Imputation of Restricted Data

Applications to Business Surveys

Caren Tempelman

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
`http://www.cbs.nl`

Key figure: I-76
ISBN: 978-90-357-1439-7
Product code: 6998207001

RIJKSUNIVERSITEIT GRONINGEN

Imputation of Restricted Data

Applications to Business Surveys

Proefschrift

ter verkrijging van het doctoraat in de
Economische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
maandag 19 februari 2007
om 14.45 uur

door

Dina Catharina Geertruida Tempelman
geboren op 17 maart 1978
te Meppel

Promotor:	Prof.dr. R.H. Koning
Copromotor:	Dr. J. Pannekoek

Beoordelingscommissie:	Prof.dr. T.J. Wansbeek
	Prof.dr. H.H. van Ark
	Dr.ir. P. Kooiman

Acknowledgements

When Statistics Netherlands presented me with the opportunity to do this research I did not hesitate and accepted the offer. As I have thoroughly enjoyed doing research this has proved to be a good decision.

A dreadful event during this research period was the time my esteemed promotor Ton Steerneman fell terribly ill and passed away. Ton has been a true inspirator and a highly involved promotor, both professionally as well as personally. I am very grateful for his enthusiasm, insightful comments and excellent ideas; he is deeply missed.

Many thanks are due to Ruud Koning, who took on the job of promotor in such a difficult time. Ruud has been fully committed, very knowledgeable and highly personally involved as well. Our vivid discussions and his valuable remarks and ideas most certainly improved this thesis.

I also would like to thank Statistics Netherlands for providing me with the means to carry out this research. In particular Ton de Waal, who initiated this project and has thoroughly read several chapters of this thesis. His comments were highly appreciated. Furthermore, I owe much to my copromoter Jeroen Pannekoek, who was always available for discussion and willing to reflect on any issue that surfaced. Jeroen has provided numerous helpful suggestions and insightful solutions.

Furthermore, I am grateful to the members of the reading committee, Bart van Ark, Peter Kooiman and Tom Wansbeek for carefully reading the manuscript.

Finally, I thank my friends and family, in particular my 'paranimfen' Rick and Maartje and of course Rinze, who has contributed next to nothing but was invaluable nonetheless.

Voorburg, January 2007

Contents

Chapter 1

An Introduction to Imputation and Editing

Statistics Netherlands collects and publishes information about all kinds of aspects of Dutch society. This information is based on data provided by persons, households, businesses, and so on. The economic figures published are of great importance to policy makers, researchers and businesses as they can be used to forecast or monitor economic quantities and to make informed business decisions. Some examples of a variety of economic statistics are the measurement of gross domestic product, use of energy, employment statistics, and investments of financial institutions. These statistics are based on survey estimates, which unfortunately are potentially subject to error. In this chapter we will discuss different sources of error and strategies that have been developed to deal with them.

1.1 Potential sources of error in survey estimates

Differences between survey estimates and the actual population parameters are caused by sampling as well as nonsampling errors. Sampling error is the error that arises due to surveying only a subset of the population rather than conducting a census on all businesses. Sampling errors would vanish if all businesses were sampled, and can be controlled by using a correct sampling design.

Nonsampling errors, however, can still occur when all businesses are surveyed. These errors arise during the data collection process and can be subdivided into the following categories, see for example Lessler and Kalsbeek (1992).

- *Frame error*
  The businesses that are to be surveyed are drawn from a sample frame, which is usually a business register. Frame errors arise due to discrepancies between the frame and the actual population.
  A large part of frame errors are caused by over- and undercoverage of businesses. Overcoverage occurs when businesses, that do not belong to the target population, are in the frame and undercoverage refers to businesses, belonging to the target population, that are not in the frame and therefore will never be sampled. Since business populations change rapidly due to births, deaths, and organisational changes, the updating of business registers is crucial in order to prevent large frame errors.
  Errors occurring in auxiliary variables, that are recorded in the business register such as size class or branch of business, are another type of frame error.

- *Nonresponse error*
  If the sampled business does not provide answers, we are confronted with nonresponse. In general two types of nonresponse are distinguished: unit and item nonresponse. Unit nonresponse refers to the case where the contacted business does not respond at all and item nonresponse refers to businesses that do respond, but omit answering some of the questions.
  There are several causes for nonresponse. First of all, some respondents refuse or are unable to provide data for a particular item or items, or sometimes they simply miss a particular question. Secondly, the interviewers may fail to record data items or data entry clerks may omit keying the data item. Besides, questionnaires may get lost or sent to the wrong address.
  If nonrespondents differ significantly from respondents in their answers, the error due to nonresponse can become large which may lead to biased estimates.

- *Measurement error*
  Measurement error is defined as the difference between the reported value and the true value. In general we distinguish between three sources of measurement error. First of all the true value may be unknown or difficult

to obtain. Businesses may keep their accounting data according to different definitions than those used by the statistical agency. For example, the reference period used at businesses may differ from the reference period requested, e.g. financial year versus calendar year. Moreover, businesses may not even keep the information requested by the agency. In this case the effort required to obtain an answer may cause the respondent to guess the value or leave it blank. Note that this implies that high item nonresponse rates may be indicative of measurement errors. Secondly, questions may be misunderstood or misread by the respondent. The respondent is, for example, asked to report in thousands of Euros, but he or she actually reports in Euros, leading to a huge bias in the survey estimates if not corrected. Finally, the firms accounting system itself may contain errors.
When measurement errors are present the difficulty is how to determine whether data are erroneous or not. Some mistakes are obvious, such as negative reported values for variables that should be non-negative, but others, such as differences in definitions that are used by respondents and statistical agencies, may be much harder to find.

- *Processing error*
  After the data have been collected they pass several processes, such as keying, coding, editing, weighting and tabulation. Errors that arise during these processes are referred to as processing errors.
  At Statistics Netherlands all business surveys are mail surveys, which means that the data need to be keyed onto a computer. Obviously this is subject to error as responses need to be interpreted. The numbers 1 and 7, for example, are easily mistaken for each other.
  Other processing errors that may occur are mistakes in the software or adjustments of item values that appear to be in error but actually are correct.

In this thesis we will focus our attention on how to deal with missing data, this means that we will first elaborate on (item) nonresponse. As there is a strong relationship between dealing with missing data and handling erroneous data, the latter will also be discussed in more detail in this chapter.

1.2 Missing data

Missing values not only mean less efficient estimates because of the reduced size of the data set, but also that standard complete data methods cannot

straightforwardly be used to analyse the data. Moreover, possible bias exists because respondents are often systematically different from nonrespondents. If they do not differ, analysis of the responding items is sufficient to obtain valid inference. However, if they are systematically different analysis restricted to the respondents may result in serious bias. Assumptions about the selectivity of nonresponse are formalised by the missing data mechanism.

1.2.1 The missing data mechanism

The missing data mechanism concerns the reasons why values are missing, and in particular whether these reasons relate to values in the data set. Any analysis of data involving item nonresponse requires some assumptions about the missing data mechanism.

This mechanism can be formalised as follows, see Rubin (1976) and Little and Rubin (2002). Consider an $n \times k$ data matrix $\mathbf{X}$, where X_{ij} is the value of variable j for respondent i and the joint distribution of $\mathbf{X}$ depends on a parameter vector $\boldsymbol{\theta}$. The main interest of the researcher is usually to obtain inference about the parameter vector $\boldsymbol{\theta}$ that determines the joint distribution of $\mathbf{X}$. We introduce a missing data indicator matrix $\mathbf{M}$ corresponding to the data matrix $\mathbf{X}$, where $M_{ij} = 1$ if X_{ij} is observed and zero otherwise (so $\mathbf{M}$ and $\mathbf{X}$ are of the same dimensions). Furthermore, let $\mathbf{X}_i$ denote the ith row and $\mathbf{X}_j$ the jth column of the data matrix $\mathbf{X}$, $\mathbf{M}_i$ and $\mathbf{M}_j$ are defined similarly. The data vector $\mathbf{X}_i$ can be partitioned in a missing and an observed part: $\mathbf{X}_i = (\mathbf{X}'_{i,mis}, \mathbf{X}'_{i,obs})'$. Partition $\mathbf{M}_i$ accordingly. This notation is somewhat sloppy as the missingness pattern may vary across respondents. However, as this notation is convenient and generally used and accepted, we will make use of it throughout this thesis as well.

In general the density of the missing data indicator is written as

$$f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}), \tag{1.1}$$

where we assume the parameter vector $\boldsymbol{\phi}$ to be distinct from $\boldsymbol{\theta}$. When the density of $\mathbf{M}_i$ can be simplified to depend solely on the parameter $\boldsymbol{\phi}$ and not the data, i.e. $f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}) = f(\mathbf{m}_i \mid \boldsymbol{\phi})$, then $\mathbf{M}_i$ is independent of $\mathbf{X}_i$. This means that the distribution of the indicator does not depend on the observed nor the missing data and the data are said to be *missing completely at random* (MCAR).

If the data are not MCAR it is important to establish whether the differences between nonrespondents and respondents can be explained by other reported or

auxiliary variables, such as business size or branch of business. If the conditional distribution of $\mathbf{M}_i$ given $\mathbf{X}_i$ does depend on the observed but not on the missing values of $\mathbf{X}_i$, such that $f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}) = f(\mathbf{m}_i \mid \mathbf{x}_{i,obs}, \boldsymbol{\phi})$, then the missing values are *missing at random* (MAR). An example of this is when the missingness in certain variables depends on business size. For instance when small companies have a higher nonresponse rate for a certain variable than large companies.

Finally, if missingness depends on both the observed and the missing values of $\mathbf{X}_i$, that is if the distribution of $\mathbf{M}_i$ given in (1.1) cannot be simplified, the data are said to be *not missing at random* (NMAR). In other words the fact that the respondent answers a certain question depends on the actual item value for that question. In social surveys this is quite common, especially when sensitive information, such as income specifics or sexual preferences, is gathered. Respondents with either a relatively low or a relatively high income, for example, are more likely not to provide an answer to questions about their income than respondents with an income closer to the average. In business surveys we do not expect this kind of effect as the requested information is likely not to be perceived as sensitive by the person who is filling in the questionnaire. Besides, businesses are obliged by law to respond to questionnaires sent out by Statistic Netherlands.

If the data are MCAR or MAR the missing data mechanism is said to be ignorable and valid estimates can be obtained without explicitly modelling the missing data mechanism. If the data are NMAR, however, the missing data mechanism is nonignorable and needs to be modelled when estimating $\boldsymbol{\theta}$. In general most methods for handling nonresponse in the survey literature assume that the missing data are MAR. Throughout this thesis we will also make this assumption.

1.2.2 Strategies for handling nonresponse

Three general strategies can be distinguished for dealing with missing data.

1. *Direct analysis of the incomplete data.*
   Most of the time cases with missing values are simply discarded. This is also referred to as complete case analysis. An advantage of complete case analysis is the ease of implementation, but a serious drawback is the rejection of information in the incomplete cases. Besides, most users are unaware of the fact that complete case analysis is only valid when the data are MCAR, which is not very likely.

For univariate analysis all cases where the variable of interest is observed can be included, which is referred to as available case analysis. An advantage is that all available information is used, but a disadvantage is that the sample base changes from variable to variable.
There are also more elaborate methods that model the incomplete data, such as the Expectation Maximisation (EM) algorithm, developed by Dempster, Laird and Rubin (1977). The EM algorithm obtains maximum likelihood estimates in the presence of nonresponse. We will discuss the EM algorithm and its underlying theory in detail in chapter 2.

2. *Weighting.*
In this case the nonrespondents are removed from the data set and weights are assigned to the respondents, based on auxiliary information from the sample frame. The survey estimates are calculated based on the respondents and their assigned weights. Weighting is often used to deal with unit nonresponse, where the sample frame contains the only available information. Weighting leads to valid inference if the missing data are MAR with respect to the auxiliary variables in the frame.

3. *Imputation.*
A third strategy is to replace the missing values by an estimated value that can be constructed from the sample frame and the observed responses. This is referred to as imputation. More information is available when faced with item nonresponse as opposed to unit nonresponse, as now both information from the sample frame as well as information from the responses to other survey items is present. Therefore imputation is mostly used when dealing with item nonresponse. As opposed to weighting, imputation can also lead to valid inference if the missing data are MAR with respect to other variables in the survey.

As we are dealing with item nonresponse in this thesis, our focus will be on imputation of the missing data items. Imputation has several desirable features, see for example Kalton (1983). First of all if used correctly, the nonresponse bias will be reduced. In many surveys no compensation is made for missing data. If the nonrespondents are significantly different from the respondents, this will lead to seriously biased estimates; see Bethlehem and Kersten (1986). Secondly, imputation of the missing items will lead to a complete data set, which will make it easier to carry out statistical analyses. Finally, the results obtained from different analyses are bound to be consistent, a feature which need not apply with

an incomplete data set. Leaving the imputation to the users may lead to conflicting estimates due to different imputation methods employed. Moreover, at the statistical agency a wealth of external information about businesses is available in administrative records such as tax records, which are not accessible to the data users for reasons of confidentiality.

Imputation also has some serious drawbacks, however. First of all, imputation does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set. Biases could arise depending on the imputation procedure, the actual missing data mechanism and the form of the estimate. Besides, there is also the risk that the analysts treat the completed data set as if they were all actual responses, and thereby underestimating the uncertainty caused by imputation. We will elaborate on this in section 1.2.4. First an overview of the most common imputation methods will be given.

1.2.3 An overview of imputation methods

There is considerable literature discussing imputation methods; see for example Kalton and Kasprzyk (1982, 1986), Kalton (1983), Sande (1982), Rubin (1987), Little (1988) and Kovar and Whitridge (1995). We will give an overview of the most common methods. Note that most imputation strategies make use of several imputation methods, depending on for example the type of variable or the available auxiliary information.

- *Logical (or deductive) imputation.*
  In this case the missing item value can be established with certainty from the other items. For example, only one component of a total is missing and can be easily deduced by subtraction, or some logical restrictions that must hold constrain the value to one possibility. That is, if the total operating expenses equal zero, and the costs of energy use are a part of total operating expenses and non-negative, then the energy costs must be equal to zero. Clearly this is the best imputation method possible, as the information needed can be derived with certainty from the observed data. This method should therefore always be applied first, before using any of the other imputation methods.

- *Mean imputation.*
  This method replaces the missing value by the mean of the responding units in a certain class for that item. The units are divided into classes according to their auxiliary variables, such as business size or type. Although respondent means are preserved, distributions of variables and

relationships between variables are seriously distorted if mean imputation is used. This is due to the fact that a peak is created at the average value of the variable and therefore the variation will be strongly underestimated.

- *Cold deck imputation.*
  This method substitutes the missing item value with a value taken from an external source. In business surveys, often data from a previous period are used. Obviously, trend and change will not be accounted for by using the value of the previous period, so extensions of this method adjust the values by modelling some sort of trend based on the reported items for that unit or other units in the sample. Other external sources could be administrative data such as tax records, which can be matched to the survey data.

- *Hot deck imputation.*
  In this instance the missing item values will be replaced with actual values from respondents in the present sample, which will be referred to as donors. This method is referred to as hot deck as the imputations are drawn from the present sample and not external data sources, which was the case for cold deck imputation. By using these donors the distribution of the population as represented by the sample is preserved. Another attractive feature of this method is that it is nonparametric, so no strong model assumptions need to be made in order to estimate the individual values. The only assumption that is made, is that the data are MAR with regard to the auxiliary variables. In general there are three common hot deck procedures:

  - *Sequential hot deck imputation.*
    The missing item value, that is filled in, is based on the value from the last responding unit preceding it in the data file. Usually, the data file is sorted according to business size or geographical location or in the case of qualitative auxiliary variables divided into subclasses based on these auxiliary variables. A problem that may occur is the multiple use of donors when there are a lot of missing values for a certain item, a feature that contributes to lowering the precision of survey estimates and underestimating variance. Another problem is that the imputed values depend completely on the order of the data set.
  - *Random hot deck imputation.*
    With this method respondents are divided into imputation classes

based on auxiliary data, so that elements in the same class are considered similar. An item value of a randomly selected respondent within an imputation class will be assigned to the missing item value. Due to the stochastic nature of this method the variances will be better preserved.

- *Nearest neighbour imputation.*
  If there are several quantitative auxiliary variables, then the use of imputation classes may have undesirable effects. First, one has to choose the boundaries for the imputation classes. Secondly, two units may be matched whereas the first unit may be near the upper bound of a certain class and the second unit may be near the lower bound of that class. To overcome this problem some sort of distance function can be used to find the nearest record. The missing item value will then be replaced by the item value of this respondent. To avoid one single respondent being used as a donor several times, a component can be included in the distance function that reduces the multiple use of donors.
  This is a deterministic procedure, but can be randomized by finding several nearest neighbours and randomly choosing a donor from these. Furthermore, the variables need to be standardized before the distance between units is calculated in order to assign equal weights to each variable.

For further reading on hot deck procedures see Ford (1983) and Sande (1983).

- *Regression imputation.*
  This method replaces the missing values of a certain item by predicted values from a regression of that variable on (some of) the other variables in the survey and the auxiliary variables in the sample frame. The regressors may be both quantitative and qualitative. The latter can be incorporated by dummy variables.
  An advantage of regression imputation is that information from a previous period can be easily exploited as well by simply adding predictor variables, hereby increasing the available information for imputation. The cold deck method, mentioned previously, that uses a trend to adjust missing values can be seen as a special case of regression imputation. Furthermore, unlike hot deck methods, where it may be difficult to find a suitable donor when imputing a somewhat outlying record, the regression approach will always

produce a replacement value.
A disadvantage is that the conditional mean is imputed in this case and therefore this method corresponds to mean imputation mentioned earlier, which means it has the same undesirable properties with regard to variance estimation. This can be overcome to some extent by adding a random component.

- *Random regression imputation.*
  This procedure is the stochastic version of regression imputation. The missing item value will be replaced by the predicted value plus a random residual term. There are several possibilities for choosing residuals, such as assuming that the residuals are normally distributed with mean zero and variance equal to the residual variance of the regression. Another method would be to find a respondent with a similar predicted value and add this residual to the predicted value of the missing item. See Kalton (1983) for more suggestions. Using random regression imputation will correct somewhat for the underestimation of variance in the case of regular regression imputation.
  A problem that may occur with regression imputation (both deterministic and random) is that infeasible item values are imputed, such as for example negative predicted values for non-negative variables. One simple solution to this would be to set the imputed value equal to zero, but this could seriously distort the distribution of the residuals. A better solution would be to make use of (nonlinear) transformations, for example by taking the natural logarithm.
  Another disadvantage is the fact that regression imputation is strongly dependent on a model and consequently may be sensitive to model misspecification.

- *Predictive mean matching.*
  A method that combines both regression and hot deck imputation is predictive mean matching. This method was developed by Little (1988). First perform a regression of the variable that needs to be imputed on a set of predictor variables. Match the predictive mean of the missing item to the closest predictive mean of the responding records and then impute the actual item value of that respondent. This method is actually similar to nearest neighbour imputation, using the differences between predicted mean values as a distance measure. Its advantage over regression imputation is that only feasible values are imputed.

- *Imputation using the Expectation-Maximisation (EM) algorithm.*
  The EM algorithm was originally designed to obtain maximum likelihood estimates in the presence of nonresponse, but it can also be used for imputation purposes. The EM algorithm consists of two steps. In the E-step the expected complete data loglikelihood is calculated based on the observed data and the current parameter values. Subsequently in the M-step, the parameter values are updated by maximising the expected complete data loglikelihood. This process is iterated until convergence. Imputations can be generated deterministically by using the expected values for the missing data as imputations, which are often, but not always, a by-product of the E-step. Stochastic imputations are obtained by randomly drawing values from the specified distribution using the maximum likelihood estimates as parameters.

Example 1.2.3. The EM algorithm in SPSS
The EM algorithm is being used extensively. A result of this is that well-known statistical software, such as SPSS, provide procedures to apply the EM algorithm. Care should be taken, however, as SPSS does not allow the user to specify starting values for the algorithm, it automatically uses the available cases estimates. But if the fraction of missing data is high, one may want to evaluate the behaviour of the loglikelihood by starting from different values. ■

The fact that every variable in the survey is potentially subject to missing data complicates the imputation task, as the variables that will be conditioned on may contain missing values. In this case the EM algorithm is an attractive and straightforward option, especially for high-dimensional data, as all missing items are imputed simultaneously. The donor methods, however, will now probably lead to different donors for each missing item, which distorts associations between items. If these associations are important, for example in multivariate analysis, it may be wise to use one donor to fill in all the missing items of a responding unit. The problem with this strategy is that it makes the imputation process more complex. First of all, there will be fewer donors available as the donors are not allowed to have any missing items for all variables that need to be imputed. So the danger arises that one particular donor will be used multiple times. Secondly, it will be difficult to create imputation classes that are homogeneous with respect to all missing items. Similar problems arise for regression imputation as standard regression techniques cannot deal with covariates that contain missing values. In this case one could use the EM algorithm for normally distributed data, which is an iterative form of regression imputation. Addition-

ally, adjustments are made to the covariance matrix in each iteration to correct for the fact that predicted means are imputed.

A serious defect of all imputation methods is that they invent data. More specifically, a single imputed value cannot represent all of the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty. Moreover the usual estimates of variance are inadequate since they do not include the error due to estimation.

1.2.4 Variance estimation in the presence of imputation

Imputing the average value of a variable X will lead to reasonably accurate estimates for the population mean or total of that variable if the data are MCAR or MAR. If the data are MCAR, the observed units simply are a random subset of the sample. Imputing the average does have an effect on the estimation of the population variance, however. Let μ denote the population mean and σ^2 the population variance. We will draw a sample of size n, which contains r respondents and m nonrespondents. The set of respondents is denoted by $\mathcal{R}$ and the set of nonrespondents by $\mathcal{M}$. The estimated population mean will be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(\sum_{i\in\mathcal{R}} x_i + \sum_{i\in\mathcal{M}} x_i).$$

Imputing the respondent mean $\bar{x}_r$ for all missing items leads to: $\hat{\mu} = \bar{x}_r$. Then

$$\mathrm{E}[\hat{\mu}] = \mathrm{E}[\bar{X}_r] = \mathrm{E}[\frac{1}{r}\sum_{i\in\mathcal{R}} X_i] = \mathrm{E}[X] = \mu,$$

if the data are MCAR, and $\hat{\mu}$ is an unbiased estimator of the population mean. Note that if the data are MAR imputing the respondent mean within imputation classes will lead to the same results.

Now the estimated population variance is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_r)^2 = \frac{1}{n-1}\sum_{i\in\mathcal{R}}(x_i - \bar{x}_r)^2.$$

So

$$\mathrm{E}[\hat{\sigma}^2] = \frac{r}{n-1}\mathrm{E}[(X - \bar{X}_r)^2] = \frac{r}{n-1}\frac{r-1}{r}\sigma^2 = \frac{r-1}{n-1}\sigma^2 < \sigma^2.$$

This means that the estimate of the population variance is calculated solely based on the variance of the respondents. The missing item values that were

imputed are not accounted for, which leads to an underestimation of the true population variance. Note that inflating the estimated variance by the factor $(n-1)/(r-1)$ will provide an unbiased estimate of the population variance.

Another important issue is the precision of these estimates, which is overstated if the imputed values are treated as actual observed values. As a result the confidence intervals of the estimates will be too short and consequently this may lead to erroneous conclusions. Stochastic imputation methods, such as random hot deck or random regression imputation, were introduced to somewhat correct for this. Although the variance of the estimate will be closer to the true variance, this still does not completely account for the extra component of variation introduced by the imputation process.

In general the total true variance consists of two components: the ordinary sampling variance and the nonresponse variance. Sampling variance is the variance that occurs due to surveying only a subset of the population and would vanish if the whole population was sampled. Nonresponse variance is introduced by the nonresponse in the sample, so if the sample would be completely observed no nonresponse variance would arise. In the literature this is mostly referred to as imputation variance, as imputation is used to deal with nonresponse most of the time. Since we are also employing incomplete data procedures, such as the EM algorithm and available case analysis, we prefer to use the term nonresponse variance.

The main purpose of the methods treated in this section is to obtain a variance estimate for the total variance of a parameter. This will produce a measure of data quality and will make valid inference possible. Additionally, knowledge of the nonresponse portion of the total variance reveals the impact of nonresponse (and consequently imputation) on precision. In the context of repeated surveys, resources can then be better allocated between a larger sample or increased follow-up and more advanced imputation procedures. In general three types of methods have been developed to calculate variance in the presence of nonresponse and imputation.

1. *The analytical method*
   According to Särndal (1992) the total variance can be represented as

   $$V_{tot} = V_{sam} + V_{imp} + V_{mix}$$

   where V_{sam} is the sampling variance, V_{imp} the imputation variance and V_{mix} a covariance between the sampled and the imputed variables, which is often negligible. V_{sam} is the sampling variance in case of a complete sample. This means that a correction needs to be made for the imputed

part of the data.
These variances are now estimated separately. The problem with this approach is that Särndal only derived these quantities for simple random sampling with mean or ratio imputation. For more complex survey designs or imputation algorithms this method becomes analytically difficult.

2. *Replication and resampling methods*
Rao and Shao (1992) developed the adjusted jackknife technique for variance estimation after imputation. The adjusted jackknife variance estimate of an estimator is calculated as follows. Each sampled element i, $i = 1, \ldots, n$ is removed from the sample once and the missing items are imputed using the $n-1$ remaining respondents. Next the estimator of interest is calculated and the adjusted jackknife variance estimate is based on the n values for the estimator.
Another well-known variance estimation procedure that falls in this category is the bootstrap technique (Shao and Sitter, 1996). In this case a random sample with replacement of size n is drawn from the observed sample n. This means that in the bootstrap sample some respondents may be present more than once and others may not be present at all. Next the bootstrap sample is imputed and the estimator of interest is calculated. The bootstrap variance estimate is based on the different values for the estimator for some bootstrap samples.
A major advantage of these methods is the fact that the variance of complicated estimators can be calculated relatively easy without the theoretical derivation of variance formulas as is the case with the analytical method.

3. *Multiple imputation*
Another way to estimate the variance due to imputation is to impute several, say m, times and calculate the variance based on a combination of the within and the between variance of these m datasets. This is referred to as multiple imputation, which was developed by Rubin (1978, 1987). This method extremely simplifies variance estimation in the presence of imputation. However, the drawback is that the data user needs to incorporate m datasets in his analyses.

The main difference between the first two methods and multiple imputation, is the fact that in the first case missing data items are imputed only once, referred to as single value imputation, whereas with multiple imputation m imputations are generated for one missing data item. The actual consideration

between single value and multiple imputation is whether the benefit of an immediate relatively simple variance estimation outweighs the simplicity of imputing only once.

The obvious appeal of single value imputation is that it allows a straightforward use of standard complete data methods and software. Besides it is easy to implement and understand for data analysts. However, users may perceive the imputed dataset as a truly observed dataset, calculating variances using regular variance estimators. Unfortunately, the true variance of estimators can only be calculated using more advanced procedures, such as the ones described above.

Multiple imputation is intuitively attractive since it incorporates the idea that imputations have a certain variability. Although theoretically appealing, multiple imputation is not used very often in (large) surveys or at statistical agencies because of the practical implications. Multiple imputation requires maintaining and storing multiple complete data sets, which is operationally difficult. Besides, data dissemination such as the tabulation of data is seriously complicated by multiple imputation. Moreover, in order to obtain valid inference with multiple imputation, the imputation method used needs to be proper. That is, the imputations should satisfy conditions 1-3 in Rubin (1987, p 118-119). In words, a proper method is a method that has enough variability between replicates to provide appropriate variance estimates. As noted by Rao (1996) some commonly used imputation methods, including random hot deck and random regression imputation, are improper because these draws do not represent the full uncertainty in estimating the data for purposes of variance estimation with multiple imputation. Another consequence of this is that multiple imputation can only be used for random imputation methods, as there will be no variability with deterministic procedures. Note that this means that the popular nearest neighbour method cannot be used. Taking all this into account, statistical agencies prefer the use of single value imputation and as valid variance estimates can be obtained based on single imputation we will not recommend using multiple imputation.

1.2.5 Concluding remarks on imputation

Although imputation is a commonly used and convenient method to deal with missing data it is crucial that information about imputation is provided along with each dataset, in order to inform data users about data quality. First of all imputed values should always be flagged, so that the user will be able to distinguish between observed and constructed data. Furthermore, the user should be informed about what imputation techniques were used and additional informa-

tion should be given for certain techniques. For instance, if regression imputation was used the regressors should be specified, if nearest neighbour imputation was used the distance function and the number of times certain donors were used should be given, and if cold deck imputation was used the external sources should be described. Additionally nonresponse percentages and counts of the number of records that required at least one imputation are also informative.

In addition to this the statistical agency should evaluate the effects of imputation and if possible provide an estimate of the variance after imputation in order to give the user an insight in the precision of the data. Furthermore it should also be made clear to the user that relations between variables may be attenuated through imputation.

1.3 Erroneous data

As we mentioned in section 1.1, besides errors due to nonresponse, survey data can also be subject to measurement and processing error. In the case of nonresponse a missing data mechanism was developed, this can be done similarly for erroneous data in order to distinguish between stochastic and systematic errors.

1.3.1 The error mechanism

Stochastic errors are errors that are randomly introduced in the sample, for instance by a writing error of the respondent, by the respondent accidentally misreading or misunderstanding a question, or by a keying error made at the statistical agency. A distinction can be made between random errors throughout the sample and random errors within classes based on auxiliary data such as branch of business or size class. One may expect small companies to make more errors than large companies as their accounting system may be less well-kept.

On the other hand systematic errors are those errors that arise because respondents do not understand or misinterpret concepts, definitions or questions being asked, or because of faults in concepts or procedures of data collection and processing. Systematic errors have a greater potential to affect the quality of survey estimates than random errors; if a large number of respondents misinterpret a question in the same way, a bias will be introduced in the estimates. Examples are when gross values are reported instead of net values, and when values are reported in Euros instead of the requested thousands of Euros. Most of the time systematic errors are hard to locate as records will not seem suspicious or implausible compared to other respondents, who made the same

mistake. Besides, if a respondent reports all variables in gross values instead of net values, accounting identities will not be violated and the record will seem correct. In order to get some insight in systematic errors questionnaires should be tested and the observed data should be compared to several external sources, such as values from previous periods and tax records.

The most difficult issue with regard to erroneous data is how to determine whether data (items) are incorrect or not and subsequently how to correct data items once they are found to be in error. Several approaches have been used to deal with data that contain errors. These approaches will be discussed in the next section.

1.3.2 Strategies for dealing with errors

In general there are three ways to deal with errors in survey data.

1. *Direct analysis of the erroneous data*
   One option is to do nothing. This means that one assumes that the data do not contain errors. In the case of large systematic errors this approach may lead to a large bias in the estimates. If only small random errors are present in the data, the estimates will probably be reasonably accurate.

2. *Measurement error models*
   Another way to deal with erroneous data is to use measurement error models, also referred to as errors-in-variables models, see for example Wansbeek and Meijer (2000). A simple form of a classical measurement model is
   $$\mathbf{X} = \mathbf{X}^* + \mathbf{e},$$
   where $\mathbf{X}$ is the measured value and $\mathbf{X}^*$ the true value, $\mathbf{e}$ represents the measurement error and is mostly assumed to be normally distributed with mean zero and covariance $\mathbf{\Sigma}$. Measurement error models also assume that the errors are independent of the actual value, which means that they are mostly only suitable for data with stochastic errors that are approximately normally distributed.
   The aim of measurement error modelling is to obtain survey estimates that are free from error. It, however, does not attempt to provide a complete correct dataset. Besides, the assumption that errors are normally distributed may be quite unlikely even for random errors.

3. *Data editing*
   A third option is to use data editing. Data editing is a procedure that loc-

ates individual errors using pre-specified rules, these errors will be removed and subsequently the data item will be imputed. The main advantage of data editing as opposed to measurement error models is the fact that data editing will provide us with a corrected and consistent dataset, which can then be analysed or released for publication. If the data published by a statistical agency still contain obvious mistakes, even if they are small, the agency will surely loose credibility.

As our main interest is to obtain a complete multi-purpose dataset our focus will be on data editing. Editing is becoming increasingly important. As we said before, without editing data intended for tabulation, publication and research may be spurious, which could harm the perceived reliability of the statistical agency.

1.3.3 Data editing

Editing is the localisation of erroneous or suspicious values in the data, which is done based on checks. These checks are referred to as edit rules, edit restrictions or simply edits.

In general two types of editing are distinguished. First of all editing can be applied as a *validating* procedure. In this case editing is used to detect inconsistencies and errors within a certain record. Examples are checking whether the parts of a sum add up to the reported total, and whether a ratio of values, such as wages divided by the number of employees, is within certain bounds. Other consistency edits are of the 'if-then' type, for example if a business is of a certain size then the costs of wages need to be at a certain level. For economic data most of the consistency edits are accounting definitions which must hold. However, some of the consistency edits are constructed by subject-matter specialists based on previous experience or specific knowledge.

Secondly, editing can be applied as a *statistical* procedure. In this case editing is used to detect errors or inconsistencies across records. The edits are mostly based on statistical analysis of the data. For example, by detecting outliers based on the sample, or by using models to obtain edit limits for a certain variable using a trend based on previous values of that record and the other responding units. In business surveys outliers are common and difficult to treat because the outlier may very well be a legitimate value and therefore should not be changed. However, outliers should always be identified and excluded from the imputation model as they may have a large effect on the imputed values.

Another classification of edits is in fatal and query edits. The first refers to

item values that do not satisfy edits, and therefore are erroneous with certainty, such as accounting identities that do not hold. Examples are the fact that profit has to equal turnover minus operating expenses and the fact that the purchasing price of goods should not exceed total operating expenses. If the record does not satisfy these restrictions it is certain that some items are incorrect.

Query edits concern item values that are highly unlikely, but may in fact be true. An example of a query edit is the fact that a value is expected to be within certain bounds based on other variables in the survey, but it need not be. Caution needs to be taken in the case of query edits, as the danger arises that the data are over-edited by using too many query edits or bounds that are too strict. Over-editing refers to the fact that the impact of editing on survey estimates is negligible and therefore the editing is unnecessary and thus needlessly expensive. Over-editing can, however, also refer to the practice that the data are edited too much in order to fit the expectations of the editor, resulting in possible bias. For further reading on this subject see Granquist (1997).

A last important distinction that is often used in this thesis is the difference between balance and inequality edit restrictions. Balance edits refer to equality restrictions on the data and inequality edits refer to inequality restrictions. Note that balance edits are always fatal edits, but that inequality edits can be both fatal and query edits. Fatal inequality edits are, for instance, the fact that turnover should be non-negative or that the number of employees should exceed the number of employees in full time equivalent. A query inequality edit is the fact that turnover is expected to be larger than the number of employees multiplied by a constant.

1.3.4 The editing process at Statistics Netherlands

The traditional editing process at statistical agencies was generally as follows. Erroneous or suspicious data were located by means of edit rules and then subject-matter specialists corrected these values by recontacting the respondent or using their expert knowledge and external information to obtain imputations. In order to speed up the editing process and lower the associated costs, nowadays editing is also done automatically using some sort of model and a computer.

Currently at Statistics Netherlands a combination of manual and automatic editing is used, which is referred to as selective editing. In this case only the most influential and implausible records are reviewed manually, the other records are edited automatically. Each reporting unit is given a certain score based on several aspects, such as business size, sampling weight, relative importance of erro-

neous items and so on. Records that have a score above a certain predetermined value are reviewed manually. Clearly this results in a far more efficient editing process, especially since it has been recognised that it is not necessary to correct all data in every detail in order to obtain valid survey estimates, see Granquist (1997) and Granquist and Kovar (1997). Results published by a statistical office are usually aggregated data, such as totals or means, so small (random) errors will often cancel out when aggregated. Moreover, in this instance the measurement and processing error will become negligible with respect to the sampling error.

The automatic editing at Statistics Netherlands is done using validity editing. This means that if a record fails the stated validity edit restrictions it is considered to be in error. Once an erroneous record is encountered, the incorrect item values within that record need to be found. The widely used Fellegi-Holt method is employed for this purpose. Fellegi and Holt (1976) developed a procedure that locates the erroneous items within a failed record by assuming that the number of errors made are as few as possible. This means that the number of errors occurring in a record equals the minimum number of values that have to be changed within this record such that it satisfies all edit restrictions.

Once an item is found to be erroneous it is set to missing and needs to be imputed, which leads to the main subject of this thesis as these items need to be imputed taking the linear edit restrictions into account.

1.3.5 Concluding remarks on editing

As we have emphasized before, data users should always be aware of the fact that the data have been edited. This means that imputed items should be flagged. We would recommend different flags for items that were truly missing and items that were imputed because they were found to be in error, such that the user will be able to distinguish between them. Additionally, other numbers on editing should be provided along with the dataset. The number of edits that were violated for each record and the number of times a certain variable is present in violated edits will give the user insight in data quality. Furthermore, the number of times a variable has been changed and the relative change that has been made is also informative.

Currently editing is only used as a procedure to clean up data. It can, however, be much more useful. Editing provides crucial information about the quality of the data and the survey process and can therefore suggest valuable changes for future surveys, for example, by identifying nonsampling error sources.

1.4 The relationship between editing and imputation

There is a strong relationship between editing and imputation as the data that were found to be incorrect need to be imputed and additionally the imputed values need to satisfy the edit constraints. This seriously complicates the imputation process, especially for economic data which are subject to a large number of linear edit restrictions. The general imputation procedures that were described in section 1.2.3 do not take edit restrictions into account, which means that imputed values are likely not to satisfy the restrictions on the data.

Consider hot deck imputation as an example. In this instance the missing data items are replaced by responses from a similar donor record. If some of the operating expenses, which need to add up to the total operating expenses reported, are missing the imputed donor values will almost certainly not be consistent with this balance edit. This means that the imputed values need to be adjusted afterwards, in order to satisfy the restrictions. Although this can be done quite straightforwardly using an optimisation algorithm and requiring minimal change in the imputed values, the effects can be harmful as changing the imputed values will distort the distribution of the imputed values and consequently the distribution of the final completed dataset.

Besides, the edits provide the imputer with valuable information as they restrict the possible outcomes of the imputed values. For instance, if the total operating expenses is reported but some of the components that constitute this total are missing, the imputer does know the sum of the variables that need to be imputed. This information should definitely be incorporated in the imputation model. Therefore it is desirable to be able to generate imputations that are properly distributed while simultaneously satisfying all edit restrictions on the data.

Despite the importance of economic data no general procedures have been developed for the imputation of missing data items such that the edit constraints are utilised and satisfied. The main focus of this thesis will therefore be on the development of imputation methods that generate imputations satisfying all edit restrictions and preserving the distribution of the data simultaneously.

Example 1.2.3 continued. The EM algorithm in SPSS
The fact that the starting values of the EM algorithm cannot be specified in SPSS does not lead to serious problems in common well-behaved missing data situations. If we are dealing with economic data that are subject to balance

Table 1.1: *Maximum likelihood estimates obtained by the EM algorithm.*

	EM using AC estimates (SPSS)		EM using CC estimates	
Costs of sales	44344	(1)	44335	(0)
Labour costs	5037	(15)	5086	(0)
Other personnel costs	839	(0)	860	(0)
Costs of transportation	509	(0)	532	(0)
Costs of energy	169	(0)	179	(0)
Costs of housing	630	(-1)	623	(0)
Costs of machinery and equipment	240	(0)	251	(0)
Sales expenses	1081	(2)	1077	(0)
Communication expenses	98	(n/a)	96	(n/a)
Costs of rendering of services	434	(0)	433	(0)
Other company expenses	596	(n/a)	234	(n/a)
Depreciation on fixed assets	770	(n/a)	770	(n/a)
Total operating expenses	54483	(-264)	54476	(0)
			$\times$1000	Euro

restrictions, however, the available cases (AC) starting values supplied by SPSS lead to imputations that do not satisfy the balance restrictions on the data. Whereas using starting values that do satisfy the balance restrictions, such as the complete cases (CC) estimates, does provide imputations that satisfy these restrictions.

This is illustrated by Table 1.1, where we have calculated the maximum likelihood estimates of the population means of different operating expenses in the trade industry by means of the EM algorithm, assuming normality. The data items are subject to a considerable amount of balance edits. First of all, the different operating expenses need to add up to the total operating expenses. Secondly, most of these operating expenses represent a total of an underlying balance edit restriction. The amount of violation of the underlying balance edit is shown between brackets. If there is no underlying balance edit, this is indicated by n/a (not applicable).

A formal proof of the fact that the EM estimates satisfy the balance restrictions if the starting values are chosen such that the restrictions are satisfied as

well will be given in chapter 4. Note that for now we have ignored the fact that the variables also need to satisfy non-negativity restrictions. ■

1.4.1 Linear balance and inequality restrictions

As we mentioned earlier a distinction often is made between balance and inequality restrictions. As these two types of restrictions have different implications for the imputation process, we will treat these restrictions separately as well. In this section we will give a more formal definition of the restrictions that need to hold throughout this thesis.

Consider an $n \times k$ data matrix $\mathbf{X}$ and let $\mathbf{X}_i$ denote the ith row of $\mathbf{X}$, corresponding to the ith respondent. Now define a $p \times k$ restriction matrix $\mathbf{A}$ that contains all balance restrictions on the data items, such that $\mathbf{AX}_i = \mathbf{0}$. We will assume that there are no redundant balance restrictions, which means that the number of balance restrictions cannot exceed the number of variables and $\mathbf{A}$ is of full row rank. Some examples of balance restrictions that occur are the fact that different company expenses add up to the total operating expenses, that financial result equals financial income minus financial expenses, or that the total number of employees equals the number of employees on the company's payroll plus personnel lent out to other businesses.

Now define an $r \times k$ restriction matrix $\mathbf{B}$ that contains all inequality restrictions on the data and the set $G := \{\mathbf{X}_i \in \mathbb{R}^k : \mathbf{l} \leq \mathbf{BX}_i \leq \mathbf{u}\}$, which defines all possible outcomes of the data vector $\mathbf{X}_i$. The upper and lower bounds $\mathbf{u}$ and $\mathbf{l}$ may equal plus or minus infinity respectively. Note that in the case of non-negativity restrictions $\mathbf{B}$ will be the identity matrix, $\mathbf{l} = \mathbf{0}$ and $\mathbf{u} = \boldsymbol{\infty}$. Also note that $\mathbf{B}$ need not be of full row rank as the data may be subject to non-negativity restrictions as well as all kinds of other inequality restrictions, so the number of restrictions r may exceed the number of variables k. Some examples of inequality restrictions are the fact the total number of employees is larger than the total number of employees in full time equivalent or the fact that the number of employees on a company's payroll is smaller than the total labour costs (in thousands of Euros) and larger than 0.001 times the total labour costs.

So, the data completed by imputation need to satisfy $\mathbf{AX}_i = \mathbf{0}$ and $\mathbf{l} \leq \mathbf{BX}_i \leq \mathbf{u}$. Due to the linear nature of these restrictions it is not possible to use nonlinear transformations, such as Box-Cox transformations, in a multivariate model as the edit structure would be lost in that case. For example, taking natural logarithms of the variables in the balance restriction $X_1 + X_2 + \cdots + X_{k-1} = X_k$ does not imply $\ln X_1 + \ln X_2 + \cdots + \ln X_{k-1} = \ln X_k$. This means that we cannot establish restrictions for the transformed data and consequently the im-

puted data are still likely not to satisfy the linear restrictions.

1.5 Overview of this thesis

The focus of this thesis is on the imputation of (economic) data that are subject to different types of linear restrictions. A strong need for imputation models that can incorporate restrictions has arisen at statistical agencies as this results in data for publication that do not contain obvious mistakes, i.e. violations of the linear restrictions, which could seriously harm the credibility of a statistical agency. In this study several imputation procedures are developed and analysed in order to provide the imputer with a set of models that can be used for varying types of restriction structures and datasets.

In chapters 3, 4, 5 and 6 we develop imputation methods that deal with restriction structures of varying complexity. For an overview of which restriction structures can be handled by the imputation models that are discussed in these chapters see Table 1.2, where $\mathbf{A}$ denotes the $p \times k$ matrix containing the balance restrictions and $\mathbf{B}$ the $r \times k$ matrix containing the inequality restrictions. Note that $\mathbf{B}$ can represent non-negativity as well as other inequality constraints.

