}{+nE[d(X) \mid X_{obs} = x_{obs}, \eta^{(t)}] + ng(\eta)}.
$$

In the M-step we maximise \( Q(\eta \mid \eta^{(t)}) \) with respect to \( \eta \) in order to obtain \( \eta^{(t+1)} \)

$$
\frac{\partial Q(\eta \mid \eta^{(t)})}{\partial \eta} = nE[t(X) \mid X_{obs} = x_{obs}, \eta^{(t)}] + n\frac{\partial g(\eta)}{\partial \eta} = 0 \\
\Rightarrow E[t(X) \mid X_{obs} = x_{obs}, \eta^{(t)}] = -\frac{\partial g(\eta)}{\partial \eta}.
$$

This means that we only need to calculate \( t^{(t+1)}(X) = E[t(X) \mid X_{obs} = x_{obs}, \eta^{(t)}] \) in the E-step and calculate \( \eta^{(t+1)} \) as a solution to \( \frac{\partial g(\eta)}{\partial \eta} = -t^{(t)}(X) \) in the M-step. Recall that the loglikelihood function is strictly concave and therefore the EM algorithm is guaranteed to converge to a global maximum.

### Example 2.3.4 The EM algorithm for univariate normally distributed data.

Assume that \( X_i \) is normally distributed: \( X_i \sim \mathcal{N}(\mu, \sigma^2) \), where \( X_i, i = 1, \ldots, r \).
is observed and $X_i$, $i = r + 1, \ldots, n$ is missing. The density of a univariate normal in the exponential family form is

$$f(x_i \mid \mu, \sigma) = \exp\left(-\ln \sqrt{2\pi} - \ln \sigma - \frac{x_i^2}{2\sigma^2} + \frac{x_i\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right).$$

This can be reparameterised by

$$f(x_i \mid \theta_1, \theta_2) = \exp\left(-\ln \sqrt{2\pi} + \theta_1 x_i^2 + \theta_2 x_i + \frac{1}{4} \theta_1^{-1}\theta_2^2 - \frac{1}{2} \ln(-\frac{1}{2}\theta_1^{-1})\right),$$

where

$$\theta_1 = \frac{1}{2} \sigma^2 \quad \text{and} \quad \theta_2 = \frac{\mu}{\sigma^2}$$

are the natural parameters,

$$t_1(x_i) = x_i^2 \quad \text{and} \quad t_2(x_i) = x_i$$

the sufficient statistics and

$$g(\theta_1, \theta_2) = \frac{1}{4} \theta_1^{-1}\theta_2^2 - \frac{1}{2} \ln(-\frac{1}{2}\theta_1^{-1}).$$

Then

$$\mathbb{E}[t_1(X)] = -\frac{\partial g(\theta_1, \theta_2)}{\partial \theta_1} = \frac{1}{4} \theta_2^2 \theta_1^{-2} - \frac{1}{2} \theta_1^{-1} = \mu^2 + \sigma^2$$

$$\mathbb{E}[t_2(X)] = -\frac{\partial g(\theta_1, \theta_2)}{\partial \theta_2} = \frac{1}{2} \theta_1^{-1} \theta_2 = \mu.$$

So, the EM consists of calculating

$$\mu^{(t+1)} = \frac{1}{n} \left( \sum_{i=1}^{r} X_i + (n-r)\mu^{(t)} \right)$$

$$(\sigma^{(t+1)})^2 = \frac{1}{n} \left( \sum_{i=1}^{r} X_i^2 + (n-r)((\sigma^{(t)})^2 + (\mu^{(t)})^2) - (\mu^{(t+1)})^2 \right),$$

until convergence.

\section*{2.4 Monte Carlo integration}

Monte Carlo integration is essentially the estimation of an integral by sampling randomly in the domain of integration from the function to be integrated.
2.4.1 Classical Monte Carlo

Suppose we wish to compute

$$I(h) = \int_a^b h(x)dx.$$ 

Now decompose $h(x)$ into a function $g(x)$ and a probability density function $p$ with positive density on the interval $[a, b]$, then note that

$$\int_a^b h(x)dx = \int_a^b g(x)p(x)dx = \mathbb{E}[g(X)].$$

This means that the integral $I(h)$ can be expressed as an expectation of $g(X)$ with respect to the density $p$. So, if we have a large number of independent drawings $X_1, \ldots, X_n$ from the density $p$, then

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

This approach is referred to as the Monte Carlo method, following Metropolis and Ulam (1949). For a complete description of Monte Carlo integration we refer to Gentle (2002), Robert and Casella (1999) and Rubinstein (1981).

