Workpackage 3

Imputation of Knowledge Economy Indicators

Deliverable 3.2

2008

List of contributors:

Tobias Enderle and Luis Huergo, University of Tübingen, Ralf Münnich, University of Trier.

Main responsibility:

Ralf Münnich, University of Trier.

CIS8–CT–2004–502529 KEI

The project is supported by European Commission by funding from the Sixth Framework Programme for Research.

`http://europa.eu.int/comm/research/index_en.cfm`

`http://europa.eu.int/comm/research/fp6/ssp/kei_en.htm`

`http://www.cordis.lu/citizens/kick_off3.htm`

`http://kei.publicstatistics.net/`

Preface

A major goal of the KEI project was to develop a composite indicator for the Knowledge Economy, based on the political framework, adequate economic definitions, as well as on state-of-the-art methodology. The study examines 125 (single) indicators with data from 25 European countries plus the US and Japan. The main focus was laid on recent data coming from the time period 2001 to 2004.

As one important output of the project, particular attention was paid to the research of analysis of aggregation issues and the behavior of the resulting composite indicator. In order to adequately build a composite indicator or the KEI composite indicator (cf. Saisana and Munda, 2008) the availability of reliable data sources and especially of a complete dataset is an essential basis for the study. Knowledge indicator data, however, in general come from many sources, possibly conducted in different years and hence yield datasets with many unobserved or even unobservable values.

One major aim of the research within deliverable 3.2 of the KEI project was to investigate means to adequately handle the large amount of missing values containing the above mentioned challenging structure. In order to accommodate this structure and to provide the necessary methods for accuracy measurement while accounting for the variability of the missingness mechanisms within the imputation process, special multiple imputation methods were investigated and implemented for handling the missing values within the KEI dataset properly.

The main work of this deliverable was based on theoretical findings which were developed and discussed in Huergo (2008). The implementation of the procedures on the dataset was part of the work shown in Enderle (2008). One additional output of the work within this deliverable was a multiply imputed KEI dataset which was used in the research of the workpackages 5 and 7.

The authors would like to thank the KEI team for many valuable comments which helped to improve the generation of the final multiply imputed dataset.

Contents

List of Figures

List of Tables

List of Symbols

The most important abbreviations and symbols, sorted by topic and in order of their appearance in the text:

Chapters 2 and 4

Y	Data matrix
Y_{obs}	Observed data
Y_{mis}	Missing data
$P(\cdot \mid \theta)$	Probability function
$f(\cdot \mid \theta)$	Density function
R	Indicator matrix
ξ	Parameter of the missingness mechanism
θ	Parameter of the data model
$L(\theta \mid \cdot)$	Likelihood-function
$l(\theta \mid \cdot)$	Loglikelihood-function
Θ	Parameter vector of a bivariate normally distributed random variable
G, H	Parameter matrices
SWP	Sweep operator
$RSWP$	Reverse sweep operator
z	Random variable
μ, Σ	Expectation vector, covariance matrix
T	Sufficient matrices
$\mathcal{O}(s)$	Column information about observed values of pattern S
$\mathcal{M}(s)$	Column information about missing values of pattern S
$\mathcal{I}(s)$	Information vector of pattern S
A	Swept θ matrix
$W^{-1}(m, \Lambda)$	Normal inverted Wishart distribution
τ, μ_0, m, Λ	Parameter of the Wishart distribution

Chapter 3

y^*	Transformed variable
θ	Power transformation parameter
z	Z-score
$\overline{m}$	Vector of moment conditions
v	Dataset with p indicators

Chapter 5

Ψ	Scale matrix
μ	Location vector
ν	Degrees of freedom
q	Unobserved scalars
w	Weighting vector
d^2	Mahalanobis distance
δ	Probability of contamination
λ	Separation parameter of contamination