Every chapter on imputation models is written according to the following structure. First, the different types of restrictions that can be incorporated in the imputation model are discussed. Next the suggested imputation model is treated and, subsequently, estimation procedures that provide parameter estimates for the model of interest are derived. Finally, deterministic and stochastic imputation procedures are discussed and applications to empirical data are presented.

First, in chapter 2 we discuss maximum likelihood estimation as throughout this thesis this procedure is used to obtain parameter estimates. Furthermore, the EM algorithm that was developed to obtain maximum likelihood estimates in the presence of nonresponse is treated extensively as well and as we may need to estimate high-dimensional integrals in the EM algorithm, Markov chain Monte Carlo methods are also be dealt with. The aim of this chapter is to provide a theoretical background that is needed for a complete understanding of the material of the later chapters in this thesis.

In chapter 3 we develop an imputation method that uses the Dirichlet distribution to model the data. This method is convenient because of its flexibility. This procedure can impute data items that are non-negative and subject to one linear balance restriction.

The Dirichlet method cannot incorporate multiple balance restrictions, however. Therefore, in chapter 4, we suggest the use of the multivariate singular

Table 1.2: Overview of the restriction structures that are dealt with.

	Inequality restrictions		
Balance restrictions	None	Non-negativity	$\mathbf{l} \leq \mathbf{BX}_i \leq \mathbf{u}$, with $\mathbf{B}$ any $r \times k$ matrix
$\mathbf{AX} = \mathbf{0}$, $p = 1$	chapter 4	chapter 3	chapter 5, chapter 6
$\mathbf{AX} = \mathbf{0}$, $p > 1$	chapter 4	chapter 5, chapter 6	chapter 5, chapter 6

normal distribution to deal with multiple balance restrictions instead of only one. It is found that the EM algorithm for multivariate normally distributed data can be extended such that singular normal data can be managed as well. This leads to an imputation procedure that is easy to implement and whose properties are well-known.

As inequality restrictions are not incorporated in the singular normal model, there is still the need for a general purpose method that can handle all sorts of balance and inequality restrictions. With this objective, the multivariate singular normal density is truncated to the region defined by the inequality restrictions in chapter 5.

This truncated singular normal distribution consists of high-dimensional integrals and consequently leads to complex modelling issues, therefore a completely different approach is investigated in chapter 6. In this chapter the joint model is split into a sequence of univariate conditional distributions. These univariate conditional models are used to sequentially impute each variable. This method can incorporate both balance and inequality restrictions simultaneously as well.

Finally, in chapter 7 conclusions, directions for future research and examples of possible applications in other fields of interest are given.

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 ith 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 $f(\cdot \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$.

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

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

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

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

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 $\boldsymbol{\Theta}$ is compact and the loglikelihood function is continuous on $\boldsymbol{\Theta}$, then the maximum likelihood estimate exists. In addition to this, if $\boldsymbol{\Theta}$ is convex and the loglikelihood function is strictly concave on $\boldsymbol{\Theta}$, then the maximum likelihood estimate is unique.

Assuming that

- the maximum is in the interior of the parameter space $\boldsymbol{\Theta}$,
- the range of the random variable $\mathbf{X}_i$ is independent of the parameter $\boldsymbol{\theta}$,
- $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

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

$$\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{2.1}$$

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 $\ell(\boldsymbol{\theta} \mid \mathbf{x})$ near $\boldsymbol{\theta}^{(t)}$, where $\boldsymbol{\theta}^{(t)}$ is an initial guess

$$\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}} \simeq \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} + \mathbf{H}(\boldsymbol{\theta} \mid \mathbf{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}),$$

where $\mathbf{H}(\boldsymbol{\theta} \mid \mathbf{x})$ is the Hessian: $\frac{\partial^2 \ell(\boldsymbol{\theta}|\mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$. Since (2.1) holds at a maximum we solve for $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t+1)}$

$$\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} + \mathbf{H}(\boldsymbol{\theta} \mid \mathbf{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) = \mathbf{0}.$$

This results in the Newton-Raphson algorithm

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\theta} \mid \mathbf{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}, \quad t = 1, 2, \ldots,$$

where $\mathbf{I}(\boldsymbol{\theta} \mid \mathbf{x})$, which is the negative of the Hessian, is the observed information matrix. Note that in order to find a maximum the Hessian needs to be negative definite, which is not guaranteed. Furthermore, if the starting value is not too far from the maximum likelihood estimate, the Taylor series expansion is accurate and the updated value will be closer to the maximum likelihood estimate. If the starting value is far from the maximum likelihood estimate, however, the Taylor series expansion is not that accurate and the algorithm may not converge to the maximum likelihood estimate.

Fisher scoring, which is a variation to Newton-Raphson, uses

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathcal{I}^{-1}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}, \quad t = 1, 2, \ldots,$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix, which is the expectation of the observed information matrix: $\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{I}(\boldsymbol{\theta} \mid \mathbf{X})]$. The advantage of this method lies in the fact that (under certain regularity conditions)

$$\mathrm{E}\Big[\frac{\partial^2 \ell(\boldsymbol{\theta} \mid \mathbf{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big] = -\mathrm{E}\Big[\sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\theta} \mid \mathbf{X}_i)}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i(\boldsymbol{\theta} \mid \mathbf{X}_i)}{\partial \boldsymbol{\theta}'}\Big], \tag{2.2}$$

where $\ell_i(\boldsymbol{\theta} \mid \mathbf{x}_i) = \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})$. 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

$$\frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{f(\mathbf{x}_i \mid \boldsymbol{\theta})} \frac{\partial f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and therefore

$$\frac{\partial f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i \mid \boldsymbol{\theta}). \tag{2.3}$$

For a probability density function the following holds by definition

$$\int f(\mathbf{x}_i \mid \boldsymbol{\theta})\, d\mathbf{x}_i = 1. \tag{2.4}$$

The density $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ is regular if the range of the random variable $\mathbf{X}_i$ is independent of the parameter vector $\boldsymbol{\theta}$ and if the first and second order derivatives of $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are bounded by integrable functions of $\mathbf{X}_i$ (see Fomby et al., 1984). Assuming that the density $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ is regular, differentiating (2.4) with respect to $\boldsymbol{\theta}$ leads to

$$\int \frac{\partial f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\mathbf{x}_i = \mathbf{0}.$$

Using equation (2.3), this results in

$$\int \frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i \mid \boldsymbol{\theta})\, d\mathbf{x}_i = \mathbf{0}, \tag{2.5}$$

which means that

$$\mathrm{E}\Big[\frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big] = \mathbf{0}. \tag{2.6}$$

Differentiating (2.5) with respect to $\boldsymbol{\theta}'$ yields

$$\int \Big(\frac{\partial^2 \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{x}_i \mid \boldsymbol{\theta}) + \frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big)\, d\mathbf{x}_i = \mathbf{0}.$$

Again using the equality in (2.3), we obtain

$$\int \Big(\frac{\partial^2 \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\mathbf{x}_i \mid \boldsymbol{\theta}) + \frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{x}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(\mathbf{x}_i \mid \boldsymbol{\theta})\Big)\, d\mathbf{x}_i = \mathbf{0},$$

which leads to

$$\mathrm{E}\Big[\frac{\partial^2 \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big] = \mathbf{0}.$$

This results in the information equality given in equation (2.2) as follows

$$\begin{aligned}
\mathrm{E}\Big[\frac{\partial^2 \ell(\boldsymbol{\theta} \mid \mathbf{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big] &= \mathrm{E}\Big[\frac{\partial^2 \sum_{i=1}^n \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big] \\
&= \mathrm{E}\Big[\sum_{i=1}^n \frac{\partial^2 \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big] \\
&= -\mathrm{E}\Big[\sum_{i=1}^n \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big].
\end{aligned}$$

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

$$\mathrm{E}\Big[\sum_{i=1}^n \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big] = n\mathrm{Var}\Big(\frac{\partial \ln f(\mathbf{X}_i \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big) \tag{2.7}$$

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 $\hat{\boldsymbol{\theta}}$ if the loglikelihood function is concave and unimodal. Moreover, both methods yield an approximation to $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ 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: $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}^{-1}(\boldsymbol{\theta}))$. These regularity conditions require that the domain of the random variables does not depend on the parameter vector $\boldsymbol{\theta}$ and that the density $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ possesses derivatives of at least third order with respect to $\boldsymbol{\theta}$, which are bounded by integrable functions of $\mathbf{X}_i$ (Fomby et al., 1984). Now consider the Taylor series expansion of the first order derivative of $\ell(\hat{\boldsymbol{\theta}} \mid \mathbf{x})$ about the true parameter value $\boldsymbol{\theta}$

$$\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \simeq \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}} + \mathbf{H}(\boldsymbol{\theta} \mid \mathbf{x})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

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

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \simeq -\mathbf{H}^{-1}(\boldsymbol{\theta} \mid \mathbf{x}) \frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{x})}{\partial \boldsymbol{\theta}}.$$

Multiply both sides by $\sqrt{n}$

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) &= -\mathbf{H}^{-1}(\boldsymbol{\theta}\mid\mathbf{x})\sqrt{n}\frac{\partial\ell(\boldsymbol{\theta}\mid\mathbf{x})}{\partial\boldsymbol{\theta}}\\ &= -\big(\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}\mid\mathbf{x})\big)^{-1}\sqrt{n}\frac{1}{n}\frac{\partial\ell(\boldsymbol{\theta}\mid\mathbf{x})}{\partial\boldsymbol{\theta}}.\end{aligned}$$

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

$$\begin{aligned}\mathrm{Var}\Big(\sqrt{n}\frac{1}{n}\frac{\partial\ell(\boldsymbol{\theta}\mid\mathbf{X})}{\partial\boldsymbol{\theta}}\Big) &= \mathrm{Var}\Big(\frac{\partial\ln f(\mathbf{X}_i\mid\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big)\\ &= \frac{1}{n}\mathcal{I}(\boldsymbol{\theta}),\end{aligned}$$

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

$$-\big(\frac{1}{n}\mathbf{H}(\boldsymbol{\theta}\mid\mathbf{x})\big)^{-1}\sqrt{n}\frac{1}{n}\frac{\partial\ell(\boldsymbol{\theta}\mid\mathbf{x})}{\partial\boldsymbol{\theta}}\xrightarrow{d}\big(\frac{1}{n}\mathcal{I}(\boldsymbol{\theta})\big)^{-1}\mathbf{Z},$$

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

$$\big(\frac{1}{n}\mathcal{I}(\boldsymbol{\theta})\big)^{-1}\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\big(\frac{1}{n}\mathcal{I}(\boldsymbol{\theta})\big)^{-1}).$$

This means that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}(\mathbf{0},\big(\frac{1}{n}\mathcal{I}(\boldsymbol{\theta})\big)^{-1}),$$

where $\mathcal{I}^{-1}(\boldsymbol{\theta})$ can be estimated consistently by $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ or $I^{-1}(\hat{\boldsymbol{\theta}})$, 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 $\mathbf{X}_i$ denote the complete data vector for respondent i, which can be partitioned into a missing

and an observed part, $\mathbf{X}_i' = (\mathbf{X}_{i,mis}', \mathbf{X}_{i,obs}')$.

The complete data now have a density function $f(\mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta})$. Let $\mathbf{M}_i = (m_{ij})$ 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 $\mathcal{L}(\mathbf{M}_i \mid \mathbf{X}_{i,mis}, \mathbf{X}_{i,obs}, \boldsymbol{\phi})$ is the missing data mechanism, where $\boldsymbol{\phi}$ is a parameter vector. The MCAR assumption states that this distribution does not depend on the data at all, so

$$f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}) = f(\mathbf{m}_i \mid \boldsymbol{\phi}).$$

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

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

$$f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}) = f(\mathbf{m}_i \mid \mathbf{x}_{i,obs}, \boldsymbol{\phi}).$$

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

The joint density of the data vector $\mathbf{X}_i$ and the missing data indicator $\mathbf{M}_i$ is

$$f(\mathbf{m}_i, \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}) f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}).$$

The joint density of the actual observed data is therefore

$$f(\mathbf{m}_i, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f(\mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}) f(\mathbf{m}_i \mid \mathbf{x}_{i,mis}, \mathbf{x}_{i,obs}, \boldsymbol{\phi}) d\mathbf{x}_{i,mis}.$$

If the data are MAR this reduces to

$$\begin{aligned} f(\mathbf{m}_i, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) &= \int f(\mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}) f(\mathbf{m}_i \mid \mathbf{x}_{i,obs}, \boldsymbol{\phi}) d\mathbf{x}_{i,mis} \\ &= f(\mathbf{m}_i \mid \mathbf{x}_{i,obs}, \boldsymbol{\phi}) \int f(\mathbf{x}_{i,mis}, \mathbf{x}_{i,obs} \mid \boldsymbol{\theta}) d\mathbf{x}_{i,mis} \\ &= f(\mathbf{m}_i \mid \mathbf{x}_{i,obs}, \boldsymbol{\phi}) f(\mathbf{x}_{i,obs} \mid \boldsymbol{\theta}). \end{aligned}$$

The likelihood of the observed data under the MAR assumption can thus be factorised into two factors where the first factor depends on $\boldsymbol{\phi}$ and the second one on $\boldsymbol{\theta}$. 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 $\boldsymbol{\theta}$ can be based solely on the observed data loglikelihood, $\ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs})$.

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 $\tilde{\boldsymbol{\theta}}$. Next re-estimate $\boldsymbol{\theta}$ 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 $f(\cdot \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$. Again partition $\mathbf{X}_i$ into a missing and an observed part, $\mathbf{X}_i' = (\mathbf{X}_{i,mis}', \mathbf{X}_{i,obs}')$. If the complete data vector $\mathbf{X}_i$ were observed, we would compute the maximum likelihood estimates of $\boldsymbol{\theta}$ 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

$$\begin{aligned} f(\mathbf{x}_i \mid \boldsymbol{\theta}) &= f(\mathbf{x}_{i,obs}, \mathbf{x}_{i,mis} \mid \boldsymbol{\theta}) \\ &= g(\mathbf{x}_{i,obs} \mid \boldsymbol{\theta}) h(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\theta}). \end{aligned}$$

From which it follows that

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}) = \ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) + \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}), \tag{2.8}$$

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

In general the EM algorithm operates as follows:

Expectation Step
Find the conditional expectation of $\ell(\boldsymbol{\theta} \mid \mathbf{X})$ given $\mathbf{X}_{obs}$ at $\boldsymbol{\theta}^{(t)}$:

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

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

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

Repeat these two steps until the estimates converge, that is until $\ell(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{x}) - \ell(\boldsymbol{\theta}^{(t)} \mid \mathbf{x}) < \varepsilon$, with ε 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 $\mathbf{X}_{mis}$, they are the results of the estimated function of $\mathbf{X}_{mis}$ appearing in the complete data loglikelihood function.

If the complete data loglikelihood is linear in the data $\mathbf{X}$, the E-step is equivalent to filling in the missing data $\mathbf{X}_{mis}$ by their expectation $\mathrm{E}[\mathbf{X}_{mis} \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}]$ and the M-step is equivalent to finding $\boldsymbol{\theta}^{(t+1)}$ 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

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) = \ell(\boldsymbol{\theta} \mid \mathbf{x}) - \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}).$$

Taking the expectation with respect to $\mathbf{X}_{mis}$ conditionally given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ leads to

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$

where

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{\theta} \mid \mathbf{x}) h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis}$$

and

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}) h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis}.$$

Now calculate

$$\begin{aligned}
&H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \\
&\qquad = \int \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}) h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis} \\
&\qquad\quad - \int \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis} \\
&\qquad = \int \ln \frac{h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta})}{h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})} h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis} \\
&\qquad = \mathrm{E}[\ln \frac{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta})}{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})} \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}].
\end{aligned}$$

Note that this is the negative of the Kullback-Leibler information number defined by

$$K\big(h(\mathbf{x} \mid \boldsymbol{\theta}^{(t)}), h(\mathbf{x} \mid \boldsymbol{\theta})\big) = \int h(\mathbf{x} \mid \boldsymbol{\theta}^{(t)}) \ln \frac{h(\mathbf{x} \mid \boldsymbol{\theta}^{(t)})}{h(\mathbf{x} \mid \boldsymbol{\theta})} \, d\mathbf{x},$$

which is always non-negative and therefore $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \leq 0$. We will show this using Jensen's inequality.

Recall that Jensen's inequality states that $\mathrm{E}[f(X)] \leq f(\mathrm{E}[X])$ if f is concave. Take $f(X) = \ln X$. Then

$$\begin{aligned}
&\mathrm{E}[\ln \frac{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta})}{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})} \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}] \\
&\qquad \leq \ln \mathrm{E}[\frac{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta})}{h(\mathbf{X}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})} \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}] \\
&\qquad = \ln \int \frac{h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta})}{h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})} h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis} \\
&\qquad \leq \ln \int h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}) d\mathbf{x}_{mis} \\
&\qquad = \ln 1 = 0.
\end{aligned}$$

This means that

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \leq H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \quad \text{for any } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

The difference in loglikelihood is

$$\begin{aligned}\ell(\boldsymbol{\theta}^{(t+1)} \mid \mathbf{x}_{obs}) - \ell(\boldsymbol{\theta}^{(t)} \mid \mathbf{x}_{obs}) &= Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) + \\ &\quad -Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}),\end{aligned}$$

which is non-negative, as for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

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

and

$$H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \leq H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}).$$

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 $\boldsymbol{\theta}$. 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 $\boldsymbol{\theta}$ 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 $\mathbf{M}(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)}$, $t = 0, 1, \ldots$. Suppose that $\mathbf{M}$ is differentiable. The parameter $\boldsymbol{\theta}^{(t)}$ converges to $\boldsymbol{\theta}^*$, so $\boldsymbol{\theta}^* = \mathbf{M}(\boldsymbol{\theta}^*)$, which means that $\boldsymbol{\theta}^*$ is a fixed point of $\mathbf{M}$. Consider the first order Taylor series expansion of $\boldsymbol{\theta}^{(t+1)} = \mathbf{M}(\boldsymbol{\theta}^{(t)})$ about $\boldsymbol{\theta}^*$

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

This leads to

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

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

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

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

2.2.6 The missing information principle

Recall that we can write the observed data loglikelihood as follows

$$-\ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) = -\ell(\boldsymbol{\theta} \mid \mathbf{x}) + \ln h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}).$$

Differentiating this equation twice with respect to $\boldsymbol{\theta}$ leads to

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

which results in

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

where $\mathbf{I}_{obs}(\boldsymbol{\theta} \mid \mathbf{x}_{obs})$ is the observed information matrix and $\mathbf{I}_{com}(\boldsymbol{\theta} \mid \mathbf{x})$ the complete information matrix. Taking the expectation over $\mathbf{X}_{mis}$ conditionally

given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ of the left-hand side results in the observed information matrix since it does not depend on $\mathbf{X}_{mis}$. Taking the expectation over $\mathbf{X}_{mis}$ conditionally given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ of the right-hand side and assuming that the order of integration and differentiation can be interchanged we obtain

$$\mathbf{I}_{obs}(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}),$$

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

$$\mathbf{I}_{obs}(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) = \mathcal{I}_{com}(\boldsymbol{\theta} \mid \mathbf{x}_{obs}) - \mathcal{I}_{mis}(\boldsymbol{\theta} \mid \mathbf{x}_{obs}),$$

where $\mathcal{I}_{com}(\boldsymbol{\theta} \mid \mathbf{x}_{obs})$ is the expected complete information matrix and $\mathcal{I}_{mis}(\boldsymbol{\theta} \mid \mathbf{x}_{obs})$ 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 $\hat{\boldsymbol{\theta}}_{mle}$ is usually estimated by the inverse of the observed information matrix, $\mathbf{I}^{-1}_{obs}(\hat{\boldsymbol{\theta}}_{mle} \mid \mathbf{x}_{obs})$. 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

$$\frac{\partial \mathbf{M}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \mathcal{I}^{-1}_{com}(\boldsymbol{\theta}^* \mid \mathbf{x}_{obs})\mathcal{I}_{mis}(\boldsymbol{\theta}^* \mid \mathbf{x}_{obs}).$$

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 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, in other words we look for a $\boldsymbol{\theta}^{(t+1)}$ such that

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

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

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

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

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

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 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 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 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 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:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \Big[\frac{\partial^2 Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\Big]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.$$

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

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 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. 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 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ by simulation, which results in the following general simulated EM algorithm.

S-step
Simulate values $\mathbf{x}_{mis}$ from the conditional distribution of $\mathbf{X}_{mis}$ given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Calculate the complete data loglikelihood based on the simulated data and estimate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ by $\hat{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.

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

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 $\mathbf{x}_{mis}^{(1)}, \ldots, \mathbf{x}_{mis}^{(M)}$ from the conditional distribution of the missing data. Then estimate the complete data loglikelihood by calculating

   $$\hat{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} \ell(\boldsymbol{\theta} \mid \mathbf{x}_{obs}, \mathbf{x}_{mis}^{(m)}).$$

   If M goes to infinity this quantity converges to $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 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 $\boldsymbol{\theta}^{(t)}$ against the

iteration number. When convergence is indicated by the stabilisation of the process with random fluctuations about $\hat{\boldsymbol{\theta}}$, 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 $\boldsymbol{\theta}$. The sequence $\tilde{\boldsymbol{\theta}}^{(j)}$, $j = 1, \ldots, t$ forms a Markov chain and the estimator $\tilde{\boldsymbol{\theta}}^{(t)}$ 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:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = a(\boldsymbol{\theta}) b(\mathbf{x}) \exp\big(\mathbf{c}(\boldsymbol{\theta})' \mathbf{t}(\mathbf{x})\big),$$

where a, b and $\mathbf{t}$ are known functions and $\mathbf{c}(\boldsymbol{\theta})$ is often called the natural parameter. If $\mathbf{t}(\mathbf{x}) = \mathbf{x}$, 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 $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \ldots, t_k(\mathbf{x}))$ are sufficient for $\boldsymbol{\theta}$.

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

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \exp\big(\boldsymbol{\eta}' \mathbf{t}(\mathbf{x}) + d(\mathbf{x}) + g(\boldsymbol{\eta})\big), \tag{2.9}$$

where $\boldsymbol{\eta} = \mathbf{c}(\boldsymbol{\theta})$ is the natural parameter. If $\boldsymbol{\eta}$ and $\mathbf{t}(\mathbf{x})$ are linearly independent and $\{\boldsymbol{\eta} : \boldsymbol{\eta} = (c_1(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is an open set, then the family is

a regular exponential family, which means that the sufficient statistic $\mathbf{t}(\mathbf{x})$ 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

$$\int f(\mathbf{x} \mid \boldsymbol{\eta}) d\mathbf{x} = 1. \tag{2.10}$$

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

$$\int \frac{\partial f(\mathbf{x} \mid \boldsymbol{\eta})}{\partial \eta_j} d\mathbf{x} = 0. \tag{2.11}$$

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

$$\int \frac{\partial^2 f(\mathbf{x} \mid \boldsymbol{\eta})}{\partial \eta_j \partial \eta_l'} d\mathbf{x} = 0, \quad j, l = 1, \ldots, k. \tag{2.12}$$

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

$$\int \frac{\partial f(\mathbf{x} \mid \boldsymbol{\eta})}{\partial \eta_j} d\mathbf{x} = \int f(\mathbf{x} \mid \boldsymbol{\eta}) \Big(t_j(\mathbf{x}) + \frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j} \Big) d\mathbf{x} = 0, \quad j = 1, \ldots, k,$$

which results in

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

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

$$\int f(\mathbf{x} \mid \boldsymbol{\eta}) \left(\Big(t_j(\mathbf{x}) + \frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j} \Big) \Big(t_l(\mathbf{x}) + \frac{\partial g(\boldsymbol{\eta})}{\partial \eta_l} \Big) + \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_l} \right) d\mathbf{x} = 0.$$

Using the result from (2.13) this leads to

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

and

$$\begin{aligned}\mathrm{Cov}(t_j(\mathbf{X}), t_l(\mathbf{X})) &= \mathrm{E}[t_j(\mathbf{X})t_l(\mathbf{X})] - \mathrm{E}[t_j(\mathbf{X})]\mathrm{E}[t_l(\mathbf{X})] \\ &= -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_l}, \quad j,l = 1,\ldots,k.\end{aligned}$$

So the covariance matrix of $\mathbf{t}(\mathbf{X})$ will be

$$\mathrm{Var}\big(\mathbf{t}(\mathbf{X})\big) = \begin{pmatrix} -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_1} & \cdots & -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_k} \\ \vdots & \ddots & \vdots \\ -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_1} & \cdots & -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_k} \end{pmatrix}.$$

2.3.3 Maximum likelihood estimation for exponential families

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

$$\ell(\boldsymbol{\eta} \mid \mathbf{x}) = \boldsymbol{\eta}' \sum_{i=1}^{n} \mathbf{t}(\mathbf{x}_i) + \sum_{i=1}^{n} d(\mathbf{x}_i) + n g(\boldsymbol{\eta}). \tag{2.14}$$

The first order derivative is

$$\frac{\partial \ell(\boldsymbol{\eta} \mid \mathbf{x})}{\partial \eta_j} = \sum_{i=1}^{n} t_j(\mathbf{x}_i) + n\frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j}, \quad j = 1,\ldots,k,$$

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

$$\frac{\partial^2 \ell(\boldsymbol{\eta} \mid \mathbf{x})}{\partial \eta_j \partial \eta_l} = n\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_l}, \quad j,l = 1,\ldots,k.$$

This means that the Hessian is

$$\mathbf{H}(\boldsymbol{\eta}) = n \begin{pmatrix} \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_1 \partial \eta_k} \\ \vdots & \ddots & \vdots \\ -\frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_1} & \cdots & \frac{\partial^2 g(\boldsymbol{\eta})}{\partial \eta_k \partial \eta_k} \end{pmatrix} = -n\mathrm{Var}\big(\mathbf{t}(\mathbf{X})\big).$$

Since $\mathrm{Var}\big(\mathbf{t}(\mathbf{X})\big)$ 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 $Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(t)})$ in the E-step

$$\begin{aligned} Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(t)}) &= \mathrm{E}[\ell(\boldsymbol{\eta} \mid \mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}] \\ &= n\boldsymbol{\eta}'\mathrm{E}[\mathbf{t}(\mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}] + \\ &\quad + n\mathrm{E}[d(\mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}] + ng(\boldsymbol{\eta}). \end{aligned}$$

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

$$\begin{aligned} &\frac{\partial Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(t)})}{\partial \boldsymbol{\eta}} = n\mathrm{E}[\mathbf{t}(\mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}] + n\frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = 0 \\ &\Rightarrow \mathrm{E}[\mathbf{t}(\mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}] = -\frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}. \end{aligned}$$

This means that we only need to calculate $\mathbf{t}^{(t)}(\mathbf{X}) = \mathrm{E}[\mathbf{t}(\mathbf{X}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\eta}^{(t)}]$ in the E-step and calculate $\boldsymbol{\eta}^{(t+1)}$ as a solution to $\frac{\partial g(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = -\mathbf{t}^{(t)}(\mathbf{X})$ 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

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

This can be reparameterised by

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

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

$$\begin{aligned} \mathrm{E}[t_1(X)] &= -\frac{\partial g(\theta_1,\theta_2)}{\partial\theta_1} = \frac{1}{4}\theta_2^2\theta_1^{-2} - \frac{1}{2}\theta_1^{-1} = \mu^2 + \sigma^2 \\ \mathrm{E}[t_2(X)] &= -\frac{\partial g(\theta_1,\theta_2)}{\partial\theta_2} = -\frac{1}{2}\theta_1^{-1}\theta_2 = \mu. \end{aligned}$$

So, the EM consists of calculating

$$\begin{aligned} \mu^{(t+1)} &= \frac{1}{n}(\sum_{i=1}^{r} X_i + (n-r)\mu^{(t)}) \\ (\sigma^{(t+1)})^2 &= \frac{1}{n}\Big(\sum_{i=1}^{r} X_i^2 + (n-r)\big((\sigma^{(t)})^2 + (\mu^{(t)})^2\big)\Big) - (\mu^{(t+1)})^2, \end{aligned}$$

until convergence. ■

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

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

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

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

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

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

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

$$\mathrm{Var}\big(\hat{I}(h)\big) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\Big) = \frac{\sigma^2}{n},$$

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

$$\widehat{\mathrm{Var}}\big(g(X)\big) = \frac{1}{n-1}\sum_{i=1}^{n} g(X_i)^2 - \frac{1}{(n-1)n}\Big(\sum_{i=1}^{n} g(X_i)\Big)^2.$$

If the second moment is finite, the Central Limit Theorem states that

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

So the error estimate of Monte Carlo integration is of order $O(n^{-\frac{1}{2}})$. 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

$$\int_a^b h(x)dx = \int_a^b \frac{h(x)}{p(x)} p(x)dx = \mathrm{E}\Big[\frac{h(X)}{p(X)}\Big]$$

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

$$\mathrm{Var}\big(\hat{I}(h)\big) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{h(X_i)}{p(X_i)}\Big) = \frac{1}{n}\mathrm{Var}\Big(\frac{h(X)}{p(X)}\Big),$$

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

$$\widehat{\mathrm{Var}}\Big(\frac{h(X)}{p(X)}\Big) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{h(X_i)}{p(X_i)}\Big)^2 - \frac{1}{(n-1)n}\Big(\sum_{i=1}^{n}\frac{h(X_i)}{p(X_i)}\Big)^2.$$

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 $\mathrm{Var}(h(X)/p(X))$ 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

$$\begin{aligned} I(h) &= \int\cdots\int_G h(\mathbf{x})d\mathbf{x} = \int\cdots\int_G g(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \mathrm{E}[g(\mathbf{X})] \simeq \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i) = \hat{I}(h). \end{aligned} \tag{2.15}$$

If we draw $\mathbf{X}_i$ uniformly from G the integral will be estimated by

$$\hat{I}(h) = \frac{V}{n}\sum_{i=1}^{n} g(\mathbf{X}_i),$$

where V is the volume of the region G.

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

$$\mathrm{Var}\big(\hat{I}(h)\big) = \mathrm{Var}\big(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i)\big) = \frac{1}{n}\mathrm{Var}\big(g(\mathbf{X})\big).$$

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 π. 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

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

$$\begin{aligned} K(\mathbf{x}^{(t)}, A) &\equiv \Pr(\mathbf{X}^{(t+1)} \in A \mid \mathbf{X}^{(t)} = \mathbf{x}^{(t)}, \ldots, \mathbf{X}^{(0)} = \mathbf{x}^{(0)}) \\ &= \Pr(\mathbf{X}^{(t+1)} \in A \mid \mathbf{X}^{(t)} = \mathbf{x}^{(t)}), \quad \mathbf{x}^{(t)} \in S, \quad A \subset S, \end{aligned}$$

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

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

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

In the continuous case, the transition kernel is

$$K(\mathbf{x}, A) = \int_A k(\mathbf{x}, \mathbf{y}) d\mathbf{y},$$

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

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

$$\pi(A) = \int K(\mathbf{x}, A) \pi(\mathbf{x}) d\mathbf{x},$$

for all sets A. This condition states that if $\mathbf{X}^{(t)}$ is distributed according to π, then all subsequent elements of the chain are also distributed as π.

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 π. 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 π-irreducible if for each set A with $\pi(A) > 0$ there exists a $t \geq 1$ so that $\Pr(\mathbf{X}^{(t)} \in A \mid \mathbf{X}^{(0)} = \mathbf{x}) > 0$ for all $\mathbf{x} \in S$. 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 $k(\mathbf{x}, \cdot)$ is positive in a neighbourhood of $\mathbf{x}$, 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: $T_A = \min\{t : \mathbf{X}^{(t)} \in A \mid \mathbf{X}^{(0)} \in A\}$. Then A is positive recurrent if $\mathrm{E}[T_A] < \infty$.

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 η_A be the number of times $\mathbf{X}^{(t)}$ is in A, then $\Pr_{\mathbf{x}}(\eta_A = \infty)$ is the probability of visiting A infinitely often starting from $\mathbf{x}$. A set A is Harris recurrent if $\Pr_{\mathbf{x}}(\eta_A = \infty) = 1$ for all $\mathbf{x} \in S$. A Markov chain is Harris recurrent if it is π-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 $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(N)})$ are draws generated by an ergodic, Harris recurrent Markov chain, from a target distribution π, then the expectation of any function $h(\mathbf{x})$ under π can be estimated by the average

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

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

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

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

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

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

In Metropolis-Hastings algorithms this transition density is constructed as

$$k(\mathbf{x},\mathbf{y}) = q(\mathbf{x},\mathbf{y})\alpha(\mathbf{x},\mathbf{y}), \quad \mathbf{x} \neq \mathbf{y} \in S,$$

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

Hastings suggested an acceptance probability of the form

$$\alpha(\mathbf{x},\mathbf{y}) = \begin{cases} \min\left\{\frac{\pi(\mathbf{y})q(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x},\mathbf{y})},1\right\} & \text{if } \pi(\mathbf{x})q(\mathbf{x},\mathbf{y}) > 0 \\ 1 & \text{otherwise} \end{cases}. \tag{2.17}$$

Note that the detailed balance equation holds for every Metropolis-Hastings chain generated by the transition kernel $k(\mathbf{x}, \mathbf{y})$ with the acceptance probability defined as (2.17). This means that π is a stationary distribution of this chain.

In practice the Metropolis-Hastings algorithm works as follows.

Metropolis-Hastings algorithm

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

A sufficient condition for the Metropolis-Hastings chain to be aperiodic is that the algorithm allows to retain the present value: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$, 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 π. Moreover, the Metropolis-Hastings chain is Harris recurrent if it is π-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 π.

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 $q(\mathbf{x}, \mathbf{y}) = q_1(\mathbf{y} - \mathbf{x})$, where q_1 is a multivariate density. The candidate is calculated as $\mathbf{y} = \mathbf{x} + \mathbf{z}$, where $\mathbf{z}$ 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 $\mathbf{z}$. 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 $\mathbf{x}$: $q(\mathbf{x}, \mathbf{y}) = q_2(\mathbf{y})$. 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 $q_2(\mathbf{y}) = \pi(\mathbf{y})$ 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 π. 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 tth 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

$$\begin{array}{lll} \text{draw} & X_1^{(t+1)} & \text{from } f_1(x_1 \mid x_2^{(t)}, \ldots, x_k^{(t)}) \\ \text{draw} & X_2^{(t+1)} & \text{from } f_2(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_k^{(t)}) \\ & \vdots & \\ \text{draw} & X_k^{(t+1)} & \text{from } f_k(x_k \mid x_1^{(t+1)}, \ldots, x_{k-1}^{(t+1)}). \end{array}$$

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 $q(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y} \mid \mathbf{x})$ be the Metropolis-Hastings proposal. In this case the acceptance probability α given in (2.17) is always one as

$$\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})} \frac{\pi(\mathbf{x} \mid \mathbf{y})}{\pi(\mathbf{y} \mid \mathbf{x})} = \frac{\pi(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x}, \mathbf{y})} = 1,$$

which means that we accept every move.

The Gibbs Markov chain is aperiodic since $q(\mathbf{x}, \mathbf{x}) > 0$ for all $\mathbf{x} \in S$. 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 $\mathbf{X}$ is distributed according to a truncated multivariate normal distribution

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

where

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

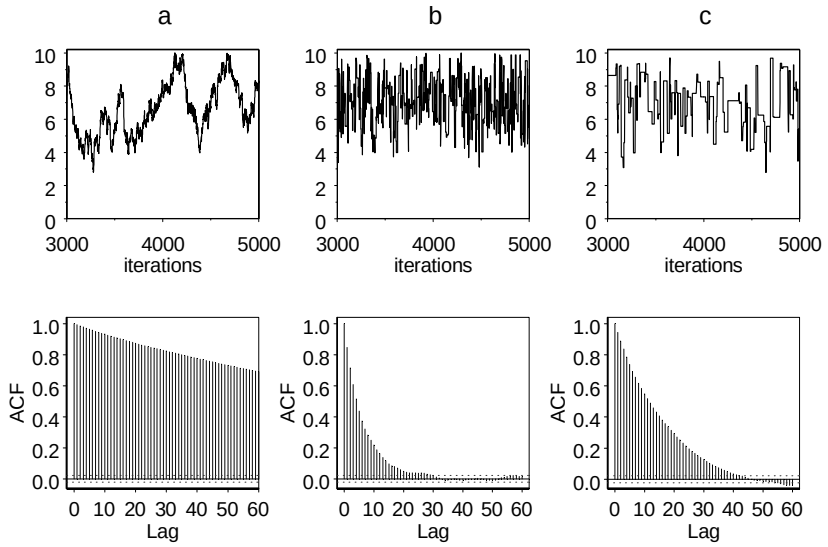

Figure 2.1: *Trace and autocorrelation plots of x_3 for a truncated normal Markov chain with a) $c^2 = 0.01$, b) $c^2 = 1.88$ and c) $c^2 = 8$.*

and the data are truncated to the region

$$\begin{aligned} x_1 + x_2 &\leq x_3 \\ x_3 &\leq 10 \\ x_1, x_2 &\geq 0. \end{aligned}$$

We will generate a Markov chain with a truncated normal stationary distribution that satisfies these restrictions with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The chain is generated using the random walk Metropolis-Hastings algorithm, with a multivariate normal proposal distribution, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, c^2\boldsymbol{\Sigma})$. As we mentioned earlier the choice of the variance of $\mathbf{Z}$ 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 $c^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ 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.

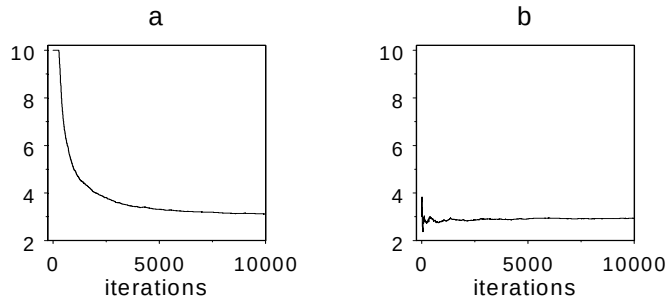

Figure 2.2: *Trace plots of the running mean for variable* x_2*, using as starting value a)* $\mathbf{x}^{(0)} = (0, 10, 10)'$ *and b)* $\mathbf{x}^{(0)} = (2, 3, 7)'$.

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

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

and

$$W = \frac{1}{m} \frac{1}{N-1} \sum_{i=1}^{m} \sum_{j=1}^{N} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i.})^2,$$

where $\bar{\mathbf{X}}_{i.}$ is the average of chain i, and $\bar{\mathbf{X}}_{..}$ 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 σ^2 as the individual chains have not had time to visit the whole target distribution. The Gelman-Rubin diagnostic is

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

where $\hat{R}$ 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 overdispersed 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_{min} that is needed when the sample is identically and independently distributed, can be determined. Now we wish to estimate $\Pr(U(\mathbf{X}) \leq u)$ to within $q \pm r$ with probability s. Calculate $U^{(t)} = U(\mathbf{X}^{(t)})$ where $\mathbf{X}^{(t)}$ 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 $\hat{\mathbf{P}}$ of $\{Z_k^{(t)}\}$ is estimated by

$$\hat{P}_{ij} = \frac{\#\{t : Z_k^{(t)} = j, Z_k^{t-1} = i\}}{\#\{t : Z_k^{t-1} = i\}}, \quad i, j = 0, 1.$$

Then set $\alpha = \hat{P}_{01}$ and $\beta = \hat{P}_{10}$. The burn-in length is $M = m^*k$, with

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

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

$$n^* = \frac{\frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^2}}{(\frac{r}{\Phi(\frac{1}{2}(1+s))})^2},$$

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.

Chapter 3

Imputation of Data Subject to One Balance Restriction

Common well-known imputation methods provide imputations that are likely not to satisfy the linear balance restrictions on economic data. Adjusting these values in order to satisfy the restrictions will lead to a distortion of the distribution of the imputed data and consequently to a distortion of the distribution of the completed data. Besides, valuable information provided by the balance restrictions is not used. In this chapter we suggest an imputation procedure that generates imputations for non-negative data items subject to one linear balance restriction. Our suggestion is to use the Dirichlet distribution to model the data.

3.1 Introduction

In chapter 1 an elaborate overview was given of frequently used imputation methods. We established that imputation methods can be either deterministic or stochastic. Deterministic methods determine imputed values uniquely, which means that when the imputation process is repeated the same value will be imputed. Stochastic methods depend on some sort of randomness, which means that when the process is repeated, other values may be imputed. Deterministic imputation methods avoid the loss in precision associated with the added randomness as opposed to stochastic methods. Therefore these methods are well suited to estimate means or totals. However, the variance will be underestimated and the shape of the distribution will be distorted. So for the creation of

general purpose datasets, stochastic imputation is preferred. We prefer to have an imputation method that can be both deterministic as well as stochastic so that they are suitable for both situations.

A property of economic data is that there are many logical constraints on the data items, such as the fact that total operating expenses must be the added total of all operating expenses reported, such as housing costs, depreciation, personnel costs and so on. Commonly used imputation methods such as hot deck and (random) regression imputation mostly do not provide imputations that satisfy these linear balance restrictions. This means that imputations need to be adjusted in order to satisfy the balance restrictions on the data. Although this can be done rather straightforwardly using linear optimisation techniques such as the simplex algorithm, the adjustment of imputed values inadvertently leads to a distortion of the distribution of the imputed values and therefore to a distortion of the distribution of the final, completed dataset. Besides, these balance restrictions provide the imputer with valuable information on what values (not) to impute and therefore it would be desirable to incorporate the restrictions in the imputation process.

In this chapter we suggest the use of an imputation scheme to obtain imputations for data items that are subject to one linear balance restriction and that need to be non-negative. In section 3.2 the edit constraints are described and in section 3.3 the Dirichlet distribution and some of its properties are discussed. Subsequently parameter estimation and the Expectation-Maximisation (EM) algorithm that is used to obtain the maximum likelihood estimates in the presence of nonresponse are treated in sections 3.4 and 3.5. In section 3.6 imputation is discussed and in section 3.7 some results are presented on the performance of this imputation method, compared to other missing data procedures, on empirical data. Finally, in section 3.8 some concluding remarks will be made.

3.2 The edit constraint

Consider a balance edit restriction of the following shape

$$\mathbf{c}'\mathbf{X} = X_k,$$

where $\mathbf{X}$ is a vector of order $(k-1) \times 1$ which contains solely non-negative elements and the c_j's, $j = 1, \ldots, k-1$, are known, positive constants. This restriction, for example, could refer to different types of operating expenses $(X_1, \ldots, X_{k-1})$ that need to add up to the total operating expenses (X_k), where

$c_j = 1$, for $j = 1, \ldots, k-1$. This restriction can be transformed by dividing the different parts by the total X_k in order to restrict the domain of the resulting variables $Y_1, \ldots, Y_{k-1}$ to the simplex:

$$\begin{array}{ccccc} \frac{c_1 X_1}{X_k} & + \cdots + & \frac{c_{k-1} X_{k-1}}{X_k} = 1, & X_k > 0 \\ Y_1 & + \cdots + & Y_{k-1} & = 1. \end{array} \tag{3.1}$$

Note that we assume that the total, X_k, is known. This is plausible for two reasons. First of all since X_k is an aggregate the nonresponse rate will probably be low. And secondly if it is indeed missing we expect to be able to estimate this value quite accurately based on the other variables in the survey, whereas subtotals are far more difficult to estimate this way. For notational convenience define $l = k - 1$, then the data that need to be modelled and imputed are the ratios $Y_1, \ldots, Y_l$.

A well-known and popular imputation scheme is the hot deck method. In the current situation the traditional hot deck method needs to be slightly altered in order to make sure the imputed items satisfy the linear balance restriction. In the adjusted hot deck method not the actual responses of a donor but their ratios with respect to the total X_k are used as imputations. We refer to this method as the ratio hot deck method, see e.g. Pannekoek and de Waal (2005). A disadvantage of this method is the fact that we need to find a suitable donor, which can be complicated for economic data, especially if the nonresponse rates are substantial. One can either select a donor randomly or use some sort of distance function to find the donor that is most similar to the record with missing data items. The latter procedure will probably lead to better results but is also more difficult as one needs to establish what variables should be present in the distance function, i.e. on what variables should the donor search be based. Furthermore, these variables need to be standardised in order to have equal weights. Finally, for large datasets this method can become quite slow as distances need to be calculated between the record with missing items and all available donors.

The method we suggest based on the Dirichlet distribution offers a flexible, model-based, alternative to ratio hot deck methods.