The strong Law of Large Numbers ensures that the Monte Carlo estimate converges to the true value of the integral almost surely

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = I(h)\right) = 1.$$ 

This means that if the sample becomes large enough, $\hat{I}(h)$ will almost surely converge to the correct answer.

The variance of $\hat{I}(h)$ is

$$\text{Var}(\hat{I}(h)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} g(X_i)\right) = \frac{\sigma^2}{n},$$

where $\sigma^2 = \text{Var}(g(X))$, which can be estimated from the data by

$$\hat{\text{Var}}(g(X)) = \frac{1}{n-1} \sum_{i=1}^{n} g(X_i)^2 - \frac{1}{(n-1)n} \left(\sum_{i=1}^{n} g(X_i)\right)^2.$$
If the second moment is finite, the Central Limit Theorem states that

$$\sqrt{n}(\hat{I}(h) - I(h)) \xrightarrow{d} N(0, \sigma^2), \quad \text{as } n \to \infty.$$ 

So the error estimate of Monte Carlo integration is of order $O(n^{-1/2})$. The main advantage of Monte Carlo integration, as opposed to deterministic numerical integration methods, is the fact that the error estimate is independent of the dimension of the integral. A drawback is that the Monte Carlo estimate for the integral converges relatively slowly to the true value. For instance, to gain another digit of accuracy, the number of points that are sampled must be increased by a factor of 100. Therefore the Monte Carlo method is usually not employed for numerical integration of univariate integrals. Other methods, such as Gaussian quadrature, are known to perform much better in univariate situations. See Press et al. (1986) for an elaborate description of numerical integration techniques. The advantage of Monte Carlo integration becomes apparent in multidimensional situations, as quadrature based methods can become quite difficult to implement for integrals of more than three dimensions (Geweke, 1996). In contrast to this, Monte Carlo methods can be straightforwardly applied to integrals of high dimensions, which will be treated in section 2.4.3.

First we will discuss a variance reduction technique that has been developed to increase the accuracy of the Monte Carlo estimate, which is referred to as importance sampling.

### 2.4.2 Importance sampling

In importance sampling one generates random samples from a non-uniform distribution that put more probability mass on the important parts of the integration region in order to reduce the variance of the Monte Carlo estimate.

Mathematically, importance sampling corresponds to a change of integration variables. Consider a probability density function $p$ on the interval $[a, b]$, then

$$\int_a^b h(x)dx = \int_a^b \frac{h(x)}{p(x)}p(x)dx = \mathbb{E}\left[\frac{h(X)}{p(X)}\right]$$

and

$$\hat{I}(h) = \frac{1}{n} \sum_{i=1}^n \frac{h(X_i)}{p(X_i)},$$

where the $X_i$ are random drawings from the density function $p$ and $p(X_i) \neq 0$ for any $X_i \in [a, b]$ for which $h(X_i) \neq 0$. 

To see how such a change in the algorithm can lead to an advantage, consider the resulting variance

\[
\text{Var}(\hat{I}(h)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} \frac{h(X_i)}{p(X_i)}\right) = \frac{1}{n} \text{Var}\left(\frac{h(X)}{p(X)}\right),
\]

where \(\text{Var}(h(X)/p(X))\) is estimated by

\[
\hat{\text{Var}}\left(\frac{h(X)}{p(X)}\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{h(X_i)}{p(X_i)}\right)^2 - \frac{1}{(n-1)n} \left(\sum_{i=1}^{n} \frac{h(X_i)}{p(X_i)}\right)^2.
\]

Therefore it will be advantageous to choose \(p\) as close in shape to \(h\) as possible, since then the variance and hence the error can be greatly reduced. It has been shown (see Rubinstein, 1981, p. 122-123) that the variance \(\text{Var}(h(X)/p(X))\) is minimised when \(p(x) \propto |h(x)|\).