Chapter 6

v	Imputation model: dataset with p indicators
D_v	Distance matrix of the v-th imputation model
ρ_v	Correlation matrix of the v-th imputation model
m01-m46	Imputation models within the KEI project
$\overline{r}_v$	Averaged correlation of the w-th model
κ_v	Model score
α	Weight for the convex combination
k	Number of multiple imputations

Chapter 1

Introduction

1.1 Indicators and Imputation

Nowadays, indicators are applied in many areas, e.g. as the foundation for economic and political decisions. In general, two goals are connected with the use of indicates, measuring performance and development of different units of interest, like countries. In order to fulfil these tasks, a complete and reliable database is an essential input. In practice, however, datasets of this size contain missing values which have to be treated adequately. Improper treatment of missing values may yield wrong values of indicators and especially of composite indicators.

The project's dataset consists of 125 indicators from 25 European countries plus US and Japan. The time period of interest was set to 2001 – 2004. The entire dataset, which is based on macrodata, contains approximately 42% missing values to which we refer as *NA*. Partially, this results from the fact that some surveys are not conducted every year which per se yields a large amount of missing values. A closer description of the dataset is given in Appendix A.1.

The treatment of missing values generally leads to either weighting or imputation methods. Whereas weighting rules are generally applied to cases where complete units of interest are unobserved in micro data, imputation methods can be applied in many cases. The idea of imputation is to fill in the missing value with a sensible value. The major question is dedicated to the term sensible.

On macro level data, last value carried forward is often applied. This could be viewed as a naive forecast or nowcast. A statistically plausible advance might be the application of time series methods. In the KEI context, this will hardly be possible due to the high dimensional sophisticated interactions based on a very short time horizon.

One method to overcome these difficulties which also enables the user to measure the accuracy of the outcome of the research on indicators is the multiple imputation approach developed by Rubin in the 1970ies (cf. Rubin, 1987, or Little and Rubin, 2002). Thus, dealing with missing data and especially with multiple imputation is a key issue of the project.

1.2 Justification of the Methods Used

Some of the most established and practically most relevant methods that occur in literature (e.g. KIM and CURRY, 1977; ROTH, 1994) when dealing with incomplete datasets are discussed at first. Due to the bad properties they come along with, alternative but more complex methods that have become increasingly popular will be introduced. And within the KEI project several extensions and improvements have to be developed to cope with the adverse data situation.

The *listwise deletion method* is still the most commonly used method, not least because of its simple applicability. All observations with incomplete values are excluded from the dataset which results in a shrunk but completely observed dataset. In doing so, much utilizable information is lost. Especially in the case of multivariate datasets the remaining observations tend to shrink enormously. However, the damage of this loss in efficiency caused by the cutback of the data is not the only one. If the mechanism of the missingness in the data is related to the parameters of interest, then deleting observations with missing values can cause a remarkable systematic bias. In addition to that, in the KEI setting, deleting rows (i.e. countries) is all but impossible.

The *omitted variable method* differs just insofar, as instead of observations this time variables with missing values are excluded. The deletion of columns (i.e. KEI indicators) is also not desired.

A further approach is the single imputation of missing values, whereas in most cases this is done in a relatively subjective manner. There are several ways to carry out a single imputation: The missing values can be imputed by an ad hoc value (*ad hoc imputation*), by the corresponding mean of the observed values (*mean imputation*) unconditionally or conditionally given some observed covariates, or by searching for some observed data which correspond to the missing data according to certain criteria and using them as proxies for the missing values (*proxy or hot deck imputation*). In spite of their simplicity the use of these methods is not advisable since they tend to yield invalid inferences (KOFMAN and SHARPE, 2003).

In order to circumvent this problem, the missing values should be imputed with certain restrictions given by LITTLE and RUBIN (2002, p. 72). According to them, imputations should be:

(a) Conditional on observed variables

As already featured by Buck's method (BUCK, 1960), which can be seen as a precursor of the single imputation methods applied here, conditional imputes on observed values improve the imputation of missing values. In doing so, following improvements will be achieved:

- reduced bias due to non-response,
- improved precision and
- preserved association between missing and observed variables.