3.3 A statistical distribution of economic data

Models for distributions are often chosen on the basis of the range of the random variable. For a variable constrained between zero and one the beta distribution

has proved useful. The beta distribution is defined by the probability density function

$$f(y \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},\ 0 < y < 1,\ \alpha, \beta > 0,$$

where $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(y) = \int_0^\infty u^{y-1}e^{-u}du$.

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$ and asymmetric otherwise. Besides it can be hump-shaped ($\alpha, \beta > 1$) or U-shaped ($\alpha, \beta < 1$). Note that it reduces to the uniform distribution if $\alpha = \beta = 1$. Also note that if $Y \sim \text{beta}(\alpha, \beta)$ then $1 - Y \sim \text{beta}(\beta, \alpha)$.

An extension of the beta distribution is the so-called Dirichlet distribution (see Johnson and Kotz, 1972), also referred to as the multivariate beta. Its pdf is

$$f(y_1, \ldots, y_l \mid \alpha_1, \ldots, \alpha_l) = \frac{\Gamma(\sum_{j=1}^l \alpha_j)}{\prod_{j=1}^l \Gamma(\alpha_j)} \prod_{j=1}^l y_j^{\alpha_j - 1}, \tag{3.2}$$

where

$$y_j \geq 0,\ \alpha_j > 0,\ j = 1, \ldots, l,\ \sum_{j=1}^l y_j = 1.$$

Note that since $\sum_{j=1}^l y_j = 1$, this is actually a $(l-1)$-dimensional distribution; one variable can be obtained with certainty from the others. Consequently, the pdf is sometimes written as

$$f(y_1, \ldots, y_{l-1} \mid \alpha_1, \ldots, \alpha_{l-1}; \alpha_l) = \frac{\Gamma(\sum_{j=1}^l \alpha_j)}{\prod_{j=1}^l \Gamma(\alpha_j)} \prod_{j=1}^{l-1} y_j^{\alpha_j - 1} \big(1 - \sum_{j=1}^{l-1} y_j\big)^{\alpha_l - 1}. \tag{3.3}$$

We will refer to the Dirichlet distribution given by (3.2) with $\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_l)$ and the Dirichlet distribution given by (3.3) with $\text{Dir}_{l-1}(\alpha_1, \ldots, \alpha_{l-1}; \alpha_l)$. Note that in the case of $l = 2$ the Dirichlet reduces to the beta distribution. The Dirichlet is a convenient distribution on the simplex, as the family of Dirichlet distributions is an exponential family and has complete sufficient statistics of the form $\ln Y_j$, $j = 1, \ldots, l$.

The restriction that the data of interest lie in the simplex appears in several disciplines, such as geology, medicine and biology as well as in economics. This type of data is often referred to as compositional data. For a detailed description of applications and statistical analysis methods for compositional data, see

Aitchison (1986). Aitchison and Shen (1980) introduce the logistic normal distribution as a framework for the analysis of compositional data as an alternative to the Dirichlet distribution. This technique assumes multivariate normality of additive log-ratio transformed data, which means that the inference tools for multivariate normal data can be applied to the transformed compositions. The additive log-ratio transformation of a vector $\mathbf{Y}$ in the $(l-1)$-dimensional simplex to $\mathbb{R}^{l-1}$ is

$$W_j = \ln\Big(\frac{Y_j}{Y_l}\Big), \quad j = 1, \ldots, l-1.$$

The (unrestricted) vector $\mathbf{W}$ is then modelled with a $(l-1)$-dimensional multivariate normal distribution. However, if we are confronted with missing data, the fact that $\mathbf{W}$ is unrestricted will most likely result in imputations that do not satisfy the composition after transforming $\mathbf{W}$ back to $\mathbf{Y}$: $Y_j = \exp\{W_j\}Y_l$. Furthermore, although Aitchison (1986) shows that all statistical procedures are invariant to the choice of component used as the denominator, this poses difficulties in an imputation context as this component needs to be observed for all records, which is highly unlikely. Additionally, this method is not readily applicable if some of the components are zero. Because of this and the fact that draws from the Dirichlet yield compositions immediately, the Dirichlet distribution seems more appropriate for the imputation of compositional data. For different applications of the Dirichlet distribution to compositional data see, for example, Jeuland et al. (1980), DeSarbo et al. (1993) and Haas and Formery (2002).

In Figure 3.1 the flexibility of the Dirichlet distribution is illustrated by some examples of Dirichlet density plots for the parameterisation given in (3.3) and $l = 3$. Figure 1 (a1) shows the density for $\boldsymbol{\alpha} = (1, 1; 2)$ and (a2) shows the density for $\boldsymbol{\alpha} = (1, 1; 6)$. Two humpshaped densities are shown in Figure 1 (b1) and (b2) with the parameters $\boldsymbol{\alpha} = (2, 2; 2)$ and $\boldsymbol{\alpha} = (2, 2; 6)$ respectively. Finally in Figure 1 (c1) and (c2) two U-shaped densities are shown with parameters $\boldsymbol{\alpha} = (0.2, 0.2; 0.2)$ and $\boldsymbol{\alpha} = (0.2, 0.2; 0.6)$, respectively.

The first and second order moments of the Dirichlet distribution are

$$\begin{aligned} \mathrm{E}[Y_j] &= \frac{\alpha_j}{\alpha}, & j = 1, \ldots, l \\ \mathrm{Var}(Y_j) &= \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}, & j = 1, \ldots, l, \end{aligned}$$

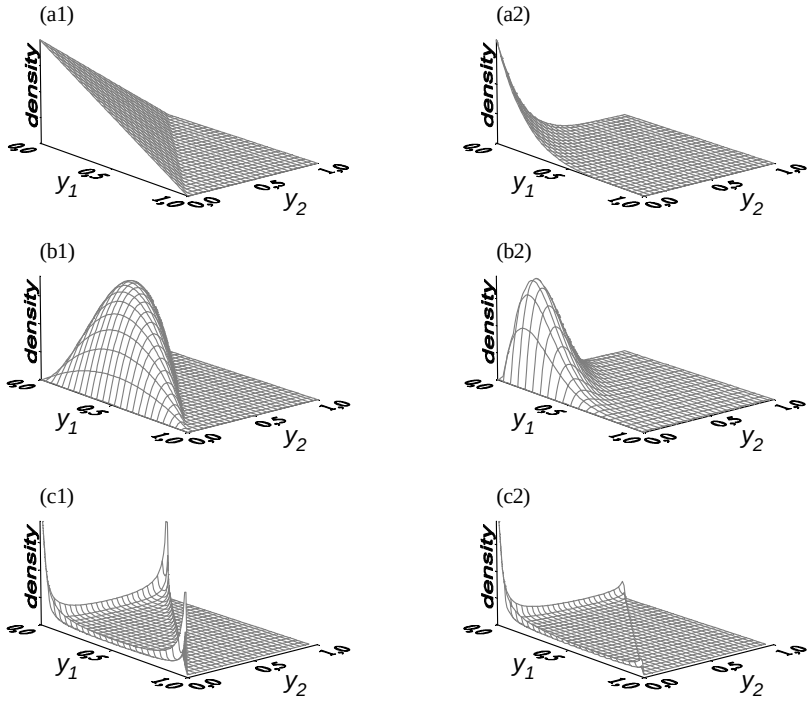

Figure 3.1: *Some bivariate Dirichlet density plots for* (a1) $\boldsymbol{\alpha} = (1, 1; 2)$, (a2) $\boldsymbol{\alpha} = (1, 1; 6)$, (b1) $\boldsymbol{\alpha} = (2, 2; 2)$, (b2) $\boldsymbol{\alpha} = (2, 2; 6)$, (c1) $\boldsymbol{\alpha} = (0.2, 0.2; 0.2)$ *and* (c2) $\boldsymbol{\alpha} = (0.2, 0.2; 0.6)$.

where $\alpha = \sum_{j=1}^{l} \alpha_j$ and the covariances between Y_j and Y_h are

$$\mathrm{Cov}(Y_j, Y_h) = \frac{-\alpha_j \alpha_h}{\alpha^2(\alpha + 1)}, \quad j, h = 1, \ldots, l, \ j \neq h.$$

Notice that if the means are held constant but α is allowed to increase, the variances and covariances decrease. For this reason α can be regarded as some sort of precision parameter: as α increases the distribution becomes more tightly concentrated about the mean.

The following theorems apply (see Wilks, 1962, for a derivation of these theorems).

Theorem 1 (Marginal Dirichlet)
If $\mathbf{Y} = (Y_1, \ldots, Y_l)$ *is a random variable vector having the* $(l-1)$*-variate Dirichlet distribution* $\mathrm{Dir}_{l-1}(\alpha_1, \ldots, \alpha_{l-1}; \alpha_l)$*, then the marginal distribution of* $\mathbf{Y}_1 = (Y_1, \ldots, Y_{l_1})$*, with* $l_1 < l$ *is the* $(l_1 - 1)$*-variate Dirichlet distribution* $\mathrm{Dir}_{l_1-1}(\alpha_1, \ldots, \alpha_{l_1-1}; \alpha_{l_1} + \cdots + \alpha_l)$*.*

Theorem 2 (Conditional Dirichlet)
If $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')' \sim \mathrm{Dir}_{l-1}(\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$ *where* $\mathbf{Y}_1$ *and* $\boldsymbol{\alpha}_1$ *consist of* l_1 *elements and* $\mathbf{Y}_2$ *and* $\boldsymbol{\alpha}_2$ *consist of* l_2 *elements and* $l = l_1 + l_2$*, then*

$$\mathbf{Y}_1^* \mid \mathbf{Y}_2 \sim \mathrm{Dir}_{l_1-1}(\boldsymbol{\alpha}_1'),$$

where $\mathbf{Y}_1^* = (1 - \mathbf{Y}_2' \boldsymbol{\iota}_{l_2})^{-1} \mathbf{Y}_1$*, with* $\boldsymbol{\iota}_{l_2}$ *a vector of ones of length* l_2*. In this way the vector* $\mathbf{Y}_1$ *is rescaled such that the elements of* $\mathbf{Y}_1^*$ *sum up to one.*

This second theorem is useful in the context of imputation as the missing data, that need to be imputed, conditional on the observed data still follow a Dirichlet distribution.

3.4 Parameter estimation

3.4.1 The method of moments estimator

The parameters $\alpha_1, \ldots, \alpha_l$ of the Dirichlet distribution can be estimated by a method of moments (MM) estimator, which is consistent. Recall that the first and second order moments of the Dirichlet distribution are

$$\mu_j = \frac{\alpha_j}{\alpha}, \qquad j = 1, \ldots, l \tag{3.4}$$

$$\sigma_j^2 = \frac{\alpha_j(\alpha - \alpha_j)}{\alpha^2(\alpha + 1)}, \quad j = 1, \ldots, l, \tag{3.5}$$

where $\alpha = \sum_{j=1}^{l} \alpha_j$. Rewrite (3.4) as $\alpha = \frac{\alpha_j}{\mu_j}$ and substitute this in equation (3.5). Then

$$\begin{aligned} \sigma_j^2 &= \frac{\alpha_j\left(\frac{\alpha_j}{\mu_j} - \alpha_j\right)}{\left(\frac{\alpha_j}{\mu_j}\right)^2\left(\frac{\alpha_j}{\mu_j} + 1\right)} \\ \left(\frac{\alpha_j}{\mu_j} + 1\right)\sigma_j^2 &= (1 - \mu_j)\mu_j. \end{aligned}$$

Solving for α_j gives the moments estimator

$$\hat{\alpha}_{MM,j} = \hat{\mu}_j\big(\frac{\hat{\mu}_j}{\hat{\sigma}_j^2}(1-\hat{\mu}_j)-1\big), \quad j=1,\ldots,l,$$

where $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}$, $j=1,\ldots,l$.

Although the method of moments is straightforward, estimation based on the method of moments generally is not statistically efficient. That is, the asymptotic covariance matrix of the estimators is usually larger than the inverse of the information matrix. However, estimation based on the method of moments can serve as an excellent initial guess to start iterations in the Newton-Raphson algorithm, which we use to maximise the likelihood function.

3.4.2 Maximum likelihood estimation

In order to find a consistent estimator for $\boldsymbol{\alpha}$, that is statistically efficient, maximum likelihood estimation can be applied, see e.g. Narayanan (1991) and Ronning (1989). The likelihood function for the $(l-1)$-variate Dirichlet distribution is

$$L(\boldsymbol{\alpha} \mid \mathbf{y}) = \frac{\Gamma^n(\sum_{j=1}^{l}\alpha_j)}{\prod_{j=1}^{l}\Gamma^n(\alpha_j)}\prod_{i=1}^{n}\prod_{j=1}^{l} y_{ij}^{\alpha_j-1}.$$

Taking the natural logarithm leads to the following loglikelihood function

$$\ell(\boldsymbol{\alpha} \mid \mathbf{y}) = n\ln\Gamma(\sum_{j=1}^{l}\alpha_j) - n\sum_{j=1}^{l}\ln\Gamma(\alpha_j) + \sum_{i=1}^{n}\sum_{j=1}^{l}(\alpha_j-1)\ln y_{ij}.$$

In order to obtain the maximum likelihood estimates we need to set the first order derivatives of the (log)likelihood equal to zero. These first order derivatives, for $j=1,\ldots,l$, are

$$S_j(\boldsymbol{\alpha}) = \frac{\partial\ell(\boldsymbol{\alpha} \mid \mathbf{y})}{\partial\alpha_j} = n\Psi(\sum_{h=1}^{l}\alpha_h) - n\Psi(\alpha_j) + \sum_{i=1}^{n}\ln y_{ij}, \tag{3.6}$$

where $\Psi(\cdot)$ is the digamma function defined by $\Psi(y) = \frac{\partial\ln\Gamma(y)}{\partial y} = \frac{\Gamma'(y)}{\Gamma(y)}$.

As these equations cannot be solved analytically, we need some iterative scheme to find the optimal parameter values. Commonly used methods are the Newton-Raphson method and the Fisher scoring technique. In case of the Dirichlet distribution the negative of the Hessian matrix of second derivatives is

equal to the expected information matrix, which we will show momentarily, and therefore the Newton-Raphson and Fisher scoring algorithms are the same, thus

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(t)}}\mathbf{S}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(t)}}, \qquad t = 1, 2, \ldots,$$

where $\mathbf{I}(\boldsymbol{\alpha})$ is the observed (or expected) information matrix and $\mathbf{S}(\boldsymbol{\alpha})$ is the score vector, which contains the first order derivatives of the loglikelihood with respect to $\boldsymbol{\alpha}$. The Hessian, $\mathbf{H}(\boldsymbol{\alpha})$, is the matrix of second order derivatives of $\ell(\boldsymbol{\alpha} \mid \mathbf{y})$, with elements

$$\begin{aligned}
\frac{\partial^2 \ell(\boldsymbol{\alpha} \mid \mathbf{y})}{\partial \alpha_j^2} &= n\Psi'(\sum_{h=1}^{l} \alpha_h) - n\Psi'(\alpha_j), \quad j = 1, \ldots, l \\
\frac{\partial^2 \ell(\boldsymbol{\alpha} \mid \mathbf{y})}{\partial \alpha_j \partial \alpha_p} &= n\Psi'(\sum_{h=1}^{l} \alpha_h), \qquad\qquad j, p = 1, \ldots, l,\ j \neq p,
\end{aligned}$$

where $\Psi'(\cdot)$ is known as the trigamma function. The observed information matrix is therefore

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\alpha}) = -\mathbf{H}(\boldsymbol{\alpha}) &= n[\mathrm{diag}\{\Psi'(\alpha_1), \ldots, \Psi'(\alpha_l)\} - \Psi'(\sum_{j=1}^{l} \alpha_j)\boldsymbol{\iota}_l \boldsymbol{\iota}_l'] \\
&= n[\mathbf{D} - c\boldsymbol{\iota}_l \boldsymbol{\iota}_l'],
\end{aligned}$$

where $\boldsymbol{\iota}_l$ denotes a vector of ones of length l. Clearly the observed information is independent of the random variable $\mathbf{Y}$, meaning that $\mathrm{E}[\mathbf{I}(\boldsymbol{\alpha})] = \mathbf{I}(\boldsymbol{\alpha})$, so in this case the expected information equals the observed information matrix.

The inverse of $\mathbf{I}(\boldsymbol{\alpha})$ can be easily calculated using a well-known matrix inversion lemma, also referred to as the Sherman-Morrison formula (see Sherman and Morrison, 1950). Then

$$\mathbf{I}^{-1}(\boldsymbol{\alpha}) = \frac{1}{n}\Big[\mathbf{D}^{-1} + \frac{c\mathbf{D}^{-1}\boldsymbol{\iota}_l \boldsymbol{\iota}_l' \mathbf{D}^{-1}}{1 - c\boldsymbol{\iota}_l' \mathbf{D}^{-1} \boldsymbol{\iota}_l}\Big].$$

Under some regularity conditions the likelihood function for exponential families is strictly concave and the maximum likelihood estimate exists and is unique. Since the Dirichlet distribution belongs to an exponential family and these regularity conditions are met, this also holds true for the Dirichlet distribution. A direct proof has been given by Ronning (1989).

Unfortunately, when we encounter missing item values, $\mathbf{Y}$ is not completely

observed and the loglikelihood cannot be maximised directly. In order to obtain the maximum likelihood estimates in the presence of nonresponse the EM (Expectation Maximisation) algorithm was developed by Dempster, Laird and Rubin (1977).

3.5 The EM algorithm

As the Dirichlet distribution is a member of an exponential family, the EM algorithm consists of iteratively calculating the expected sufficient statistics and using these quantities to calculate the maximum likelihood estimates until convergence, as was shown in chapter 2.

The expected sufficient statistics can be easily computed from the natural parameterisation of the exponential family representation of the Dirichlet distribution. The density function of the Dirichlet for $\mathbf{Y} = (Y_1, \ldots, Y_l)'$ can be written in this form as follows

$$f(\mathbf{y} \mid \boldsymbol{\alpha}) = \exp\{\ln\Gamma(\sum_{j=1}^{l}\alpha_j) - \sum_{j=1}^{l}\ln\Gamma(\alpha_j) + \sum_{j=1}^{l}(\alpha_j - 1)\ln y_j\}.$$

The natural parameter of the Dirichlet is $\eta_j = \alpha_j - 1$ and the sufficient statistic is $t_j(\mathbf{y}) = \ln y_j$, $j = 1, \ldots, l$. In section 2.3.2 in chapter 2 we established that the negative of the derivative of the function $g(\boldsymbol{\eta})$ with respect to the natural parameter $\boldsymbol{\eta}$ is equal to the expectation of the sufficient statistic, i.e.

$$\mathrm{E}[t_j(\mathbf{Y})] = -\frac{\partial g(\boldsymbol{\eta})}{\partial \eta_j}, \quad j = 1, \ldots, l.$$

In case of the Dirichlet distribution the function $g(\boldsymbol{\eta})$ is

$$g(\boldsymbol{\eta}) = \ln\Gamma(\sum_{j=1}^{l}\eta_j + l) - \sum_{j=1}^{l}\ln\Gamma(\eta_j + 1).$$

So, for $j = 1, \ldots, l$,

$$\mathrm{E}[\ln Y_j \mid \boldsymbol{\alpha}] = \Psi(\eta_j + 1) - \Psi(\sum_{h=1}^{l}\eta_h + l) = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{l}\alpha_h).$$

The expectation of the sufficient statistic of the missing data items conditional on the observed values can be easily calculated as the conditional distribution of Dirichlet distributed variables is also a Dirichlet (see Theorem 2).

Consider a data vector $\mathbf{Y}$ that is distributed according to a Dirichlet distribution: $\mathbf{Y} \sim \mathrm{Dir}_{l-1}(\alpha_1, \ldots, \alpha_l)$. For each record i partition $\mathbf{Y}_i$ into a missing and an observed part, $\mathbf{Y}_i' = (\mathbf{Y}_{i,mis}', \mathbf{Y}_{i,obs}')$. Let m_i denote the number of missing items in record i. Then

$$\mathbf{Y}_{i,mis}^* \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha} \sim \mathrm{Dir}_{m_i-1}(\alpha_1, \ldots, \alpha_{m_i}),$$

where

$$\mathbf{Y}_{i,mis}^* = (1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i})^{-1} \mathbf{Y}_{i,mis}.$$

The expectation step of the EM algorithm now consists of calculating the expected sufficient statistics for $i = 1, \ldots, n$ and $j = 1, \ldots, l$, which are

$$\hat{\mathrm{E}}[\ln Y_{ij} \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}] = \begin{cases} \ln y_{ij} & \text{if } Y_{ij} \text{ is observed} \\ \mathrm{E}[\ln Y_{ij} \mid \mathbf{y}_{i,obs}, \boldsymbol{\alpha}] & \text{if } Y_{ij} \text{ is missing} \end{cases} . \quad (3.7)$$

We know that

$$\mathrm{E}[\ln Y_{ij}^* \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}] = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h).$$

From this it follows that

$$\mathrm{E}[\ln(1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i})^{-1} Y_{ij} \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}] = \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h)$$

and therefore

$$\mathrm{E}[\ln Y_{ij} \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}] = \ln(1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i}) + \Psi(\alpha_j) - \Psi(\sum_{h=1}^{m_i} \alpha_h).$$

Now calculate $\hat{\mathrm{E}}[\ln Y_{ij} \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}]$ given in equation (3.7) and plug these values into equation (3.6). The maximisation step then consists of finding the parameter value for $\boldsymbol{\alpha}$ for which the expected loglikelihood is maximal. Next these updated parameter values are used to re-estimate the expectation and so on. This process is iterated until the estimates converge.

Several criteria can be used to assess whether the algorithm has converged. For instance, when the changes in the parameter estimates are sufficiently small: $\max |\boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha}^{(t-1)}| < \varepsilon$, or when the first order conditions are sufficiently close to zero: $\max |\mathbf{S}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(t)}}| < \varepsilon$.

3.6 Imputation of missing data items

Now recall the edit constraint given in equation (3.1), where the data items $Y_1, \ldots, Y_l$ are restricted to the simplex. Assume that $\mathbf{Y}$ is Dirichlet distributed with parameters $\alpha_1, \ldots, \alpha_l$. Partition the data vector for respondent i into a missing and an observed part: $\mathbf{Y}_i' = (\mathbf{Y}_{i,mis}', \mathbf{Y}_{i,obs}')$ and partition $\boldsymbol{\alpha}$ accordingly. Again m_i denotes the number of missing items for respondent i. It holds that

$$\mathbf{Y}_{i,mis}^* \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs} \sim \mathrm{Dir}_{m_i-1}(\boldsymbol{\alpha}_{i,mis}), \tag{3.8}$$

for $\mathbf{Y}_{i,mis}^* = (1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i})^{-1} \mathbf{Y}_{i,mis}$. Imputations for $\mathbf{Y}_{i,mis}$ can be obtained either deterministically or stochastically.

Deterministic imputation can be done by using the expected values for the missing items given the observed items, based on the parameter estimates from the EM algorithm, as imputations. Note that, as opposed to distributions where the loglikelihood is linear in the data, the calculations in the E-step cannot be used for imputation as $\mathrm{E}[\ln Y] \neq \ln \mathrm{E}[Y]$. We will therefore use the first order moment to calculate the expected values

$$\mathrm{E}\Big[(1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i})^{-1} \mathbf{Y}_{i,mis} \mid \mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}, \boldsymbol{\alpha}_{i,mis}\Big] = \frac{\boldsymbol{\alpha}_{i,mis}}{\alpha_i},$$

where $\alpha_i = \sum_{j=1}^{m_i} \alpha_j$ and thus

$$\mathbf{Y}_{i,imp} = (1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i}) \frac{\boldsymbol{\alpha}_{i,mis}}{\alpha_i}.$$

Stochastic imputation can be done by generating random draws from the Dirichlet distribution given in equation (3.8). Recall that if $U_1 \sim \mathrm{gamma}(\alpha, \lambda)$ and $U_2 \sim \mathrm{gamma}(\beta, \lambda)$ then $Z = \frac{U_1}{U_1+U_2} \sim \mathrm{beta}(\alpha, \beta)$. This can be generalised to the Dirichlet distribution, see for example Wilks (1962). Suppose that $U_1, \ldots, U_l$ are independent random variables having gamma distributions $\mathrm{gamma}(\alpha_1, \lambda), \ldots, \mathrm{gamma}(\alpha_l, \lambda)$. For $j = 1, \ldots, l$, let $Z_j = \frac{U_j}{U_1+\cdots+U_l}$, then $\mathbf{Z}$ has the $(l-1)$-variate Dirichlet distribution $\mathrm{Dir}_{l-1}(\alpha_1, \ldots, \alpha_l)$. Thus random values from the Dirichlet distribution can be obtained by drawing independently from gamma distributions. Let $\mathbf{V}$ denote a random draw from (3.8), then

$$\mathbf{Y}_{i,imp} = (1 - \mathbf{y}_{i,obs}' \boldsymbol{\iota}_{l-m_i}) \mathbf{V}.$$

Finally, the imputed proportions can be transformed to imputations for the original data $\mathbf{X}_i$ by multiplying them with the total for that respondent, X_{ik}:

$$\mathbf{X}_{i,imp} = \mathbf{Y}_{i,imp} X_{ik}.$$

3.7 Imputation performance

In order to assess the performance of this imputation method on empirical data, we will use it to impute data that have been gathered by Statistics Netherlands on a part of the wholesale industry for businesses with more than 10 employees. The effects on estimation of population parameters as well as the imputation performance with respect to individual values will be assessed.

3.7.1 Description of the data

The questioned businesses provide information about their company profits, sales and expenses. Additionally, data is provided on employees, such as the number of employees, the number of temporary employees, their wages and so on.

The first dataset that will be used concerns labour costs, which consists of the following variables

$$\begin{aligned} X_{11} &= \text{gross wages and salaries} \\ X_{12} &= \text{social security costs} \\ X_{13} &= \text{pension charges} \\ X_{14} &= \text{other social costs} \\ X_{1t} &= \text{total labour costs,} \end{aligned}$$

which are all non-negative and for which the following balance restriction holds

$$X_{11} + X_{12} + X_{13} + X_{14} = X_{1t}.$$

The second dataset concerns costs of third party rendering of services, consisting of the variables

$$\begin{aligned} X_{21} &= \text{costs of banking} \\ X_{22} &= \text{insurance premiums} \\ X_{23} &= \text{costs of accounting, tax or legal advice} \\ X_{24} &= \text{costs of automisation} \\ X_{25} &= \text{costs of waste processing} \\ X_{26} &= \text{other costs of third party services} \\ X_{2t} &= \text{total costs of third party rendering of services.} \end{aligned}$$

These variables again are all non-negative and satisfy the balance restriction

$$X_{21} + X_{22} + X_{23} + X_{24} + X_{25} + X_{26} = X_{2t}.$$

3.7.2 Estimation of population parameters

As we already mentioned in chapter 1, estimates of population parameters are potentially subject to the following sources of variation: sampling variance and nonresponse variance. Sampling variance is the variance that occurs due to surveying only a subset of the population and would vanish if the whole population was sampled. Nonresponse variance is introduced by the nonresponse in the sample, so if the sample would be completely observed no nonresponse variance would arise. We are currently mostly interested in the nonresponse component of the total variance of the population estimate. A simulation study will be conducted to assess the magnitude of this nonresponse variance for different imputation and incomplete data procedures.

3.7.3 Generation of missing data items

In order to be able to assess the effects of the imputation procedures we need to have completely observed data, i.e. data without any missing items, at our disposal. The cases with missing items will therefore be removed from the datasets on labour and third party rendering of services costs, and only the complete cases will be used for analysis as sufficiently large datasets will still remain. Next data items can be removed from these completely observed datasets, which means that for each dataset we will have completely observed data and corresponding data that contain missing values. In order to assess the variance that arises due to nonresponse and imputation several samples will be taken from these complete datasets, with randomly assigned missing data items. These different samples containing missing values will subsequently be used for parameter estimation.

The missing data items will be generated assuming that the missing data are MCAR (missing completely at random). This means that the probability that an item is missing does not depend on the value of that item or other data items in the survey. The missing data are generated using Bernoulli draws with parameter p. The value of p is chosen such that the percentage of generated nonresponse is similar to the percentage of nonresponse that was observed in the original survey. Although MCAR may be a somewhat unrealistic assumption it does give us insight in the performance of the different imputation methods

regarding point estimation of the parameters and the sensitivity of the outcomes with respect to the realised set of missing items. After generation of the missing data, we will first employ deductive imputation to obtain the imputations that can be derived with certainty based on the observed items in the survey. The remaining datasets are used for imputation.

3.7.4 The effects of imputation on parameter estimation

The main target of this survey is to estimate the population parameters on mean and dispersion: $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. To assess both the ability of imputation methods to produce accurate point estimates as well as the effects of nonresponse and imputation on the nonresponse variance component of the total variance of these population parameter estimates we will compare parameter estimates based on the imputed datasets as well as parameter estimates based on the incomplete data with the true parameter estimates for several randomly realised sets of missing data.

Procedures employed for parameter estimation using the incomplete data are the complete cases (CC) approach, which discards all records that contain missing responses, and the available cases (AC) method, which removes missing items on a variable-by-variable basis. The datasets with missing items will consequently be imputed after which the parameters are estimated as well. First of all we will use the Dirichlet approach, where either the expected values (Dir) or random draws (DirR) are used as imputations. Secondly, a ratio hot deck (RHD) approach is used where donors are found at random. Finally, ratios of the nearest neighbours (RNN) are used for imputation. The distance between records is calculated as the difference between the totals (X_{1t} for the first dataset and X_{2t} for the second dataset). Note that the Dirichlet method with expectations imputed and the nearest neighbour ratio imputation method are both deterministic methods, meaning that given a certain set of missing items these methods will always generate the same imputations. The Dirichlet method with random draws imputed and the random ratio hot deck method on the other hand are stochastic procedures, where the imputed values will vary given a set of missing values. These latter two procedures are developed to obtain better point estimates for $\boldsymbol{\sigma}$ and to preserve the distribution of the data better. This will, however, create a larger nonresponse variance. The process of generating and imputing missing items and subsequently estimating the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ is iterated 100 times.

In Table 3.1 the results on the generation of missing item values are presented for both datasets, the standard errors are shown between brackets. The

Table 3.1: *Generation of missing items.*

	Labour costs			*Third party rendering of services*		
sample size		597			663	
# of complete cases		286	(13)		227	(10)
# of missing cases	X_{11}	186	(10)	X_{21}	207	(12)
	X_{12}	185	(11)	X_{22}	214	(12)
	X_{13}	187	(13)	X_{23}	221	(11)
	X_{14}	186	(13)	X_{24}	207	(11)
				X_{25}	201	(12)
				X_{26}	203	(12)

parameter estimates and their 95% confidence intervals, which represent the nonresponse variance, for the dataset on labour costs are presented in Figures 3.2 and 3.3 and the parameter estimates and 95% confidence intervals for the dataset on the costs of third party rendering of services are presented in Figures 3.4 and 3.5. The solid line refers to the true parameter estimate, which is calculated based on the completely observed data.

Although the missing data mechanism is MCAR, using the complete cases estimates will lead to inaccurate parameter estimates. This is due to the high rate of nonresponse, which leads to a relatively small complete sample, especially for the dataset on third party rendering of services as can be seen in Table 3.1. Using the available cases results in nonresponse rates of about 30% for each variable in both datasets and therefore leads to somewhat more accurate point estimates, which do not satisfy the linear balance restriction however. In both cases the 95% confidence intervals are unacceptably large most of the time, meaning that the nonresponse variance is quite substantial. Consequently the accuracy of the parameter estimates will be highly dependent on the realised set of missing values, which is undesirable. Besides, both methods seriously underestimate the standard deviations in both datasets (Figures 3.3 and 3.5). So, if we are dealing with a considerable number of missing records the complete and available cases estimates cannot be employed.

For the parameter estimates that are based on the imputed datasets concerning labour costs (Figures 3.2 and 3.3) we observe the following. The nearest neighbour method produces accurate point estimates, but sometimes the confid-

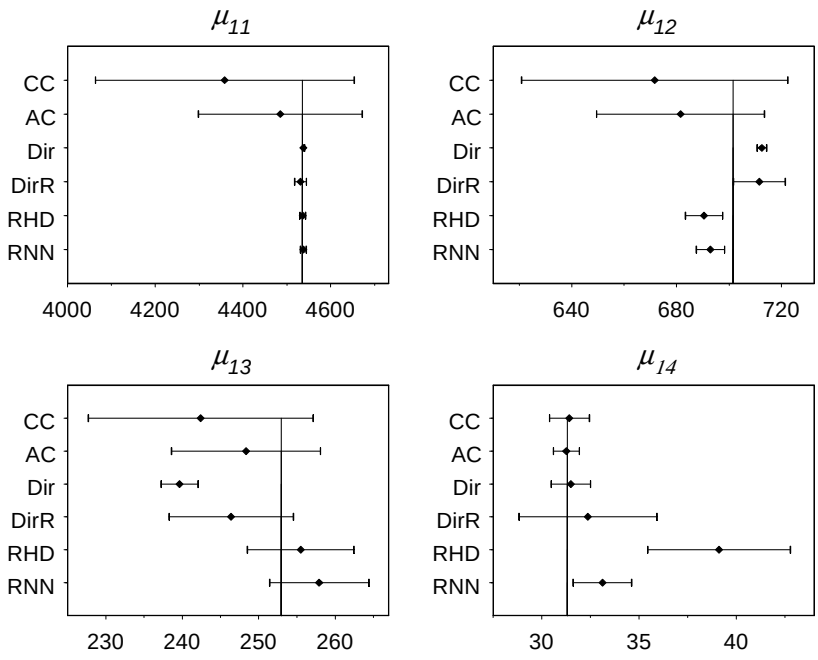

Figure 3.2: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

ence intervals remain quite large, which reduces precision. The Dirichlet method, where the expectation is imputed, generates acceptable point estimates as well, with higher precision. As was expected the Dirichlet method with random imputations yields similar point estimates, but with less precision due to the added randomness. Beforehand we believed that stochastic imputation methods are necessary to provide non-biased point estimates for the standard deviations. From Figure 3.3 it becomes clear that this is not always the case as reasonable estimates for the standard deviations, with small confidence intervals, are obtained using the Dirichlet method with expectations imputed. This is probably due to the fact that one conditions on the observed data, which show quite some variation and therefore the imputed values will vary as well. Besides, the model parameters $\boldsymbol{\alpha}$ of the Dirichlet distribution need to be estimated, introducing even more variation. Apparently, these two sources of variation result in a good approximation of the actual variation. Furthermore, using a stochastic imputa-

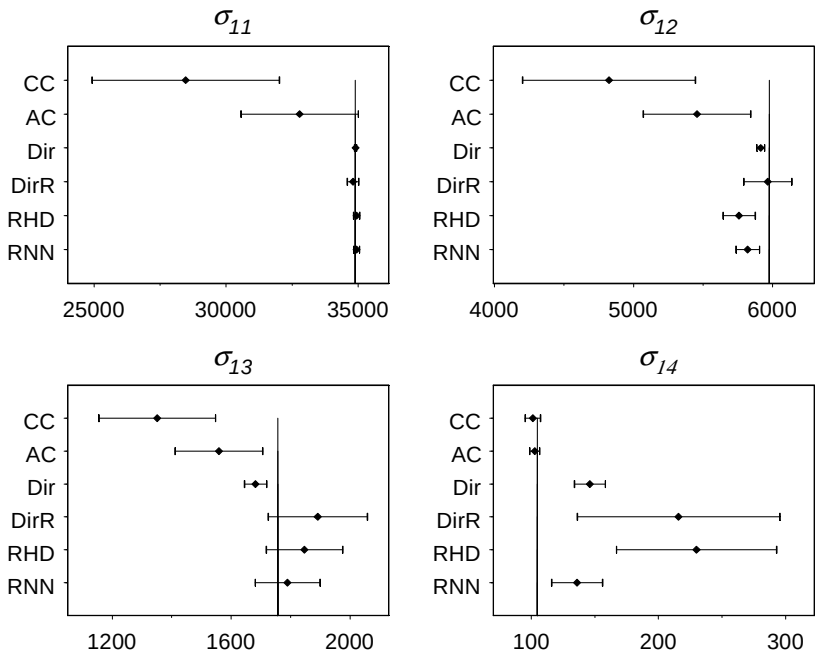

Figure 3.3: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

tion procedure results in considerably more nonresponse variance. Especially the random ratio hot deck method produces point estimates with quite large confidence intervals, in particular for the variable X_{14}. This method therefore seems to be the worst imputation method for this type of data if one is mainly interested in aggregate values.

The effects of nonresponse and imputation on parameter estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for the data on third party rendering of services (Figures 3.4 and 3.5) are less clear-cut. First of all we find that all the imputation methods yield similar estimates for μ_{21}, μ_{22}, μ_{24} and μ_{25}. The nearest neighbour method produces much more accurate point estimates for μ_{23} and μ_{26}, however. With respect to the nonresponse variance, on average, the nearest neighbour method and the random Dirichlet display the largest variation and the Dirichlet method with expectations imputed results in the smallest nonresponse variance. The estimates of $\boldsymbol{\sigma}$ show a somewhat different pattern. Again, all imputation methods

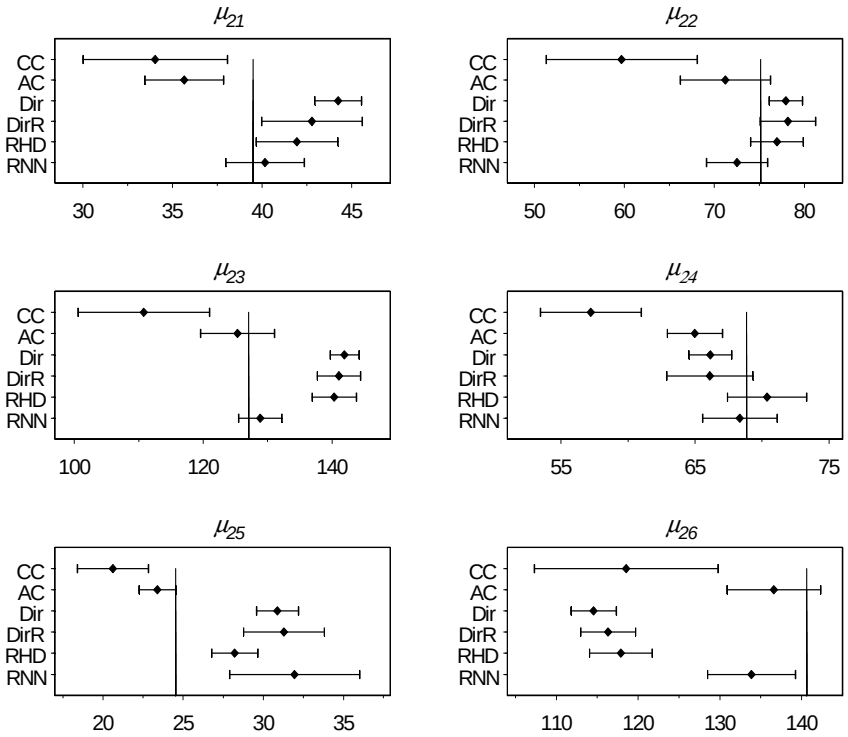

Figure 3.4: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

yield similar estimates for σ_{12} and σ_{22}, but for σ_{24} the random Dirichlet and the random hot deck method and for σ_{25} the random Dirichlet together with the nearest neighbour approach clearly perform worse than the other methods. Furthermore, as opposed to the differences in performances for the estimation of μ_{23}, all methods result in similar deviation of the estimates of σ_{23} from the true value. Finally, for the estimation of σ_{26} the nearest neighbour method remains the best choice. In general, similar to the results obtained from the previous dataset, the nearest neighbour method provides quite reasonable point estimates for both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for almost all variables. The confidence intervals, however, can be substantial especially for the estimates of $\boldsymbol{\sigma}$, indicating a considerable nonresponse variance. The Dirichlet method with expectations imputed

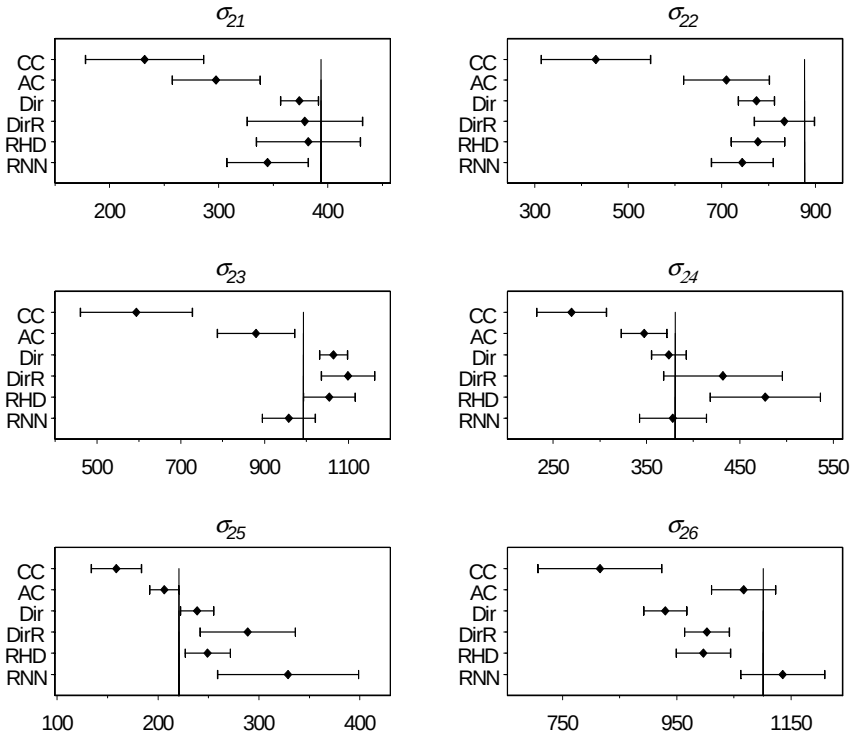

Figure 3.5: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

produces less accurate but mostly quite acceptable point estimates with higher precision, meaning that this method is much less dependent on the realised set of missing data. Again it is also observed that no stochastic imputation procedures are needed to obtain reasonable estimates for $\boldsymbol{\sigma}$.

In conclusion, it is difficult to state which imputation method performs best, as their performances differ across variables and datasets. Furthermore, a balance needs to be established between required accuracy and precision. A natural candidate to measure the performance of imputation methods, that incorporates both the bias and the variance of an estimate, is the mean squared error (MSE) of an estimate. Unfortunately, the MSE only provides a measure to compare imputation methods within variables and not across variables. Consequently, a

procedure remains to be developed that combines information from the MSE's for each method across variables in order to determine which method performs best.

An important result that can be derived from Figures 3.2, 3.3, 3.4 and 3.5 is the fact that some methods produce much larger confidence intervals in comparison to the other methods. As we mentioned earlier, large confidence intervals indicate a high nonresponse variance, which means that the imputation or incomplete data method in question is highly dependent on the actual set of missing data. In reality only one of the possible sets of missing data actually is drawn and it is therefore desirable to use a procedure that will result in similar parameter estimates, irrespective of the realised set of missing data. In this sense, the Dirichlet method appears to be a promising alternative with respect to ratio hot deck imputation techniques.

3.7.5 The performance of the imputation methods on item level

If the main interest of the imputer is to obtain accurate population estimates as well as a general purpose dataset, other desirable properties arise. Ideally, we would like the results of any statistical analysis on the imputed data to lead to the same conclusions as the same analysis on the true data. Obviously it is impossible to define performance measures that can assess this. The imputation methods will therefore be judged based on their ability to preserve the true values (predictive accuracy) and the distribution of the true values (distributional accuracy). With respect to the preservation of the true values, the average absolute deviation of the imputed data from the true data is calculated. To assess whether the distribution of the imputed data with respect to the true data is preserved, the Kolmogorov-Smirnov statistic, which measures the maximal distance between the empirical distribution functions of the imputed and the true data, is calculated. The results are presented in Table 3.2 for the dataset concerning labour costs and in Table 3.3 for the dataset on costs of third party rendering of services, standard errors are given between brackets.