### 2.4.3 Multivariate Monte Carlo integration

Now consider the multivariate case

\[
I(h) = \int \cdots \int_{G} h(x_1, \ldots, x_k) dx_1 \cdots dx_k,
\]

where \(G\) is a given \(k\)-dimensional region. The Monte Carlo method for approximating the high-dimensional integral \(I(h)\) is a straightforward extension of the univariate case. So find a probability density function \(p\) on \(G\) and it can be written that

\[
I(h) = \int \cdots \int_{G} h(x) dx = \int \cdots \int_{G} g(x)p(x) dx = E[g(X)] \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \hat{I}(h).
\]

If we draw \(X_i\) uniformly from \(G\) the integral will be estimated by

\[
\hat{I}(h) = \frac{V}{n} \sum_{i=1}^{n} g(X_i),
\]

where \(V\) is the volume of the region \(G\).

The variance of \(\hat{I}(h)\) in (2.15) is

\[
\text{Var}(\hat{I}(h)) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^{n} g(X_i)\right) = \frac{1}{n} \text{Var}(g(X)).
\]
This means that the error remains of order $O(n^{-\frac{1}{2}})$. So, as we mentioned before, the convergence rate does not depend on the dimensionality of the integral. This is why Monte Carlo integration works well for high-dimensional integrals.

### 2.4.4 Concluding remarks

An issue in Monte Carlo integration concerns the choice of the random variables $X_i$. This can become a problem when very large numbers of random numbers are required. Most random number generators are subject to periodicities and other nonrandom behaviour after a certain number of selections have been made. Any nonrandom behaviour will destroy the probabilistic nature of the Monte Carlo scheme and thereby limit the accuracy of the answer. However, there are problems that can be solved by Monte Carlo methods that defy solution by any other method, such as the calculation of high-dimensional integrals.

Another important issue is the fact that in most applications directly generating samples from a certain target distribution is infeasible. Especially in the multivariate case where regions may become rather complex.

One solution is to use Acceptance-Rejection sampling and draw values from a region $H$ that contains the complex region $G$, then accept values that are in $G$ and reject values that are not. The region $H$ should, however, be close to $G$ otherwise this procedure may become very inefficient. So for complex regions, this method will not be useful.

Fortunately, we can also resort to dependent sampling through Markov chains.

### 2.5 Markov chain Monte Carlo

In the previous sections we saw that we do not necessarily need to draw values from the density $h$ to approximate the integral $I(h)$. We can also estimate the integral using random draws from another density defined on the integration region. However, in the multivariate case, complex regions may arise, from which it is difficult to draw independently.

Recently different strategies have been developed that obtain a dependent rather than an independent sample from a density defined on the integration region without directly simulating from that density. The basic principle is to use an ergodic Markov chain with stationary distribution $\pi$. This Markov chain can then be used for Monte Carlo computations. These methods are referred to as Markov chain Monte Carlo (MCMC) methods, which is in essence Monte Carlo
2.5. Markov chain Monte Carlo

integration using Markov chains. The two most popular Markov chain Monte Carlo methods are the Metropolis Hastings algorithm and Gibbs sampling. We will address these methods later, first we will focus on Markov chains in general.

2.5.1 Markov chains

A Markov chain describes a system whose state changes over time. The changes are governed by probability distributions. These probability distributions incorporate a simple sort of dependence structure where the distribution of a future state of the system depends only on the present state. That is, in predicting the future of the system only the present state is relevant, and not the path by which the system got into this state.

The resulting sequence of random variables is the Markov chain. Assuming some regularity conditions are met, proceeding along the sequence, the distribution of the elements will stabilise to the distribution of interest, referred to as the stationary distribution. This stationary distribution does not depend on the initial value with which the Markov chain starts. This means that these values can be used to calculate the integral by means of the Monte Carlo method. For further reading on Markov chains see, for example, Robert and Casella (1999).

The evolution of a Markov chain is governed by the transition kernel $K$

\[ K(x^{(t)}, A) = \Pr(X^{(t+1)} \in A \mid X^{(t)} = x^{(t)}, \ldots, X^{(0)} = x^{(0)}) \]

\[ = \Pr(X^{(t+1)} \in A \mid X^{(t)} = x^{(t)}), \quad x^{(t)} \in S, \quad A \subset S, \]

where $S$ denotes the state space. When $S$ is discrete the transition kernel simply is a transition matrix $P$ with elements

\[ p_{xy} = \Pr(X^{(t+1)} = y \mid X^{(t)} = x), \quad x, y \in S, \]

which satisfies $p_{xy} \geq 0$ and $\sum_{y \in S} p_{xy} = 1$.

In the continuous case, the transition kernel is

\[ K(x, A) = \int_A k(x, y)dy, \]

where the transition density $k(x, \cdot)$ denotes the conditional density of $X^{(t+1)}$ given $X^{(t)} = x$.