(b) Multivariate

Imputations in multivariate settings make sure that associations between missing variables will be preserved.

(c) Draws from the predictive distribution rather than means

Indeed, imputing means from a predictive distribution yields consistent estimates but tends to systematically underestimate the variability and leads to invalid inferences. Thus, the imputations should be random draws from a predictive distribution *to provide valid estimates of a wide range of estimands* (Little and Rubin, 2002, p. 72).

(d) Multiple

Because inferences achieved by single imputation methods don't account for imputation uncertainty, values have to be imputed multiply.

To cope with these requirements the used imputations rely on two powerful groups of methods called *Expectation Maximization* Algorithm (EM) and *Markov Chain Monte Carlo* methods (MCMC). The use of these techniques ensures that the listed conditions are met.

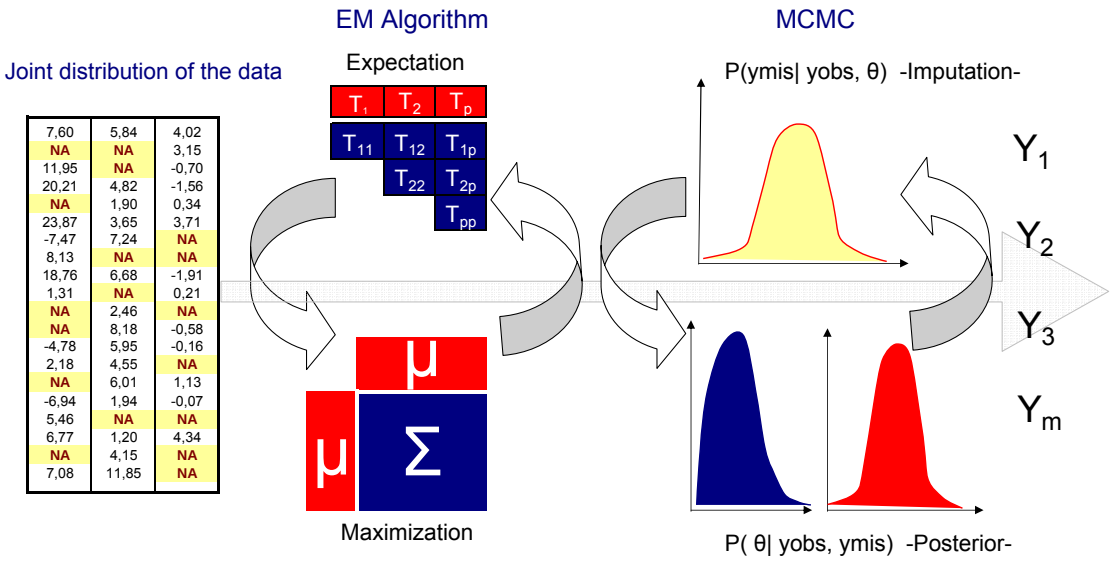

Figure 1.1: General imputation schema.

While the EM algorithm is used to search for the optimal parameters of a multivariate dataset with missing values, the MCMC algorithm yields m imputes for each missing value. (Source: Huergo, 2008.)

Since the KEI dataset consists of continuous indicators the proposed imputation models are based upon the multivariate normal model. In this respect Schafer states:

> *The most common probability model for continuous multivariate data is the multivariate normal distribution. [...] Moreover, the classical techniques of linear regression [...] assume conditional normality of the response variables given linear functions of the predictors, which is the conditional distribution*

> *implied by a multivariate normal model for all the variables. Because statistical methods motivated by assumptions of normality are in such widespread use, it is natural to seek general techniques for inference from incomplete normal data.* (SCHAFER, 1997, p. 147)

However, due to the problems regarding the bad underlying data situation, some adjustments to the base model have to be discussed. Frequently occurring problems are:

1. The presence of outliers (because of the varying construction of the indicators between countries)
2. Important departures from the Normal Distribution
3. Presence of Variables with a strictly positive domain
4. Non-linear relationships between variables
5. Small sample sizes
6. High proportion of missing values (NAs)

Here are some proposals on how to handle these problems (HUERGO, 2008):

1. **The presence of outliers:** Various further developments of the base model for multivariate normal data, created by different authors (LIU and RUBIN, 1995; LANGE et al., 1989; LITTLE, 1988; LIU, 1995): $t-$*model, adaptive* $t-$*Model* and *contaminated normal model*, can be implemented to deal with outliers. By means of a selective weighting of the observations, these methods are often able to yield robust parameter estimates of the multivariate normal distribution.