In Table 3.2 we find that with regard to the predictive accuracy of the imputation methods, the Dirichlet method with expectations imputed performs best. This is not surprising as the expectation seems to be the best prediction. With respect to the preservation of the distribution, this Dirichlet method performs worse than the other methods, particularly for the variable X_{14}. Note, however that the Dirichlet method did preserve μ_{14} and σ_{14} well. In this case it seems that nonparametric methods are much better at preserving the dis-

tribution of the variable. This is probably due to the fact that X_{14} is a semi-continuous variable with a high proportion of records with a fixed value (at zero) and a small continuous part. The donor methods will impute much more zero item values in this case, which the Dirichlet with expectations imputed obviously will not. In chapter 6 models are discussed and employed to deal with this particular type of variable. For the other three variables the Kolmogorov-Smirnov distance is relatively small for all methods, indicating that the imputed and the true data are from equal distributions. The random hot deck and nearest neighbour methods, however, do preserve the distributions better as the Kolmogorov-Smirnov distance is smaller for these methods. This means that if the imputer is solely interested in the distribution of the data, the random hot deck and nearest neighbour method are preferred for this dataset.

The data on third party rendering of services yield similar results, which are presented in Table 3.3. The Dirichlet method, with expectations imputed, outperforms the other methods with respect to predictive accuracy, but is outperformed with regard to distributional accuracy. As this dataset consists of large proportions of zeroes in most of these variables, the nonparametric methods perform much better with respect to distributional accuracy. Again, this will be further investigated in chapter 6.

Finally, it is observed that the Dirichlet method with expectations imputed results in relatively small standard errors for both datasets. This means that also with respect to predictive and distributional accuracy the Dirichlet method displays a small variation due to nonresponse.

3.8 Concluding remarks

In this chapter we have developed an imputation method that models the data through a Dirichlet distribution in order to impute non-negative data items that are part of a balance restriction from which the total is assumed to be known. As a part of this imputation procedure the EM algorithm for the Dirichlet distribution is derived such that maximum likelihood estimates can be obtained in the presence of nonresponse.

The results in section 3.7 on empirical data show that the choice of an appropriate imputation method strongly depends on the main interest of the data user. If preserving aggregates such as averages, totals and variances is most important to the data user, the Dirichlet method with expectations imputed seems a good approach with high precision. Another good option would be to use the nearest neighbour imputation method. When the main interest of the data user

Table 3.2: *Predictive and distributional accuracy of the imputed data on labour costs.*

	Average deviation				*Kolmogorov-Smirnov distance*			
	Dir	DirR	RHD	RNN	Dir	DirR	RHD	RNN
X_{11}	110	266	185	173	0.03	0.03	0.03	0.03
	(36)	(198)	(97)	(83)	(0.00)	(0.01)	(0.01)	(0.01)
X_{12}	95	228	165	137	0.11	0.08	0.06	0.06
	(27)	(144)	(104)	(76)	(0.02)	(0.02)	(0.02)	(0.01)
X_{13}	110	186	176	167	0.12	0.17	0.08	0.08
	(37)	(126)	(102)	(81)	(0.02)	(0.04)	(0.03)	(0.02)
X_{14}	46	58	72	54	0.54	0.28	0.10	0.09
	(15)	(58)	(55)	(22)	(0.03)	(0.04)	(0.04)	(0.03)

Table 3.3: *Predictive and distributional accuracy of the imputed data on third party rendering of services costs.*

	Average deviation				*Kolmogorov-Smirnov distance*			
	Dir	DirR	RHD	RNN	Dir	DirR	RHD	RNN
X_{21}	51	73	64	66	0.36	0.20	0.11	0.09
	(22)	(41)	(35)	(32)	(0.03)	(0.03)	(0.03)	(0.03)
X_{22}	66	86	87	78	0.18	0.21	0.09	0.09
	(35)	(47)	(54)	(47)	(0.02)	(0.04)	(0.03)	(0.03)
X_{23}	78	105	99	92	0.09	0.14	0.10	0.07
	(34)	(53)	(50)	(49)	(0.02)	(0.03)	(0.02)	(0.02)
X_{24}	68	87	89	86	0.33	0.21	0.11	0.08
	(18)	(46)	(45)	(31)	(0.03)	(0.03)	(0.03)	(0.04)
X_{25}	53	61	49	65	0.58	0.30	0.16	0.10
	(23)	(43)	(26)	(65)	(0.03)	(0.05)	(0.05)	(0.04)
X_{26}	113	124	130	127	0.57	0.31	0.20	0.10
	(44)	(53)	(56)	(64)	(0.03)	(0.05)	(0.05)	(0.04)

is to preserve the true values of the data this Dirichlet method seems to be the best approach, but if the data user is mainly interested in the distribution of the data using the random hot deck and nearest neighbour method may be wiser,

especially if the dataset contains a large number of semi-continuous variables.

Another important issue is the fact that the nonresponse variance varies strongly between imputation methods. This means that some imputation methods are much more dependent on the actual set of missing data that is realised. It is preferred to use an imputation method that produces as little nonresponse variance as possible, while still producing accurate parameter estimates, imputed values and so on. In this sense, the Dirichlet method appears to be the best option.

In this chapter the Dirichlet approach has been applied to two relatively small datasets concerning company expenses gathered by a certain business survey. Note, however, that the procedure is applicable to all datasets that consist of continuous non-negative variables that add up to a given total. The Dirichlet method can therefore also be used to impute data on employment or company turnover variables. Furthermore, this model can be straightforwardly extended to other business surveys as almost all business surveys at Statistics Netherlands are based on a similar structure. In social surveys this method may be of use as well, for example for the imputation of household income variables.

Further research is needed with respect to the linear restrictions that can be handled by the imputation procedure. The advantage that the Dirichlet method, as well as the ratio imputation methods, satisfies non-negativity restrictions also has a downside. Some financial variables, such as profit, financial result or exceptional result are allowed to be negative and therefore cannot be imputed using this approach. Besides, economic data usually need to satisfy many linear restrictions. As this method deals with data that are subject to non-negativity constraints and one linear balance restriction only, there is still a need for an imputation method that can take several linear balance and other types of inequality restrictions into account. In the next chapter we will therefore develop an imputation method that can cope with multiple linear balance restrictions.

Chapter 4

Imputation of Data Subject to Multiple Balance Restrictions

Economic data are subject to several linear restrictions. Common well-known imputation methods provide imputations that are likely not to satisfy the linear balance restrictions on economic data. In the previous chapter we discussed imputation procedures for variables that need to satisfy non-negativity constraints and one linear balance restriction. The method that was suggested, based on Dirichlet distribution, is however not capable of incorporating multiple linear balance restrictions. In this chapter we will therefore consider imputation methods for data items that need to satisfy multiple linear balance restrictions. We suggest using the singular normal distribution to model these data.

4.1 Introduction

The flexibility of the Dirichlet distribution, that was described in chapter 3, is an attractive feature for imputation purposes. Unfortunately, the Dirichlet method can only deal with one balance restriction at a time. If we are confronted with a more complex balance edit structure, where variables are present in more than one balance restriction, the Dirichlet distribution cannot be utilised.

This will be illustrated in the following example. The total operating ex-

penses of a company can be subdivided into several items, such as the purchasing price of goods, labour costs, costs of housing, depreciation and so on. Most of these items are themselves also the total of another subdivision. Labour costs can, for example, be subdivided into gross wages, social security costs, pension charges and other social costs. Now assume we want to model a dataset consisting of the following variables: X_{11}, X_{12}, X_{13}, X_{14}, X_{1t}, X_{2t}, X_{3t}, X_t, for which it holds that

$$X_{11} + X_{12} + X_{13} + X_{14} = X_{1t} \quad \text{and} \quad X_{1t} + X_{2t} + X_{3t} = X_t.$$

The Dirichlet distribution can be employed to model the data if these two restrictions are combined into one balance restriction, e.g.

$$X_{11} + X_{12} + X_{13} + X_{14} + X_{1t} + 2X_{2t} + 2X_{3t} = 2X_t.$$

Note, however, that in this case the fact that $X_{11} + X_{12} + X_{13} + X_{14} = X_{1t}$ is not taken into account. This means that the imputed data may very well not satisfy this linear restriction. Consequently, the Dirichlet method cannot be applied in this manner. Another option is to use the Dirichlet distribution in conjunction with hierarchical imputation. In this instance X_{1t}, X_{2t}, X_{3t} are modelled first and subsequently X_{1t} is used to model X_{11}, X_{12}, X_{13} and X_{14}. In an imputation context this procedure can, however, lead to inconsistencies as the observed values for X_{11}, X_{12}, X_{13} and X_{14} are not taken into account when imputing X_{1t}. This means that the sum of the observed values may exceed the imputed total for X_{1t}, which is not allowed.

So the Dirichlet distribution cannot be straightforwardly extended to the presence of multiple balance restrictions with interdependencies and therefore the need arises to develop an imputation method that can handle any type of linear balance edit structure. In this chapter we will discuss the use of a singular normal distribution for this purpose.

First the edit structure will be discussed in section 4.2. Next in section 4.3 the probability density function of the singular normal is derived, maximum likelihood estimation for completely observed data and in the presence of missing data is treated in sections 4.4 and 4.5 respectively. In section 4.6 it is shown that the maximum likelihood estimates and the imputed data concur with the balance restrictions. Deterministic and stochastic imputation methods are given in section 4.7 and in section 4.8 some results of these imputation procedures on empirical data are shown. Finally, in section 4.9 some concluding remarks will be given.

4.2 Balance edit restrictions

Consider an $n \times k$ data matrix $\mathbf{X}$ and a $p \times k$ restriction matrix $\mathbf{A}$, with p the number of linear balance restrictions, where it holds that $\mathbf{AX}' = \mathbf{0}$. For convenience we assume that every variable will be present in at least one linear balance restriction, this is not necessary however. Furthermore we assume that there are no redundant balance restrictions, which means that $\mathbf{A}$ is of full row rank.

Note that the solution to this system of linear equations for respondent i is

$$\mathbf{X}_i = (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{Z}_i,$$

where $\mathbf{I}$ is an $k \times k$ identity matrix and $\mathbf{Z}_i$ is an arbitrary vector of order $k \times 1$, see e.g. Rao (1973). The matrix $\mathbf{A}^{-}$ denotes a generalised inverse of $\mathbf{A}$. A generalised inverse of the matrix $\mathbf{A}$ is a matrix for which it holds that $\mathbf{AA}^{-}\mathbf{A} = \mathbf{A}$. If $\mathbf{A}$ is square, and consequently nonsingular, there is only one solution for $\mathbf{X}_i$: $\mathbf{X}_i = \mathbf{0}$.

The multivariate singular normal distribution is useful for modelling data subject to several linear balance restrictions as the singularity, that has arisen due to the restrictions, is immediately taken into account. This method can also be applied for the case where $p = 1$, i.e. if there is only one linear balance restriction present.

Non-negativity restrictions, however, are not taken into account as the singular normal is defined on an affine subspace of $\mathbb{R}^k$. Using natural logarithmic transformations of the data items would seem to be a straightforward solution to this, as $\ln X \in \mathbb{R}$, for $X > 0$. Unfortunately, if we take logarithms or employ some other nonlinear transformation (e.g. Box-Cox transformations) the singular structure of the data will be lost as $\ln X_1 + \ldots \ln X_{k-1} \neq \ln X_k$. This means that the imputed data will not satisfy the linear restrictions, which rules out the use of nonlinear transformations.

4.3 Multivariate singular normal distribution

Nonsingular distributions are often defined by specifying the density function of the distribution with respect to Lebesgue measure on $\mathbb{R}^k$. For a singular distribution, the rank of the covariance matrix is: $\text{rank}(\mathbf{\Sigma}) = q < k$, so the inversion of the covariance matrix in the probability density of the normal distribution is not possible and therefore no explicit determination of the density function with respect to Lebesgue measure in $\mathbb{R}^k$ is possible. The density function, however,

does exist on a subspace.

Assume that $\mathbf{X}_i \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\text{rank}(\boldsymbol{\Sigma}) = q$. Let L be the null space (or kernel) of $\boldsymbol{\Sigma}$: $\text{Null}(\boldsymbol{\Sigma}) = \{\mathbf{y} \in \mathbb{R}^k : \boldsymbol{\Sigma}\mathbf{y} = \mathbf{0}\}$. The dimension of L is $k - q$, which is equal to the number of linear balance restrictions p. The subspace $L^\perp$ is the orthogonal complement of L, that is the set of all vectors in $\mathbb{R}^k$ that are orthogonal to L. Now $\mathbf{X}_i$ has the following probability density function (e.g. Khatri, 1968)

$$\varphi(\mathbf{x}_i) = (2\pi)^{-\frac{q}{2}} |\boldsymbol{\Sigma}|_q^{-\frac{1}{2}} \exp\big(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^+(\mathbf{x}_i - \boldsymbol{\mu})\big), \tag{4.1}$$

for $\mathbf{x}_i \in \boldsymbol{\mu} + L^\perp$, which is an affine subspace. The pseudo-determinant $|\boldsymbol{\Sigma}|_q$ is defined as $\prod_{j=1}^q \lambda_j(\boldsymbol{\Sigma})$, that is $|\boldsymbol{\Sigma}|_q$ equals the product of the nonzero eigenvalues of $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{\Sigma}^+$ denotes the Moore-Penrose inverse of $\boldsymbol{\Sigma}$. The Moore-Penrose inverse $\mathbf{B}^+$ of a matrix $\mathbf{B}$ is a unique matrix for which it holds that

$$\begin{array}{lllll} 1. & \mathbf{B}\mathbf{B}^+\mathbf{B} & = & \mathbf{B} \\ 2. & \mathbf{B}^+\mathbf{B}\mathbf{B}^+ & = & \mathbf{B}^+ \\ 3. & (\mathbf{B}\mathbf{B}^+)' & = & \mathbf{B}\mathbf{B}^+ \\ 4. & (\mathbf{B}^+\mathbf{B})' & = & \mathbf{B}^+\mathbf{B}. \end{array}$$

This situation is illustrated graphically in a 3-dimensional setting by Figure 4.1, where $\boldsymbol{\mu} + L^\perp$ represents the affine subspace in which the $\mathbf{X}_i$ lie. The circles represent the contour lines of the (bivariate) normal density.

The density function of $\mathbf{X}_i$ in (4.1) can be derived as follows (see Khatri, 1968). Let $\mathbf{C}$ be the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}$. Now partition $\mathbf{C}$ as follows

$$\mathbf{C} = (\mathbf{C}_1 \ \mathbf{C}_2),$$

where $\mathbf{C}_1$ is the matrix of eigenvectors corresponding to the nonzero eigenvalues of $\boldsymbol{\Sigma}$. This means that $\mathbf{C}_2$ is the matrix of eigenvectors corresponding to the zero eigenvalues of $\boldsymbol{\Sigma}$ and therefore $\boldsymbol{\Sigma}\mathbf{C}_2 = \mathbf{0}$. Note that the columns of $\mathbf{C}_2$ span L, the null space of $\boldsymbol{\Sigma}$, and the columns of $\mathbf{C}_1$ span its orthogonal complement $L^\perp$. Another matrix that spans the null space is $\mathbf{A}'$ as it is of full rank and

$$\begin{aligned} \boldsymbol{\Sigma}\mathbf{A}' &= (\mathrm{E}[\mathbf{X}_i\mathbf{X}_i'] - \boldsymbol{\mu}\boldsymbol{\mu}')\mathbf{A}' \\ &= \mathrm{E}[\mathbf{X}_i(\mathbf{A}\mathbf{X}_i))'] - \boldsymbol{\mu}(\mathbf{A}\boldsymbol{\mu})' \\ &= \mathbf{0}, \end{aligned}$$

as $\mathbf{A}\mathbf{X}_i = \mathbf{A}\boldsymbol{\mu} = \mathbf{0}$.

The matrix $\boldsymbol{\Sigma}$ can be decomposed by means of an eigenvalue decomposition

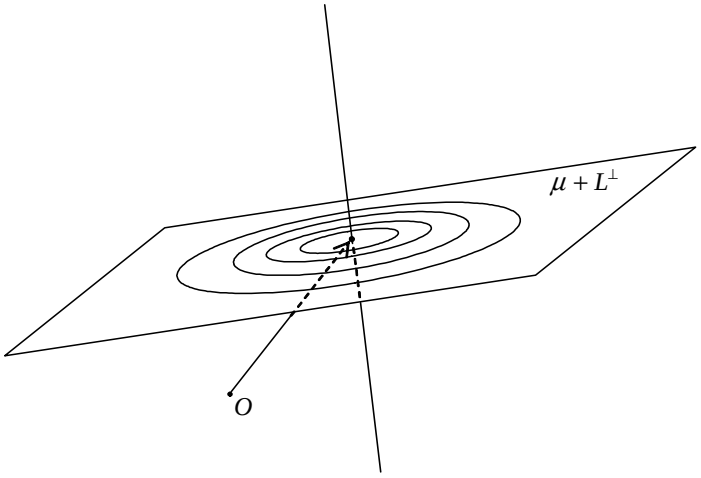

Figure 4.1: *Plot of the affine subspace in which* $\mathbf{X}$ *lies, the curves represent contour lines of the normal density.*

in $\mathbf{\Sigma} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$, where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_q, 0, \ldots, 0\}$ is the matrix of eigenvalues. Therefore it holds that $\mathbf{\Sigma} = \mathbf{C}_1\mathbf{\Lambda}_1\mathbf{C}_1'$, where $\mathbf{\Lambda}_1 = \mathrm{diag}\{\lambda_1, \ldots, \lambda_q\}$ and the Moore-Penrose inverse of $\mathbf{\Sigma}$ is $\mathbf{\Sigma}^+ = \mathbf{C}_1\mathbf{\Lambda}_1^{-1}\mathbf{C}_1'$.

Now consider the following transformation

$$\begin{aligned} \mathbf{Y}_i^{(1)} &= \mathbf{C}_1'\mathbf{X}_i \\ \mathbf{Y}_i^{(2)} &= \mathbf{C}_2'\mathbf{X}_i. \end{aligned}$$

Then

$$\mathrm{E}[\mathbf{Y}_i^{(2)}] = \mathbf{C}_2'\boldsymbol{\mu} \quad \text{and} \quad \mathrm{Var}(\mathbf{Y}_i^{(2)}) = \mathbf{C}_2'\mathbf{\Sigma}\mathbf{C}_2 = \mathbf{0}.$$

This means that $\mathbf{Y}_i^{(2)} = \mathbf{C}_2'\boldsymbol{\mu}$ with probability 1. Note that $\mathbf{Y}_i^{(1)}$ is normally distributed

$$\mathbf{Y}_i^{(1)} \sim \mathcal{N}_q(\mathbf{C}_1'\boldsymbol{\mu}, \mathbf{C}_1'\mathbf{\Sigma}\mathbf{C}_1),$$

where the covariance matrix is nonsingular as $\mathbf{C}_1'\mathbf{\Sigma}\mathbf{C}_1 = \mathrm{diag}\{\lambda_1, \ldots, \lambda_q\}$. Therefore $\mathbf{Y}_i^{(1)}$ has the probability density

$$(2\pi)^{-\frac{q}{2}}|\mathbf{C}_1'\mathbf{\Sigma}\mathbf{C}_1|^{-\frac{1}{2}} \exp\big(-\frac{1}{2}(\mathbf{y}_i^{(1)} - \mathbf{C}_1'\boldsymbol{\mu})'(\mathbf{C}_1'\mathbf{\Sigma}\mathbf{C}_1)^{-1}(\mathbf{y}_i^{(1)} - \mathbf{C}_1'\boldsymbol{\mu})\big). \quad (4.2)$$

So the density of $\mathbf{X}_i$ or similarly the density of $(\mathbf{Y}_i^{(1)}, \mathbf{Y}_i^{(2)})$ is determined by (4.2) and $\mathbf{Y}_i^{(2)} = \mathbf{C}_2'\boldsymbol{\mu}$. Note that

$$\begin{aligned}(\mathbf{Y}_i^{(1)} - \mathbf{C}_1'\boldsymbol{\mu})'(\mathbf{C}_1'\boldsymbol{\Sigma}\mathbf{C}_1)^{-1}(\mathbf{Y}_i^{(1)} - \mathbf{C}_1'\boldsymbol{\mu}) \\ &= (\mathbf{C}_1'\mathbf{X}_i - \mathbf{C}_1'\boldsymbol{\mu})'(\mathbf{C}_1'\boldsymbol{\Sigma}\mathbf{C}_1)^{-1}(\mathbf{C}_1'\mathbf{X}_i - \mathbf{C}_1'\boldsymbol{\mu}) \\ &= (\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{C}_1(\mathbf{C}_1'\boldsymbol{\Sigma}\mathbf{C}_1)^{-1}\mathbf{C}_1'(\mathbf{X}_i - \boldsymbol{\mu}) \\ &= (\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{C}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1'(\mathbf{X}_i - \boldsymbol{\mu}) \\ &= (\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{+}(\mathbf{X}_i - \boldsymbol{\mu}).\end{aligned}$$

Hence the density in (4.2) becomes

$$(2\pi)^{-\frac{q}{2}}(\prod_{j=1}^{q}\lambda_j)^{-\frac{1}{2}}\exp\big(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{+}(\mathbf{x}_i - \boldsymbol{\mu})\big), \quad \text{where } \mathbf{x}_i \in \boldsymbol{\mu} + L^{\perp},$$

with $\mathbf{C}_2'\mathbf{X}_i = \mathbf{C}_2'\boldsymbol{\mu}$.

4.4 Maximum likelihood estimation for the singular normal distribution

Khatri (1968) also shows that this singular distribution leads to the same maximum likelihood estimates as the nonsingular model. From nonsingular theory the maximum likelihood estimates of $\mathbf{C}_1'\boldsymbol{\mu}$ and $\mathbf{C}_1'\boldsymbol{\Sigma}\mathbf{C}_1$ are

$$\begin{aligned}\mathbf{C}_1'\hat{\boldsymbol{\mu}} &= \mathbf{C}_1'\bar{\mathbf{x}} \\ \mathbf{C}_1'\hat{\boldsymbol{\Sigma}}\mathbf{C}_1 &= \mathbf{C}_1'\mathbf{S}\mathbf{C}_1,\end{aligned}$$

where $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. So $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$. We established before that $\mathbf{C}_2'\mathbf{X}_i = \mathbf{C}_2'\boldsymbol{\mu}$ with probability 1. This leads to

$$\mathbf{C}_2'\hat{\boldsymbol{\mu}} = \mathbf{C}_2'\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_2'\mathbf{x}_i = \mathbf{C}_2'\boldsymbol{\mu}$$

and

$$\mathbf{C}_2'\hat{\boldsymbol{\Sigma}}\mathbf{C}_2 = \mathbf{C}_2'\mathbf{S}\mathbf{C}_2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{C}_2'\mathbf{x}_i - \mathbf{C}_2'\bar{\mathbf{x}})(\mathbf{C}_2'\mathbf{x}_i - \mathbf{C}_2'\bar{\mathbf{x}})' = \mathbf{0} = \mathbf{C}_2'\boldsymbol{\Sigma}\mathbf{C}_2.$$

So the singular case yields the same maximum likelihood estimates as the nonsingular case.

4.4.1 Maximum likelihood estimation and linear balance restrictions

These maximum likelihood estimates concur with the linear balance restrictions on the data. First of all

$$\mathbf{A}\hat{\boldsymbol{\mu}} = \mathbf{A}\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}\mathbf{x}_i = \mathbf{0},$$

as the balance restrictions hold for each respondent.

Furthermore, the null space of the estimated covariance matrix needs to be equal to the null space of $\boldsymbol{\Sigma}$. Recall that Null($\boldsymbol{\Sigma}$) is spanned by $\mathbf{A}'$, as was established in section 4.3. So postmultiply $\hat{\boldsymbol{\Sigma}}$ with $\mathbf{A}'$

$$\hat{\boldsymbol{\Sigma}}\mathbf{A}' = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{A}' = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{A}\mathbf{x}_i - \mathbf{A}\bar{\mathbf{x}})' = \mathbf{0}.$$

This means that Null($\hat{\boldsymbol{\Sigma}}$) = Null($\boldsymbol{\Sigma}$). The maximum likelihood estimates are therefore in concurrence with the linear balance restrictions.

4.5 The EM algorithm applied to singular normal data

Now suppose we encounter missing data. The EM algorithm can be used to find maximum likelihood estimates in the presence of nonresponse. Again suppose that $\mathbf{X}_i$, $i = 1, \ldots, n$, of order $k \times 1$ contains the item responses for record i, and the complete $n \times k$ data matrix is $\mathbf{X}$, with elements X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$. For each record i partition the items into a missing and an observed part, thus $\mathbf{X}_i' = (\mathbf{X}_{i,mis}', \mathbf{X}_{i,obs}')$ and partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly. Let m_i denote the number of missing items for respondent i.

It is well known that, if $\mathbf{X}_i$ is normally distributed with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the following holds (see e.g. Anderson, 1984, for a proof)

$$\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}_{m_i}(\boldsymbol{\mu}_{mis.obs}, \boldsymbol{\Sigma}_{mis,mis.obs}),$$

where

$$\begin{aligned} \boldsymbol{\mu}_{mis.obs} &= \boldsymbol{\mu}_{mis} + \boldsymbol{\Sigma}_{mis,obs}\boldsymbol{\Sigma}_{obs,obs}^{-}(\mathbf{x}_{i,obs} - \boldsymbol{\mu}_{obs}) \\ \boldsymbol{\Sigma}_{mis,mis.obs} &= \boldsymbol{\Sigma}_{mis,mis} - \boldsymbol{\Sigma}_{mis,obs}\boldsymbol{\Sigma}_{obs,obs}^{-}\boldsymbol{\Sigma}_{obs,mis}. \end{aligned}$$

Note that the matrix $\boldsymbol{\Sigma}^{-}_{obs,obs}$, as opposed to the Moore-Penrose inverse, is not unique. Pringle and Rayner (1971) establish the invariance of $\boldsymbol{\mu}_{mis.obs}$ and $\boldsymbol{\Sigma}_{mis,mis.obs}$ under any choice of $\boldsymbol{\Sigma}^{-}_{obs,obs}$.

The matrix $\boldsymbol{\Sigma}_{obs,obs}$ is singular if all items in at least one linear balance restriction are observed. If $\boldsymbol{\Sigma}_{obs,obs}$ is nonsingular the generalised inverse equals the regular inverse. Note that the dependence of the partitioned mean vector and the partitioned covariance matrix on i is left out, this is done for ease of notation. It is, however, important to keep in mind that the missingness patterns and consequently $\boldsymbol{\mu}_{mis.obs}$ and $\boldsymbol{\Sigma}_{mis,mis.obs}$ vary across respondents. Also note that $\boldsymbol{\Sigma}_{mis,mis.obs}$ will always be singular as all variables are incorporated in at least one linear restriction, so one item value can always be derived from the other items. A formal proof will be given in subsection 4.7.1.

For the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we will use the EM algorithm, which calculates the maximum likelihood estimates in the presence of missing data. For a detailed description of the EM algorithm in general see chapter 2 of this thesis. See Schafer (1997), and Little and Rubin (2002) for a description of the EM algorithm for normally distributed data.

In short, the EM algorithm for data that are distributed according to a singular normal is as follows. Since the singular normal distribution is a member of an exponential family, the E-step consists of estimating the expected sufficient statistics

$$\begin{aligned} \mathrm{E}[T_1(\mathbf{X})] &= \mathrm{E}[\sum_{i=1}^{n} \mathbf{X}_i] = \sum_{i=1}^{n} \mathrm{E}[\mathbf{X}_i] \\ \mathrm{E}[T_2(\mathbf{X})] &= \mathrm{E}[\mathbf{X}'\mathbf{X}] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[\mathbf{X}_i \mathbf{X}_i']. \end{aligned}$$

Recall that

$$\begin{aligned} &\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\qquad = \boldsymbol{\mu}^{(t)}_{mis} + \boldsymbol{\Sigma}^{(t)}_{mis,obs} (\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-} (\mathbf{x}_{i,obs} - \boldsymbol{\mu}^{(t)}_{obs}) \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} &\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \\ &\qquad = \boldsymbol{\Sigma}^{(t)}_{mis,mis} - \boldsymbol{\Sigma}^{(t)}_{mis,obs} (\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-} \boldsymbol{\Sigma}^{(t)}_{obs,mis}. \end{aligned} \tag{4.4}$$

Let X^*_{ij} denote the elements of $\mathbf{X}_{i,mis}$ estimated by equation (4.3). Then for $i = 1, \ldots, n$, and $j, l = 1, \ldots, k$,

$$\hat{\mathrm{E}}[X_{ij} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \begin{cases} x_{ij} & \text{if } X_{ij} \text{ is observed} \\ X^*_{ij} & \text{if } X_{ij} \text{ is missing} \end{cases}$$

and

$$\hat{\mathrm{E}}[X_{ij}X_{il} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]$$
$$= \begin{cases} x_{ij}x_{il} & \text{if } X_{ij} \text{ and } X_{il} \text{ are both observed} \\ X^*_{ij}x_{il} & \text{if } X_{ij} \text{ is missing and } X_{il} \text{ is observed} \\ x_{ij}X^*_{il} & \text{if } X_{ij} \text{ is observed and } X_{il} \text{ is missing} \\ X^*_{ij}X^*_{il} + \text{covariances} & \text{if } X_{ij} \text{ and } X_{il} \text{ are both missing} \end{cases},$$

where covariances $= \mathrm{Cov}(X^*_{ij}, X^*_{il} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$, which are the elements of the $m_i \times m_i$ covariance matrix of the missing items estimated by (4.4). The covariances $\mathrm{Cov}(x_{ij}, x_{il} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$, $\mathrm{Cov}(X^*_{ij}, x_{il} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ and $\mathrm{Cov}(x_{ij}, X^*_{il} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ all are equal to zero since at least one of the X's is observed and thus regarded as fixed. Now create a $k \times k$ matrix $V^{(t)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ with the following elements for $j, l = 1, \ldots, k$

$$V^{(t)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})_{jl}$$
$$= \begin{cases} \mathrm{Cov}(X^*_{ij}, X^*_{il} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) & \text{if } X_{ij} \text{ and } X_{il} \text{ are both missing} \\ 0 & \text{otherwise} \end{cases}.$$

Now $\mathbf{X}'_i = (\mathbf{X}'_{i,mis}, \mathbf{X}'_{i,obs})$ is replaced by its expectation: $\mathbf{X}'_i = (\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]', \mathbf{x}'_{i,obs})$ and then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are re-estimated by

$$\begin{aligned} \boldsymbol{\mu}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \\ \boldsymbol{\Sigma}^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}^{(t+1)})(\mathbf{X}_i - \boldsymbol{\mu}^{(t+1)})' + \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} V^{(t)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{X}_i\mathbf{X}'_i + V^{(t)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})\right) + \\ &\quad - \boldsymbol{\mu}^{(t+1)}(\boldsymbol{\mu}^{(t+1)})'. \end{aligned}$$

Next the updated parameter estimates are used to re-estimate equations (4.3) and (4.4), which in their turn are used to update the parameter estimates. This process is iterated until convergence. That is, until the change in parameter estimates is sufficiently small.

4.6 EM estimates and linear balance restrictions

It is crucial that the estimates generated by the EM algorithm concur with the linear balance restrictions on the data, in order to obtain imputations that satisfy these restrictions. This means that $\mathbf{A}\boldsymbol{\mu}^{(t)}$ should equal zero and Null($\boldsymbol{\Sigma}^{(t)}$) should equal Null($\boldsymbol{\Sigma}$).

For $\mathbf{A}\boldsymbol{\mu}^{(t)}$ to be equal to zero it is sufficient if it holds that

$$\mathbf{A}\boldsymbol{\mu}^{(t)} = \mathbf{A}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}\mathbf{X}_i = \mathbf{0},$$

for $\mathbf{X}_i$ at iteration t. We know that this holds for the completely observed data and need to show that this also holds for the data with missing items. Again partition $\mathbf{X}_i$ into a missing and an observed part and partition $\mathbf{A}$ accordingly. This means we need to show for the estimated data that

$$\mathbf{A}_{i,mis}\hat{\mathbf{X}}_{i,mis} = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$$

holds at every iteration t. Let $\mathcal{R}$ denote the set of respondents without any item nonresponse and let r denote the number of respondents without any item nonresponse.

In the EM algorithm the missing items for respondent i will be replaced by their expected values:

$$\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = \boldsymbol{\mu}^{(t)}_{mis} + \boldsymbol{\Sigma}^{(t)}_{mis,obs}(\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-}(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(t)}_{obs}).$$

First iteration

We will start at $t = 0$. Assuming there are sufficient complete cases, the complete cases mean and covariance matrix will be used as starting values, which means that $\mathbf{A}\boldsymbol{\mu}^{(0)} = \mathbf{0}$ and Null($\boldsymbol{\Sigma}^{(0)}$) = Null($\boldsymbol{\Sigma}$) as $\boldsymbol{\Sigma}^{(0)}\mathbf{A}' = \mathbf{0}$ which we found in subsection 4.4.1.

Expectation step
The updated $\boldsymbol{\mu}^{(1)}$ will satisfy the linear balance restrictions if the imputed items in this iteration satisfy the balance restrictions, so it has to hold that

$$\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}.$$

Now obtain

$$\begin{aligned}\mathbf{A}_{i,mis}&\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}] \\ &= \mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} + \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,obs}(\boldsymbol{\Sigma}^{(0)}_{obs,obs})^{-}(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}),\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,obs} &= \mathbf{A}_{i,mis}\frac{1}{r}\sum_{h\in\mathcal{R}}(\mathbf{X}_{h,mis} - \boldsymbol{\mu}^{(0)}_{mis})(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})' \\ &= \frac{1}{r}\sum_{h\in\mathcal{R}}(\mathbf{A}_{i,mis}\mathbf{X}_{h,mis} - \mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis})(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})' \\ &= \frac{1}{r}\sum_{h\in\mathcal{R}}(-\mathbf{A}_{i,obs}\mathbf{X}_{h,obs} + \mathbf{A}_{i,obs}\boldsymbol{\mu}^{(0)}_{obs})(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})' \\ &= -\mathbf{A}_{i,obs}\frac{1}{r}\sum_{h\in\mathcal{R}}(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})' \\ &= -\mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(0)}_{obs,obs}.\end{aligned}$$

So

$$\begin{aligned}\mathbf{A}_{i,mis}&\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}] \\ &= \mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} - \mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(0)}_{obs,obs}(\boldsymbol{\Sigma}^{(0)}_{obs,obs})^{-}(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}). \quad (4.5)\end{aligned}$$

If $\boldsymbol{\Sigma}^{(0)}_{obs,obs}$ is nonsingular, the generalised inverse becomes the regular inverse and it immediately follows that $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$.

However, if $\boldsymbol{\Sigma}^{(0)}_{obs,obs}$ is singular the proof is less straightforward. Decompose $\boldsymbol{\Sigma}^{(0)}_{obs,obs}$ by means of an eigenvalue decomposition

$$\boldsymbol{\Sigma}^{(0)}_{obs,obs} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}' = \mathbf{C}\begin{pmatrix}\boldsymbol{\Lambda}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{C}',$$

where $\mathbf{C}$ is the orthogonal matrix of eigenvectors, and $\mathbf{\Lambda}_1$ a diagonal matrix with the nonzero eigenvalues of $\mathbf{\Sigma}^{(0)}_{obs,obs}$ on the diagonal. Since $\mathbf{\Sigma}^{(0)}_{obs,obs}$ is symmetric, its generalised inverse is of the form (Pringle and Rayner, 1971)

$$(\mathbf{\Sigma}^{(0)}_{obs,obs})^{-} = \mathbf{C} \begin{pmatrix} \mathbf{\Lambda}_1^{-1} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{pmatrix} \mathbf{C}',$$

where $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are arbitrary matrices. Partition the matrix $\mathbf{C}$ into $\mathbf{C} = (\mathbf{C}_1 \ \mathbf{C}_2)$, where $\mathbf{C}_1$ is the matrix of eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{\Sigma}^{(0)}_{obs,obs}$. This means that $\mathbf{\Sigma}^{(0)}_{obs,obs}\mathbf{C}_2 = \mathbf{0}$, as $\mathbf{C}_2$ corresponds to the zero eigenvalues of $\mathbf{\Sigma}^{(0)}_{obs,obs}$.

Now recall that the marginal distribution of the observed data for respondent i in this iteration is

$$\mathbf{X}_{i,obs} \sim \mathcal{N}(\boldsymbol{\mu}^{(0)}_{obs}, \mathbf{\Sigma}^{(0)}_{obs,obs}).$$

Transforming $\mathbf{X}_{i,obs}$ leads to

$$\mathbf{C}_2'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_2'\mathbf{\Sigma}^{(0)}_{obs,obs}\mathbf{C}_2),$$

where $\mathbf{C}_2'\mathbf{\Sigma}^{(0)}_{obs,obs}\mathbf{C}_2 = \mathbf{0}$, since $\mathbf{\Sigma}^{(0)}_{obs,obs}\mathbf{C}_2 = \mathbf{0}$. So $\mathbf{C}_2'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}) = \mathbf{0}$ with probability one.

The right-hand side of equation (4.5) now becomes

$$\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} - \mathbf{A}_{i,obs}\mathbf{C} \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{C}'\mathbf{C} \begin{pmatrix} \mathbf{\Lambda}_1^{-1} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{pmatrix} \mathbf{C}'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}),$$

which reduces to

$$\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} - \mathbf{A}_{i,obs}\mathbf{C} \begin{pmatrix} \mathbf{I}_{k-m_i-p_i} & \mathbf{\Lambda}_1\mathbf{U} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{C}'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}),$$

since $\mathbf{C}'\mathbf{C} = \mathbf{I}_{k-m_i}$, where p_i denotes the number of nonredundant balance restrictions on the missing items for respondent i. This can then be written as

$$\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} - \mathbf{A}_{i,obs}(\mathbf{C}_1\mathbf{C}_1' + \mathbf{C}_1\mathbf{\Lambda}_1\mathbf{U}\mathbf{C}_2')(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}).$$

We already established that $\mathbf{C}_2'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}) = \mathbf{0}$ with probability one, so this equation results in

$$\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(0)}_{mis} - \mathbf{A}_{i,obs}\mathbf{C}_1\mathbf{C}_1'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}). \tag{4.6}$$

Finally, note that the following holds

$$\mathbf{I}_{k-m_i} = \mathbf{C}\mathbf{C}' = (\mathbf{C}_1 \ \mathbf{C}_2)\begin{pmatrix} \mathbf{C}_1' \\ \mathbf{C}_2' \end{pmatrix} = \mathbf{C}_1\mathbf{C}_1' + \mathbf{C}_2\mathbf{C}_2'.$$

Post-multiplying this equation with $\mathbf{X}_{i,obs} - \boldsymbol{\mu}_{obs}^{(0)}$ leads to

$$\begin{aligned} \mathbf{X}_{i,obs} - \boldsymbol{\mu}_{obs}^{(0)} &= (\mathbf{C}_1\mathbf{C}_1' + \mathbf{C}_2\mathbf{C}_2')(\mathbf{X}_{i,obs} - \boldsymbol{\mu}_{obs}^{(0)}) \\ &= \mathbf{C}_1\mathbf{C}_1'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}_{obs}^{(0)}). \end{aligned}$$

Now (4.6) becomes

$$\mathbf{A}_{i,mis}\boldsymbol{\mu}_{mis}^{(0)} - \mathbf{A}_{i,obs}\mathbf{X}_{i,obs} + \mathbf{A}_{i,obs}\boldsymbol{\mu}_{obs}^{(0)} = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}.$$

This means that $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$ for both singular and nonsingular $\boldsymbol{\Sigma}_{obs,obs}^{(0)}$.

Note that the above line of reasoning not only holds for a singular $\boldsymbol{\Sigma}_{obs,obs}$ at iteration $t = 0$, but for a singular $\boldsymbol{\Sigma}_{obs,obs}$ at any iteration t.

Maximisation step

Now proceed to the M-step. The mean, $\boldsymbol{\mu}^{(1)}$ is calculated based on the observed and estimated data in the expectation step and as for each respondent in the imputed dataset it holds that $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$, it will also hold that $\mathbf{A}_{i,mis}\boldsymbol{\mu}_{mis}^{(1)} = -\mathbf{A}_{i,obs}\boldsymbol{\mu}_{obs}^{(1)}$. Next $\boldsymbol{\Sigma}$ will be re-estimated using an updated $\mathbf{X}_i = (\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}]', \mathbf{X}_{i,obs}')'$:

$$\boldsymbol{\Sigma}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i - \boldsymbol{\mu}^{(1)})(\mathbf{X}_i - \boldsymbol{\mu}^{(1)})' + \mathrm{V}^{(0)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})\Big).$$

For each record i it holds that $\mathbf{A}\mathrm{V}^{(0)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) = \mathbf{0}$. This can be established as follows. Again partition $\mathbf{A}$ into a missing and an observed part and partition the matrix $\mathrm{V}^{(0)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$ accordingly. Now for any record i

$$\begin{aligned} &\mathbf{A}\mathrm{V}^{(0)}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) \\ &\qquad = (\mathbf{A}_{i,mis} \ \mathbf{A}_{i,obs})\begin{pmatrix} \mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &\qquad = (\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) \quad \mathbf{0}), \end{aligned}$$

where

$$\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) = \boldsymbol{\Sigma}^{(0)}_{mis,mis} - \boldsymbol{\Sigma}^{(0)}_{mis,obs}(\boldsymbol{\Sigma}^{(0)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(0)}_{obs,mis}.$$

Now

$$\begin{aligned}
&\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) \\
&\quad = \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} - \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,obs}(\boldsymbol{\Sigma}^{(0)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(0)}_{obs,mis} \\
&\quad = \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} + \mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(0)}_{obs,obs}(\boldsymbol{\Sigma}^{(0)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(0)}_{obs,mis}. \qquad (4.7)
\end{aligned}$$

If $\boldsymbol{\Sigma}^{(0)}_{obs,obs}$ is nonsingular this leads to

$$\begin{aligned}
\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} + \mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(0)}_{obs,mis} \\
&= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} - \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} \\
&= \mathbf{0}.
\end{aligned}$$

However, if $\boldsymbol{\Sigma}^{(0)}_{obs,obs}$ is singular, equation (4.7) becomes

$$\begin{aligned}
&\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) \\
&\quad = \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} + \mathbf{A}_{i,obs}(\mathbf{C}_1\mathbf{C}_1' + \mathbf{C}_1\boldsymbol{\Lambda}_1\mathbf{U}\mathbf{C}_2')\boldsymbol{\Sigma}^{(0)}_{obs,mis}. \qquad (4.8)
\end{aligned}$$

Since

$$\boldsymbol{\Sigma}^{(0)}_{obs,mis} = \frac{1}{r}\sum_{h\in\mathcal{R}}(\mathbf{X}_{h,obs} - \boldsymbol{\mu}^{(0)}_{obs})(\mathbf{X}_{h,mis} - \boldsymbol{\mu}^{(0)}_{mis})'$$

and

$$\begin{aligned}
\mathbf{C}_2'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}) &= \mathbf{0} \\
\mathbf{C}_1\mathbf{C}_1'(\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}) &= (\mathbf{X}_{i,obs} - \boldsymbol{\mu}^{(0)}_{obs}),
\end{aligned}$$

equation (4.8) reduces to

$$\begin{aligned}
\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}) &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(0)}_{mis,mis} + \mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(0)}_{obs,mis} \\
&= \mathbf{0},
\end{aligned}$$

and consequently $\mathbf{A}\mathrm{V}^{(0)}(\mathbf{X}_{i,mis}) = \mathbf{0}$.

The null space of $\boldsymbol{\Sigma}^{(1)}$ can now be established by postmultiplying $\boldsymbol{\Sigma}^{(1)}$ with

$\mathbf{A}'$

$$\begin{aligned}\boldsymbol{\Sigma}^{(1)}\mathbf{A}' &= \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i-\boldsymbol{\mu}^{(1)})(\mathbf{X}_i-\boldsymbol{\mu}^{(1)})'+\mathrm{V}^{(0)}(\mathbf{X}_{i,mis})\Big)\mathbf{A}' \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i-\boldsymbol{\mu}^{(1)})(\mathbf{A}\mathbf{X}_i-\mathbf{A}\boldsymbol{\mu}^{(1)})'+\mathrm{V}^{(0)}(\mathbf{X}_{i,mis})\mathbf{A}'\Big) \\ &= \mathbf{0}.\end{aligned}$$

This means that the null space of $\boldsymbol{\Sigma}^{(1)}$ is equal to the null space of $\boldsymbol{\Sigma}$. We have now established that the parameter estimates obtained in the first iteration of the EM algorithm concur with the linear balance restrictions.