If a limiting distribution $\pi$ exists it must be invariant with respect to the transition kernel $K$, therefore

\[ \pi(A) = \int K(x, A)\pi(x)dx, \]
for all sets \( A \). This condition states that if \( X^{(t)} \) is distributed according to \( \pi \), then all subsequent elements of the chain are also distributed as \( \pi \).

The fundamental theorem of Markov chains says that if a Markov chain is ergodic, the probability distribution of the state space converges to a unique distribution \( \pi \). A Markov chain is ergodic if it is irreducible, aperiodic and positive recurrent (or finite).

- **Irreducibility**
  The property of irreducibility is a measure of the sensitivity of the Markov chain to the initial condition. It is crucial in MCMC algorithms since it leads to a guarantee of convergence. The Markov chain is \( \pi \)-irreducible if for each set \( A \) with \( \pi(A) > 0 \) there exists a \( t \geq 1 \) so that \( \Pr(X^{(t)} \in A \mid X^{(0)} = x) > 0 \) for all \( x \in S \). That is, all states communicate with each other, as one can always go from any state to any other state with positive probability.

- **Aperiodicity**
  A chain is said to be aperiodic when the chain is not forced into some cycle of fixed length between certain states. A sufficient condition for aperiodicity is that the transition density \( k(x, \cdot) \) is positive in a neighbourhood of \( x \), so the chain can remain in this neighbourhood before visiting any set \( A \).

- **Positive recurrence**
  The irreducibility property ensures that every measurable set \( A \) will be visited by the Markov chain, but this property does not ensure that the chain visits \( A \) often enough. This leads to the notion of recurrence. An irreducible Markov chain is said to be recurrent, with regard to a given set \( A \), if the expected number of returns to \( A \) in the limit is equal to infinity. The Markov chain is positive recurrent if the mean time to return to \( A \) is bounded. More technically, let \( T_A \) denote the return time to the set \( A \):
  \[
  T_A = \min\{t : X^{(t)} \in A \mid X^{(0)} \in A\}.
  \]
  Then \( A \) is positive recurrent if \( \mathbb{E}[T_A] < \infty \).

If we were solely dealing with discrete or finite bounded state spaces, this form of recurrence would be sufficient to obtain convergence. With unbounded, continuous state spaces it is however necessary to use a stricter definition of recurrence, i.e. Harris recurrence. If a chain is Harris recurrent, the chain has the same limiting behaviour for every starting value, which means that the chain is guaranteed to converge from every starting point. Harris recurrence is defined
as follows. Let \( \eta_A \) be the number of times \( X^{(i)} \) is in \( A \), then \( \Pr_x(\eta_A = \infty) \) is the probability of visiting \( A \) infinitely often starting from \( x \). A set \( A \) is Harris recurrent if \( \Pr_x(\eta_A = \infty) = 1 \) for all \( x \in S \). A Markov chain is Harris recurrent if it is \( \pi \)-irreducible and every measurable set \( A \) for which \( \pi(A) > 0 \) is Harris recurrent. So every set is reachable infinitely often from every starting point and the stationary distribution does not depend on the initial state. This means that the initial state can be chosen rather arbitrarily, it is, however, advantageous if the initial state belongs to the centre of the distribution since convergence is faster in that instance.

2.5.2 Convergence of Markov chain Monte Carlo methods

The fact that Markov chains produce dependent draws causes no substantive complications in summarising the target distribution. If \( (X^{(1)}, \ldots, X^{(N)}) \) are draws generated by an ergodic, Harris recurrent Markov chain, from a target distribution \( \pi \), then the expectation of any function \( h(x) \) under \( \pi \) can be estimated by the average

\[
\mathbb{E}[h(X)] \simeq \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)}),
\]

similarly as in the case of random samples. Since the Law of Large numbers can be used to show that for an ergodic, Harris recurrent Markov chain

\[
\Pr \left( \frac{1}{N} \sum_{i=1}^{N} h(X^{(i)}) \to \mathbb{E}[h(X)] \right) = 1, \quad \text{as } N \to \infty,
\]

regardless of the chosen initial value.

Expectations can thus be approximated by empirical averages just as for ordinary Monte Carlo. The correlation in the Markov chain however implies that the size of the Markov chain sample needs to be greater than when independent simulations are used in ordinary Monte Carlo, in order to obtain a given level of accuracy.