2.-3. **Departures from the Normal Distribution:** Numerous indicators do not seem to follow a normal distribution. Furthermore, many of them are defined on $\mathbb{R}^+$, i.e. negative values are not valid. To adjust the empirical distributions to approximative normality and to avoid imputing invalid values, the EM algorithm was expanded by an adaptive power transformation, which is based upon a *Generalized Method of Moments* (GMM). This power transformation is often able to transform arbitrarily distributed variables into a symmetric, bell shaped distribution, and thus to provide a better estimation of the multivariate model.

4. **Non-linear relationships between indicators:** Non-linear relationships between the indicators can cause biases when imputing linearly, as done by the proposed algorithms. To deal with this problem, the power transformation has to be expanded such that an approximation to a *multivariate* normal distribution can be achieved. Because of the linear dependence structure of the multivariate normal distribution (SCHAICH and MÜNNICH, 2001, p. 78, Def. 2.-21), linear imputation procedures yield correct imputed values.

5. **Small sample sizes:** Because of the already mentioned small sample size of the KEI dataset, previous observations of preliminary years have been taken into account. The dataset has thus a panel structure which has to be accounted for.

   Different imputation methods of how to deal with panel data are discussed in the literature (e.g. Nijman and Verbeek, 1992). State-of-the-art methods are commonly based upon the so called *Mixed Effects* models (Schafer and Yucel, 2002). These typically bayesian methods imply the existence of a joint normal distribution, which, as already mentioned, cannot be assumed for the KEI dataset. Further developments of this panel methods to account for departures from normality turn out to be complicated and it is not clear whether the underlying small sample size can justify this complexity.

   The proposed implementation of the panel structure can be realized by a simple extension of the models proposed by Schafer (1997) and Little and Rubin (2002). In this extension, a dummy variable will be included for each year and will be treated and modeled as an additional *design* variable. This extension retains the robustness of the methods, namely the possibility of neutralizing outliers via selective weighting.

   Of course, the marginal distributions of these dummy variables deviate from the normal distribution. The fact that such a strategy does not cause any estimation problems is stated by Schafer (1997, p. 35):

   - these dummy variables are totally observed (i.e. no imputation of them will be necessary) and
   - under the existence of a joint normal distribution of the remaining variables, all conditional distributions, given these dummy variables, are normal. This is ensured by the linearity of the regression curves.

   With this extension of the model, the time dimension can be taken into account in a very parsimonious manner.

   At times, modeling of time effects via dummy variables has been contested. An application of the proposed methods to dummy variables is regarded as adverse since standard errors are slightly diluted (Bollen, 1989, pp. 375). But as mentioned above, is the question justified, how much sophistication in the modeling of the time effects can be compatible with the observed data (especially with regard to the small sample sizes).

   Furthermore, the trivial task of inserting and implementing dummy variables within linear regression models turns out to be challenging for the MCMC methods being used.

6. **High proportion of NAs:** There are a lot of countries in the dataset that provide no values for certain indicators. Furthermore, some years, values for a given indicator are missing for all countries. Therefore, applying approaches such as *repeated measures models* (Schafer, 1997, pp. 379), which follow every observation in the course of time, is not possible. Hence, models based upon the i.i.d. assumption are chosen to ensure feasible imputations.