Subsequent iterations

Next consider the case where we are at $t = 1$. The parameter estimates now depend not only on the observed data but on the imputed data as well.

Expectation step
Impute for nonrespondent i once again, but now based on the new estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}^{(1)}]=\boldsymbol{\mu}^{(1)}_{mis}+\boldsymbol{\Sigma}^{(1)}_{mis,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^{(1)}_{obs}).$$

Then

$$\begin{aligned}&\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}^{(1)}]= \\ &\qquad\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(1)}_{mis}+\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(1)}_{mis,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^{(1)}_{obs}).\end{aligned}$$

From $\mathbf{A}\boldsymbol{\Sigma}^{(1)}=\mathbf{0}$, it follows that $\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(1)}_{mis,obs}=-\mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(1)}_{obs,obs}$. So

$$\begin{aligned}&\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}^{(1)}]= \\ &\qquad\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(1)}_{mis}-\mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(1)}_{obs,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^{(1)}_{obs}).\end{aligned}$$

This immediately leads to $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis}\mid\mathbf{X}_{i,obs},\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}^{(1)}]=-\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$ if $\boldsymbol{\Sigma}^{(1)}_{obs,obs}$ is nonsingular. If $\boldsymbol{\Sigma}^{(1)}_{obs,obs}$ is singular we can use the same line of reasoning as in the first iteration to find that

$$\boldsymbol{\Sigma}^{(1)}_{obs,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}(\mathbf{X}_{i,obs}-\boldsymbol{\mu}^{(1)}_{obs})=\mathbf{X}_{i,obs}-\boldsymbol{\mu}^{(1)}_{obs}$$

and thus: $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$.

Maximisation step
The M-step now estimates $\boldsymbol{\mu}^{(2)}$. For both the estimated and the observed data at this iteration, we found that $\mathbf{A}_{i,mis}\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}] = -\mathbf{A}_{i,obs}\mathbf{X}_{i,obs}$, so it will also hold that $\mathbf{A}_{i,mis}\boldsymbol{\mu}^{(2)}_{mis} = -\mathbf{A}_{i,obs}\boldsymbol{\mu}^{(2)}_{obs}$.

Next the covariance matrix will be re-estimated by

$$\boldsymbol{\Sigma}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\Big((\mathbf{X}_i - \boldsymbol{\mu}^{(2)})(\mathbf{X}_i - \boldsymbol{\mu}^{(2)})' + \mathrm{V}^{(1)}(\mathbf{X}_{i,mis})\Big).$$

Now

$$\mathbf{A}V^{(1)}(\mathbf{X}_{i,mis}) = (\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}) \;\; \mathbf{0}),$$

where

$$\begin{aligned}\mathbf{A}_{i,mis}\mathrm{Var}(\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}) &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(1)}_{mis,mis} + \\ &\quad -\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(1)}_{mis,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(1)}_{obs,mis} \\ &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(1)}_{mis,mis} + \\ &\quad +\mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(1)}_{obs,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(1)}_{obs,mis}.\end{aligned}$$

Once again the same line of reasoning as in the first iteration can be used to establish that $\boldsymbol{\Sigma}^{(1)}_{obs,obs}(\boldsymbol{\Sigma}^{(1)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(1)}_{obs,mis} = \boldsymbol{\Sigma}^{(1)}_{obs,mis}$ for a singular $\boldsymbol{\Sigma}^{(1)}_{obs,obs}$ and therefore that $\mathbf{A}V^{(1)}(\mathbf{X}_{i,mis}) = \mathbf{0}$. So $\boldsymbol{\Sigma}^{(2)}\mathbf{A}' = \mathbf{0}$, which means that $\mathrm{Null}(\boldsymbol{\Sigma}^{(2)}) = \mathrm{Null}(\boldsymbol{\Sigma})$.

This can be straightforwardly extended to the subsequent iterations $t = 2, 3, \ldots$. So at each iteration in the EM algorithm the maximum likelihood estimates concur with the linear balance restrictions, that is $\mathbf{A}\boldsymbol{\mu}^{(t)} = \mathbf{0}$ and the null space of $\boldsymbol{\Sigma}^{(t)}$ is equal to the null space of $\boldsymbol{\Sigma}$ and this will result in imputed values that satisfy the balance restrictions.

4.6.1 Starting values

The EM algorithm requires a starting value for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The mean vector and covariance matrix can, for example, be calculated from the completely observed data or by using the available cases for each variable. Using the complete cases provides consistent estimates of the parameters if the data are MCAR and if there are at least $k + 1$ observations. In general the choice of the starting value

is not crucial, unless the fraction of missing data is very high (Schafer, 1997).

In our case, however, the choice of the starting value is crucial in the sense that if the starting values do not correspond with the balance restrictions, i.e. $\mathbf{A}\boldsymbol{\mu}^{(0)} \neq \mathbf{0}$ or $\text{Null}(\boldsymbol{\Sigma}^{(0)}) \neq \text{Null}(\boldsymbol{\Sigma})$, the final estimates of the parameters will not either. This rules out the use of available cases estimates for data subject to linear balance restrictions.

4.7 Imputation

4.7.1 The singularity of $\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs}$

Previously we mentioned that the covariance matrix $\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs}$, that is used to generate imputations, is always singular if all variables are present in at least one balance restriction. We will present a formal proof here. Recall that for each respondent i at iteration t it holds that

$$\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs} = \boldsymbol{\Sigma}^{(t)}_{mis,mis} - \boldsymbol{\Sigma}^{(t)}_{mis,obs}(\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(t)}_{obs,mis}.$$

We leave out the dependence on i for ease of notation. Premultiply $\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs}$ with $\mathbf{A}_{i,mis}$, the restriction matrix on the missing items of respondent i

$$\begin{aligned}\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs} &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(t)}_{mis,mis} - \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(t)}_{mis,obs}(\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(t)}_{obs,mis} \\ &= \mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(t)}_{mis,mis} + \mathbf{A}_{i,obs}\boldsymbol{\Sigma}^{(t)}_{obs,obs}(\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(t)}_{obs,mis},\end{aligned}$$

since $\mathbf{A}\boldsymbol{\Sigma}^{(t)} = \mathbf{0}$, which was shown in section 4.6.

If $\boldsymbol{\Sigma}^{(t)}_{obs,obs}$ is nonsingular, it immediately follows that $\mathbf{A}_{i,mis}\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs}$ equals zero. We will use the fact that $\boldsymbol{\Sigma}^{(t)}_{obs,obs}(\boldsymbol{\Sigma}^{(t)}_{obs,obs})^{-}\boldsymbol{\Sigma}^{(t)}_{obs,mis} = \boldsymbol{\Sigma}^{(t)}_{obs,mis}$ (see section 4.6), if $\boldsymbol{\Sigma}^{(t)}_{obs,obs}$ is singular. Hence $\boldsymbol{\Sigma}^{(t)}_{mis,mis.obs}\mathbf{A}'_{i,mis} = \mathbf{0}$ for both singular and nonsingular $\boldsymbol{\Sigma}^{(t)}_{obs,obs}$. This means that the columns of $\mathbf{A}'_{i,mis}$ are in the null space of $\boldsymbol{\Sigma}_{mis,mis.obs}$. Since every variable is present in at least one linear balance restriction and since there are no redundant balance restrictions, the restriction matrix $\mathbf{A}$ is of full rank. Therefore it holds that $\text{rank}(\mathbf{A}_{i,mis}) \geq 1$, from which it follows that $\boldsymbol{\Sigma}_{mis,mis.obs}$ is always singular with rank $m_i - \text{rank}(\mathbf{A}_{i,mis})$.

4.7.2 Imputation of missing data items

The missing data items can now be imputed using either deterministic or stochastic imputation. Deterministic imputation can be done by using the expected

missing data items that were calculated in the E-step for imputation. Using expectations, however, may lead to an underestimation of the true variance and may attenuate relationships between variables. If one is mostly interested in the estimation of means and totals this method is expected to work well. This means that the imputed values are

$$\hat{\mathbf{X}}_{i,imp} = \mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}],$$

which, as we established earlier, will satisfy the balance restrictions.

If the aim of the imputation procedure is to obtain a general purpose dataset, it would be better to use a stochastic imputation method. For example by using draws from the singular normal, with parameter estimates from the EM algorithm, as imputations. In this case the true variances of the variables are better represented in the imputed dataset. The missing data items have the following distribution

$$\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}_{m_i}(\boldsymbol{\mu}_{mis.obs}, \boldsymbol{\Sigma}_{mis,mis.obs}).$$

The $m_i \times m_i$ matrix $\boldsymbol{\Sigma}_{mis,mis.obs}$ is singular (see subsection 4.7.1), which means that we can decompose it by means of an eigenvalue decomposition into $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$, where $\mathbf{C}$ is an orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}$ on the main diagonal. So $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_{m_i - p_i}, 0, \ldots, 0\}$, assuming that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m_i - p_i} > 0$, where λ_i is the i-th eigenvalue, again m_i denotes the number of missing items and p_i denotes the number of nonredundant balance restrictions on the missing items for respondent i. Then $\mathbf{X}_{i,mis}$ can be transformed as follows

$$\mathbf{C}'(\mathbf{X}_{i,mis} - \boldsymbol{\mu}_{mis.obs}) \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}_{m_i}(\mathbf{0}, \boldsymbol{\Lambda}).$$

Generate $Z_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, m_i - p_i$ and take Z_j equal to zero otherwise. Then calculate $\tilde{\mathbf{Z}} = \boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{Z}$ so that $\tilde{\mathbf{Z}} \sim \mathcal{N}_{m_i}(\mathbf{0}, \boldsymbol{\Lambda})$. Finally, impute the missing data items by

$$\hat{\mathbf{X}}_{i,imp} = \mathbf{C}\tilde{\mathbf{Z}} + \boldsymbol{\mu}_{mis.obs}.$$

4.8 Imputation performance

Similarly to chapter 3 we will use this method to impute data that have been gathered by Statistics Netherlands on a part of the wholesale industry for businesses with more than 10 employees, in order to assess the performance of this method on empirical data. The effects on estimation of population parameters will be analysed.

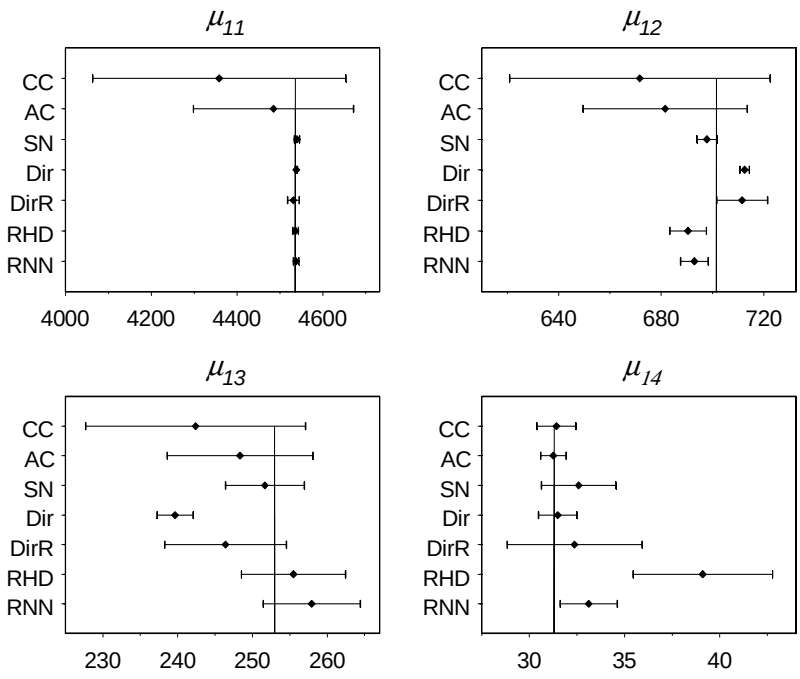

Figure 4.2: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

4.8.1 Estimation of population parameters

The effects of imputation will be investigated with respect to estimation of the population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. As we mentioned in chapter 3, estimates of population parameters are potentially subject to sampling variance and nonresponse variance. Nonresponse variance is introduced by the nonresponse in the sample and is currently our main interest. First of all, this method is incorporated in the simulation study that was executed in chapter 3 for the datasets on labour costs and costs of third party rendering of services, which are both subject to one balance restriction. Next this method will be used for data that are subject to multiple linear balance restrictions.

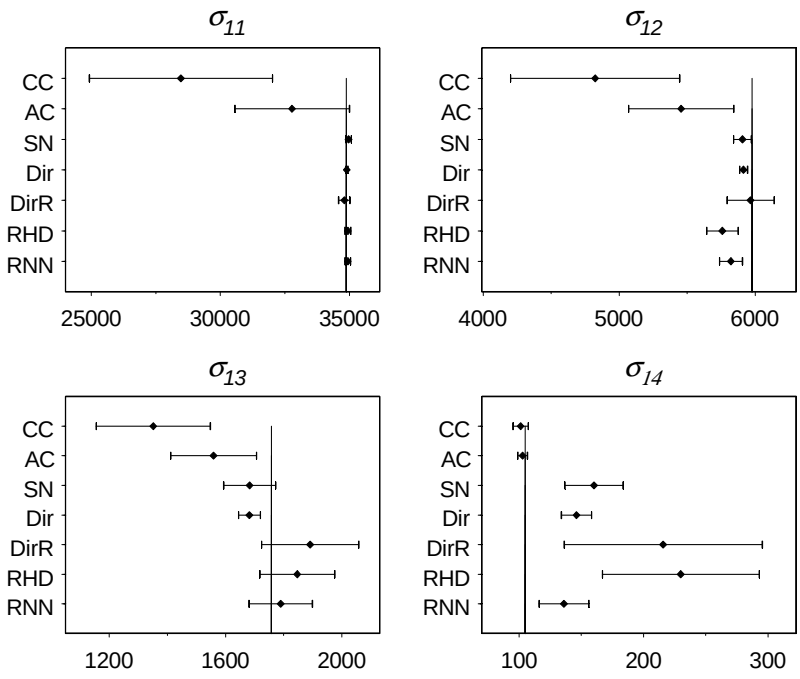

Figure 4.3: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

4.8.2 One linear balance restriction

Clearly, the singular normal distribution can also be applied to data that are subject to only one linear balance restriction ($p = 1$). This means that we can compare the results for this imputation method with the results that were obtained in chapter 3. Recall, however, that the methods in chapter 3 also included non-negativity constraints, which is not the case for the singular normal model. Consequently in this instance the individual imputed values may violate the non-negativity constraints. For parameter estimation this will probably not pose any problems as it is highly unlikely that the parameter estimates will become negative.

Two datasets were used, the first dataset dealt with labour costs and the second with costs of third party rendering of services. Then 100 random samples of missing data were generated and next the remaining datasets containing miss-

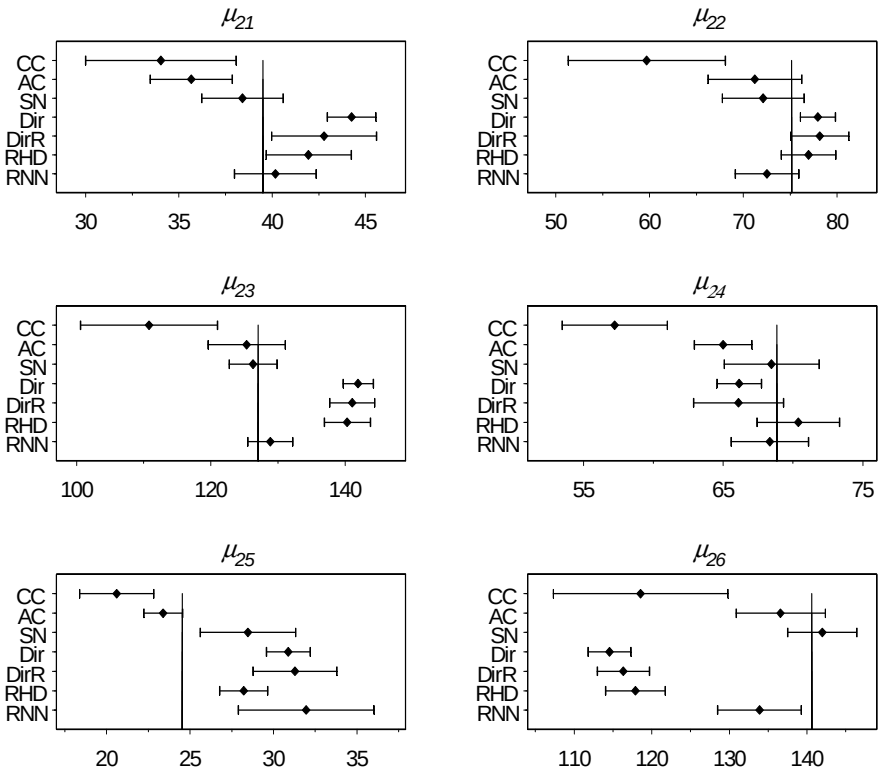

Figure 4.4: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

ing data items were used for parameter estimation, either by using the incomplete data or data completed by imputation. The averaged parameter estimates and their 95% confidence intervals that are provided for the incomplete datasets using the EM algorithm for singular normal data (SN) are added to the results from chapter 3, which is presented in Figures 4.2, 4.3, 4.4 and 4.5.

For the data on labour costs (Figures 4.2 and 4.3) it is observed that the parameter estimates generated by the EM algorithm for the singular normal distribution are quite accurate point estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The nonresponse variance is, however, somewhat larger than for example the nonresponse variance produced by the Dirichlet method with expectations imputed (Dir).

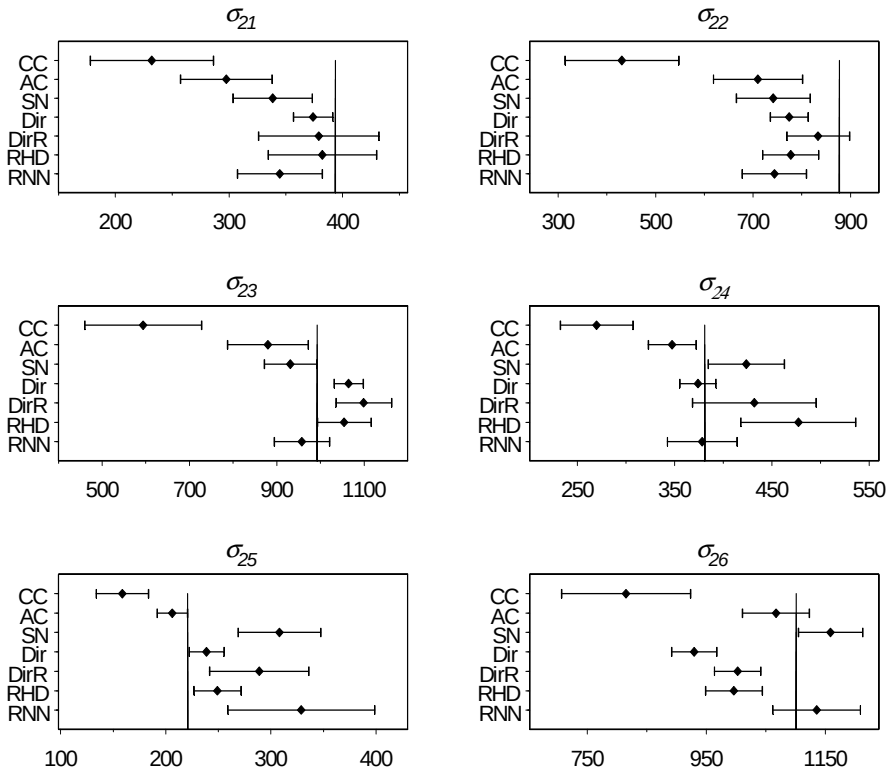

Figure 4.5: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

With respect to the data on costs of third party rendering of services (Figures 4.4 and 4.5) the EM algorithm for singular normal data produces very accurate point estimates for $\boldsymbol{\mu}$, again with a somewhat larger nonresponse variance than the Dirichlet method. This method is less capable of finding accurate point estimates for $\boldsymbol{\sigma}$, however, and in this case the nonresponse variance is again quite substantial.

In conclusion, the EM algorithm for singular normal data appears to be a promising approach. For the mean parameters accurate point estimates are obtained, but on the other hand the dispersion parameters are less well estimated. Besides, the nonresponse variance can be substantial in some cases, meaning

that the parameter estimates depend on the realised set of missing data. Furthermore, as the EM algorithm for singular normal data converges quite slowly the procedure can become time-consuming, in which case the other imputation methods are probably preferred.

The main advantage of this method, however, is that it is able to impute data that are subject to any set of multiple linear balance restrictions. Besides, in this case we do not need to assume that the variable, which represents a certain total, is completely observed. In the next section the effects on parameter estimation for a dataset subject to multiple linear balance restrictions are analysed.

4.8.3 Multiple linear balance restrictions

4.8.3.1 Description of the data

The dataset that will be examined concerns the cost structure of companies in part of the wholesale industry with more than 10 employees. The variables X_{1t} and X_{2t}, representing total labour costs and total costs of third party rendering of services respectively, that were used in the previous analyses are in fact part of the total operating expenses of a business. The total operating expenses (X_t) is composed of

$$\begin{aligned}
X_{1t} &= \text{total labour costs} \\
X_{2t} &= \text{total costs of rendering of services} \\
X_{3t} &= \text{total costs of sales} \\
X_{4t} &= \text{total other personnel costs} \\
X_{5t} &= \text{total costs of transportation} \\
X_{6t} &= \text{total costs of energy} \\
X_{7t} &= \text{total costs of housing} \\
X_{8t} &= \text{total costs of machinery and equipment} \\
X_{9t} &= \text{total sales expenses} \\
X_{10t} &= \text{communication expenses} \\
X_{11t} &= \text{other company expenses} \\
X_{12t} &= \text{depreciation on fixed assets.}
\end{aligned}$$

The variables $X_{11}, \ldots, X_{14}$ and $X_{21}, \ldots, X_{26}$, used in chapter 3 and the previous section are also present in the dataset. The following balance restrictions hold:

$$X_{1t} = X_{11} + \cdots + X_{14} \tag{4.9}$$
$$X_{2t} = X_{21} + \cdots + X_{26} \tag{4.10}$$
$$X_t = X_{1t} + \cdots + X_{12t}. \tag{4.11}$$

As these variables represent some sort of operating expense, they all are non-negative. Moreover, most of the variables $X_{3t}, \ldots, X_{12t}$ also represent the total of an underlying balance restriction, which are left out for now in order to reduce the complexity of the model.

Again records with missing items are removed from the original survey and missing values are generated in this dataset based on the MCAR assumption, using Bernoulli draws with parameter p chosen such that the nonresponse rate is similar to the rate observed in the original survey. As most of these variables are aggregate values, the observed nonresponse rates for the variables $X_{1t}, \ldots, X_{12t}, X_t$ are much lower (about 17%) than the nonresponse rates that were found for the variables $X_{11}, \ldots, X_{14}$ and $X_{21}, \ldots, X_{26}$ (about 31%). After generation of the missing items deductive imputation is applied to impute the data items that can be derived with certainty, based on the edit constraints given by (4.9), (4.10), (4.11) and the non-negativity restrictions.

Parameter estimates are calculated based on the remaining dataset. The complete cases (CC), available cases (AC) and the EM algorithm for the singular normal model (SN) are used to obtain parameter estimates. Due to the fact that there are no other imputation methods that can straightforwardly yield imputed values or parameter estimates that satisfy multiple linear balance restrictions, no other imputation methods can be used for comparison purposes. This means that we can only assess the performance of the singular normal model with respect to the complete and available cases estimates. In chapter 3 and section 4.8.2 we found that these estimates strongly depend on the realised set of missing data. As the amount of missing values is reduced in this dataset we expect that the incomplete data procedure that uses the available cases will result in accurate point estimates with lower nonresponse variance.

4.8.3.2 The effects on parameter estimation

In Figures 4.6 and 4.7 the parameter estimates and their 95% confidence intervals for respectively $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are provided. Due to the high number of variables present in this dataset, only the results for the aggregate variables, $X_{1t}, \ldots, X_{12t}$

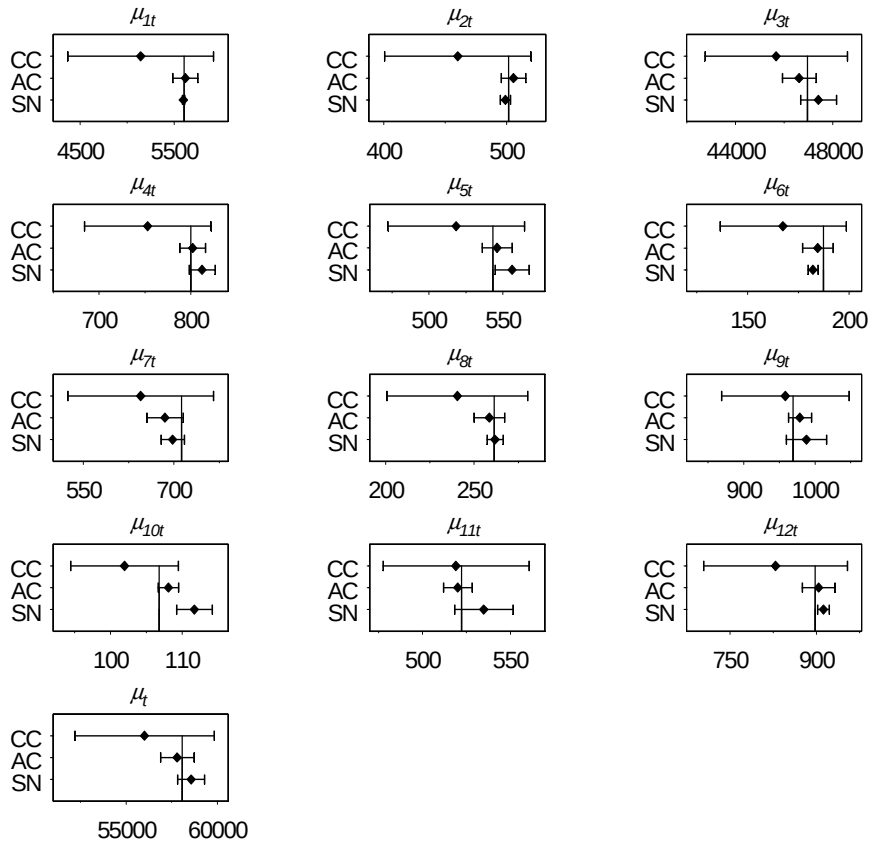

Figure 4.6: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

and X_t are shown. It is clear that the method based on the available cases now produces good results with respect to accuracy as well as precision for the estimation of both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The balance restrictions that need to hold are, however, violated. The average absolute violation relative to the variable representing the total of that restriction is 8%, 9% and 8% for restrictions (4.9), (4.10) and (4.11) respectively.

For the mean parameter, the estimates that are obtained with the EM al-

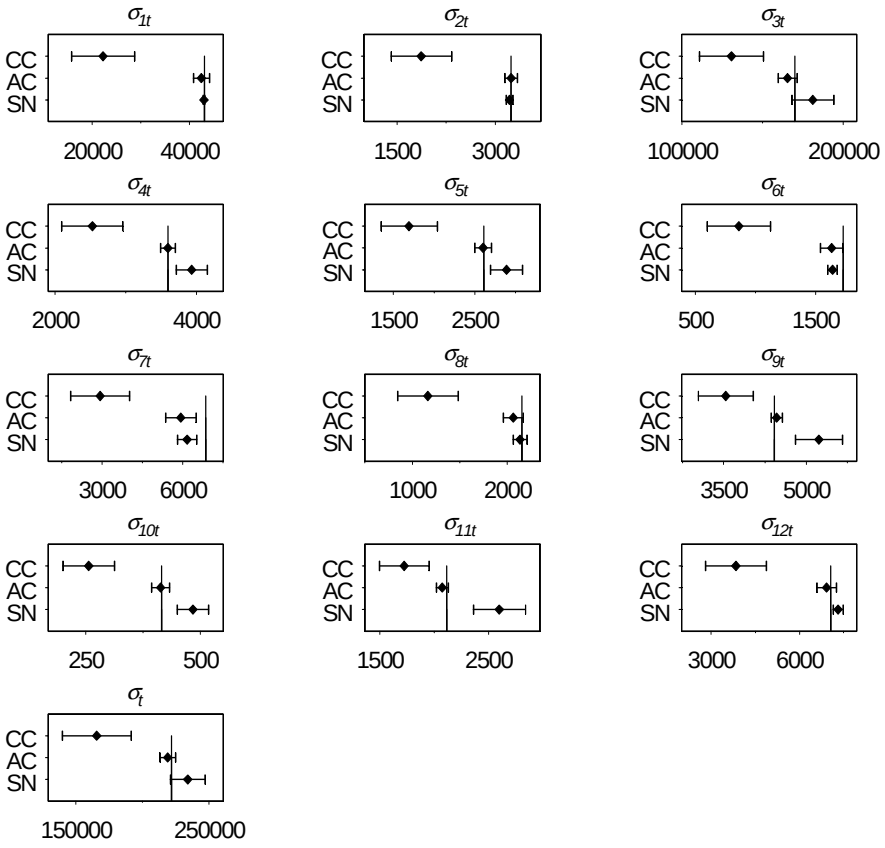

Figure 4.7: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

gorithm for singular normal data closely resemble the available cases estimates, so both methods appear quite capable of preserving the true parameter estimates. The confidence intervals, indicating the sensitivity of the results with respect to the realised set of missing data, are quite acceptable as well for both methods. A striking result is the fact that the parameter estimates obtained with the EM algorithm for μ_{1t} and μ_{2t} are very close to the true value with a relatively small confidence interval. This is probably due to the fact that the vari-

ables X_{1t} and X_{2t} are aggregates of the variables $X_{11}, \ldots, X_{14}$ and $X_{21}, \ldots, X_{26}$ respectively, which are also taken into account in the imputation process. This means that, although these variables contain missing values as well, more information is available for the imputation of X_{1t} and X_{2t}, which clearly provides better estimates. This is an encouraging result as most of the other variables $X_{3t}, \ldots, X_{12t}$ are also the aggregate of an underlying balance restriction. These variables were left out in this analysis as the dataset would have become quite substantial in that case, meaning that a simulation study would take up too much computing time. This will certainly be a topic of further research, however.

With regard to the estimation of $\boldsymbol{\sigma}$, the available cases estimates appear to have a slight advantage over the estimates obtained through the EM algorithm for singular normal data. The latter shows on average less accuracy and a larger nonresponse variance. Moreover, the EM estimates appear to have a tendency to somewhat overestimate $\boldsymbol{\sigma}$. The major disadvantage of the available cases method is of course that the covariance matrix produced is not singular. With respect to the variables X_{1t} and X_{2t} again we observe that the estimates produced by the EM algorithm are very accurate, with small confidence intervals. This means that also for the estimation of $\boldsymbol{\sigma}$ it appears quite beneficial to incorporate information from underlying balance restrictions. Also note that for these variables the EM estimates do not overestimate $\boldsymbol{\sigma}$. In summary, we have proposed a promising new method to take multiple linear balance restrictions into account.

4.9 Concluding remarks

In this chapter we have discussed the use of the singular normal distribution to generate imputations for data that are subject to multiple linear restrictions. We have shown that, by choosing appropriate starting values, the EM algorithm for the singular normal distribution will produce maximum likelihood estimates for the mean and dispersion parameters that concur with the balance restrictions. This means that imputations can be obtained that satisfy multiple linear balance restrictions, which, to our knowledge, is not possible with any other imputation technique developed so far.

Another advantage of this model is the fact that the balance restrictions need not be specified as they are embedded in the singularity of the covariance matrix, which is very convenient.

In section 4.8 on empirical data we found that this method is quite capable of

obtaining accurate parameter estimates for the population parameters on mean and dispersion for both data subject to one restriction as well as data subject to multiple balance restrictions. For the latter case the results indicate that the performance of the singular normal model strongly improves if data items are present in multiple balance restrictions, which is not surprising as more information for imputation is available in that instance. This area could be a topic for future research on this imputation method.

The performance of this procedure on individual level has not been investigated as there were no alternative imputation procedures available that could impute data subject to multiple restrictions, immediately satisfying these restrictions. If more procedures become available, this could also be a subject for further research.

The procedure is generally applicable to continuous data as the singular normal model can incorporate both variables that are present in a balance restriction as well as variables that are not. This means that the imputation method can be used for a large part of a business survey at once. So variables concerning company expenses, company turnover, company profits and employment can be simultaneously imputed using all available balance restrictions. A downside however, reducing its general applicability, is the fact that non-negativity restrictions cannot be incorporated. This means that the singular normal distribution will mostly be of use for the imputation of data representing aggregate variables that are far from zero.

Sometimes economic data need to satisfy other inequality restrictions as well. Clearly, the singular normal distribution does not take inequality restrictions into account. This means that there is still a need for an imputation procedure that can deal with both balance and inequality restrictions simultaneously. This will be the main topic of the next chapter.

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 $\mathbf{X}_i$ denote the data vector for respondent i. Define the set $G = \{\mathbf{X}_i \in \mathbb{R}^k : \mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}\}$, which defines all possible outcomes of the variables. The matrix $\mathbf{B}$ is an $r \times k$ matrix containing r inequality restrictions on the data, $\mathbf{B}$ 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 $\mathbf{u}$ and $\mathbf{l}$ 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 $\mathbf{A}$, with p the number of linear balance restrictions, for which it holds that $\mathbf{A}\mathbf{X}_i = \mathbf{0}$. We assume that every variable is present in at least one balance restriction and that there are no redundant balance restrictions, which means that $\mathbf{A}$ is of full row rank.

So, the data $\mathbf{X}_i$, $i = 1, \ldots, n$, which are subject to $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$ and $\mathbf{A}\mathbf{X}_i = \mathbf{0}$ 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 $\mathbf{\Sigma}$ 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 beliefs 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

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

where $\psi(\mathbf{x}_i \mid \boldsymbol{\theta})$ is a multivariate nontruncated normal density function with parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\theta}' = (\boldsymbol{\mu}', \text{vec}(\mathbf{\Sigma})')$. The normalising probability $\Pr(\mathbf{X}_i \in F) = \int\cdots\int_F \psi(\mathbf{x}_i \mid \boldsymbol{\theta})\, d\mathbf{x}_i$ 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 μ, as X is truncated from below. Moreover, the sample standard deviation, s, underestimates the parameter σ 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 $\mathbf{X}_i$ be distributed according to a

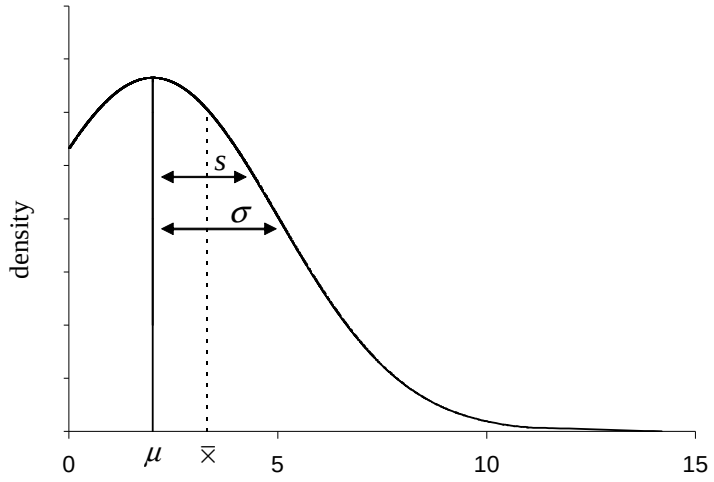

Figure 5.1: *Density plot for a univariate truncated normal density.*

truncated normal distribution

$$\mathbf{X}_i \sim \mathcal{N}_k^T(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \mathbf{X}_i \in G.$$

Multiplying the vector $\mathbf{X}_i$ with a $k \times k$ matrix $\mathbf{D}$ of full rank leads to

$$\mathbf{D}\mathbf{X}_i \sim \mathcal{N}_k^T(\mathbf{D}\boldsymbol{\mu}, \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'), \quad \mathbf{X}_i \in G.$$

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

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y}))\text{abs}|\mathbf{J}|$$

is the probability density function of $\mathbf{Y}$, where $\mathbf{J}$ is the matrix of partial derivatives of $g^{-1}(\mathbf{y})$. Let $\mathbf{Y}_i = \mathbf{D}\mathbf{X}_i$, $\mathbf{Y}_i \in T$, with $T := \{\mathbf{D}\mathbf{X}_i,\ \mathbf{X}_i \in G\}$. Thus

$g^{-1}(\mathbf{y}_i) = \mathbf{D}^{-1}\mathbf{y}_i$ and $\frac{\partial g^{-1}(\mathbf{y}_i)}{\partial \mathbf{y}_i} = (\mathbf{D}^{-1})'$. This means that

$$\begin{aligned} f_{\mathbf{Y}_i}(\mathbf{y}_i) &= f_{\mathbf{X}_i}(\mathbf{D}^{-1}\mathbf{y}_i)\text{abs}|(\mathbf{D}^{-1})'| \\ &= \frac{1}{\Pr(\mathbf{X}_i \in G)}\psi(\mathbf{D}^{-1}\mathbf{y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\text{abs}|\mathbf{D}^{-1}| \\ &= \frac{(2\pi)^{-k/2}}{\Pr(\mathbf{X}_i \in G)}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(\frac{1}{2}(\mathbf{D}^{-1}\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{D}^{-1}\mathbf{y}_i - \boldsymbol{\mu})\big)\text{abs}|\mathbf{D}^{-1}|. \end{aligned}$$

Note that

$$(\mathbf{D}^{-1}\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{D}^{-1}\mathbf{y}_i - \boldsymbol{\mu}) = (\mathbf{y}_i - \mathbf{D}\boldsymbol{\mu})'(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}')^{-1}(\mathbf{y}_i - \mathbf{D}\boldsymbol{\mu})$$

and

$$|\mathbf{D}^{-1}| = |\mathbf{D}|^{-1} = |\mathbf{D}|^{-\frac{1}{2}}|\mathbf{D}|^{-\frac{1}{2}} = |\mathbf{D}|^{-\frac{1}{2}}|\mathbf{D}'|^{-\frac{1}{2}}.$$

Then

$$\begin{aligned} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\text{abs}|\mathbf{D}^{-1}| &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\text{abs}(|\mathbf{D}|^{-\frac{1}{2}}|\mathbf{D}'|^{-\frac{1}{2}}) \\ &= \text{abs}(|\mathbf{D}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\mathbf{D}'|^{-\frac{1}{2}}) = \text{abs}(|\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'|^{-\frac{1}{2}}), \end{aligned}$$

as $\boldsymbol{\Sigma}$ is positive definite and therefore the determinant of $\boldsymbol{\Sigma}$ is positive. This also implies that $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'$ is always positive definite and therefore $|\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'| > 0$, so $\text{abs}|\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'|^{-1/2} = |\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'|^{-1/2}$. This means that the density function for $\mathbf{Y}_i$ becomes

$$f_{\mathbf{Y}_i}(\mathbf{y}_i) = \frac{(2\pi)^{-k/2}}{\Pr(\mathbf{X}_i \in G)}|\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'|^{-\frac{1}{2}} \exp\big(\frac{1}{2}(\mathbf{y}_i - \mathbf{D}\boldsymbol{\mu})'(\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}')^{-1}(\mathbf{y}_i - \mathbf{D}\boldsymbol{\mu})\big).$$

Since $\Pr(\mathbf{Y}_i \in T) = \Pr(\mathbf{X}_i \in G)$, the probability density function of $\mathbf{Y}_i$ is truncated normal as well, with mean vector $\mathbf{D}\boldsymbol{\mu}$ and covariance matrix $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'$. Note that this will only hold if the matrix $\mathbf{D}$ is nonsingular.

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

$$\mathbf{D} = \begin{pmatrix} \mathbf{I}_m & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{k-m} \end{pmatrix},$$

where m is the length of data vector $\mathbf{X}_{i,1}$. Now calculate

$$\mathbf{D}(\mathbf{X}_i - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{X}_{i,1} - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_{i,2} - \boldsymbol{\mu}_2) \\ \mathbf{X}_{i,2} - \boldsymbol{\mu}_2 \end{pmatrix}.$$

This will still be truncated multivariate normal with zero mean vector and covariance matrix $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}'$. Note that

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

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

$$\mathbf{X}_{i,1} \mid \mathbf{X}_{i,2} \sim \mathcal{N}_m^T(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_{i,2} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}),\ \mathbf{X}_{i,1} \in G_1(\mathbf{X}_{i,2}),$$

with $G_1(\mathbf{X}_{i,2}) = \{\mathbf{Y}_i \in \mathbb{R}^m : \mathbf{l} - \mathbf{B}_2\mathbf{X}_{i,2} \leq \mathbf{B}_1\mathbf{Y}_i \leq \mathbf{u} - \mathbf{B}_2\mathbf{X}_{i,2}\}$.

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

$$f_{\mathbf{X}_{i,2}}(\mathbf{x}_{i,2}) = \frac{f_{\mathbf{X}_{i,1},\mathbf{X}_{i,2}}(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})}{f_{\mathbf{X}_{i,1}|\mathbf{X}_{i,2}}(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})}.$$

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

$$\begin{aligned} f_{\mathbf{X}_{i,2}}(\mathbf{x}_{i,2}) &= \frac{\psi(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})}{\int \cdots \int \psi(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1} d\mathbf{x}_{i,2}} \frac{\int \cdots \int \psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1}}{\psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})} \\ &= \frac{\psi(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})}{\psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})} \frac{\int \cdots \int \psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1}}{\int \cdots \int \psi(\mathbf{x}_{i,1}, \mathbf{x}_{i,2}) d\mathbf{x}_{i,1}\, d\mathbf{x}_{i,2}} \\ &= \frac{\psi(\mathbf{x}_{i,2})}{\int \cdots \int \psi(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1}\, d\mathbf{x}_{i,2}} \int \cdots \int \psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1}, \end{aligned}$$

where

$$\psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}_{11.2}|^{-\frac{1}{2}} \exp\big(-\frac{1}{2}(\mathbf{x}_{i,1} - \boldsymbol{\mu}_{1.2})' \boldsymbol{\Sigma}_{11.2}^{-1}(\mathbf{x}_{i,1} - \boldsymbol{\mu}_{1.2})\big),$$

with

$$\begin{aligned}\boldsymbol{\mu}_{1.2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_{i,2} - \boldsymbol{\mu}_2)\\ \boldsymbol{\Sigma}_{11.2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.\end{aligned}$$

So the term $\int \cdots \int \psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1}$ is still dependent on $\mathbf{X}_{i,2}$ after integration, which means that the marginal distribution of $\mathbf{X}_{i,2}$ is not truncated multivariate normal in general. Only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, the marginal distributions of $\mathbf{X}_{i,1}$ and $\mathbf{X}_{i,2}$ are truncated normal, as in this case it holds that

$$\int \cdots \int \psi(\mathbf{x}_{i,1} \mid \mathbf{x}_{i,2})\, d\mathbf{x}_{i,1} = \int \cdots \int \psi(\mathbf{x}_{i,1})\, d\mathbf{x}_{i,1} = 1.$$

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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ 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 $f(\cdot \mid \boldsymbol{\theta})$ is

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid \boldsymbol{\theta}) = \frac{\prod_{i=1}^{n} \psi(\mathbf{x}_i \mid \boldsymbol{\theta})}{(\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\theta})\, d\mathbf{x}_i)^n}, \quad \mathbf{x}_i \in G,$$

where $\psi(\cdot \mid \boldsymbol{\theta})$ is the nontruncated density. This results in the following loglikelihood

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

and therefore the loglikelihood for the truncated normal distribution is

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{x}) &= -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \\ &\quad -n \ln \int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\mathbf{x}_i.\end{aligned}$$

This loglikelihood is difficult to compute and maximise as it contains a multidimensional 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 Hajivassiliou 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 (Hajivassiliou 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 Hajivassiliou 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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$, since it is simpler to differentiate the loglikelihood with respect to the parameters of $\boldsymbol{\Sigma}^{-1}$, and note that the following holds true

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

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

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

Then the first order derivatives are

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \mathbf{x})}{\partial \boldsymbol{\mu}} &= \boldsymbol{\Gamma}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}) - n\frac{\int\cdots\int_G \frac{\partial \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})}{\partial \boldsymbol{\mu}} d\mathbf{x}_i}{\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \\ &= \boldsymbol{\Gamma}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}) - n\frac{\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\boldsymbol{\Gamma}(\mathbf{x}_i - \boldsymbol{\mu})\, d\mathbf{x}_i}{\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \\ &= \boldsymbol{\Gamma}\Big(\sum_{i=1}^{n}\mathbf{x}_i - n\frac{\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\mathbf{x}_i\, d\mathbf{x}_i}{\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i}\Big) \\ &= \boldsymbol{\Gamma}\Big(\sum_{i=1}^{n}\mathbf{x}_i - n\mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}]\Big) \end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \mathbf{x})}{\partial \boldsymbol{\Gamma}} &= \frac{n}{2}\boldsymbol{\Gamma}^{-1} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' - n\frac{\int \cdots \int_G \frac{\partial \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})}{\partial \boldsymbol{\Gamma}}\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \\
&= \frac{n}{2}\boldsymbol{\Gamma}^{-1} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' + \\
&\quad -n\frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\left(\frac{1}{2}\boldsymbol{\Gamma}^{-1} - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\right) d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \\
&= -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' + \\
&\quad +\frac{n}{2}\frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \\
&= -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' + \\
&\quad +\frac{n}{2}\mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}].
\end{aligned}$$

Recall that $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, 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

$$\mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}] = \frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\mathbf{x}_i\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i} \tag{5.1}$$

and

$$\begin{aligned}
&\mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}] \\
&\qquad = \frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})\, d\mathbf{x}_i}.
\end{aligned} \tag{5.2}$$

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 $\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \mathbf{x}_i \, d\mathbf{x}_i$, $\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' \, d\mathbf{x}_i$ and $\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \, d\mathbf{x}_i$. Note that the resulting estimate may be biased, as unbiased estimation of $\int \cdots \int \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \, d\mathbf{x}_i$ does not imply unbiased estimation of $1/\int \cdots \int \psi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Gamma}) \, d\mathbf{x}_i$. 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 Keane (1994). This simulator is convenient as it has been investigated thoroughly and its properties are well known.