2.5.3 The burn-in period

A last issue we need to address is the burn-in period. The burn-in period refers to the first \( m \) iterations of the Markov chain that are often discarded in order to get reliable estimates. Although it is almost always mentioned in papers on MCMC methods, formally there is no need for burn-in. In MCMC methods the sample
average of the Markov chain is used to approximate the expectation with respect to the stationary distribution of that chain. As we just mentioned for an ergodic, Harris recurrent Markov chain it is guaranteed that the average converges to the expectation with probability one regardless of the chosen starting value and its distribution.

The burn-in period was developed to correct for starting values that were unlikely because they are in the tail of the target distribution. This is in fact a state that the chain should visit if it is run long enough, so we will only encounter a problem if the chain is too short. The solution is to disregard the first \( m \) iterations that are needed to get close enough to the centre of the distribution. Obviously, it would be wiser and more efficient to start in the centre of the distribution or at least in a point with high probability and not discard any iterations. If such a point is unknown or unsure we can always resort to the burn-in approach.

### 2.5.4 The Metropolis-Hastings algorithm

The oldest Markov chain Monte Carlo method was developed by Metropolis et al. (1953) and generalised by Hastings (1970). For a detailed explanation of the Metropolis-Hastings algorithm we refer to Chib and Greenberg (1995).

A Markov chain, with transition kernel \( K \), satisfies the detailed balance equation if

\[
\pi(x)k(x, y) = \pi(y)k(y, x), \quad \forall \, x, y \in S, \tag{2.16}
\]

where \( k \) is the conditional density defined as \( K(x, A) = \int_A k(x, y)dy \). If this detailed balance property is satisfied, then \( \pi \) is the stationary distribution of the chain and the chain is reversible. So finding a Markov chain with stationary distribution \( \pi \), means deriving a transition density from (2.16).

In Metropolis-Hastings algorithms this transition density is constructed as

\[
k(x, y) = q(x, y)\alpha(x, y), \quad x \neq y \in S,
\]

where \( q \) is the proposal density, also referred to as the candidate generating density, and \( \alpha \) is the acceptance probability. A new value is generated using \( q \) and accepted with probability \( \alpha \).

Hastings suggested an acceptance probability of the form

\[
\alpha(x, y) = \begin{cases} 
\min \left\{ \frac{\pi(y)q(y|x)}{\pi(x)q(x|y)}, 1 \right\} & \text{if } \pi(x)q(x,y) > 0 \\
1 & \text{otherwise}
\end{cases}. \tag{2.17}
\]
Note that the detailed balance equation holds for every Metropolis-Hastings chain generated by the transition kernel \( k(x, y) \) with the acceptance probability defined as (2.17). This means that \( \pi \) is a stationary distribution of this chain.

In practice the Metropolis-Hastings algorithm works as follows.

**Metropolis-Hastings algorithm**

1. Choose an initial value \( x^{(0)} \) and set \( t = 0 \).
2. Generate \( y \) from \( q(x^{(t)}, \cdot) \) and \( u \) from \( U(0, 1) \).
3. Calculate \( \alpha \) defined by (2.17). Then if \( u \leq \alpha(x^{(t)}, y) \) set \( x^{(t+1)} = y \), otherwise set \( x^{(t+1)} = x^{(t)} \). Let \( t := t + 1 \) and go back to step 2.
4. Return the values \( \{x^{(0)}, x^{(1)}, \ldots, x^{(N)}\} \).

A sufficient condition for the Metropolis-Hastings chain to be aperiodic is that the algorithm allows to retain the present value: \( x^{(t+1)} = x^{(t)} \), this means that the probability of such an event is nonzero. The irreducibility condition is satisfied if \( q \) has a positive density on the same support as \( \pi \). Moreover, the Metropolis-Hastings chain is Harris recurrent if it is \( \pi \)-irreducible (Robert and Casella, 1999). This means that the Markov chain generated by the Metropolis-Hastings algorithm satisfying the conditions mentioned above will converge to the stationary distribution \( \pi \).

An important issue is the choice of a candidate generating density, especially, since Metropolis-Hastings will work for many choices of \( q \). One family of candidate generating densities is given by \( q(x, y) = q_1(y - x) \), where \( q_1 \) is a multivariate density. The candidate is calculated as \( y = x + z \), where \( z \) is drawn from \( q_1 \). Since the candidate is equal to the current value plus noise, this is called the random walk Metropolis-Hastings chain. Possible choices for \( q_1 \) include a multivariate normal or a uniform density. One has to be careful, however, in choosing the variance of \( z \). If it is too large, many candidates will be rejected, and as a consequence the chain may remain stuck at a certain value for many iterations. On the other hand, if the variance is too small, hardly any candidates will be rejected and the chain will make small moves, leading to inefficient move through the support of the target distribution. In both cases this may lead to high autocorrelations.