1.3 Remaining Problems

Although the proposed solutions expand the basic imputation method to account for the underlying data situation, there are some remaining problems:

1. **The i.i.d. assumption:** The i.i.d. assumption is not very realistic for macro data like the KEI dataset, because the countries normally exhibit regional dependences.
2. **Transformation:** The power transformation is based upon some higher moments of a distribution and thus is suitable for large sample sizes (HAYASHI, 2000, p. 215). The dataset for the construction of composite indicators has by definition a small sample size.
3. **Selective weighting:** The robust methods don't allow for selective weighting of the rows' elements (i.e. different indicators).
4. **No random effects:** The heterogeneity of the countries cannot be modeled separately.

However, even when the proposed methods exhibit these shortcomings, they are designed to yield consistent estimates under adverse conditions.

1.4 Implicit Methods and Convexcombination

This work favors methods which are based upon a solid statistical theory, which is evidenced by the implementation and adaptation of MCMC methods. However, the situation in the KEI project is unique insofar as many first-time studies concerning the indicators' behavior had to be carried out on *reconstructed* data. Therefore, it seems to be more important to deliver robust estimates of the missing values rather than to make sure that their standard errors are underestimated (which can be seen as the typical bone of contention of single imputation methods (LITTLE and RUBIN, 2002, p. 61)).

Because of this reason, an additional imputation method which belongs to the implicit or ad hoc methods was implemented (LITTLE and RUBIN, 2002, p. 60). This conservative and model-free method carries out (1) a *spline interpolation* of those missing values that are surrounded by observed values, and (2) a *Last Value Carried Forward* (LVCF) extension to those that have no coterminous observation in the previous or successive year.

The method is conservative insofar as the data are not extrapolated. Thus, merely values which are either observed for a certain combination of indicator and country or which are a result of a spline interpolation of observed values will be imputed.

It has to be mentioned, that this method is not able to provide values for all missing values. The difficulty is that various countries have delivered absolutely no values for certain indicators in the considered time slice, thus the i.i.d. assumption to enhance the possibility of imputations.

The results of the different imputation methods will be aggregated by a convex combination, whose weights can be seen as a function of the number of observations per indicator and country. These weights reflect the trustworthiness of an imputation based on the observed data for one country and indicator.

1.5 Overview

This work is organized as follows. Chapter 2 begins with an introduction to the necessary basic assumptions and the background on the EM and MCMC methods. Chapter 4 presents a proposed approach on how to transform the KEI indicators, so that Chapter 3 can present the application of these methods to multivariate normal data. But due to the presence of outliers, these models must be expanded by robust extensions, which will be discussed in Chapter 5. Finally, Chapter 6 presents the final imputation round.

1.6 Remarks

The procedures and algorithms presented and applied in this work were implemented with R, a language and environment for statistical computing and graphics. A comprehensive overview including examples appears in the books of Ligges (2006) and Rizzo (2008). Applications of Bayesian computing are presented by Albert (2007); an exposition on the R graphical system by Murrell (2005).

Chapter 2

EM and Data Augmentation

2.1 The Complete Data Model

One very important issue when dealing with missing values is the study of the underlying mechanism which causes the data not to be observed. Indeed, the missingness can occur for different reasons, for example non-sampling entry errors. A problem that often arises is that certain questions (items) in a survey are left unanswered. This is referred to *item non-response*. A typical example is income. So for instance it is assumed that people with a high monthly wage fail to name their income because of tax implications.

To formalize the treatment of the subject, let define the $n \times p$ data matrix $Y = (y_{ij})$, where the rows stand for n observations ($i = 1, ..., n$) and the columns represent the p variables ($j = 1, ..., p$). If all data were available and under the assumption that the rows are independent and identically distributed (i.i.d.), the joint probability function of the data could be written as follows:

$$P(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \ . \tag{2.1}$$

In the presence of missing data, the probability function cannot be stated like that any more. Hence, the data matrix Y has to be split into two components: The observed data Y_{obs} and the missing data Y_{mis}, with $Y = (Y_{obs}, Y_{mis})$ (Schafer, 1997, pp. 10). To distinguish between the observed and missing components of the data matrix Y, a $n \times p$ indicator matrix $R = (r_{ij})$ has to be introduced. These indicator variables determine which values in Y are available and which are not:

$$R = (r_{ij})_{\substack{i \in \{1,...,n\} \\ j \in \{1,...,p\}}} \text{ with } r_{ij} = \begin{cases} 1, & \text{if missing value} \\ 0, & \text{if observed value.} \end{cases}$$