If $\mathbf{X}_i \sim \mathcal{N}^T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$, then $\mathbf{B}\mathbf{X}_i \sim \mathcal{N}^T(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$ if we assume that $\mathbf{B}$ is nonsingular. Factorise $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'$ using the Cholesky decomposition into $\boldsymbol{\Omega}\boldsymbol{\Omega}'$, where $\boldsymbol{\Omega}$ is a lower-triangular matrix. This means that

$$\begin{aligned} \int \cdots \int_{\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}} \psi(\mathbf{x}_i) \, d\mathbf{x}_i &= \Pr(\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}), \quad \mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \Pr(\mathbf{l}^* \leq \boldsymbol{\Omega}\mathbf{Z}_i \leq \mathbf{u}^*) \quad \mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \end{aligned}$$

where $\mathbf{l}^* = \mathbf{l} - \mathbf{B}\boldsymbol{\mu}$ and $\mathbf{u}^* = \mathbf{u} - \mathbf{B}\boldsymbol{\mu}$. Next draw a sequence of Z_j, for $j = 1, \ldots, k$, from the truncated standard normal distribution

$$\begin{aligned} &\text{draw } Z_1 \quad \text{from} \quad \mathcal{N}(0,1) \text{ s.t. } l_1^* \leq \omega_{11} z_1 \leq u_1^* \\ &\qquad\qquad \vdots \\ &\text{draw } Z_k \quad \text{from} \quad \mathcal{N}(0,1) \text{ s.t. } l_k^* \leq \omega_{k1} z_1 + \cdots + \omega_{kk} z_k \leq u_k^*, \end{aligned}$$

where $\omega_{ij},\ i,j = 1,\ldots,k$, denotes the ijth element of $\mathbf{\Omega}$.

Then

$$\begin{aligned} Q_1 &= \Pr\Big(\frac{l_1^*}{\omega_{11}} \leq z_1 \leq \frac{u_1^*}{\omega_{11}}\Big) \\ &\vdots \\ Q_k(z_1,\ldots,z_{k-1}) &= \Pr\Big(\frac{l_1^* - \sum_{j=1}^{k-1}\omega_{kj}z_j}{\omega_{kk}} \leq z_k \leq \frac{u_1^* - \sum_{j=1}^{k-1}\omega_{kj}z_j}{\omega_{kk}}\Big). \end{aligned}$$

The GHK simulator then becomes

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

with R the number of replications. This simulation technique provides asymptotically unbiased simulators 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 $\mathbf{B}$ 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 $\mathbf{B}$ is nonsingular. When $\mathbf{B}$ is singular, $\mathbf{B\Sigma B'}$ will be positive semidefinite and consequently $\mathbf{\Omega}$ will be positive semidefinite as well, which means that some of the ω_{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 $g(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Gamma})$ denote the truncated normal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ $(= \boldsymbol{\Sigma}^{-1})$. If we draw $\mathbf{V}_i$ from the truncated normal distribution: $\mathbf{V}_i \sim \mathcal{N}^T(\boldsymbol{\mu}, \boldsymbol{\Gamma}^{-1})$, $\mathbf{V}_i \in G$, then

$$\begin{aligned} \hat{\mathrm{E}}[\mathbf{X}_i \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}] &= \frac{1}{R}\sum_{i=1}^{R} \mathbf{V}_i \\ \hat{\mathrm{E}}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}] &= \frac{1}{R}\sum_{i=1}^{R} (\mathbf{V}_i - \boldsymbol{\mu})(\mathbf{V}_i - \boldsymbol{\mu})'. \end{aligned}$$

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 outside the admissible region. For the truncated normal this algorithm will work well if $\boldsymbol{\mu}$ lies near the centre of the region and the region is sufficiently large. However, if $\boldsymbol{\mu}$ 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 convergence issues. Obtaining the same underlying random variates in the case of AR sampling can be achieved as follows. Let $\boldsymbol{\Phi}\boldsymbol{\Phi}'$ be the Cholesky decomposition

of $\boldsymbol{\Sigma}$. Then draws from $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be generated using $\mathbf{X}_i = \boldsymbol{\Phi}\mathbf{Z}_i + \boldsymbol{\mu}$, where $\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Afterwards the draws of $\mathbf{X}_i$ 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 $\boldsymbol{\Phi}$ 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

$$\begin{aligned}
\text{draw } X_1^{(t)} &\sim g(x_1 \mid x_2^{(t-1)}, \ldots, x_k^{(t-1)}, \boldsymbol{\theta}_1) \\
\text{draw } X_2^{(t)} &\sim g(x_2 \mid x_1^{(t)}, x_3^{(t-1)}, \ldots, x_k^{(t-1)}, \boldsymbol{\theta}_2) \\
&\vdots \\
\text{draw } X_k^{(t)} &\sim g(x_k \mid x_1^{(t)}, \ldots, x_{k-1}^{(t)}, \boldsymbol{\theta}_k),
\end{aligned}$$

where g denotes the univariate conditional truncated normal density and $\boldsymbol{\theta}_j = (\mu_{j.-j}, \Sigma_{jj.-j})$. The term $-j$ refers to the set of k variables in the model with the jth element removed. After t_G such iterations, we have a draw $\mathbf{X}_i^{(t_G)} = (X_{i,1}^{(t_G)}, \ldots, X_{i,k}^{(t_G)})$, 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 $Z = \Phi^{-1}((\Phi(u^*) - \Phi(l^*))U + \Phi(l^*)) \sim \mathcal{N}^T(0,1)$, with $l^* \leq Z \leq u^*$, where $l^* = 1/\sigma(l - \mu)$ and $u^* = 1/\sigma(u - \mu)$. Consequently, $Y = \sigma Z + \mu \sim \mathcal{N}^T(\mu, \sigma^2)$, 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

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

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 $\boldsymbol{\Sigma}$ is nonsingular. Let $\mathbf{t} = \mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}]$ and $\mathbf{T} = \mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mid \mathbf{X}_i \in G, \boldsymbol{\mu}, \boldsymbol{\Gamma}]$ 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 $\mathbf{t}$ and $\mathbf{T}$, 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.*

	True	AR			Gibbs		
		sample size ($\times$ 100)					
		1	10	100	1	10	100
$\hat{t}_1$	2.625	2.605	2.631	2.628	2.639	2.621	2.627
		(0.159)	(0.049)	(0.013)	(0.152)	(0.046)	(0.012)
$\hat{t}_2$	3.463	3.453	3.466	3.464	3.471	3.441	3.461
		(0.172)	(0.054)	(0.014)	(0.167)	(0.052)	(0.015)
$\hat{t}_3$	6.415	6.416	6.424	6.413	6.426	6.394	6.411
		(0.243)	(0.062)	(0.021)	(0.234)	(0.055)	(0.018)
$\hat{T}_{11}$	2.905	2.928	2.920	2.914	2.912	2.903	2.919
		(0.423)	(0.158)	(0.048)	(0.522)	(0.149)	(0.044)
$\hat{T}_{12}$	1.390	1.359	1.411	1.390	1.422	1.353	1.400
		(0.355)	(0.088)	(0.039)	(0.354)	(0.113)	(0.035)
$\hat{T}_{13}$	1.409	1.376	1.404	1.411	1.454	1.360	1.413
		(0.459)	(0.123)	(0.043)	(0.362)	(0.086)	(0.039)
$\hat{T}_{22}$	3.259	3.180	3.239	3.255	3.217	3.190	3.261
		(0.443)	(0.132)	(0.054)	(0.466)	(0.125)	(0.049)
$\hat{T}_{23}$	2.401	2.367	2.394	2.401	2.462	2.354	2.404
		(0.446)	(0.107)	(0.043)	(0.456)	(0.147)	(0.038)
$\hat{T}_{33}$	4.487	4.455	4.477	4.492	4.623	4.451	4.483
		(0.619)	(0.159)	(0.054)	(0.752)	(0.223)	(0.064)

using the Central Limit Theorem.

Since $\mathbf{X}_i$ is truncated from below its expected value will be larger than $\boldsymbol{\mu}$ and its expected variance will be smaller than $\boldsymbol{\Sigma}$, which we already observed for the univariate case in figure 5.1. These differences will increase as $\boldsymbol{\mu}$ 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-

timating the expected complete data loglikelihood $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ defined as

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= \mathrm{E}[\ell(\boldsymbol{\theta} \mid \mathbf{X}_{mis}, \mathbf{x}_{obs}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}] \\ &= \int \cdots \int_{G(\mathbf{x}_{obs})} \ell(\boldsymbol{\theta} \mid \mathbf{x}_{mis}, \mathbf{x}_{obs}) h(\mathbf{x}_{mis} \mid \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)}) d\mathbf{x}_{mis}. \end{aligned}$$

The M-step then consists of maximising $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ to obtain an update $\boldsymbol{\theta}^{(t+1)}$. If an analytic solution to $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ cannot be found, this quantity can be estimated unbiasedly by

$$\hat{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} \ell(\boldsymbol{\theta} \mid \mathbf{X}_{mis}^{(m)}, \mathbf{x}_{obs}),$$

where $\mathbf{X}_{mis}^{(m)}$, $m = 1, \ldots, M$ are draws from the conditional distribution of the missing data given the observed data and the current value of the parameter vector; $h(\cdot \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\theta}^{(t)})$. 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 $\boldsymbol{\mu}^{(t+1)}$ and $\boldsymbol{\Gamma}^{(t+1)}$ $(= (\boldsymbol{\Sigma}^{(t+1)})^{-1})$ for which $\nabla_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}) = \mathbf{0}$ and $\nabla_{\boldsymbol{\Gamma}} Q(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}) = \mathbf{0}$, until the parameter estimates converge. As the area G is independent of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$, this means that the updated $\boldsymbol{\mu}^{(t+1)}$ and $\boldsymbol{\Gamma}^{(t+1)}$ are the parameters for which

$$\begin{aligned} \nabla_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}) &= \boldsymbol{\Gamma} \sum_{i=1}^{n} \mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] + \\ &\quad -n\boldsymbol{\Gamma} \mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in G, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \\ \nabla_{\boldsymbol{\Gamma}} Q(\boldsymbol{\mu}, \boldsymbol{\Gamma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}) &= -\frac{n}{2} \mathrm{E}[\mathbf{S} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] + \\ &\quad + \frac{n}{2} \mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mid \mathbf{X}_i \in G, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \end{aligned}$$

are equal to zero. So, similar to the EM algorithm for nontruncated normal data, we need to calculate $\mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}]$ and $\mathrm{E}[\mathbf{X}_i \mathbf{X}_i' \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}]$. Note that for each record i

$$\begin{aligned} &\mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}]' \\ &\qquad = \mathrm{E}[(\mathbf{X}_{i,mis}', \ \mathbf{x}_{i,obs}') \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \\ &\qquad = (\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}]', \ \mathbf{x}_{i,obs}') \end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}[\mathbf{X}_i\mathbf{X}_i' \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \\
&\quad = \mathrm{E}\Big[\begin{pmatrix} \mathbf{X}_{i,mis} \\ \mathbf{x}_{i,obs} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{i,mis}' & \mathbf{x}_{i,obs}' \end{pmatrix} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)} \Big] \\
&\quad = \mathrm{E}\Big[\begin{pmatrix} \mathbf{X}_{i,mis}\mathbf{X}_{i,mis}' & \mathbf{X}_{i,mis}\mathbf{x}_{i,obs}' \\ \mathbf{x}_{i,obs}\mathbf{X}_{i,mis}' & \mathbf{x}_{i,obs}\mathbf{x}_{i,obs}' \end{pmatrix} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)} \Big].
\end{aligned}$$

In this case

$$\begin{aligned}
&\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \\
&\qquad = \int \cdots \int_{G(\mathbf{x}_{i,obs})} \mathbf{x}_{i,mis} h(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)})\, d\mathbf{x}_{i,mis}
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}[\mathbf{X}_{i,mis}\mathbf{X}_{i,mis}' \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] \\
&\qquad = \int \cdots \int_{G(\mathbf{x}_{i,obs})} \mathbf{x}_{i,mis}\mathbf{x}_{i,mis}' h(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)})\, d\mathbf{x}_{i,mis}.
\end{aligned}$$

These quantities can be estimated through Monte Carlo integration, that is

$$\hat{\mathrm{E}}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] = \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{i,mis}^{(m)} \tag{5.3}$$

$$\hat{\mathrm{E}}[\mathbf{X}_{i,mis}\mathbf{X}_{i,mis}' \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Gamma}^{(t)}] = \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{i,mis}^{(m)}(\mathbf{V}_{i,mis}^{(m)})', \tag{5.4}$$

for $\mathbf{V}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}^T(\boldsymbol{\mu}_{mis.obs}^{(t)}, \boldsymbol{\Sigma}_{mis,mis.obs}^{(t)})$ with $\mathbf{V}_{i,mis} \in G(\mathbf{x}_{i,obs})$ and

$$\begin{aligned}
\boldsymbol{\mu}_{mis.obs}^{(t)} &= \boldsymbol{\mu}_{mis}^{(t)} + \boldsymbol{\Sigma}_{mis,obs}^{(t)}(\boldsymbol{\Sigma}_{obs,obs}^{(t)})^{-}(\mathbf{x}_{i,obs} - \boldsymbol{\mu}_{obs}^{(t)}) \\
\boldsymbol{\Sigma}_{mis,mis.obs}^{(t)} &= \boldsymbol{\Sigma}_{mis,mis}^{(t)} - \boldsymbol{\Sigma}_{mis,obs}^{(t)}(\boldsymbol{\Sigma}_{obs,obs}^{(t)})^{-}\boldsymbol{\Sigma}_{obs,mis}^{(t)}.
\end{aligned}$$

Random draws for $\mathbf{V}_{i,mis}$ 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 $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ denote the maximum likelihood estimates obtained through the EM algorithm. The missing data items $\mathbf{X}_{i,mis}$ can now be imputed either stochastically by drawing from: $\hat{\mathbf{X}}_{i,imp} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}^T(\hat{\boldsymbol{\mu}}_{mis.obs}, \hat{\boldsymbol{\Sigma}}_{mis,mis.obs})$, $\hat{\mathbf{X}}_{i,imp} \in G(\mathbf{x}_{i,obs})$ or deterministically by retaining the estimated values for (5.3) obtained in the E-step of the EM algorithm. So $\hat{\mathbf{X}}_{i,imp} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{i,mis}^{(m)}$.

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

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

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, $H := \{\mathbf{X}_i \in \mathbb{R}^k : \mathbf{l} \leq \mathbf{R}\mathbf{X}_i \leq \mathbf{u}\}$. 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 $\mathbf{X}_i$ is distributed according to a truncated singular normal distribution, that is $\mathbf{X}_i \sim \mathcal{N}^T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$ and where $\boldsymbol{\Sigma}$ 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

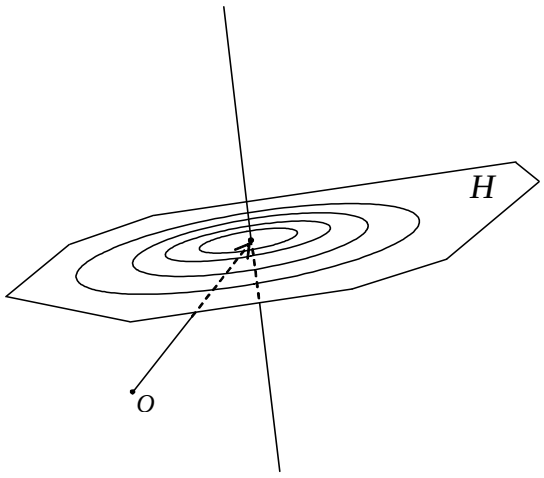

Figure 5.2: *Plot of the truncation region in which* $\mathbf{X}_i$ *lies, the curves represent contour lines of the normal density.*

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 $\boldsymbol{\Sigma}$ can be decomposed by means of an eigenvalue decomposition into $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$, where $\mathbf{C}$ is the orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_q, 0 \ldots, 0\}$ is the diagonal matrix of eigenvalues. Again let $\boldsymbol{\Lambda}_1 = \text{diag}\{\lambda_1, \ldots, \lambda_q\}$ 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 $\boldsymbol{\Sigma}$; $\boldsymbol{\Sigma}\mathbf{C}_2 = \mathbf{0}$.

Recall that the density function of the singular normal is

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

where $\prod_{j=1}^{q} \lambda_j = \det(\boldsymbol{\Lambda}_1)$ and $\boldsymbol{\Sigma}^{+} = \mathbf{C}_1 \boldsymbol{\Lambda}_1^{-1} \mathbf{C}_1'$. This means that the truncated singular normal density is defined as

$$g(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \begin{cases} 0 & \mathbf{x}_i \notin H \\ \frac{\varphi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int \cdots \int_H \varphi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\mathbf{x}_i} & \mathbf{x}_i \in H \end{cases} .$$

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

$$\mathbf{A}_c = \begin{pmatrix} \mathbf{A} \\ \mathbf{A}^* \end{pmatrix},$$

where $\mathbf{A}^*$ is a $q \times k$ matrix chosen such that $\mathbf{A}_c$ is nonsingular. Partition the restriction matrix $\mathbf{A}$ into $\mathbf{A} = (\mathbf{A}_p \ \mathbf{A}_q)$, where $\mathbf{A}_p$ is a $p \times p$ nonsingular matrix, which always exists as $\text{rank}(\mathbf{A}) = p$. Now partition $\mathbf{A}^*$ accordingly: $\mathbf{A}^* = (\mathbf{A}_p^* \ \mathbf{A}_q^*)$, where $\mathbf{A}_p^*$ is a $q \times p$ matrix. Taking $\mathbf{A}_p^* = \mathbf{0}$ and $\mathbf{A}_q^* = \mathbf{I}_q$ results in a nonsingular $\mathbf{A}^c$ as

$$\begin{aligned} \det(\mathbf{A}^c) &= \det \begin{pmatrix} \mathbf{A}_p & \mathbf{A}_q \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \\ &= \det(\mathbf{A}_p) \det(\mathbf{I}_q - \mathbf{0}\mathbf{A}_p^{-1}\mathbf{A}_q) \\ &= \det(\mathbf{A}_p) \neq 0. \end{aligned}$$

Now partition $\mathbf{B}$, $\mathbf{X}_i$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ 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

$$\mathbf{A}_c\mathbf{X}_i = \begin{pmatrix} \mathbf{A}\mathbf{X}_i \\ \mathbf{A}^*\mathbf{X}_i \end{pmatrix} \sim \mathcal{N}^T \left(\begin{pmatrix} \mathbf{A}\boldsymbol{\mu} \\ \mathbf{A}^*\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' & \mathbf{A}\boldsymbol{\Sigma}(\mathbf{A}^*)' \\ \mathbf{A}^*\boldsymbol{\Sigma}\mathbf{A}' & \mathbf{A}^*\boldsymbol{\Sigma}(\mathbf{A}^*)' \end{pmatrix} \right),$$

where $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$. As $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}\mathbf{A}' = \mathbf{0}$ (see chapter 4), it follows that

$$\begin{pmatrix} \mathbf{A}\mathbf{X}_i \\ \mathbf{A}^*\mathbf{X}_i \end{pmatrix} \sim \mathcal{N}^T \left(\begin{pmatrix} \mathbf{0} \\ \mathbf{A}^*\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^*\boldsymbol{\Sigma}(\mathbf{A}^*)' \end{pmatrix} \right), \text{ with } \mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}.$$

So $\mathbf{A}\mathbf{X}_i = \mathbf{0}$ with probability one and therefore $\mathbf{X}_{i,p} = -\mathbf{A}_p^{-1}\mathbf{A}_q\mathbf{X}_{i,q}$. Furthermore $\mathbf{A}^*\mathbf{X}_i \sim \mathcal{N}^T(\mathbf{A}^*\boldsymbol{\mu}, \mathbf{A}^*\boldsymbol{\Sigma}(\mathbf{A}^*)')$. Using the fact that $\mathbf{A}^* = (\mathbf{0} \ \mathbf{I}_q)$ we find that $\mathbf{X}_{i,q} \sim \mathcal{N}^T(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_{qq})$, with $\mathbf{l} \leq (\mathbf{B}_q - \mathbf{B}_p\mathbf{A}_p^{-1}\mathbf{A}_q)\mathbf{X}_{i,q} \leq \mathbf{u}$. This means that the maximum likelihood estimates for $\boldsymbol{\mu}_q$ and $\boldsymbol{\Gamma}_{qq}$ $(= \boldsymbol{\Sigma}_{qq}^{-1})$ can be obtained by

setting

$$\begin{aligned}\nabla_{\boldsymbol{\mu}_q}\ell(\boldsymbol{\mu}_q,\boldsymbol{\Gamma}_{qq}\mid \mathbf{x}_q) &= \boldsymbol{\Gamma}_{qq}(\sum_{i=1}^{n}\mathbf{x}_{i,q} - n\mathrm{E}[\mathbf{X}_{i,q}\mid \mathbf{X}_{i,q}\in H_q,\boldsymbol{\mu}_q,\boldsymbol{\Gamma}_{qq}])\\ \nabla_{\boldsymbol{\Gamma}_{qq}}\ell(\boldsymbol{\mu}_q,\boldsymbol{\Gamma}_{qq}\mid \mathbf{x}_q) &= -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_{i,q}-\boldsymbol{\mu}_q)(\mathbf{x}_{i,q}-\boldsymbol{\mu}_q)' + \\ &\quad +\frac{n}{2}\mathrm{E}[(\mathbf{X}_{i,q}-\boldsymbol{\mu}_q)(\mathbf{X}_{i,q}-\boldsymbol{\mu}_q)'\mid \mathbf{X}_{i,q}\in H_q,\boldsymbol{\mu}_q,\boldsymbol{\Gamma}_{qq}]\end{aligned}$$

equal to zero, where $H_q := \{\mathbf{Y}\in\mathbb{R}^q : \mathbf{l}\le(\mathbf{B}_q-\mathbf{B}_p\mathbf{A}_p^{-1}\mathbf{A}_q)\mathbf{Y}\le\mathbf{u}\}$. Once $\hat{\boldsymbol{\mu}}_q$ and $\hat{\boldsymbol{\Sigma}}_{qq}$ $(=\hat{\boldsymbol{\Gamma}}_{qq}^{-1})$ are found, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ can be derived by using the fact that $\mathbf{A}\hat{\boldsymbol{\mu}}=\mathbf{A}_p\hat{\boldsymbol{\mu}}_p+\mathbf{A}_q\hat{\boldsymbol{\mu}}_q=\mathbf{0}$ and

$$\hat{\boldsymbol{\Sigma}}\mathbf{A}' = \begin{pmatrix}\hat{\boldsymbol{\Sigma}}_{pp} & \hat{\boldsymbol{\Sigma}}_{pq}\\ \hat{\boldsymbol{\Sigma}}_{qp} & \hat{\boldsymbol{\Sigma}}_{qq}\end{pmatrix}\begin{pmatrix}\mathbf{A}_p'\\ \mathbf{A}_q'\end{pmatrix} = \mathbf{0}.$$

So

$$\begin{aligned}\hat{\boldsymbol{\mu}}_p &= -\mathbf{A}_p^{-1}\mathbf{A}_q\hat{\boldsymbol{\mu}}_q\\ \hat{\boldsymbol{\Sigma}}_{qp} &= -\hat{\boldsymbol{\Sigma}}_{qq}\mathbf{A}_q'(\mathbf{A}_p')^{-1}\\ \hat{\boldsymbol{\Sigma}}_{pq} &= -\mathbf{A}_p^{-1}\mathbf{A}_q\hat{\boldsymbol{\Sigma}}_{qq}\\ \hat{\boldsymbol{\Sigma}}_{pp} &= -\mathbf{A}_p^{-1}\mathbf{A}_q\hat{\boldsymbol{\Sigma}}_{qq}\mathbf{A}_q'(\mathbf{A}_p')^{-1}.\end{aligned}$$

This means that the estimates for $\boldsymbol{\mu}_p$, $\boldsymbol{\Sigma}_{qp}$, $\boldsymbol{\Sigma}_{pq}$ and $\boldsymbol{\Sigma}_{pp}$ 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, $\mathbf{X}_{i,p}$, may now be partly or completely observed and $\mathbf{X}_{i,q}$ may contain missing items. In that case the vector $\mathbf{X}_{i,p}$ contains information about $\mathbf{X}_{i,q}$, which is not present in $\mathbf{X}_{i,q}$ and consequently $\mathbf{X}_{i,p}$ is not completely redundant anymore. Clearly, we need to use this information as well for the maximum likelihood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This can be done as follows.

The E-step consists of calculating $Q(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$

$$\begin{aligned} &Q(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \\ &\quad = \mathrm{E}[\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{X}_{mis}, \mathbf{x}_{obs}) \mid \mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\quad = -\frac{nq}{2} \ln 2\pi - \frac{n}{2} \ln |\Lambda_1| - \frac{1}{2} \sum_{i=1}^{n} \mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{+} (\mathbf{X}_i - \boldsymbol{\mu})] + \\ &\qquad -n \ln \int \cdots \int_H \varphi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\mathbf{x}_i. \end{aligned} \tag{5.5}$$

As $\boldsymbol{\Sigma}^{+} = \mathbf{C}_1 \boldsymbol{\Lambda}_1^{-1} \mathbf{C}_1'$, this expected loglikelihood actually consists of three parameters; $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1$ and $\mathbf{C}_1$. So in order to obtain updated parameter values for $\boldsymbol{\mu}^{(t+1)}$ and $\boldsymbol{\Sigma}^{(t+1)}$, we need to differentiate (5.5) with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}_1^{-1}$ and $\mathbf{C}_1$ and set these first order conditions equal to zero. As the integration region H is independent of the parameters, these first order conditions become

$$\begin{aligned} &\nabla_{\boldsymbol{\mu}} Q(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1, \mathbf{C}_1 \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \mathbf{C}_1^{(t)}) \\ &\quad = \boldsymbol{\Sigma}^{+} \sum_{i=1}^{n} \mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] + \\ &\qquad -n \boldsymbol{\Sigma}^{+} \mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in H, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\nabla_{\boldsymbol{\Lambda}_1^{-1}} Q(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1, \mathbf{C}_1 \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \mathbf{C}_1^{(t)}) \\ &\quad = -\frac{1}{2} \sum_{i=1}^{n} \mathrm{E}[\mathbf{C}_1' (\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{C}_1 \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] + \\ &\qquad + \frac{n}{2} \mathrm{E}[\mathbf{C}_1' (\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{C}_1 \mid \mathbf{X}_i \in H, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\nabla_{\mathbf{C}_1} Q(\boldsymbol{\mu}, \boldsymbol{\Lambda}_1, \mathbf{C}_1 \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \mathbf{C}_1^{(t)}) \\ &\quad = -\sum_{i=1}^{n} \mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{C}_1 \boldsymbol{\Lambda}_1^{-1} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] + \\ &\qquad + n \mathrm{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})' \mathbf{C}_1 \boldsymbol{\Lambda}_1^{-1} \mid \mathbf{X}_i \in H, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}]. \end{aligned}$$

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

$$\begin{aligned} &\mathrm{E}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\qquad = \int \cdots \int_{H(\mathbf{x}_{i,obs})} \mathbf{x}_{i,mis} g(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \, d\mathbf{x}_{i,mis} \end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}[\mathbf{X}_{i,mis}\mathbf{X}'_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] \\ &\quad = \int \cdots \int_{H(\mathbf{x}_{i,obs})} \mathbf{x}_{i,mis}\mathbf{x}'_{i,mis} g(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})\, d\mathbf{x}_{i,mis},\end{aligned}$$

which can be done by generating draws from the truncated singular normal

$$\begin{aligned}\hat{\mathrm{E}}[\mathbf{X}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] &= \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{i,mis}^{(m)} \\ \hat{\mathrm{E}}[\mathbf{X}_{i,mis}\mathbf{X}'_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}] &= \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}_{i,mis}^{(m)}(\mathbf{V}_{i,mis}^{(m)})',\end{aligned}$$

for $\mathbf{V}_{i,mis} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}^T(\boldsymbol{\mu}^{(t)}_{mis.obs}, \boldsymbol{\Sigma}^{(t)}_{mis,mis.obs})$ with $\mathbf{V}_{i,mis} \in H(\mathbf{x}_{i,obs})$ and $\boldsymbol{\Sigma}_{mis,mis.obs}$ singular. Then $\mathbf{V}_{i,mis}$ can be obtained through either AR or Gibbs sampling.

AR sampling
In the case of AR sampling draws can be generated using the nontruncated singular normal distribution. If $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\mathbf{V}_{i,mis} = \mathbf{C}'\boldsymbol{\Lambda}^{1/2}\mathbf{Z} + \boldsymbol{\mu}_{mis.obs} \sim \mathcal{N}(\boldsymbol{\mu}_{mis.obs}, \boldsymbol{\Sigma}_{mis,mis.obs})$, where the singular $\boldsymbol{\Sigma}_{mis,mis.obs}$ is decomposed by means of an eigenvalue decomposition into $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$. Draws of $\mathbf{V}_{i,mis}$ are accepted if they fall in $H(\mathbf{x}_{i,obs})$ and rejected otherwise.

Gibbs sampling
Using the Gibbs sampler to generate draws is less straightforward, however. Recall that for singular data some variables can be derived with certainty from the other variables and the Gibbs sampler will therefore remain stuck in the initial values if all variables are taken into account 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 as was described in section 5.7.2.

First determine $p_i = \mathrm{rank}(\mathbf{A}_{i,mis})$ and remove any redundant rows from $\mathbf{A}_{i,mis}$ in order to obtain $\mathbf{A}^n_{i,mis}$, which is a $p_i \times m_i$ matrix, with m_i the number of missing items for that record. Partition $\mathbf{A}^n_{i,mis} = (\mathbf{A}^n_{i,mis,p_i}\ \mathbf{A}^n_{i,mis,q_i})$, where $\mathbf{A}^n_{i,mis,p_i}$ is nonsingular. Now partition $\mathbf{X}_{i,mis}$, $\boldsymbol{\mu}_{mis.obs}$, $\boldsymbol{\Sigma}_{mis,mis.obs}$ and the matrix with inequality restrictions on the missing items, $\mathbf{B}_{i,mis}$, accordingly. Then draw

$$\mathbf{V}_{i,mis,q_i} \sim \mathcal{N}^T(\boldsymbol{\mu}_{mis.obs,q_i}, \boldsymbol{\Sigma}_{mis,mis.obs,q_i}),$$

where

$$\mathbf{l}^* \leq (\mathbf{B}_{i,mis,q_i} - \mathbf{B}_{i,mis,p_i}(\mathbf{A}^n_{i,mis,p_i})^{-1}\mathbf{A}^n_{i,mis,q_i})\mathbf{V}_{i,mis,q_i} \leq \mathbf{u}^*,$$

with

$$\mathbf{l}^* = \mathbf{l} - \mathbf{B}_{i,obs}\mathbf{x}_{i,obs} + \mathbf{B}_{i,mis,p_i}(\mathbf{A}^n_{i,mis,p_i})^{-1}\mathbf{A}^n_{i,obs}\mathbf{x}_{i,obs}$$

and

$$\mathbf{u}^* = \mathbf{u} - \mathbf{B}_{i,obs}\mathbf{x}_{i,obs}\mathbf{B}_{i,mis,p_i}(\mathbf{A}^n_{i,mis,p_i})^{-1}\mathbf{A}^n_{i,obs}\mathbf{x}_{i,obs},$$

making use of the Gibbs sampler. Next calculate $\mathbf{V}_{i,mis,p_i}$ based on the generated sample and using the fact that $\mathbf{V}_{i,mis,p_i} = -(\mathbf{A}^n_{i,mis,p_i})^{-1}\mathbf{A}^n_{i,mis,q_i}\mathbf{V}_{i,mis,q_i}$.

Note that this method can be time-consuming as $\mathbf{A}^n_{i,mis,p_i}$ 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 $\boldsymbol{\Sigma}$ 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 $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ denote the maximum likelihood estimates. The missing data items $\mathbf{X}_{i,mis}$ can be imputed either stochastically by drawing from the truncated singular normal $\hat{\mathbf{X}}_{i,imp} \mid \mathbf{X}_{i,obs} = \mathbf{x}_{i,obs} \sim \mathcal{N}^T(\hat{\boldsymbol{\mu}}_{mis.obs}, \hat{\boldsymbol{\Sigma}}_{mis,mis.obs})$, $\hat{\mathbf{X}}_{i,imp} \in H(\mathbf{x}_{i,obs})$ or deterministically by retaining the estimated values that were obtained in the E-step of the EM algorithm. So $\hat{\mathbf{X}}_{i,imp} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{V}^{(m)}_{i,mis}$.

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 $1e05$. 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: $\mathbf{X} \sim \mathcal{N}^T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 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

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

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 $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. 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

Table 5.2: *Parameter estimates for* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$.

	μ_1	μ_2	Σ_{11}	Σ_{12}	Σ_{22}
True	50	100	1000	750	1500
Nontruncated	56	104	873	661	1350
Truncated					
Complete data	51	101	1109	843	1503
Missing data	53	102	978	739	1408
	(1.1)	(1.2)	(61.8)	(55.5)	(53.4)

that the estimates for μ_3, Σ_{13}, Σ_{23} and Σ_{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 Σ_{11} and Σ_{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 $\boldsymbol{\Sigma}$.

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 $\boldsymbol{\Sigma}$

The covariance matrix $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Phi}'$. It is clear that any $\boldsymbol{\Sigma}$ defined by $\boldsymbol{\Phi}$ is positive (semi)definite. There are several choices available for $\boldsymbol{\Phi}$ that lead to different parameterisations of $\boldsymbol{\Sigma}$. We propose using the Cholesky factorization, which defines $\boldsymbol{\Phi}$ as an lower triangular matrix. Note that, if $\boldsymbol{\Upsilon} = \boldsymbol{\Phi}^{-1}$

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

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

$$\ln|\boldsymbol{\Gamma}| = \ln|\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}| = \ln|\boldsymbol{\Upsilon}'||\boldsymbol{\Upsilon}| = \ln|\boldsymbol{\Upsilon}|^2 = 2\ln|\boldsymbol{\Upsilon}|$$

and

$$\frac{\partial|\boldsymbol{\Upsilon}|}{\partial\boldsymbol{\Upsilon}} = \frac{\partial \exp(\ln|\boldsymbol{\Upsilon}|)}{\partial\boldsymbol{\Upsilon}} = \exp(\ln|\boldsymbol{\Upsilon}|)\frac{\partial \ln|\boldsymbol{\Upsilon}|}{\partial\boldsymbol{\Upsilon}} = |\boldsymbol{\Upsilon}|(\boldsymbol{\Upsilon}')^{-1}.$$

Also note that

$$\frac{\partial \mathbf{y}'\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}\mathbf{y}}{\partial\boldsymbol{\Upsilon}} = 2\boldsymbol{\Upsilon}'\mathbf{y}\mathbf{y}'.$$

The loglikelihood can now be rewritten as

$$\begin{aligned}\ell(\boldsymbol{\mu},\boldsymbol{\Upsilon} \mid \mathbf{x}) &= -\frac{nk}{2}\ln 2\pi + n\ln|\boldsymbol{\Upsilon}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}(\mathbf{x}_i-\boldsymbol{\mu}) + \\ &\quad -n\ln\int\cdots\int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu},\boldsymbol{\Upsilon})\,d\mathbf{x}_i.\end{aligned}$$

From this it follows that the first order derivatives become

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}\ell(\boldsymbol{\mu},\boldsymbol{\Upsilon} \mid \mathbf{x}) &= \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}(\sum_{i=1}^{n}\mathbf{x}_i - n\mathrm{E}[\mathbf{X}_i \mid \mathbf{X}_i \in G,\boldsymbol{\mu},\boldsymbol{\Upsilon}]) \\ \nabla_{\boldsymbol{\Upsilon}}\ell(\boldsymbol{\mu},\boldsymbol{\Upsilon} \mid \mathbf{x}) &= -\boldsymbol{\Upsilon}'\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})' \\ &\quad +n\boldsymbol{\Upsilon}'\mathrm{E}[(\mathbf{X}_i-\boldsymbol{\mu})(\mathbf{X}_i-\boldsymbol{\mu})' \mid \mathbf{X}_i \in G,\boldsymbol{\mu},\boldsymbol{\Upsilon}].\end{aligned}$$

So instead of the unique elements of $\boldsymbol{\Gamma}$, the lower triangular elements of $\boldsymbol{\Upsilon}$ are estimated, from which $\boldsymbol{\Gamma}$ can be derived by $\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon}$.

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

$$\begin{aligned}\begin{pmatrix}\boldsymbol{\mu}^{(t+1)}\\ \mathrm{vech}(\boldsymbol{\Upsilon}^{(t+1)})\end{pmatrix} &= \begin{pmatrix}\boldsymbol{\mu}^{(t)}\\ \mathrm{vech}(\boldsymbol{\Upsilon}^{(t)})\end{pmatrix} + \\ &\quad +a\hat{\mathcal{I}}^{-1}(\boldsymbol{\mu},\boldsymbol{\Upsilon})|_{\boldsymbol{\mu}=\boldsymbol{\mu}^{(t)},\boldsymbol{\Upsilon}=\boldsymbol{\Upsilon}^{(t)}}\nabla\ell(\boldsymbol{\mu},\boldsymbol{\Upsilon} \mid \mathbf{x})|_{\boldsymbol{\mu}=\boldsymbol{\mu}^{(t)},\boldsymbol{\Upsilon}=\boldsymbol{\Upsilon}^{(t)}},\end{aligned}$$

where a denotes the step size, and $\hat{\mathcal{I}}(\boldsymbol{\mu}, \boldsymbol{\Upsilon}) = \sum_{i=1}^{n} \nabla \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x}) \nabla \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x})'$, with $\nabla \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x}) = (\nabla_{\boldsymbol{\mu}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x})', \text{vech}(\nabla_{\boldsymbol{\Upsilon}} \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x}))')'$ and $\nabla \ell(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x}) = \sum_{i=1}^{n} \nabla \ell_i(\boldsymbol{\mu}, \boldsymbol{\Upsilon} \mid \mathbf{x})$. 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: $\ell^{(t)} = \ell(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)} \mid \mathbf{x})$. Then the difference between loglikelihoods in subsequent iterations is

$$\Delta(\ell^{(t+1)}, \ell^{(t)}) = \ell_{nt}^{(t+1)} - \ell_{nt}^{(t)} - n \ln \frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})\, d\mathbf{x}_i},$$

where ℓ_{nt} denotes the loglikelihood function of the nontruncated normal distribution. Now define

$$R = \frac{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})\, d\mathbf{x}_i}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})\, d\mathbf{x}_i}.$$

The term $\ln R$ cannot be estimated without bias by $\ln \hat{R}$, as $\mathrm{E}[\ln R] \neq \ln \mathrm{E}[R]$. However, recall that Jensen's inequality states that $\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$ for a concave function g. This means that

$$\begin{aligned} \hat{\Delta}(\ell^{(t+1)}, \ell^{(t)}) &= \ell_{nt}^{(t+1)} - \ell_{nt}^{(t)} - n\mathrm{E}[\ln R] \\ &\geq \ell_{nt}^{(t+1)} - \ell_{nt}^{(t)} - n \ln \mathrm{E}[R]. \end{aligned}$$

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

$$\begin{aligned} R &= \int \cdots \int_G \frac{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})} \frac{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}\, d\mathbf{x}_i \\ &= \int \cdots \int_G \frac{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})}{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})} \frac{\psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}{\int \cdots \int_G \psi(\mathbf{x}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}\, d\mathbf{x}_i \\ &= \mathrm{E}\Big[\frac{\psi(\mathbf{X}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})}{\psi(\mathbf{X}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}\Big]. \end{aligned}$$

So if $\mathbf{V}_i$ represents draws from the truncated normal with parameters $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)} = ((\boldsymbol{\Upsilon}^{(t)})' \boldsymbol{\Upsilon}^{(t)})^{-1}$, then R can be estimated by

$$\hat{R} = \frac{1}{N} \sum_{i=1}^{N} \frac{\psi(\mathbf{V}_i \mid \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\Upsilon}^{(t+1)})}{\psi(\mathbf{V}_i \mid \boldsymbol{\mu}^{(t)}, \boldsymbol{\Upsilon}^{(t)})}.$$

An appropriate step size can now be easily found by choosing an initial step size, for example $a = 1$, and calculating the corresponding $\hat{\Delta}$. If $\hat{\Delta}$ is negative the step size needs to be reduced, for example by taking $a := a/2$ until $\hat{\Delta}$ is non-negative.

Chapter 6

Imputation of Data Using a Sequential Regression Approach

In the chapters 3, 4 and 5 we discussed imputation of data subject to different types of linear restrictions. Probability distributions are chosen such that the data can be modelled while incorporating both linear balance and inequality restrictions. The complexity of the probability models increases with the complexity of the structure of the restrictions that can be handled. The truncated singular normal appears to be the only model that can take all the linear restrictions that may occur into account. For economic data a disadvantage of this model is the dependence on the normal distribution without being able to make use of nonlinear transformations to approximate normality. In this chapter an imputation method is developed that uses the univariate full conditional distributions to model the variables separately, so that there is no need to specify a joint model.

6.1 Introduction

In the previous chapters we have developed several imputation techniques that model the distribution of the missing data conditional on the observed data, taking linear restrictions into account making use of the Dirichlet, the multivariate

singular normal or the multivariate truncated singular normal distribution. The differences between these methods are the complexity of the edit structure that can be handled and consequently the complexity of the imputation model. A property shared by all these methods is that they are multivariate and parametric. That is, the multivariate distributions of the variables are assumed to belong to a known family of probability distributions. Survey data, however, usually consist of a large number of variables, which may have several distributional forms. This means that, although ideally imputations should be draws from the conditional distribution of the missing values given the observed data, it may be difficult to find an appropriate multivariate model. In our case the search for this joint model is complicated by the fact that imputed values need to satisfy certain linear restrictions. In chapter 5 we used a truncated singular normal distribution to model data subject to restrictions. Although this method is theoretically sound, the distributional assumptions that need to be taken can be unrealistic in practice. When abandoning the dependence on the normal distribution we have, however, not been able to find an appropriate joint model that can take the linear restrictions into account.

In this chapter we will investigate an imputation method that makes use of univariate conditional distributions. The variables are modelled and imputed univariately using a sequence of regressions. This process is iterated so that the final imputed values converge to draws from the multivariate model.

First of all, in section 6.2 the different linear restrictions that may occur are discussed. Next in section 6.3 full conditional densities and problems of incompatibility are treated. The suggested imputation method is dealt with in section 6.4 and section 6.5 discusses the regression models used. Then section 6.6 deals with Box-Cox transformations to approximate normality and in section 6.7 it is explained how balance restrictions can be incorporated. Next, in section 6.8 some results for empirical data are presented and, finally, in section 6.9 some concluding remarks will be made.