Another family of frequently used candidate generating densities are the so-called independence chains. In this case the candidate generating density is independent of \( x \): \( q(x, y) = q_2(y) \). Now the Metropolis-Hastings algorithm is
similar to rejection sampling, except for the fact that the current point is retained when a rejection occurs. In this case the proposals form an independently, identically distributed sequence from the density \( q_2 \). Note that if \( q_2(y) = \pi(y) \) we are sampling directly from the target density.

In general the candidate generating density \( q \) has to be such that it guarantees sufficient move of the chain for covering the whole support of \( \pi \). It should also guarantee a good acceptance and convergence rate for the algorithm to be efficient. Moreover it should be relatively easy to simulate from \( q \).

### 2.5.5 Gibbs sampling

Gibbs sampling is an iterative Markov chain Monte Carlo method that is attractive due to its simplicity. It was introduced by Geman and Geman (1984), but the paper of Gelfand and Smith (1990) demonstrated its broad applicability.

The basic idea in Gibbs sampling is that in updating the \( t \)th sample value, \( \mathbf{X}^{(t)} \), to the next, \( \mathbf{X}^{(t+1)} \), the items in the vector are updated sequentially, using the conditional distribution of these items given the others. Suppose that \( \mathbf{X}^{(t)} = (X_1^{(t)}, \ldots, X_k^{(t)})' \) and let \( f(\mathbf{x}) \) denote the joint distribution of \( \mathbf{X} \).

So the value of \( \mathbf{X}^{(t+1)} \) is obtained sequentially as follows

\[
\begin{align*}
\text{draw} & \quad X_1^{(t+1)} \text{ from } f_1(x_1 | x_2^{(t)}, \ldots, x_k^{(t)}) \\
\text{draw} & \quad X_2^{(t+1)} \text{ from } f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_k^{(t)}) \\
& \quad \vdots \\
\text{draw} & \quad X_k^{(t+1)} \text{ from } f_k(x_k | x_1^{(t+1)}, \ldots, x_{k-1}^{(t+1)}) .
\end{align*}
\]

The sampled elements then constitute \( \mathbf{x}^{(t+1)} \). Repeat this process to obtain a sequence \( \{\mathbf{x}^{(t)} , \ t = 1, \ldots, N\} \), which forms a Markov chain.

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. This can be seen as follows. Let \( q(x, y) = \pi(y | x) \) be the Metropolis-Hastings proposal. In this case the acceptance probability \( \alpha \) given in (2.17) is always one as

\[
\frac{\pi(y) \pi(x | y)}{\pi(x) \pi(y | x)} = \frac{\pi(y, x)}{\pi(x, y)} = 1 ,
\]

which means that we accept every move.

The Gibbs Markov chain is aperiodic since \( q(x, x) > 0 \) for all \( x \in S \). The chain is irreducible if the marginal distributions \( f_i(\cdot) \) of \( X_i \) are \( f_i(x_i) > 0 \), \( i = 1, \ldots, k \), which implies that \( f(x_1, \ldots, x_k) > 0 \). If the chain is irreducible it
is also Harris recurrent (Robert and Casella, 1999).

An important application of the Gibbs sampler is the fact that values can be simulated from the truncated multivariate normal distribution, using the univariate conditional densities, which are less complicated to simulate.

2.5.6 Practical convergence of Markov chains

The theory of Markov chain Monte Carlo tells us that if the Markov chain is ergodic and Harris recurrent, it will eventually produce samples from the target distribution if we run the chain for a sufficiently long time. The difficult part is to establish how long is long enough.

Most users of MCMC methods use diagnostic tools in order to assess the convergence of the Markov chain. Cowles and Carlin (1996) and Brooks and Roberts (1998) provide overviews of the diagnostic tools that are available. In this section we will describe some of these diagnostic tools and illustrate a few with an application on data generated by means of a (random walk) Metropolis-Hastings Markov chain. It is important to keep in mind, though, that these diagnostic tools do not establish convergence but rather a lack of convergence.