That is, an observed value y_{ij} results in $r_{ij} = 0$ and a missing y_{ij} in $r_{ij} = 1$. To get a complete data set, imputation refers to the simulation of the unobserved component in Y, that is Y_{mis}.

2.2 Missing Data Mechanism

2.2.1 Classification

As mentioned in Rubin (1987), the missing data mechanism is characterized by the conditional distribution of R given Y: $f(R \mid Y, \xi)$, where ξ describes a typically unknown parameter related to the missingness mechanism. Differences in the properties of this conditional distribution of the missingness indicator R now allows to classify such data into three categories (Rubin, 1976):

(a) Missing Completely at Random (MCAR): The missing values are a random sample of all values. That means, the missingness in Y doesn't depend on Y_{obs} or Y_{mis}.

$\Longrightarrow f(R \mid Y, \xi) = f(R \mid \xi)$

(b) Missing at Random (MAR): MAR is a weaker or less restrictive assumption for missing data. The missingness here just depends on Y_{obs}, but not on Y_{mis}.

$\Longrightarrow f(R \mid Y, \xi) = f(R \mid Y_{obs}, \xi)$

(c) Not Missing at Random (NMAR): The missingness in Y depends on Y_{mis} and cannot be explained only by Y_{obs}.

$\Longrightarrow f(R \mid Y, \xi) = f(R \mid Y_{obs}, Y_{mis}, \xi)$

It must be pointed out that these definitions are not restrictions on the *pattern of missingness*, but that they describe how the missingness depends on the values of all data, both missing and observed.

Applying the MAR assumption, the relation between variables and the missingness in other variables can be used to impute missing values. Thus, finding correlations becomes a very important task in this framework. To tackle the previous income example, credit card expenditures could be used as covariate, making the MAR assumption more tenable.

Throughout this work, it will be assumed that the missing values are generated by a MAR mechanism.

2.2.2 Distinctness of the Parameters

Furthermore, it is assumed that the parameter of the data model, θ, and the parameter of the missingness mechanism, ξ , are *distinct* from each other. Distinctness means that the joint parameter space of (θ, ξ) corresponds to the cartesian product of the two parameter spaces of θ and ξ.

2.2.3 Ignorability

If both assumptions - MAR and distinctness between θ and ξ - occur, the missing data mechanism is said to be *ignorable*.

This *ignorability* assumption is of great value for the Maximum Likelihood (ML) based inferences on the parameters θ of the data matrix Y. For this purpose, one has to examine the probability function of the observed data

$$P(R, Y_{obs} \mid \theta, \xi) = \int P(R, Y \mid \theta, \xi) dY_{mis} = \int P(R \mid Y, \xi) P(Y \mid \theta) dY_{mis} \ . \tag{2.2}$$

The indicator matrix R has to be taken into account, to express the observed data. Under a valid MAR assumption, Equation (2.2) can be transformed into:

$$\begin{aligned} \int P(R \mid Y, \xi) P(Y \mid \theta) dY_{mis} &= P(R \mid Y_{obs}, \xi) \int P(Y \mid \theta) \, dY_{mis} \\ &= P(R \mid Y_{obs}, \xi) \, P(Y_{obs} \mid \theta) \ , \end{aligned}$$

that is, the probability can be presented by two factorizable parts. With the further assumption of *distinctness*, the likelihood based inference regarding θ is independent of ξ, and thus also of the former factor $P(R \mid Y_{obs}, \xi)$. That is, the missingness-data mechanism can be ignored. For the observed-data likelihood holds then