6.2 Linear inequality and balance restrictions

Let $\mathbf{X}$ be an $n \times k$ data matrix and let $\mathbf{X}_i$ denote the data vector of order $k \times 1$ for respondent i. The inequality restrictions are represented by $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$, where the matrix $\mathbf{B}$ is an $r \times k$ matrix containing r inequality restrictions on the data. The upper and lower bounds $\mathbf{u}$ and $\mathbf{l}$ may equal plus or minus infinity, which means that the variables are truncated from one side only, e.g. by non-negativity constraints.

Additionally, consider a $p \times k$ restriction matrix $\mathbf{A}$, with p the number of linear balance restrictions, where it holds that $\mathbf{AX}_i = \mathbf{0}$. We assume that there are no redundant balance restrictions, so $\mathbf{A}$ is of full row rank.

We need to model the data $\mathbf{X}_i$, $i = 1, \ldots, n$, which are subject to $\mathbf{l} \leq \mathbf{BX}_i \leq \mathbf{u}$ and $\mathbf{AX}_i = \mathbf{0}$. Due to the added complexity caused by balance restrictions, we will first consider data subject to inequality restrictions only. This model will then be extended in section 6.7 to deal with balance restrictions as well.

6.3 Full conditional distributions

As opposed to specifying a joint model, another approach is to model the data through fully conditionally specified distributions. Let $\mathbf{X}_j$ denote the jth variable and partition this into a missing, $\mathbf{X}_j^{mis}$, and an observed part, $\mathbf{X}_j^{obs}$. Note that this is different from $\mathbf{X}_{i,mis}$ and $\mathbf{X}_{i,obs}$ that was used in previous chapters, as missingness is now defined within variables and not within records. For notational convenience we assume that each variable j contains at least one missing item value. Furthermore, let $\mathbf{X}_{-j} = (\mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_k)$. Now instead of an explicit joint model $f(\mathbf{x}_i \mid \boldsymbol{\xi})$, the data are specified by separate conditional densities for each variable $\mathbf{X}_j$ conditioned on the matrix $\mathbf{X}_{-j}$ and the parameter vector $\boldsymbol{\xi}$, $f_j(\mathbf{x}_j \mid \mathbf{X}_{-j}, \boldsymbol{\xi})$, which constitute a joint model. The idea is based on MCMC techniques, where Markov chains are used to generate draws from full conditionals in order to sample from multivariate, intractable probability distributions. Under certain conditions (see chapter 2) and if the Markov chain is long enough the draws stabilise to a stationary distribution, which is the distribution of interest.

In missing data situations the Bayesian point of view considers both parameters and missing data as random variables. Information about these unknown quantities is expressed in the form of a posterior distribution. MCMC methods are often applied for exploring these posterior distributions. Consider the posterior density $p(\mathbf{X}^{mis}, \boldsymbol{\xi} \mid \mathbf{X}^{obs})$. This joint posterior can be simulated iteratively using univariate full conditional distributions, for example in combination with a Gibbs sampler. This would result in a sampler that cycles between draws for $\xi_j^{(t)}$, from

$$g_j^*(\xi_j \mid \xi_1^{(t)}, \mathbf{x}_1^{(t)}, \ldots, \xi_{j-1}^{(t)}, \mathbf{x}_{j-1}^{(t)}, \mathbf{x}_j^{(t-1)}, \xi_{j+1}^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}, \ldots, \xi_k^{(t-1)}, \mathbf{x}_k^{(t-1)})$$

and draws for $\mathbf{X}_j^{mis,(t)}$ from

$$f_j^*(\mathbf{x}_j \mid \xi_1^{(t)}, \mathbf{x}_1^{(t)}, \ldots, \xi_{j-1}^{(t)}, \mathbf{x}_{j-1}^{(t)}, \xi_j^{(t)}, \xi_{j+1}^{(t-1)}, \mathbf{x}_{j+1}^{(t-1)}, \ldots, \xi_k^{(t-1)}, \mathbf{x}_k^{(t-1)})$$

at each iteration t. This will constitute a Markov chain that converges to $p(\mathbf{X}^{mis}, \boldsymbol{\xi} \mid \mathbf{X}^{obs})$. So, imputations for $\mathbf{X}^{mis}$ can be obtained by using draws from this posterior distribution once the Markov chain has converged.

A major advantage of using these univariate full conditionals is that this approach is extremely flexible. Each variable can be modelled separately, which means that variable specific distributional properties can be taken into account. The downside of this is that it can be quite time-consuming to model each variable. Besides more structure is introduced by the univariate models, which means that they are more sensitive to model misspecification.

6.3.1 Incompatibility

Using full conditional densities without a specified joint model, unfortunately, has some drawbacks. In this instance the conditional models are specified directly, without choosing an encompassing multivariate model for the entire dataset. Due to this flexibility in the conditional distributions, there may be no implicit joint distribution underlying the imputation model. If the conditional distributions are incompatible and therefore do not constitute a joint model, there is no multivariate distribution to which the Markov chain and thus the imputed values converge.

Arnold and Press (1989) discuss necessary and sufficient conditions for the existence and uniqueness of a joint density, which cannot be easily used however, especially in multi-dimensional situations. For the bivariate case they derived that the conditional distributions $p_{Y|X}(y \mid x)$ and $p_{X|Y}(x \mid y)$ are said to be compatible if functions $u(x)$ and $v(y)$ exist, such that

$$\frac{p_{X|Y}(x \mid y)}{p_{Y|X}(y \mid x)} = u(x)v(y),$$

where $\int u(x)\,dx < \infty$.

The consequences of incompatibility are unclear and need to be investigated. There has been some research in this area, e.g. van Buuren et al. (2006) and Heeringa et al. (2002), and despite the lack of theory incompatibility does not appear to be a problem in practice. Gelman and Raghunathan (2001) state that if multivariate normality or other distributional assumptions are doubtful, it may make more sense to use separate regressions instead of a joint model. One argument is that having a joint distribution is less important than incorporating information from other variables and variable specific properties, such as variable type, semi-continuity, bounds and so on. But, although this approach seems

computationally attractive, caution needs to be applied and more research needs to be done in this area.

Another issue that arises due to incompatibility is the fact that the order of conditioning may play a role. If the univariate conditionals constitute a multivariate distribution the draws will converge to that stationary distribution irrespective of the order of conditioning. If the univariate conditionals are incompatible the distribution to which the draws converge changes with the different orders of conditioning. Note that this property can lead to an implicit test of (in)compatibility. For this purpose, generate a set of draws with different orders of conditioning and if the generated samples appear to be from different multivariate distributions this can be a sign of incompatibility.

Even in the case of compatibility some orders can still be better than others as they are more efficient. We suggest using the most efficient ordering when we are dealing with possible incompatibility. This means that the variables should be conditioned in order of increasing missingness. The missing variables will then be conditioned on as much observed variables as possible.

6.4 Sequential regression multivariate imputation

Based on the idea of using conditionally specified distributions, Van Buuren, Boshuizen and Knook (1999) and Raghunathan et al. (2001) have developed a general purpose imputation procedure for the imputation of multivariate missing data. In short, the algorithm starts with imputing an initial simple guess for the missing data items and next each variable is sequentially modelled and imputed by a regression model conditional on all the other variables. This regression model is specified for each variable separately and can therefore depend on the type of variable that is imputed and the appropriate restrictions that must hold. This sequence is repeated across cycles so that the final imputed values converge to draws from the multivariate distribution. In this way a multivariate problem is split into a series of univariate problems. This approach is referred to as sequential regressions, chained equations or regression switching and has been implemented in IVEware (Raghunathan et al., 2002) and MICE (Van Buuren and Oudshoorn, 1999).

In this chapter we will use a similar approach to impute economic data that are subject to multiple linear inequality and balance restrictions. The following imputation procedure is suggested, which is similar to the sequential regression

method, employed by Raghunathan et al. (2001) and Van Buuren, Boshuizen and Knook (1999).

$$
\begin{array}{lll}
\text{obtain} & \boldsymbol{\zeta}_1^{(t)} & \text{by regressing } \mathbf{X}_1^{(t-1)} \text{ on } \mathbf{X}_2^{(t-1)}, \ldots, \mathbf{X}_k^{(t-1)} \\
\text{draw} & \mathbf{X}_1^{mis,(t)} & \text{from } f_1(\mathbf{x}_1 \mid \mathbf{x}_2^{(t-1)}, \ldots, \mathbf{x}_k^{(t-1)}, \boldsymbol{\zeta}_1^{(t)}) \\
\text{obtain} & \boldsymbol{\zeta}_2^{(t)} & \text{by regressing } \mathbf{X}_2^{(t-1)} \text{ on } \mathbf{X}_1^{(t)}, \mathbf{X}_3^{(t-1)}, \ldots, \mathbf{X}_k^{(t-1)} \\
\text{draw} & \mathbf{X}_2^{mis,(t)} & \text{from } f_2(\mathbf{x}_2 \mid \mathbf{x}_1^{(t)}, \mathbf{x}_3^{(t-1)}, \ldots, \mathbf{x}_k^{(t-1)}, \boldsymbol{\zeta}_2^{(t)}) \\
 & \vdots & \\
\text{obtain} & \boldsymbol{\zeta}_k^{(t)} & \text{by regressing } \mathbf{X}_k^{(t-1)} \text{ on } \mathbf{X}_1^{(t)}, \ldots, \mathbf{X}_{k-1}^{(t)} \\
\text{draw} & \mathbf{X}_k^{mis,(t)} & \text{from } f_k(\mathbf{x}_k \mid \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_{k-1}^{(t)}, \boldsymbol{\zeta}_k^{(t)}).
\end{array}
$$

This process is iterated for a sufficiently long time, such that the algorithm converges and the parameter estimates stabilise. Note that compatibility is not guaranteed in this case.

In order to initialise this algorithm a complete dataset, without missing values, is needed. So the first step of the algorithm is to impute the missing items with a simple imputation method. Several imputation models can be used for this purpose, but clearly the algorithm will converge faster if 'good' imputations are generated. Furthermore, note that it is crucial that these imputed values satisfy all linear restrictions on the data in order to avoid inconsistencies. This is not straightforward, we therefore suggest to use the EM algorithm for singular normal data, that was discussed in chapter 4, to obtain imputations after which the imputed values are adjusted with a simple procedure such that the inequality restrictions are satisfied. Although the EM algorithm converges relatively slowly and therefore will result in more computing time, using the EM algorithm seems a good approach as it is likely to generate imputations that are close to the final draws and the imputed values will immediately satisfy the linear balance restrictions on the data. Furthermore, we expect that adjusting the variables such that all inequality restrictions are satisfied can be done rather straightforwardly.

In contrary to the original sequential imputation procedure $\boldsymbol{\zeta}$ will not be drawn, but estimated. This approach ignores the uncertainty in estimating $\boldsymbol{\zeta}$ and is therefore deemed improper by Rubin (1987). It does, however, lead to acceptable approximations, when the fraction of missing data is modest (Little and Raghunathan, 1997) and when the number of observations for estimating $\boldsymbol{\zeta}$ is large.

6.5 Regression models

A major advantage of using univariate conditional distributions is the fact that variable-specific properties, such as semi-continuity, non-negativity, or linear restrictions, can be taken into account as variables are modelled separately.

Semi-continuity refers to variables that take on a single discrete value (e.g. zero) with positive probability, but are continuously distributed otherwise. In general in business surveys conducted by Statistics Netherlands variables are encountered that are either continuous or semi-continuous. Examples of semi-continuous variables are costs of hired personnel, costs of research and development and advertising costs.

In addition to this, most economic quantities are non-negative, such as for example expenses, revenues and the number of employees. Variables that can take on negative values are profit, financial benefits or provisions. It is important to realise that semi-continuity does not imply non-negativity. An example of this is other income, which is semi-continuous but not non-negative.

These different types of variables can also be subject to all sorts of linear inequality restrictions other than non-negativity constraints. For instance, the number of employees in fte (full-time equivalent) may not exceed the total number of employees, salary payments in thousands of Euros must be larger than the number of employees, and so on. In this case the domain of the variables is truncated to a certain region.

An added advantage of univariate modelling is the fact that nonlinear transformations can be employed. This is impossible for multivariate data as the edit structure will be lost after applying nonlinear transformations. Note that $X_1 + \cdots + X_k \leq X_{k+1}$ does not imply $T(X_1) + \cdots + T(X_k) \leq T(X_{k+1})$ for a nonlinear transformation $T(\cdot)$. In the univariate approach this will not be a problem, as the edit restrictions can be determined for each variable separately and transformed subsequently. For example, if we want to impute X_1 and model this variable using a nonlinear transformation, the resulting edit restriction will be $T(X_1) \leq T(\tilde{X})$, with $\tilde{X} = X_{k+1} - X_k - \ldots - X_2$. A consequence of this is that the variables can be transformed to approximate normality using Box-Cox transformations, which will be discussed in more detail in section 6.6. Once these variables are (approximately) normally distributed, the following regressions models can be used.

- *Normal linear regression model*
  The classical linear regression model will be used to handle continuous, unbounded variables.

- *Truncated normal regression model*
  The truncated regression model will be used for continuous variables that are subject to restrictions and therefore lie between certain bounds.

- *Logistic regression model*
  Semi-continuous variables can be dealt with in different ways. First of all, censored regression (or Tobit) models can be used for direct estimation of the parameters. Alternatively, a two stage procedure can be applied using a logit or a probit models, where the zero/non-zero status of the variable is determined in the first stage and the imputed value is obtained in the second stage using a (truncated) normal regression model for the non-zero records.
  All semi-continuous variables in business surveys at Statistics Netherlands have a probability mass at zero. Some of these variables are non-negative and need not satisfy any other inequality restrictions, which means that they can be straightforwardly modelled using a censored regression model. Some of the semi-continuous variables are unbounded, however, and can consequently take on both positive as well as negative values in which case censored regression models are an illogical choice. Besides, if the semi-continuous variables are bounded from above the model becomes complex as we need a combination of censored and truncated regression models then. As the two stage procedure can be applied for all instances, we prefer using this method.
  This means that a choice has to be made between logit and probit models. It has been demonstrated (e.g. Chambers and Cox, 1967) that logit and probit modelling often produce similar results. Therefore we suggest using the logit model in this instance as it was used by Raghunathan et al. (2001) as well.

Note that this imputation method is well-suited to deal with other types of variables, such as categorical or count variables and other properties such as skip patterns in the survey. As we are focusing on business surveys conducted by Statistics Netherlands, where these instances do not occur, we will not discuss these models at present. It is important to realise, however, that this model can be straightforwardly extended to handle social surveys. For a discussion of regression models for these variables we refer to Raghunathan et al. (2001). An overview of variable types, that appear in the business surveys of interest, and the accompanying suggested regression models is given in table 6.1.

Table 6.1: *An overview of variable specifics and the appropriate regression models.*

Variable type	Restrictions	Regression model
Continuous	Unbounded (e.g. financial result)	Normal
Continuous	Bounded (e.g. turnover)	Truncated normal
Semi-continuous	Unbounded (e.g. other income)	Logistic and normal
Semi-continuous	Bounded (e.g. advertising costs)	Logistic and truncated normal

6.5.1 Classical linear regression model

The parameters that model $\mathbf{X}_j$ in this case are $\boldsymbol{\beta}$ and σ^2. So $\boldsymbol{\zeta}$ is defined as $\boldsymbol{\zeta} = (\boldsymbol{\beta}', \sigma)'$. Let $\mathbf{X}_j$ denote the variable that will be re-imputed and let $\mathbf{X}_{-j}$ denote the most recently updated matrix of predictors. Then

$$\mathbf{X}_j = \mathbf{X}_{-j}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(0, \sigma_j^2 \mathbf{I}_n).$$

The parameters are estimated by ordinary least squares

$$\begin{aligned} \hat{\boldsymbol{\beta}}_j &= (\mathbf{X}'_{-j}\mathbf{X}_{-j})^{-1}\mathbf{X}'_{-j}\mathbf{X}_j \\ \hat{\sigma}_j^2 &= \frac{1}{n}(\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_j)'(\mathbf{X}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_j). \end{aligned}$$

Imputations are generated by drawing from the normal density using the estimated parameters. Let m_j be the number of missing item values in variable j and let $\mathbf{X}_{-j}^{mis}$ denote the matrix of predictors corresponding to the records that have missing values in variable j. Then

$$\mathbf{X}_j^{mis} \sim \mathcal{N}(\mathbf{X}_{-j}^{mis}\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2 \mathbf{I}_{m_j}).$$

6.5.2 Truncated regression model

In the presence of linear restrictions we cannot use the classical linear regression model, as $\mathrm{E}[\boldsymbol{\varepsilon}_j \mid \mathbf{X}_{-j}] \neq 0$ in this instance. The variable X_{ij} conditional on $\mathbf{X}_{i,-j}$

is now assumed to be normally distributed truncated to a region: $\mathbf{l}-\mathbf{B}_{-j}\mathbf{X}_{i,-j} \leq \mathbf{B}_j X_{ij} \leq \mathbf{u}-\mathbf{B}_{-j}\mathbf{X}_{i,-j}$. This means that: $l_i^* \leq X_{ij} \leq u_i^*$, where

$$\begin{aligned} l_i^* &= \max_{k=1,\ldots,q} \frac{1}{b_{kj}}(\mathbf{l}-\mathbf{B}'_{-j}\mathbf{X}_{i,-j}), \quad \text{for } b_{kj} \neq 0 \\ u_i^* &= \min_{k=1,\ldots,q} \frac{1}{b_{kj}}(\mathbf{u}-\mathbf{B}'_{-j}\mathbf{X}_{i,-j}), \text{ for } b_{kj} \neq 0, \end{aligned}$$

where b_{kj} is the kjth element of $\mathbf{B}$. The variable X_{ij} is truncated normal: $X_{ij} \sim \mathcal{N}^T(\mathbf{X}'_{i,-j}\boldsymbol{\beta}_j, \sigma_j^2)$, with $l_i^* \leq X_{ij} \leq u_i^*$. The parameters of this truncated normal regression model are obtained by maximum likelihood estimation, see Amemiya (1973). The density of X_{ij} is

$$f_{X_{ij}}(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2) = \frac{\psi(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2)}{\int_{l_i^*}^{u_i^*} \psi(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2)\, dx_{ij}},$$

where $\psi(\cdot)$ is the univariate normal density. Note that

$$\psi(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2) = \frac{1}{\sigma_j}\phi\Big(\frac{x_{ij}-\mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j}\Big)$$

and

$$\begin{aligned} \int_{l_i^*}^{u_i^*} \psi(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2)\, dx_{ij} &= \int_{l_i^*}^{u_i^*} \frac{1}{\sigma_j}\phi\Big(\frac{x_{ij}-\mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j}\Big)\, dx_{ij} \\ &= \int_{l_i^{**}}^{u_i^{**}} \frac{1}{\sigma_j}\sigma_j\phi(v)\, dv \\ &= \Phi(u_i^{**}) - \Phi(l_i^{**}), \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ respectively are the density and cumulative density function of the standard normal distribution. The upper and lower bounds are $u_i^{**} = \frac{1}{\sigma_j}(u_i^* - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)$ and $l_i^{**} = \frac{1}{\sigma_j}(l_i^* - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)$. This means that the truncated density function for X_{ij} is

$$f_{X_{ij}}(x_{ij} \mid \boldsymbol{\beta}_j, \sigma_j^2) = \frac{\frac{1}{\sigma_j}\phi\Big(\frac{1}{\sigma_j}(x_{ij}-\mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)\Big)}{\Phi(u_i^{**}) - \Phi(l_i^{**})}.$$

So the loglikelihood becomes

$$\ell(\boldsymbol{\beta}_j, \sigma_j \mid \mathbf{x}_j) = -n \ln \sigma_j + \sum_{i=1}^{n} \ln \phi\Big(\frac{x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j}\Big) - n \ln \big(\Phi(u_i^{**}) - \Phi(l_i^{**})\big).$$

In order to find the maximum likelihood estimates we need to differentiate the loglikelihood with respect to the parameters $\boldsymbol{\beta}_j$ and σ_j:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}_j, \sigma_j \mid \mathbf{x}_j)}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n} \Big(\frac{x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j}{\sigma_j^2} + \frac{1}{\sigma_j} \frac{\phi(u_i^{**}) - \phi(l_i^{**})}{\Phi(u_i^{**}) - \Phi(l_i^{**})}\Big)\mathbf{x}_{i,-j} \\
\frac{\partial \ell(\boldsymbol{\beta}_j, \sigma_j \mid \mathbf{x}_j)}{\partial \sigma_j} &= \frac{-n}{\sigma_j} + \sum_{i=1}^{n} \frac{(x_{ij} - \mathbf{x}'_{i,-j}\boldsymbol{\beta}_j)^2}{\sigma_j^3} + \frac{1}{\sigma_j} \frac{u_i^{**}\phi(u_i^{**}) - l_i^{**}\phi(l_i^{**})}{\Phi(u_i^{**}) - \Phi(l_i^{**})}.
\end{aligned}$$

As these conditions cannot be solved analytically, we need to resort to iterative techniques, such as Fisher scoring. Recall that $\boldsymbol{\zeta}'_j = (\boldsymbol{\beta}'_j, \sigma_j)$, then the maximum likelihood estimates are found by iterating, until convergence, over

$$\boldsymbol{\zeta}_j^{(t+1)} = \boldsymbol{\zeta}_j^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\zeta}_j)|_{\boldsymbol{\zeta_j}=\boldsymbol{\zeta_j}^{(t)}} \nabla_{\boldsymbol{\zeta_j}} \ell(\boldsymbol{\zeta}_j \mid \mathbf{x}_j)|_{\boldsymbol{\zeta_j}=\boldsymbol{\zeta_j}^{(t)}},$$

where $\nabla_{\boldsymbol{\zeta_j}} \ell_i(\boldsymbol{\zeta}_j \mid \mathbf{x}_j)' = (\nabla_{\boldsymbol{\beta}_j} \ell_i(\boldsymbol{\beta}_j, \sigma_j \mid \mathbf{x}_j)', \nabla_{\sigma_j} \ell_i(\boldsymbol{\beta}_j, \sigma_j \mid \mathbf{x}_j))$, $\nabla_{\boldsymbol{\zeta}_j} \ell(\boldsymbol{\zeta}_j \mid \mathbf{x}_j) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j \mid \mathbf{x}_j)$, and $\mathbf{I}(\boldsymbol{\zeta}_j \mid \mathbf{x}_j) = \sum_{i=1}^{n} \nabla_{\zeta_j} \ell_i(\boldsymbol{\zeta}_j) \nabla_{\boldsymbol{\zeta}_j} \ell_i(\boldsymbol{\zeta}_j \mid \mathbf{x}_j)'$.

Imputations for X_{ij}^{mis} can then be obtained by drawing from the truncated normal density:

$$X_{ij}^{mis} \sim \mathcal{N}^T(\mathbf{X}_{i,-j}\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2),\ l_i^* \leq X_{ij}^{mis} \leq u_i^*.$$

6.5.3 Logistic regression model

For semi-continuous data, we first need to establish whether the missing item value is zero or not. Let Y_{ij} indicate the zero/non-zero status of variable X_{ij}, where 1 represents a non-zero status. Sometimes the value of Y_{ij} can be derived with certainty using information from the edit restrictions and other variables in the survey. For instance, let X_{ij} denote the costs of hired personnel. If no personnel is hired this variable equals zero, otherwise it is non-zero. This means that if this variable is missing but the number of hired personnel is observed, the zero/non-zero status can be derived with certainty. This is a form of deductive imputation and will be applied first.

For the missing values of Y_{ij} that cannot be derived in this manner we will

use a logistic regression model. Let $p_{ij} = \Pr(Y_{ij} = 1 \mid \mathbf{x}_{i,-j})$, note that this equals $\mathrm{E}[Y_{ij} \mid \mathbf{x}_{i,-j}]$. The logit model is then

$$\begin{aligned} \operatorname{logit} p_{ij} &= \boldsymbol{\delta}_j' \mathbf{x}_{i,-j} \\ \ln \frac{p_{ij}}{1-p_{ij}} &= \boldsymbol{\delta}_j' \mathbf{x}_{i,-j}. \end{aligned}$$

The logit function is the inverse of the cumulative logistic distribution function, which is $\Lambda(z) = e^z/(1+e^z)$.

The coefficients of this regression model are estimated by means of maximum likelihood, where observations are treated as draws from the Bernoulli distribution. The first order conditions are

$$\frac{\partial \ell(\boldsymbol{\delta}_j \mid \mathbf{y}_j)}{\partial \boldsymbol{\delta}} = \mathbf{X}_{-j}'(\mathbf{y}_j - \boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta}_j)),$$

where $\boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta}_j)$ is a vector of order $n \times 1$ with elements $\Lambda(\boldsymbol{\delta}_j' \mathbf{x}_{i,-j})$, $i = 1, \ldots, n$. The parameter estimates can be found using Newton-Raphson, as the Hessian can be straightforwardly derived:

$$\frac{\partial^2 \ell(\boldsymbol{\delta}_j \mid \mathbf{y}_j)}{\partial \boldsymbol{\delta}_j \partial \boldsymbol{\delta}_j'} = -\mathbf{X}_{-j}' \mathbf{W} \mathbf{X}_{-j},$$

where $\mathbf{W} = \operatorname{diag}\{\boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta}_j)(\boldsymbol{\iota}_n - \boldsymbol{\Lambda}(\mathbf{X}_{-j}\boldsymbol{\delta}_j))'\}$. Note that since Y_{ij} is not present in the Hessian, the Newton-Raphson algorithm is identical to Fisher scoring. Furthermore, also note that the Hessian is negative definite, which means that the loglikelihood has a global maximum.

Once Y_{ij} is imputed, the variable X_{ij} can be imputed by using the (truncated) linear regression model described in the previous sections, based on the records where $Y_{ij} = 1$.

6.6 Box-Cox transformations to normality

6.6.1 Normal variables

The (truncated) regression models in this chapter depend on the assumption of a normal model. As economic variables are expected to approximate normal behaviour after an appropriate transformation this assumption seems reasonable. In

their seminal paper Box and Cox (1964) developed the Box-Cox transformation, defined by

$$X(\lambda) = \begin{cases} \frac{(X+c)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(X+c) & \text{if } \lambda = 0 \end{cases} .$$

The parameter c can be used to rescale X so that it is strictly positive, for now we will use $c = 0$.

A value for the parameter λ needs to be found, so that the variable $X(\lambda)$ is as close to a normal distribution as possible. This is usually done by maximum likelihood. For a detailed description of the different estimation methods that can be used to find λ, see Spitzer (1982). Assume that the transformed item values are normally distributed: $X_i(\lambda) \sim \mathcal{N}(\mu, \sigma^2)$. The model parameters now are μ, σ and λ. The density for $X_i(\lambda)$ becomes

$$f_{X_i(\lambda)}\big(x_i(\lambda)\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i(\lambda) - \mu)^2\right).$$

Using the change of variables formula, we can obtain the density for X_i

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\Big(\frac{x_i^\lambda - 1}{\lambda} - \mu\Big)^2\right) |x_i^{\lambda-1}|.$$

This results in the following loglikelihood function

$$\ell(\mu, \sigma, \lambda \mid \mathbf{x}) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i(\lambda) - \mu)^2 + (\lambda - 1)\sum_{i=1}^{n}\ln x_i.$$

For each fixed λ this likelihood equation is proportional to the likelihood equation for estimating μ and σ if $X_i(\lambda)$ was observed. This means that the maximum likelihood estimates are

$$\begin{aligned} \hat{\mu}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda) \\ \hat{\sigma}^2(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}(x_i(\lambda) - \hat{\mu}(\lambda))^2. \end{aligned}$$

Substitute these estimates in the likelihood equation, to obtain

$$\ell(\lambda \mid \mathbf{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2} - n\ln\hat{\sigma}(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\ln x_i,$$

which is referred to as the profile or concentrated loglikelihood. This can be simplified by working with the alternative transformation

$$\tilde{X}(\lambda) = \begin{cases} \frac{(X)^{\lambda}-1}{\lambda g^{\lambda-1}} & \text{if } \lambda \neq 0 \\ \ln(X)g & \text{if } \lambda = 0 \end{cases},$$

where g is the geometric mean of the original response: $g = \prod_{i=1}^{n} x_i^{1/n}$, which equals $\exp(\frac{1}{n}\sum_{i=1}^{n} \ln x_i)$. The profile loglikelihood then becomes

$$\ell(\lambda \mid \mathbf{x}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2} - \frac{n}{2}\ln \hat{\tilde{\sigma}}^2(\lambda),$$

where $\hat{\tilde{\sigma}}^2(\lambda)$ is the variance based on $\tilde{X}(\lambda)$. This function is maximised if $\hat{\tilde{\sigma}}^2(\lambda)$ is minimised. In practice $\hat{\tilde{\sigma}}^2(\lambda)$ is evaluated for several values of λ, e.g. from -2 to 2, with increments of 0.1, and the value of λ is chosen where $\hat{\tilde{\sigma}}^2(\lambda)$ is minimal.

6.6.2 Truncated normal variables

If the original variable X is bounded, e.g. $l \leq X \leq u$, the transformed variable, $X(\lambda)$, will be bounded as well: $l(\lambda) \leq X(\lambda) \leq u(\lambda)$. This means that the Box-Cox transformation obtains variables that are distributed according to a truncated normal instead of a normal distribution, see also Poirier (1978). The parameter λ can now be obtained as follows.

Assume that $X_i(\lambda) \sim \mathcal{N}^T(\mu, \sigma^2)$, where $l_i(\lambda) \leq X_i(\lambda) \leq u_i(\lambda)$. This means that the density for $X_i(\lambda)$ is

$$f_{X_i(\lambda)}(x_i(\lambda)) = \frac{\frac{1}{\sigma}\phi(\frac{x_i(\lambda)-\mu}{\sigma})}{\Phi(\frac{u_i(\lambda)-\mu}{\sigma}) - \Phi(\frac{l_i(\lambda)-\mu}{\sigma})}.$$

Again use the change of variables formula to obtain the density of X_i

$$f_{X_i}(x_i) = \frac{\frac{1}{\sigma}\phi(\frac{x_i(\lambda)-\mu}{\sigma})}{\Phi(\frac{u_i(\lambda)-\mu}{\sigma}) - \Phi(\frac{l_i(\lambda)-\mu}{\sigma})}|x_i^{\lambda-1}|,$$

which leads to the loglikelihood

$$\begin{aligned} \ell(\mu, \sigma, \lambda \mid \mathbf{x}) \quad = \quad & -\sum_{i=1}^{n} \ln\left(\Phi\Big(\frac{u_i(\lambda)-\mu}{\sigma}\Big) - \Phi\Big(\frac{l_i(\lambda)-\mu}{\sigma}\Big)\right) - n\ln\sigma + \\ & + \sum_{i=1}^{n} \ln\phi\Big(\frac{x_i(\lambda)-\mu}{\sigma}\Big) + (\lambda-1)\sum_{i=1}^{n}\ln x_i. \end{aligned}$$

For fixed λ this is again proportional to the likelihood equation for estimating μ and σ, had we observed $X_i(\lambda)$. Unfortunately, as we are dealing with truncated normal variables the maximum likelihood estimates for this equation are not available in closed form and need to be found with an iterative algorithm, which was described in section 6.5.2. The maximum likelihood estimates will be denoted by $\hat{\mu}(\lambda)$ and $\hat{\sigma}(\lambda)$. The profile loglikelihood then becomes

$$\begin{aligned}\ell(\lambda \mid \mathbf{x}) &= -\sum_{i=1}^{n} \ln\left(\Phi\Big(\frac{u_i(\lambda)-\hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)}\Big) - \Phi\Big(\frac{l_i(\lambda)-\hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)}\Big)\right) - n\ln\hat{\sigma}(\lambda) + \\ &\quad + \sum_{i=1}^{n} \ln\phi\Big(\frac{x_i(\lambda)-\hat{\mu}(\lambda)}{\hat{\sigma}(\lambda)}\Big) + (\lambda-1)\sum_{i=1}^{n}\ln x_i.\end{aligned}$$

Again, in practice, an estimate of λ can be found by evaluating several values of λ, for example from -2 to 2 with increments of 0.1, and computing the appropriate bounds $l(\lambda)$ and $u(\lambda)$, the maximum likelihood estimates $\hat{\mu}(\lambda)$, $\hat{\sigma}(\lambda)$ and subsequently the profile loglikelihood. The optimal value for λ is the one for which the loglikelihood is maximal.

6.6.3 Conclusions

Note that since the support of the original response variable X is restricted to be positive in order to use the Box-Cox transformation, the transformed variable $X(\lambda)$ actually always follows a truncated normal distribution for $\lambda \neq 0$, where $-1/\lambda$ is the truncation point, as

$$\begin{aligned} -\infty \leq X(\lambda) \leq -\tfrac{1}{\lambda} \quad &\text{if } \lambda < 0 \\ -\tfrac{1}{\lambda} \leq X(\lambda) \leq \infty \quad &\text{if } \lambda > 0. \end{aligned}$$

This fact is usually ignored when estimating λ, as it is expected that the truncation point is in the tail of the distribution of $X(\lambda)$. However, if the mass of the variable X is near zero, then subsequently the mass of $X(\lambda)$ will be close to the truncation point as well. In this case it could be wiser to use the approach described in subsection 6.6.2 to obtain an estimate for λ. If the mass of the response variable is away from zero, this method will yield the same estimates as the method for normal variables. However, due to the increased computer time we prefer to use the nontruncated estimation procedure whenever possible. To examine the effects of both estimation methods on empirical data, we will estimate λ for several variables in the business surveys, using both methods.

Table 6.2: *Estimation of* λ.

Variable	Nontruncated	Truncated
X_{11}	0.0	0.0
X_{14}	−0.1	−0.1
X_{1t}	0.0	0.1
X_{21}	−0.2	−0.5
X_{26}	−0.1	0.2
X_{2t}	0.0	0.1
X_t	−0.1	0.0
X_r	−0.1	−0.1

Example 6.6.3. Truncated versus nontruncated estimation of λ.
In this example we will investigate the effects of truncation on the estimation of λ for empirical data. The variables used are gathered through the wholesale survey and, except for net turnover, have been used in previous chapters in the context of imputation. The following variables are examined

$$\begin{aligned} X_{11} &= \text{gross wages and salaries} \\ X_{14} &= \text{other social costs} \\ X_{1t} &= \text{total labour costs} \\ X_{21} &= \text{costs of banking} \\ X_{26} &= \text{other costs of third party services} \\ X_{2t} &= \text{total costs of third party rendering of services} \\ X_t &= \text{total operating expenses} \\ X_r &= \text{net turnover.} \end{aligned}$$

All of these variables are truncated and the following inequality restrictions need to hold

$$\begin{aligned} 0 \leq \quad & X_{11}, X_{14} \quad \leq X_{1t} \\ 0 \leq \quad & X_{21}, X_{26} \quad \leq X_{2t} \\ 0 \leq \quad & X_{1t}, X_{2t} \quad \leq X_t \\ 0 \leq \quad & X_r. \end{aligned}$$

The results are presented in table 6.2, where the nontruncated column refers to λ that is estimated using the procedure described in subsection 6.6.1 and

the truncated column refers to the estimation procedure in subsection 6.6.2. We find that the estimated values of λ for both methods are quite similar for most variables. The variables X_{21} and X_{26} produce the largest differences between estimation procedures. These variables are both semi-continuous, which could be the cause of this difference as λ is estimated based on a much smaller sample. As the number of semi-continuous variables is small and the difference between estimated λ's is still acceptable, we do not expect difficulties when using the nontruncated procedure. ■

6.7 Incorporation of linear balance restrictions

A disadvantage of this univariate conditional approach is the fact that linear balance restrictions cannot be straightforwardly incorporated in the conditional densities. That is, if a variable is a component of a balance restriction, univariately its value is known with certainty based on the other variables in that balance restriction. For the algorithm this means that variables present in a balance restriction will remain stuck in the initial values, which obviously is undesirable.

We will solve this problem by leaving out the variables that are present in the balance restrictions such that the data does not contain singularities, which was inspired by section 5.7.2. If the data were completely observed this can be done without consequence as redundant information is discarded in that case. In the presence of nonresponse, however, this may lead to a loss of information because the information in the items values that are observed may not be redundant anymore as some of the variables in that balance restriction can contain missing values. Removing variables from the dataset before imputation and deriving these variables once the imputation process is completed may result in a substantial loss of predictive and distributional accuracy for that variable. To spread this loss across all variables and because the missingness is scattered across variables, we suggest to treat each record separately and to remove variables that are already missing.

For each record i partition $\mathbf{X}_i' = (\mathbf{X}_{i,mis}' \ \mathbf{X}_{i,obs}')$ and $\mathbf{A} = (\mathbf{A}_{i,mis} \ \mathbf{A}_{i,obs})$. Then $\text{rank}(\mathbf{A}_{i,mis}) = p_i$ is the number of variables that are redundant and therefore need to be left out, which was also discussed in section 5.7.2 and 5.7.3 for the multivariate truncated singular normal. Let $\mathbf{A}_{i,mis}^n$ be the $p_i \times m_i$ matrix consisting of the nonredundant rows in $\mathbf{A}_{i,mis}$ and let $\mathbf{A}_{i,mis,p_i}^n$ be a $p_i \times p_i$ nonsingular submatrix of $\mathbf{A}_{i,mis}^n$. Note that there may be several options for $\mathbf{A}_{i,mis}^n$, that is if several variables in a balance restriction are miss-

ing each one of these could be discarded. Let m_i denote the number of missing items for record i and let $q_i = m_i - p_i$. Then, at iteration t, determine $\mathbf{A}^n_{i,mis}$ for each record i and choose the appropriate $\mathbf{A}^n_{i,mis,p_i}$ and $\mathbf{A}^n_{i,mis,q_i}$, where the matrix $\mathbf{A}^n_{i,mis,q_i}$ contains the q_i columns of $\mathbf{A}^n_{i,mis}$ that are not in $\mathbf{A}^n_{i,mis,p_i}$. Partition $\mathbf{X}'_{i,mis} = (\mathbf{X}'_{i,mis,p_i}\ \mathbf{X}'_{i,mis,q_i})$ accordingly, where $\mathbf{X}_{i,mis,p_i}$ refers to the elements in $\mathbf{X}_{i,mis}$ that need to be left out. The variables that are removed from the dataset need to be eliminated from the set of inequality restrictions as well. This can be done by using the fact that $\mathbf{X}_{i,mis,p_i} = (\mathbf{A}^n_{i,mis})^{-1}(-\mathbf{A}^n_{i,obs}\mathbf{X}_{i,obs} - \mathbf{A}^n_{i,mis,q_i}\mathbf{X}_{i,mis,q_i})$ and filling this is for the $\mathbf{X}_{i,mis,p_i}$ in $\mathbf{l} - \mathbf{B}_{i,obs}\mathbf{X}_{i,obs} \leq \mathbf{B}_{i,mis}\mathbf{X}_{i,mis} \leq \mathbf{u} - \mathbf{B}_{i,obs}\mathbf{X}_{i,obs}$, which will lead to a reduced set of inequality restrictions.

The algorithm will now be as follows. Start the algorithm at the first variable and calculate the regression parameters for this variable if it is present in $\mathbf{X}_{i,mis,q_i}$ conditioned on the other variables in $\mathbf{X}_{i,mis,q_i}$ and using the reduced set of inequality restrictions. Note that this results in several regression models as the missingness pattern varies across records. Records with similar missingness patterns, however, will obtain the same parameter estimates. Subsequently the missing items in this variable, that are present in $\mathbf{X}_{i,mis,q_i}$, are re-imputed with the appropriate regression model, the parameter estimates found and subject to the reduced set of linear inequality restrictions. After this the variables that were removed, $\mathbf{X}_{i,mis,p_i}$, need to be recalculated using the updated values for $\mathbf{X}_{i,mis,q_i}$ in order to make sure that the balance restrictions remain satisfied. Now proceed to the next variable and repeat this process until all k variables are treated.

As we stated before, leaving out data could result in a loss of information and consequently the predictive and distributional accuracy of the imputed values of $\mathbf{X}_{i,mis,p_i}$ may be reduced. As we have not found another approach to incorporate balance restrictions it appears that we cannot overcome this issue. To somewhat dampen these effects, we suggest choosing the variables that are left out for each record randomly in each cycle.

6.8 Imputation performance

In order to assess the performance of this imputation procedure on empirical data, we will use this method to impute data that have been gathered by Statistics Netherlands on a part of the wholesale industry for businesses with more than 10 employees. The effects of imputation on the estimation of population estimates as well as the ability of the imputation method to preserve individual

values and distributions are investigated.

6.8.1 Description of the data and generation of missing items

In this chapter we will use the dataset on labour costs that was introduced in chapter 3. The dataset consists of the variables

$$\begin{aligned} X_{11} &= \text{gross wages and salaries} \\ X_{12} &= \text{social security costs} \\ X_{13} &= \text{pension charges} \\ X_{14} &= \text{other social costs} \\ X_{1t} &= \text{total labour costs.} \end{aligned}$$

In chapter 3 we stated that the restrictions on this dataset are one balance restriction: $X_{11} + X_{12} + X_{13} + X_{14} = X_{1t}$, and non-negativity restrictions on all data items. In fact this is not completely true as there are more inequality restrictions that must hold, which could not be incorporated by the imputation methods used in chapter 3. These inequality restrictions are

$$\begin{aligned} X_{1t} &\geq 1.01X_{11} \\ X_{1t} &\leq 2X_{11} \\ X_{12} &\geq 0.06X_{11} \\ X_{12} &\leq 0.26X_{11}. \end{aligned}$$

The sequential imputation procedure is capable of incorporating these added inequality restrictions and therefore they will now be taken into account as well.

The missing data will be generated in the same way as was described in chapter 3. That is, assuming that the missing data are MCAR and using draws from the Bernoulli distribution to establish whether an item is missing or not. The amount of nonresponse created is also the same as in chapter 3. In this case, however, the amount of data imputed by deductive imputation may be somewhat larger as more restrictions are available. In order to be able to compare this method to other imputation procedures, we will not introduce missingness in the variable X_{1t}. This means that we can impute the data with the procedures that were described in chapter 3 (disregarding the added inequality restrictions), which will provide more insight in the performance of the sequential imputation procedure. The remaining dataset with missing values is used for imputation.

6.8.2 Imputation using the sequential regression approach

The missing data will now be imputed using the sequential regression approach that was suggested in this chapter. This means that the variables $X_{11}, \ldots, X_{14}$ are sequentially modelled and imputed. The order in which this is done is based on the amount of missingness in each variable, starting with the variable with the least number of missing items and ending with the variable with the highest amount of nonresponse. This ordering is used as it has been found to be the most efficient.

As all of these variables are restricted by linear bounds, we will use a truncated regression model for each variable. Due to the skewness of the densities of these variables, they will be transformed to approximate normality using Box-Cox transformations. The value for the transformation parameter λ is calculated based on the available cases for each variable.

As the data need to satisfy a balance restriction, one variable from the analysis needs to be eliminated as was described in section 6.7. For reasons of convenience this will not be done at random as was suggested, but using the following procedure. If a record contains missing values the first missing variable that is encountered in the order $X_{11}, \ldots, X_{14}$ is removed from the regression model. Note that this means that variable X_{11} will always be discarded if it is missing and X_{14} will always be used in the regression models. We therefore expect that the accuracy of the imputed variable X_{11} will be somewhat lower. Further research can be aimed at choosing a variable at random.

The variable that is removed from the regression model also needs to be eliminated from the inequality restrictions. This can be done using the balance restriction. For example, if X_{11} is left out, we can fill in $X_{11} = X_{1t} - X_{14} - X_{13} - X_{12}$ in the inequality restrictions and determine the appropriate inequality restrictions that must hold for the variable that needs to be imputed. If the variable X_{12} is missing this would lead to the following set of restrictions

$$\begin{aligned} X_{12} &\geq \frac{0.01}{1.01} X_{1t} - X_{14} - X_{13} \\ X_{12} &\leq \frac{1}{2} X_{1t} - X_{14} - X_{13} \\ X_{12} &\geq \frac{0.06}{1.06} (X_{1t} - X_{14} - X_{13}) \\ X_{12} &\leq \frac{0.26}{1.26} (X_{1t} - X_{14} - X_{13}) \end{aligned}$$

and of course the non-negativity restriction: $X_{12} \geq 0$. Once a variable is imputed, the removed variable needs to be recalculated using the updated values of the imputed variable and the balance restriction in order to avoid inconsistencies.