2.5.6.1 Graphical assessment of convergence

A natural first diagnostic tool is to use graphs in order to assess the behaviour of the Markov chain. First of all, it is important to assess whether the chain mixes well. That is, whether the chain does not remain stuck in a state for a long time, whether the whole support of the target distribution is explored reasonably fast and whether the dependence on previous values is not too high. This is usually done by making trace plots, where the variable value is plotted against the iteration number.

Example 2.5.6.1 A Metropolis-Hastings chain with a truncated normal stationary distribution

Assume that $X$ is distributed according to a truncated multivariate normal distribution

$$X \sim \mathcal{N}^T(\mu, \Sigma),$$

where

$$\mu = \begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$
and the data are truncated to the region

\[ \begin{align*}
  x_1 + x_2 & \leq x_3 \\
  x_3 & \leq 10 \\
  x_1, x_2 & \geq 0.
\end{align*} \]

We will generate a Markov chain with a truncated normal stationary distribution that satisfies these restrictions with parameters \( \mu \) and \( \Sigma \). The chain is generated using the random walk Metropolis-Hastings algorithm, with a multivariate normal proposal distribution, \( Z \sim \mathcal{N}(0, c^2 \Sigma) \). As we mentioned earlier the choice of the variance of \( Z \) is crucial since it determines the amount of mixing of the chain and the level of autocorrelation. In Figure 2.1 we have varied \( c^2 \) to illustrate this. In Figure 2.1.a we have set \( c^2 = 0.01 \), resulting in an acceptance rate of about 90%. This high acceptance rate leads to few rejections and small moves through the support of the target distribution, which obviously results in high autocorrelations. In 2.1.b we have used the optimal value for \( c^2 \) suggested by Gelman et al. (1996): \( c^2 = 1.88 \), resulting in an acceptance rate of
about 23%. Clearly the chain is mixing much better and the autocorrelations are much lower. Finally, in 2.1.c we have used $c^2 = 8$, which results in big jumps and therefore a large number of rejections. The autocorrelations are in fact reasonable due to the high jumps, but the algorithm is rather inefficient with an acceptance rate of about 5%.

So the variance of the proposal distribution can be seen as a tuning parameter that influences the mixing of the chain. Gelman et al. (1996) established that the optimal scaling parameter for a random walk Metropolis-Hastings algorithm using a multivariate normal proposal distribution is $c^2 \Sigma$, where $\Sigma$ is the covariance matrix of the data and $c$ is defined as $c = 2.38 / \sqrt{d}$, with $d$ equal to the dimension of the data. This leads to an acceptance rate of about 23%.

Another important point is the choice of starting values, which influences the required length of the Markov chain or creates the need for a burn-in period. Theoretically, the chain will converge to the stationary distribution regardless of the starting value, provided it is Harris recurrent. In practice the required length of the Markov chain may be strongly influenced by the starting value. Therefore it would be wise to choose the starting value as close to the centre of the distribution as possible. For example, by using an approximate maximum likelihood estimate. As a diagnostic tool it is, however, advisable to generate chains using several highly dispersed starting values in order to detect undesirable behaviour.

A tool to assess the effects of starting values is to draw a trace plot of the mean of a variable against the iteration number.

**Example 2.5.6.1 continued**

We will again generate a Markov chain with a truncated normal stationary distribution as was done in example 2.5.6.1 with similar parameter values for $\mu$ and $\Sigma$ and $c^2 = 1.88$, but now we will start from different points. In Figure 2.2.a the chain was generated using $(0,10,10)'$ as a starting value, in Figure 2.2.b the starting value was chosen at the centre of the distribution. It is clear that choosing the starting value far from the centre of the distribution leads to a much slower convergence rate of the average, which means that the chain should be run for a long time or a burn-in period should be used.

Other plots that can be informative in detecting deviant behaviour are histograms, box plots and plots of partial autocorrelations.
2.5.6.2 Quantitative convergence diagnostics

There are also several quantitative convergence diagnostics. We will treat the two most popular ones, which are those suggested by Gelman and Rubin (1992) and Raftery and Lewis (1992a, 1992b).

The first method consists of generating several, say $m$, chains starting from overdispersed starting values. This means that the starting distribution should be more variable than the target distribution. Gelman and Rubin (1992) propose a method for creating such an overdispersed starting distribution.