A problem that arises using the sequential regression approach is the fact that the dataset contains two substantial outliers. These outliers have a large impact on the parameter estimates obtained using the truncated regression model. A result of this is that some of the expected values of the other records will fall outside the interval created by their lower and upper bounds, meaning that the algorithm may sometimes be unable to generate draws that fall inside this interval. Clearly, this is an undesirable effect and the outliers will therefore be removed from the dataset. It is a disadvantage of this method, however, that it is sensitive to outlying values, even once the data have been transformed.

If all variables have been imputed, the process has passed through one cycle. In order to assess convergence and determine the number of iterations that are needed for the process to stabilise we performed the imputation process, with 500 cycles, for a dataset with randomly assigned missing data. The estimates for the population means against the number of iterations are plotted in Figure 6.1. As was found by Van Buuren, Boshuizen and Knook (1999) as well, the process appears to converge within a few iterations. We have repeated this procedure for several different realised sets of missing data, which resulted in similar behaviour. Therefore we will set the number of iterations used in the imputation process to 10.

6.8.3 The effects of imputation on parameter estimation

As the main target of our business surveys usually is to estimate the population estimates on mean and dispersion, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, the imputation procedure will be judged based on its ability to produce accurate point estimates. Another point of interest is the magnitude of the nonresponse variance component of the population parameters for this imputation procedure as this is an indication of the sensitivity of the procedure with respect to the realised set of missing data. Again we will conduct a simulation study to investigate the performance of the sequential regression imputation procedure.

In chapter 3 we found that incomplete data procedures, using the complete or available cases, result in both inaccurate point estimates and high nonresponse variance, due to the amount of missing data. These procedures will therefore not be used in this study. Another result obtained in chapter 3 is that the Dirichlet approach with expectations imputed (Dir) and the ratio nearest neighbour

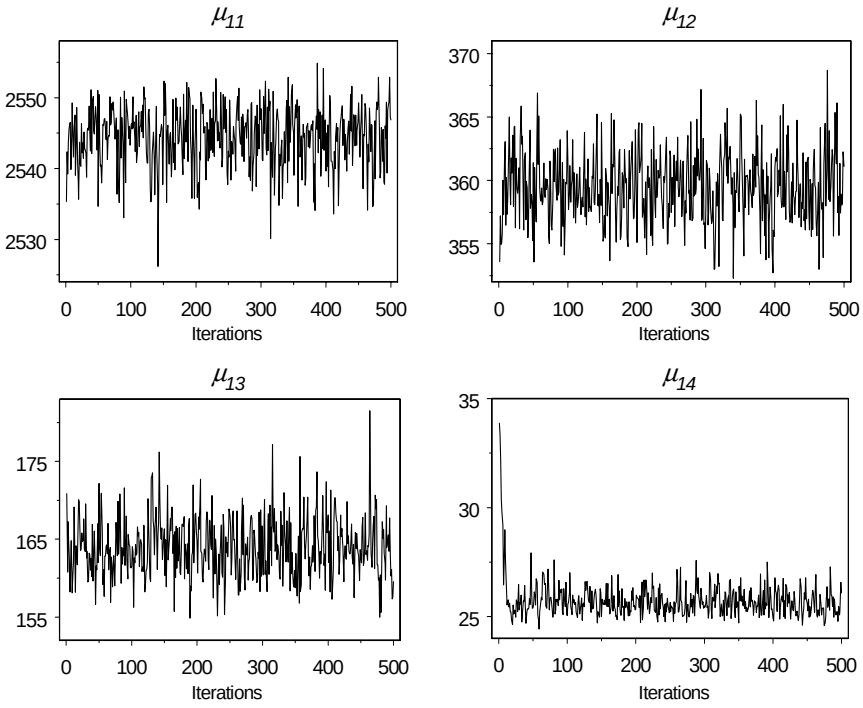

Figure 6.1: *Plots of the population mean estimates against the number of iterations.*

method (RNN) produced the best results for this dataset. We will therefore use these methods in this study as well in order to be able to assess the performance of the sequential regression procedure (Seq). Note that these methods cannot take the added linear inequality restrictions into account, so the imputed values that are produced are likely to violate these restrictions.

The process of generating and imputing missing item values and subsequently estimating population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ is iterated 100 times. The parameter estimates and their 95% confidence intervals are given in Figures 6.2 and 6.3. The sequential regression approach generates parameter estimates that are quite accurate for both $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, with nonresponse variances that are similar to those obtained by the Dirichlet and nearest neighbour method. The sequential regres-

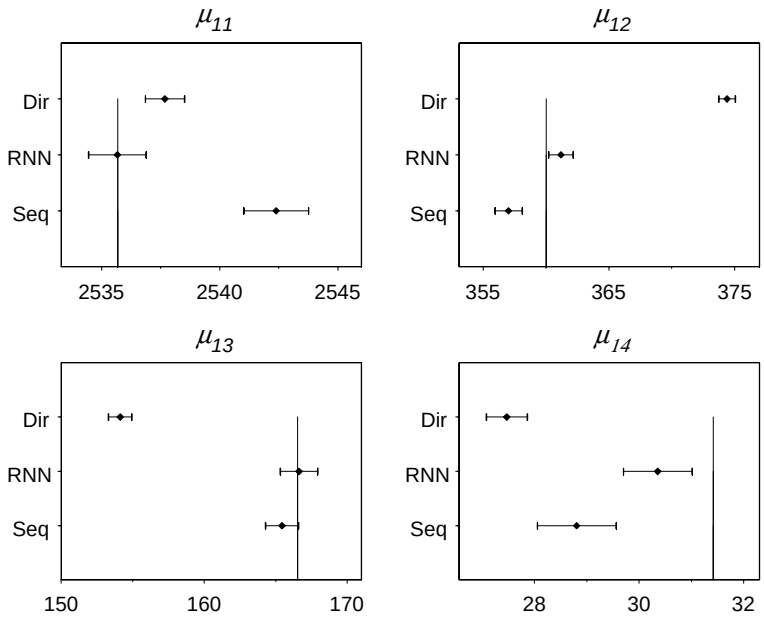

Figure 6.2: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\mu}$.

sion approach even seems to somewhat outperform the Dirichlet procedure in this case, one needs to keep in mind, however, that this latter method is sensitive to outliers which have been removed from the dataset. This seriously reduces its general applicability.

Beforehand we stated that the accuracy of the parameter estimates μ_{11} and σ_{11} will probably be lower than the accuracy of the parameters estimates of the other variables in the dataset obtained using the sequential regression imputation as the first variable is removed from the data most often. With regard to that variable we observe that the sequential regression method performs worse than the other methods when estimating μ_{11} and σ_{11}. This last quantity is strongly overestimated, which was to be expected since there will be more variability in X_{11}. This effect can be spread out across all variables by choosing the variable that needs to be removed at random, which will be a topic of further research.

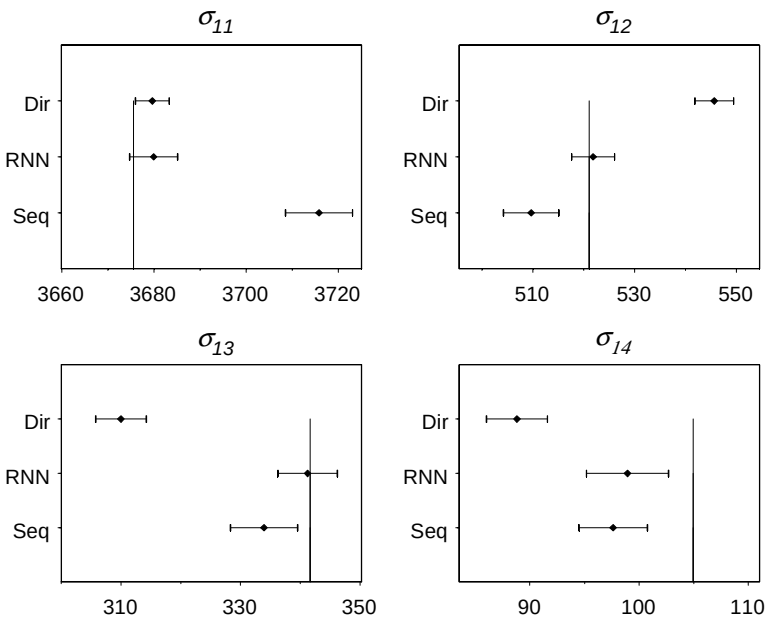

Figure 6.3: *95% confidence intervals for the parameter estimates of* $\boldsymbol{\sigma}$.

6.8.4 The performance of the imputation method on item level

Other desirable properties arise if the main interest of the imputer is to obtain a general purpose dataset as well as accurate point estimates. Similar to chapter 3, the imputation methods will be judged on predictive accuracy, i.e. the ability of the method to preserve individual true values, and distributional accuracy, i.e. the ability to preserve the distribution of the true values. The average absolute deviation of the imputed data from the true data is calculated to assess the predictive accuracy and the Kolmogorov-Smirnov statistic, that measures the maximal distance between empirical distribution functions, is used to assess distributional accuracy. The results are presented in Table 6.3.

With regard to the average deviation between the imputed and the true data, the sequential regression method appears to perform slightly worse than the nearest neighbour method and quite a bit worse than the Dirichlet method,

Table 6.3: *Predictive and distributional accuracy of the imputed data on labour costs.*

	Average deviation			*Kolmogorov-Smirnov distance*		
	Dir	RNN	Seq	Dir	RNN	Seq
X_{11}	78	105	127	0.03	0.03	0.04
	(10)	(14)	(19)	(0.01)	(0.01)	(0.01)
X_{12}	74	83	93	0.11	0.06	0.07
	(11)	(12)	(13)	(0.02)	(0.01)	(0.02)
X_{13}	75	100	113	0.12	0.08	0.08
	(11)	(12)	(15)	(0.02)	(0.02)	(0.02)
X_{14}	32	45	38	0.54	0.08	0.23
	(6)	(7)	(7)	(0.03)	(0.03)	(0.10)

which was found to preserve individual values best in chapter 3.

With respect to distributional accuracy, the sequential regression approach performs quite well, especially for the variables X_{11}, X_{12} and X_{13}. For the variable X_{14} this method has used a model that developed specifically for semi-continuous data. This does result in a lower Kolmogorov-Smirnov distance for the semi-continuous variable X_{14} as opposed to the Dirichlet method, but the K-S distance is still larger than the K-S distance of the nonparametric nearest neighbour method. So, it remains quite difficult to find an appropriate parametric model for semi-continuous data. The use of other models, such as probit or censored regression models, could also be investigated for this purpose, but we do not expect large differences in the outcomes.

6.9 Concluding remarks

In this chapter we have discussed the use of a sequential regression imputation procedure developed by Raghunathan et al. (2001) to impute data that are subject to inequality restrictions. This procedure models and imputes each variable separately using different regression models, based on variable-specific properties such as semi-continuity and linear bounds. Consequently, the procedure is extremely flexible and can deal with all kinds of data. We have extended this method in order to incorporate linear balance restrictions as well.

An important flaw of the sequential imputation procedure is the fact that the different univariate conditional models may be incompatible, that is they may not constitute a joint model. Although we did not encounter problems with convergence in section 6.8 on the application to empirical data, this should definitely be a topic for future research.

In section 6.8 we also found that the method appears to perform quite well with respect to preservation of population estimates and distributional accuracy of the imputed data. Therefore the sequential regression procedure appears to be a promising approach in the context of data imputation subject to linear restrictions. The huge advantage of being able to impute data while incorporating several restrictions is, however, toned down by the fact that the procedure is sensitive to outliers and can only incorporate balance restrictions by temporarily removing variables from the imputation procedure, resulting in less accurate parameter estimates. This method, however, remains the only developed imputation method that can impute data subject to both balance and inequality restrictions and that can be applied in practice.

This technique is very suitable for large datasets as the data are modelled with univariate distributions. This means that a large part of a business survey can be imputed simultaneously. So variables on company expenses, company turnover, company profits, employment and so on can be imputed all at once.

Another advantage of univariate modelling is that separate models can be used for different variables, which means that the variables are not required to be continuous anymore. This procedure can therefore be straightforwardly extended to impute all sorts of social surveys as well.

Chapter 7

Conclusions

In this thesis we have developed several methods that provide imputations for restricted data. An application of this is to economic data, which often need to satisfy a large amount of linear restrictions. Throughout this thesis we have distinguished between balance and inequality restrictions, as these two types of restrictions lead to different modelling issues. The complexity of the imputation model is strongly determined by the complexity of the restriction structure that can be handled. In each chapter on imputation models we have therefore first discussed the different types of restrictions that can be dealt with.

An important issue in the context of imputation is the missing data mechanism, which concerns the reasons why values are missing. The imputation procedures developed in this thesis can be used for data that are either missing completely at random or missing at random. In this case the fact that a certain data item is missing does not depend on the actual value of that data item.

Ideally, imputation procedures have several desirable features. First of all, the imputation model should preserve aggregates, such as the estimates for population means and variances. But, on the other hand, it is also desirable that the imputation methods preserve individual item values (predictive accuracy) and distributions of variables (distributional accuracy). In general, the performance of an imputation procedure will vary with respect to these three features. Imputing the expected value of an item will, for example, probably result in accurate point estimates for the population means, but may distort the distribution of that variable. Therefore it is important to know what the main interests of the users are. The main task of a statistical agency is to publish aggregate data, so in that case the principal objective will be to preserve aggregate values.

In chapter 3 we have developed an imputation method based on the Dirichlet distribution that can be used to model and impute data items that need to be non-negative and that are subject to one balance restriction. In this case not the actual data but proportions of the data items with respect to a certain total are modelled. The EM algorithm is applied to the Dirichlet distribution in order to obtain parameter estimates. We found that this method is quite capable of preserving population parameters and individual item values. With respect to distributional accuracy other (nonparametric) methods performed better. A major advantage of the Dirichlet approach is that the nonresponse variance component of the total variance of the estimates was quite small, meaning that the method is robust with respect to the realised set of missing data. Besides this, the method is easy to implement and does not require much computing time. A drawback of this method is that it can only cope with one linear balance restriction at a time.

Therefore, in chapter 4, we used the singular normal distribution to model and impute data that are subject to multiple linear balance restrictions. We derived an application of the EM algorithm that produces maximum likelihood estimates that concur with the linear balance restrictions and consequently the imputed values will always satisfy the balance restrictions. In addition to this, the restrictions need not be specified as they are embedded in the singularity of the covariance matrix, which is very convenient. This method performs quite well with respect to the preservation of estimates of population means and variances. This is especially the case if there is a large amount of restrictions on the data, which clearly is a very encouraging result.

If the data are subject to one balance restriction only, the method developed in chapter 3 is probably preferred as the EM algorithm for singular normal data produces similar results with respect to imputation performance but requires quite a bit more computing time. An advantage of the singular normal model as opposed to the Dirichlet approach is the fact that in this case it is not needed that a variable, which represents a total, is completely observed. The singular normal model cannot take non-negativity restrictions into account, however, which obviously is a drawback.

This led to the development of an imputation model, that employs the truncated singular normal distribution, in chapter 5. This model can incorporate all different types of linear restrictions that may occur, which means that it is a generic procedure. A downside of this approach is that a complex model is employed that uses high-dimensional integrals. These integrals need to be estimated and, although this can be done quite accurately using (Markov chain) Monte Carlo methods, the procedure will need a large amount of computing

time. An even greater drawback is the fact that the imputation procedure is strongly dependent on model assumptions, such as multivariate normality, that are likely not to hold in practice, reducing its general applicability.

As joint modelling of the data, while incorporating the linear restrictions with model assumptions that can be met in practice, was found to be a difficult task, a different approach was used in chapter 6. In this chapter instead of a specific multivariate model, univariate conditional distributions were determined. These conditional distributions were employed to sequentially model and impute the data. The advantage of this imputation procedure is that it is flexible and easy to implement. A disadvantage is the lack of a sound theoretical basis. The univariate conditional distributions are chosen without an encompassing multivariate model, which may lead to incompatibility issues. The effects of this are unclear and have not been thoroughly investigated. The performance of this method on empirical data with respect to distributional accuracy and the preservation of population parameters seemed promising, however.

The imputation procedures, developed throughout this thesis, were subsequently applied to subsets of an empirical dataset. In reality, the data that need to be imputed contain many more variables and therefore most of the time an imputation strategy is developed that combines several imputation methods. We think that the methods developed in this thesis can be quite beneficial in different parts of such a strategy.

For instance, aggregate variables, that need to satisfy multiple linear balance restrictions, can be modelled and imputed using the singular normal model in combination with the EM algorithm. As aggregate variables are often far from zero the risk of imputing negative values will be low. Once the aggregate values are imputed, the data that remain to be imputed can be partitioned into several datasets that need to satisfy only one linear balance restriction, e.g. adding up to one of the aggregate variables that were imputed earlier. These data items can still be subject to multiple inequality restrictions.

In this case the Dirichlet model, the sequential regression approach or the multivariate truncated singular normal model can be used to impute these data items, depending on the variable specifics. That is, if the data items only need to satisfy non-negativity constraints, the Dirichlet is a fast and robust approach. If the data items are subject to other inequality restrictions as well, the sequential regression approach can be applied, especially if there is a large amount of semi-continuous variables present. Finally, if the data are approximately normally distributed, the truncated singular normal can be employed.

Further research is needed on other types of data and using different missing data mechanisms to investigate the possibilities of these imputation procedures

and to determine an optimal imputation strategy, such that guidelines can be provided to the imputer on what imputation method to use given the data of interest and the main purpose of the user. Future work could also be aimed at further investigation of the method in chapter 6, for example with respect to the incompatibility issues mentioned or by applying the imputation procedure to larger datasets that are subject to multiple balance restrictions.

Finally, we would like to emphasise that, although imputation is a commonly used and convenient method to deal with missing data, it is crucial that information about imputation is provided along with each dataset, in order to inform users about data quality. First of all imputed values should always be flagged, so that the user will be able to distinguish between observed and constructed data. Furthermore, the user should be informed about what imputation techniques were used and additional information should be given for certain techniques. In addition to this the statistical agency should evaluate the effects of imputation and if possible provide an estimate of the variance after imputation in order to give the user an insight in the precision of the data. Furthermore it should also be made clear to the user that relations between variables may be attenuated through imputation.

The models that were developed in this thesis were originally targeted at business surveys, but may also be of use in other areas of interest of the statistical agency. First of all, these methods can also be applied to impute numeric variables subject to restrictions in social surveys, such as the household income variables that are measured in budget surveys. Furthermore, the sequential imputation procedure discussed in chapter 6 can be straightforwardly applied to all different types of variables that may occur in social surveys, such as race or sex, as it is not required that variables are numeric in this procedure.

Another area of possible interest is statistical confidentiality and the release of microdata. In this case we might need to perturb the original data in order to protect respondent confidentiality. In this instance a dataset without missing values is analysed with respect to confidentiality before publication. If there are item values present that are considered to be unsafe because they reveal sensitive information on individual respondents the data need to be perturbed. Clearly, the perturbed data still need to satisfy the linear restrictions on the data. The models described in this thesis can be employed for this purpose by generating (random) draws from the distributions used and replacing some of the observed unsafe data by these draws.

These models may, however, also be of use in areas outside the scope of the statistical agency. They can, for example, be beneficial in the banking industry for the creation of reference datasets in the context of Probability of Default

and Loss Given Default modelling as a part of Basel II. These reference datasets are subject to several accounting restrictions and are likely to consist of missing data, which means that they need to be imputed.

Another field of interest is, for instance, marketing, if market shares need to be determined in the presence of missing data. The Dirichlet model that was developed to impute proportions of variables will be suitable for this situation. Clearly, this model may be of use for the imputation of proportions in various disciplines, such as results of elections, time budgets or compositions of a certain species.

Bibliography

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, New York: Chapman & Hall.

Aitchison, J. and S.M. Shen (1980), "Logistic-normal distributions: Some properties and uses", *Biometrika*, **59**, 19–35.

Amemiya, T. (1973), "Regression analysis when the dependent variable is truncated normal", *Econometrica*, **41**, 997–1016.

Amemiya, T. (1985), *Advanced Econometrics*, Oxford: Basil Blackwell.

Arnold, B.C. and S.J. Press (1989), "Compatible conditional distributions", *Journal of the American Statistical Association*, **84**, 152–156.

Bethlehem, J.G. and H.M.P. Kersten (1986), *Werken met Non-respons*, Ph.D. thesis (in Dutch only).

Box, G.E.P. and D.R. Cox (1964), "An analysis of transformations", *Journal of the Royal Statistical Society, Series B*, **26**, 211–246.

Brooks, S.P. and G.O. Roberts (1998), "Assessing convergence of Markov chain Monte Carlo algorithms", *Statistics and Computing*, **8**, 319–335.

Brown, L.D. (1986), *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Hayward, California: Institute of Mathematical Statistics.

Celeux, G. and J. Diebolt (1985), "The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem", *Computational Statistics Quarterly*, **2**, 73–82.

Chambers, E.A. and D.R. Cox (1967), "Discrimination between alternative binary response models", *Biometrika*, **54**, 573–578.

Chib, S. and E. Greenberg (1995), "Understanding the Metropolis-Hastings algorithm", *The American Statistician*, **49**, 327–335.

Cowles, M.K. and B.P. Carlin (1996), "Markov chain Monte Carlo convergence diagnostics: A comparative study", *Journal of the American Statistical Association*, **91**, 883–904.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society B*, **39**, 1–38.

DeSarbo, W.S., V. Ramaswamy and P. Lenk (1993), "A latent class procedure for the structural analysis of two-way compositional data", *Journal of Classification*, **10**, 159–193.

Fellegi, I.P. and D. Holt (1976), "A systematic approach to automatic edit and imputation", *Journal of the American Statistical Association*, **71**, 17–35.

Fomby, T.B., R.C. Hill and S.R. Johnson (1984), *Advanced Econometric Methods*, New York: Springer-Verlag.

Ford, B.L. (1983), "An overview of hot deck procedures", in *Incomplete Data in Sample Surveys, Vol.II: Theory and Bibliographies*, eds Madow, W.G., I. Olkin and D.B. Rubin, New York: Academic Press.

Gelfand, A.E. and A.F.M. Smith (1990), "Sampling based approaches to calculating marginal densities", *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. and T.E. Raghunathan (2001), "Discussion of "Conditionally specified distributions", by Arnold, B.C. et al.", *Statistical Science*, **16**, 268–269.

Gelman, A. and D.B. Rubin (1992), "Inference from iterative simulation using multiple sequences", *Statistical Science*, **7**, 457–472.

Gelman, A., G.O. Roberts and W.R. Gilks (1996), "Efficient Metropolis jumping rules", in *Bayesian Statistics*, eds Bernardo, J.M., J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press.

Geman, S. and D. Geman (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Gentle, J.E. (2002), *Elements of Computational Statistics*, New York: Springer-Verlag.

Geweke, J. (1991), "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints", *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 571–578.

Geweke, J. (1996), "Monte Carlo simulation and numerical integration", in *Handbook of Computational Economics*, eds Amman, H.M., D.A. Kendrick and J. Rust, Amsterdam: Elsevier.

Granquist, L. (1997), "The new view on editing", *International Statistical Review*, **65**, 381–387.

Granquist, L. and J. Kovar (1997), "Editing of survey data: How much is enough", in *Survey Measurement and Process Quality*, eds Lyberg, L.E. and P. Biemer, New York: John Wiley & Sons.

Haas, A. and P. Formery (2002), "Uncertainties in facies proportion estimation I. Theoretical framework: The Dirichlet distribution", *Mathematical Geology*, **34**, 679–702.

Hajivassiliou, V.A. and D.L. McFadden (1990), "The method of simulated scores for the estimation of LDV models with an application to external debt crises", *Cowles Foundation Discussion Paper No. 967.*

Hajivassiliou, V.A. and D.L. McFadden (1998), "The method of simulated scores for the estimation of LDV models", *Econometrica*, **66**, 863–896.

Hajivassiliou, V.A. and P.A. Ruud (1994), "Classical estimation methods for LDV models using simulation", in *The Handbook of Econometrics IV*, eds Engle, R.F. and D.L. McFadden, Amsterdam: North-Holland.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, **57**, 97–109.

Heeringa, S.G., R.J.A. Little and T.E. Raghunathan (2002), "Multivariate imputation of coarsened survey data on household wealth", in *Survey Nonresponse*, eds Groves, R.M., D.A. Dillman, J.L. Eltinge and R.J.A. Little, New York: John Wiley & Sons.

Horrace, W.C. (2005), "Some results on the multivariate truncated normal distribution", *Journal of Multivariate Analysis*, **94**, 209–221.

Ibrahim, J.G. (1990), "Incomplete data in generalized linear models", *Journal of the American Statistical Association*, **85**, 765–769.

Jeuland, A.P., F.M. Bas and G.P. Wright (1980), "A multibrand stochastic model compounding heterogeneous Erlang timing and multinomial choice processes", *Operations Research*, **28**, 255–277.

Johnson, N.L. and S. Kotz (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, New York: John Wiley & Sons.

Kalton, G. (1983), *Compensating for Missing Survey Data*, Ann Arbor: Survey Research Center, University of Michigan.

Kalton, G. and D. Kasprzyk (1982), "Imputing for missing survey responses", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22–31.

Kalton, G. and D. Kasprzyk (1986), "The treatment of missing survey data", *Survey Methodology*, **12**, 1–16.

Keane, M. (1994), "A computationally practical simulation estimator for panel data", *Econometrica*, **62**, 95–116.

Khatri, C.G. (1968), "Some results for the singular normal multivariate regression models", *Sankhyā, Series A*, **30**, 267–280.

Kovar, J. and P. J. Whitridge (1995), "Imputation of business survey data", in *Business Survey Methods*, eds Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, New York: John Wiley & Sons.

Lessler, J.T. and W.D. Kalsbeek (1992), *Nonsampling Error in Surveys*, New York: John Wiley & Sons.

Little, R.J.A. (1988), "Missing-data adjustments in large surveys", *Journal of Business and Economic Statistics*, **6**, 287–297.

Little, R.J.A. and T.E. Raghunathan (1997), "Should imputation of missing data condition on all observed variables?", *Proceedings of the Survey Research Methods Section, American Statistical Association*, 617–622.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data*, Second Edition, New York: John Wiley & Sons.

Louis, T.A. (1982), "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

Magnus, J.R. and H. Neudecker (1988), *Matrix Differential Calculus*, New York: John Wiley & Sons.

McLachlan, G.J. and T. Krishnan (1997), *The EM Algorithm and Extensions*, New York: John Wiley & Sons.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), "Equations of state calculations by fast computing machines", *Journal of Chemical Physics*, **21**, 1087–1092.

Metropolis, N. and S. Ulam (1949), "The Monte Carlo method", *Journal of the American Statistical Association*, **44**, 335–341.

Narayanan, A. (1991), "Maximum likelihood estimation of the parameters of the Dirichlet distribution", *Applied Statistics*, **40**, 365–374.

Orchard, T. and M.A. Woodbury (1972), "A missing information principle: Theory and applications", *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, **1**, 697–715.

Pannekoek, J. and T. de Waal (2005), "Automatic editing and imputation for business surveys: The Dutch contribution to the EUREDIT project", *Journal of Official Statistics*, **21**, 257–286.

Poirier, D.J. (1978), "The use of the Box-Cox transformation in limited dependent variable models", *Journal of the American Statistical Association*, **73**, 284–287.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge: Cambridge University Press.

Raftery, A.E. and S.M. Lewis (1992a), "How many iterations in the Gibbs sampler", in *Bayesian Statistics*, eds Bernardo, J.M., J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford University Press.

Raftery, A.E. and S.M. Lewis (1992b), "One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo", *Statistical Science*, **7**, 493–497.

Raghunathan, T.E., J.M. Lepkowski, J. van Hoewyk and P. Solenberger (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey Methodology*, **27**, 85–95.

Raghunathan, T.E., P. Solenberger and J. van Hoewyk (2002), "IVEware: Imputation and variance estimation software", *http://www.isr.umich.edu/src/smp/ive*.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons.

Rao, J.N.K. (1996), "On variance estimation with imputed survey data", *Journal of the American Statistical Association*, **91**, 499–506.

Rao, J.N.K. and J. Shao (1992), "Jackknife variance estimation with survey data under hot deck imputation", *Biometrika*, **79**, 811–822.

Robert, P.R. and G. Casella (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Ronning, G. (1989), "Maximum likelihood estimation of Dirichlet distributions", *Journal of Statistical Computation and Simulation*, **32**, 215–221.

Rosenthal, J.S. (1995), "Minorization conditions and convergence rates for Markov chain Monte Carlo", *Journal of the American Statistical Association*, **90**, 558–566.

Rubin, D.B. (1976), "Inference and missing data", *Biometrika*, **63**, 581–590.

Rubin, D.B. (1978), "Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20–34.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.

Rubinstein, R.Y. (1981), *Simulation and the Monte Carlo Method*, New York: John Wiley & Sons.

Sande, I.G. (1982), "Imputation in surveys: Coping with reality", *The American Statistician*, **36**, 145–152.

Sande, I.G. (1983), "Hot deck imputation procedures", in *Incomplete Data in Sample Surveys, Vol.III: Proceedings of the Symposium*, eds Madow, W.G. and I. Olkin, New York: Academic Press.

Särndal, C.-E. (1992), "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, **18**, 241–252.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Shao, J. and R.R. Sitter (1996), "Bootstrap for imputed survey data", *Journal of the American Statistical Association*, **91**, 1278–1288.

Sherman, J. and W.J. Morrison (1950), "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix", *Annals of Mathematical Statistics*, **21**, 124–127.

Spitzer, J.J. (1982), "A primer on Box-Cox estimation", *Review of Economics and Statistics*, **64**, 307–313.

Stuart, A. and J.K. Ord (1991), *Kendall's Advanced Theory of Statistics*, Vol 2 (5th edition), New York: Oxford University Press.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press.

Van Buuren, S., H.C. Boshuizen and D.L. Knook (1999), "Multiple imputation of missing blood pressure covariates in survival analysis", *Statistics in Medicine*, **18**, 681–694.

Van Buuren, S., J.P.L. Brand, C.G.M. Groothuis-Oudshoorn and D.B. Rubin (2006), "Fully conditional specification in multivariate imputation", *Journal of Statistical Computation and Simulation*, forthcoming.

Van Buuren, S. and C.G.M. Oudshoorn (1999), "Flexible multivariate imputation by MICE", *TNO-report PG 99.054*, TNO Prevention and Health, Leiden.

Wansbeek, T.J. and E. Meijer (2000), *Measurement Errors and Latent Variables in Econometrics*, Advanced Textbooks in Economics, 37, Amsterdam: North-Holland.

Wei, G.C.G. and M.A. Tanner (1990), "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm", *Journal of the American Statistical Association*, **85**, 699–704.

Wilks, S.S. (1962), *Mathematical Statistics*, New York: John Wiley & Sons.

Samenvatting (Summary in Dutch)

Dit proefschrift bespreekt verscheidene modellen die zijn ontwikkeld om bedrijfsgegevens, die veelal onderhevig zijn aan verschillende soorten lineaire beperkingen, te imputeren. Voorbeelden van beperkingen die voorkomen zijn de verschillende bedrijfslasten die moeten optellen tot het totaal bedrijfslasten, het feit dat het aantal medewerkers omgerekend in vte (voltijdequivalenten) kleiner moet zijn dan het totale aantal medewerkers en het feit dat de meeste financiële variabelen niet negatief mogen zijn. Standaard imputatiemethoden zoals 'hot deck'- en regressie-imputatie houden geen rekening met deze verschillende beperkingen in het imputatiemodel met als gevolg dat de geïmputeerde waarden hoogstwaarschijnlijk niet zullen voldoen aan de geldende beperkingen. Er is op het Centraal Bureau voor de Statistiek (CBS) een sterke behoefte aan consistente gegevens, i.e. gegevens die aan alle beperkingen voldoen, omdat inconsistente gegevens met zichtbare onjuistheden de geloofwaardigheid en dus het imago van het CBS negatief kunnen beïnvloeden. Het is hierom dus noodzakelijk dat de geïmputeerde gegevens aan de verschillende beperkingen voldoen. Bovendien bevatten deze beperkingen waardevolle informatie over de te imputeren waarden, in het geval van een optelling weet men nu immers wat het totaal van de geïmputeerde waarden moet zijn. Het is wenselijk om deze additionele informatie mee te nemen in het imputatiemodel.

In dit proefschrift wordt onderscheid gemaakt tussen lineaire gelijkheids- en ongelijkheidsbeperkingen. Het eerste geval betreft gelijkheden die moeten gelden, zoals de bovengenoemde optelling van de bedrijfslasten. Deze gelijkheidsbeperkingen vormen in dit proefschrift de $p \times k$ matrix $\mathbf{A}$, waarvoor geldt dat $\mathbf{A}\mathbf{X}_i = \mathbf{0}$ met $\mathbf{X}_i$ de data vector met bedrijfsgegevens van respondent i. De tweede situatie verwijst naar ongelijkheden waaraan moet worden voldaan, zoals

de bovengenoemde niet-negativiteits beperkingen en het aantal medewerkers dat omgerekend in vte niet groter mag zijn dan het totaal aantal medewerkers. De ongelijkheidsbeperkingen worden in dit proefschrift gerepresenteerd door de $r \times k$ matrix $\mathbf{B}$, waarvoor geldt dat $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$, met $\mathbf{l}$ en $\mathbf{u}$ respectievelijk de onder- en bovengrenzen die gelijk mogen zijn aan -$\boldsymbol{\infty}$ en $\boldsymbol{\infty}$.

In de hoofdstukken 3 t/m 6 zijn imputatiemodellen ontwikkeld en geanalyseerd waarmee gegevens geïmputeerd kunnen worden die moeten voldoen aan bepaalde lineaire beperkingen. Deze modellen variëren met betrekking tot de complexiteit van de beperkingen die zij aankunnen. Al deze hoofdstukken zijn geschreven volgens eenzelfde structuur. Allereerst worden de verschillende typen beperkingen besproken waarmee het imputatiemodel rekening kan houden. Vervolgens wordt het model zelf uitgebreid behandeld en daarna worden schattingsprocedures beschreven zodat de parameters van het desbetreffende model kunnen worden bepaald. Ten slotte zullen dan deterministische en stochastische imputatieprocedures worden afgeleid. Een deterministische imputatieprocedure levert een geïmputeerde waarde op die niet verandert als het proces wordt herhaald. Een stochastische procedure bevat echter een toevallige component waardoor de geïmputeerde waarde varieert als het proces wordt herhaald. De eerste methode heeft de voorkeur als men voornamelijk geïnteresseerd is in puntschattingen, zoals totalen en gemiddelden. Een stochastische imputatiemethode is wenselijk als men de verdeling van de data en de onderlinge relaties tussen variabelen zo goed mogelijk wil behouden.

In hoofdstuk 2 wordt allereerst een aantal theoretische onderwerpen besproken met als doel de lezer van de benodigde informatie te voorzien voor een volledig begrip van de latere hoofdstukken in dit proefschrift. Zo wordt er eerst dieper ingegaan op de methode der grootste aannemelijkheid ('maximum likelihood'), omdat deze methode gebruikt wordt om de parameters van de verschillende imputatiemodellen te schatten. Vervolgens behandelen we het 'Expectation-Maximisation' (EM) algoritme, dat is ontwikkeld om de aannemelijkheidsfunctie te maximaliseren in het geval van incomplete waarnemingen. Op deze manier kunnen parameterschattingen verkregen worden als er sprake is van ontbrekende gegevens. Het EM algoritme leidt in sommige gevallen tot de behoefte om meer-dimensionale integralen te schatten. Daarom wordt in dit hoofdstuk eveneens ingegaan op (Markov-keten) Monte Carlo methoden.

In hoofdstuk 3 ontwikkelen we een imputatiemethode die gebruik maakt van de zogeheten Dirichlet verdeling. Deze verdeling is gedefinieerd op het simplex, wat betekent dat de verschillende elementen van de data vector $\mathbf{Y}_i$ niet negatief zijn en optellen tot 1. Dit betekent dat de te imputeren bedrijfsgegevens, $\mathbf{X}_i$, zodanig zullen moeten worden getransformeerd dat zij eveneens in het sim-

plex liggen. Dit doet men door de verschillende variabelen in een gelijkheidsbeperking te delen door het totaal in die beperking. Zo worden bijvoorbeeld in het geval van de bedrijfslasten alle verschillende bedrijfslasten, zoals personeelskosten, vervoerskosten en de inkoopwaarde van goederen gedeeld door het totaal bedrijfslasten. We maken gebruik van deze verdeling omdat zij zeer flexibel is en bovendien tot een exponentiële familie behoort, waardoor het EM algoritme eenvoudig toe te passen is. In dit hoofdstuk laten we zien hoe parameterschattingen voor het model verkregen kunnen worden met behulp van het EM algoritme. Vervolgens stellen we zowel een deterministisch als een stochastisch imputatiemodel op. Uit de resultaten op basis van empirische gegevens blijkt dat de methode kan concurreren met niet-parametrische methoden en op sommige punten zelfs betere resultaten oplevert.

Er kleven echter twee nadelen aan deze aanpak. Allereerst moet de variabele die het totaal representeert altijd volledig geobserveerd zijn om zo de gegevens naar het simplex te kunnen transformeren. Verder kan er slechts één gelijkheidsbeperking per keer gemodelleerd worden, terwijl de gegevens vaak aan meerdere, onderling afhankelijke, gelijkheidsbeperkingen moeten voldoen.

Om ook te beschikken over imputatiemodellen die gelijktijdig met meerdere gelijkheidsbeperkingen rekening kunnen houden, onderzoeken we daarom in hoofdstuk 4 het gebruik van de singuliere normale verdeling. In dit geval veronderstellen we dat de (ongetransformeerde) data vector $\mathbf{X}_i$ normaal verdeeld is met een singuliere covariantie matrix: $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, met $\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$. In dit hoofdstuk leiden we een toepassing van het EM algoritme af voor de singuliere normale verdeling zodat de parameterschattingen die door het EM algoritme worden verkregen nog steeds overeenkomen met de geldende beperkingen. Dit betekent dat de geschatte verwachting $\hat{\boldsymbol{\mu}}$ aan de gelijkheidsbeperkingen voldoet ($\mathbf{A}\hat{\boldsymbol{\mu}} = \mathbf{0}$) en dat de geschatte covariantiematrix $\hat{\boldsymbol{\Sigma}}$ dezelfde nulruimte heeft als $\boldsymbol{\Sigma}$ ($\hat{\boldsymbol{\Sigma}}\mathbf{A} = \mathbf{0}$). Vervolgens stellen we wederom zowel een deterministisch als stochastisch imputatiemodel op. Een groot voordeel van deze methode is dat de gelijkheden, waaraan de gegevens moeten voldoen, niet gespecificeerd hoeven te worden omdat deze verankerd zijn in de singulariteit van de covariantiematrix. Dit betekent dat de methode met complexe gelijkheidsbeperkingen vol onderlinge afhankelijkheden om kan gaan.

Ook deze methode is getest op verschillende empirische datasets en presteert goed met betrekking tot het verkrijgen van accurate puntschattingen. Dit is in het bijzonder het geval als de gegevens aan een groot aantal gelijkheidsbeperkingen moeten voldoen. Een nadeel van deze methode is echter dat het singuliere normale model geen rekening kan houden met ongelijkheidsbeperkingen. Omdat bij bedrijfsgegevens deze beperkingen vaak in combinatie met gelijkheids-

beperkingen voorkomen is er behoefte aan een model dat simultaan rekening kan houden met zowel gelijkheids- als ongelijkheidsbeperkingen.

Daarom wordt in hoofdstuk 5 de singuliere normale verdeling, die gebruikt is in hoofdstuk 4, afgeknot tot het gebied dat gedefinieerd wordt door de ongelijkheidsbeperkingen. Dit betekent dat we veronderstellen dat de data vector $\mathbf{X}_i$ een afgeknotte singuliere normale verdeling volgt: $\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, waarbij $\boldsymbol{\Sigma}\mathbf{A} = \mathbf{0}$ en $\mathbf{l} \leq \mathbf{B}\mathbf{X}_i \leq \mathbf{u}$, met $\mathbf{u}$ en $\mathbf{l}$ respectievelijk de boven- en ondergrenzen. Dit imputatiemodel kan nu met allerlei verschillende beperkingen omgaan en is dus een generieke procedure. We leiden wederom een schattingsprocedure af. Het afknotten van een verdeling heeft helaas als resultaat dat de schattingsprocedure nu gebruik maakt van verschillende meer-dimensionale integralen. Ondanks het feit dat deze vrij accuraat geschat kunnen worden met (Markov-keten) Monte Carlo methoden zal de procedure hierdoor veel rekentijd kosten, wat de generaliteit reduceert. Bovendien introduceert het gebruik van het EM algoritme ook nog enkele meer-dimensionale integralen, waardoor de rekentijd van de procedure nog meer wordt verhoogd.

Een voordeel van deze methode is het feit dat de afgeknotte singuliere normale verdeling tot dezelfde parameterschattingen leidt als het niet-singuliere geval, wat we in dit hoofdstuk aantonen. Dit betekent dat deze aanpak eenvoudig te gebruiken is in het geval van zowel gelijkheids- als ongelijkheidsbeperkingen.

Bij toepassing op empirische gegevens blijkt echter dat deze imputatiemethode sterk afhankelijk is van modelveronderstellingen, zoals multivariate normaliteit, die in de praktijk vaak niet zullen gelden. Hierdoor is de methode, ondanks het feit dat er rekening kan worden gehouden met allerlei typen beperkingen, verminderd toepasbaar.

Om deze reden wordt in hoofdstuk 6 het gebruik van een sequentiële imputatieprocedure onderzocht. In dit geval worden, in plaats van één multivariaat model, verschillende univariate, conditionele verdelingen gebruikt om de gegevens sequentieel te modelleren en te imputeren. Deze univariate, conditionele verdelingen zijn gebaseerd op regressiemodellen. Dit is een iteratief proces dat wordt herhaald tot de parameterschattingen stabiliseren. In dat geval worden de imputaties verondersteld trekkingen uit de bijbehorende multivariate verdeling te hebben benaderd. Omdat er voor iedere variabele een verschillend regressiemodel opgesteld kan worden, is de procedure heel flexibel en kunnen ongelijkheidsbeperkingen en andere variabele-specifieke eigenschappen goed gemodelleerd worden.

Een nadeel van deze aanpak is echter het gebrek aan een theoretische basis. De univariate, conditionele verdelingen worden gekozen zonder deze te baseren op een multivariaat model. Het kan dus zijn dat trekkingen uit deze verschil-

lende verdelingen niet zullen convergeren tot trekkingen uit een multivariate verdeling, een verschijnsel dat men incompatibiliteit noemt. Omdat momenteel nog niet volledig duidelijk is wat de gevolgen van eventuele incompatibiliteit zijn, is voorzichtigheid geboden wanneer men deze methode toe wil passen.

Een andere belangrijke kwestie is het feit dat dit imputatiemodel niet zonder meer rekening kan houden met gelijkheidsbeperkingen aangezien gegevens die aan een gelijkheidsbeperking moeten voldoen op univariaat niveau vastliggen. Dit wordt opgelost door in iedere iteratie één variabele in een gelijkheidsbeperking buiten beschouwing te laten en de overgebleven gegevens te imputeren volgens de bovengenoemde imputatieprocedure. Vervolgens wordt de weggelaten variabele afgeleid op basis van de imputaties die zijn verkregen in deze iteratie.

Uit de toepassing op empirische gegevens blijkt dat deze methode tot goede resultaten leidt. Dus ondanks het feit dat een volledige theoretische basis ontbreekt lijkt deze methode veelbelovend. Temeer omdat dit de enige ontwikkelde imputatiemethode is die rekening kan houden met alle geldende beperkingen én in de praktijk toepasbaar is.

De imputatiemethoden die in dit proefschrift zijn ontwikkeld bieden de gebruiker een set van mogelijkheden die van nut zullen zijn om gegevens te imputeren die aan allerlei beperkingen moeten voldoen. In de praktijk zal het zo zijn dat er gebruik wordt gemaakt van een imputatiestrategie waarbij voor een bepaalde dataset niet één, maar een aantal verschillende imputatiemethoden worden gebruikt. Wij denken dat de besproken methoden op verschillende plaatsen in zo'n strategie van pas zullen komen.