Once suitable starting values are obtained the Markov chains are run for $2N$ iterations. The first $N$ iterations will be discarded in order to avoid burn-in problems. Next the between-sequence and within-sequence variance is calculated for the $m$ Markov chains of length $N$, defined by

$$B = \frac{N}{m} \sum_{i=1}^{m} (\bar{X}_i - \bar{X}_{-i})^2$$

and

$$W = \frac{1}{m} \frac{1}{N-1} \sum_{i=1}^{m} \sum_{j=1}^{N} (X_{ij} - \bar{X}_i)^2,$$
where $X_i$ is the average of chain $i$, and $X$ is the overall average.

The method is based on the idea that after a certain period the chains have
forgotten their starting value and should be similar, that is the variance within
the chains should be the same as the variance across the chains. The target
variance is estimated by

$$\hat{\sigma}^2 = \frac{N-1}{N}W + \frac{1}{N}B,$$

which would be an unbiased estimate of the true variance if the starting points
of the sequences were drawn from the target density, but overestimates the true
variance if the starting values are overdispersed. Besides for a finite $n$, $W$ should
be less than $\sigma^2$ as the individual chains have not had time to visit the whole
target distribution. The Gelman-Rubin diagnostic is

$$\sqrt{R} = \sqrt{\frac{N-1}{N} + \frac{1}{N} \frac{B}{W}},$$

where $\sqrt{R}$ is referred to as the potential scale reduction factor. Convergence has
been reached if this statistic is near one. A drawback of this method is that it
requires multiple chains to be generated, which may result in large computing
times. Besides it relies on the ability to find a starting distribution that is over-
dispersed with respect to the target distribution.

Raftery and Lewis (1992b) developed another popular quantitative convergence
diagnostic. Their diagnostic is based on the problem of calculating the
number of iterations that are necessary to estimate a posterior quantile from a
single run of a Markov chain. In advance the minimum number of iterations,
$N_{\text{min}}$, that is needed when the sample is identically and independently distributed, can be determined. Now we wish to estimate $\Pr(U(X) \leq u)$ to within
$q \pm r$ with probability $s$. Calculate $U(t) = U(X(t))$ where $X(t)$ is the state of
the Markov chain at state $t$. Then form $Z(t) = I_{U(t) \leq u}$, where $I$ is an indicator function. The sequence $\{Z(t)\}$ is a binary process derived from a Markov
chain. It is, however, not a Markov chain itself. The subsequence $\{Z_k(t)\}$, where
$Z_k(t) = Z^{1+(t-1)k}$, will approach a Markov chain for large enough $k$. Raftery and
Lewis (1992a) determine $k$ by making use of BIC ratios. After establishing $k$
the approximate transition matrix $\hat{P}$ of $\{Z_k(t)\}$ is estimated by

$$\hat{P}_{ij} = \frac{\#\{t : Z_k(t) = j, Z_{k-1} = i\}}{\#\{t : Z_{k-1} = i\}}, \quad i, j = 0, 1.$$
Then set $\alpha = \hat{P}_{01}$ and $\beta = \hat{P}_{10}$. The burn-in length is $M = m^* k$, with

$$m^* = \frac{\ln \left( \frac{\epsilon(\alpha+\beta)}{\max(\alpha,\beta)} \right)}{\ln(1 - \alpha - \beta)},$$

where $\epsilon$ represents a convergence tolerance (e.g. $\epsilon = 0.001$). The required length of the chain is $N = n^* k$, with

$$n^* = \frac{\alpha \beta (2-\alpha-\beta)}{(\alpha+\beta)^2 \Phi \left( \frac{\epsilon}{\sqrt{2(1+\epsilon)}} \right)^2},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The Raftery-Lewis convergence diagnostic then becomes

$$I = \frac{M + N}{N_{\min}},$$

this statistic measures the increase in the number of iterations due to the dependence in the Markov chain. Values of $I$ exceeding 5 are indicative of problems with convergence.

The strength of this method lies in the focus on an accurate estimation of quantiles. However, the convergence rate of the quantile of interest may not be the convergence rate of the chain. Furthermore different quantiles will lead to different estimates of $M$ and $N$.

### 2.5.6.3 Concluding remarks on convergence diagnostics

The quantitative diagnostics that were described in the previous subsection are often used to assess the convergence of Markov chains. Their popularity, however, is mainly due to ease of implementation and interpretation. In fact each of the diagnostic tools that are available have their own drawbacks, which means that there is no globally best diagnostic method. It is therefore wise to use several different diagnostic tools. Finally, one should always keep in mind that the diagnostics do not establish convergence, just a lack of it.