$$L(\theta \mid Y_{obs}) \propto P(Y_{obs} \mid \theta) \ , \tag{2.3}$$

the following likelihood can be maximized

$$L_{ign}(\theta \mid Y_{obs}) = P(Y_{obs} \mid \theta) \ .$$

Whereas the complete likelihood is defined to be

$$L_{full}(\theta, \xi \mid Y_{obs}) = P(Y_{obs}, R \mid \theta, \xi) \ .$$

Due to the ignorability assumption, the inferences, which are based upon L_{ign}, are equivalent to the ML estimation, which are based upon L_{full}. Thus, the parameters for the whole data matrix $Y = (Y_{obs}, Y_{mis})$ can be computed in a much easier manner.

Inference in case of multiple imputations (i.e. the Bayesian framework) will be based on the posterior distribution when the missing-data mechanism is ignorable. This is the case if the missing data are MAR and (according to the assumption of distinctness)

$$P(\theta, \xi) = P(\theta) P(\xi) \ ,$$

that is, the parameters θ and ξ are a priori independent.

2.2.4 Problems

Restricting to a method generates problems, such as violations of the met assumptions. Here three of them are listed and discussed:

(a) The MAR assumption

It is said that if the data are missing because of the design of the research (*missing by design*) the data satisfy the MAR assumption because the researcher didn't intend to collect all data. When there are missing data, it is not always reasonable to apply a missing-data method, because the missingness doesn't necessarily imply that values are merely missing. And in the literature (e.g. Krosnick et al., 2000) a wide range of reasons are given why respondents fail to answer.

When the missingness is noncontrollable, it isn't safe to say if the MAR assumption is appropriate. To do a formal test, the missing data, or at least a sample of it, has to be available externally. If that is not feasible, all findings are based upon heuristic assumptions.

When the missing data are NMAR the missing-data mechanism cannot be ignored and the likelihood must be properly included in the analysis.

(b) The i.i.d. assumption

For the probability function in Equation (2.1), the rows need to be independent and identically distributed. The following chapters will describe an algorithm, which is intended for multivariate normal data and which assumes that observations fulfill the i.i.d. assumption.

(c) The ignorability assumption

A violation of the ignorability assumption does not automatically imply a break down of missing-data methods, especially not in case of multivariate settings. Schafer agrees with David et al. (1986) and draws the conclusion, that

> *improvements in missing-data procedures would probably come from better modeling of the multivariate structure of the data, not from nonignorable modeling* (Schafer, 1997, p. 27).

For more details see Schafer (1997, pp. 20), who extensively treats violations of these assumptions.

2.3 Maximum Likelihood Method

Figure 2.1 depicts the following situation: A certain phenomenon has to be described as best as possible with a specific statistical distribution. For this purpose a sample of data, which are the result of a random experiment, is available (see the red crosses on the abscissa). Although the functional form of the distribution is assumed to be known, their parameters are not. The task is thus the estimation of these parameters.

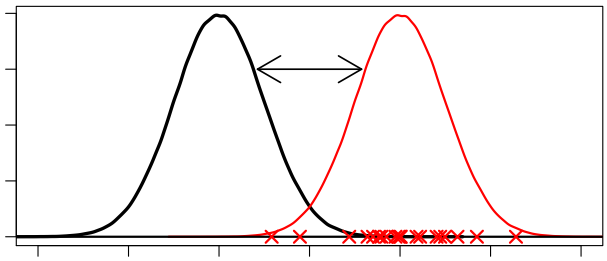

Figure 2.1: Maximum Likelihood schema.

For a sample y of observed data from a distribution known up to a parameter (vector) θ the Maximum Likelihood method estimates the unknown parameters in a way that these estimates assign maximal likelihood to the observed data (see the horizontal adjustment in Figure 2.1).

Since the data are assumed to be independent and identically distributed, their likelihood can be written as

$$L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta) \ .$$

The Maximum likelihood estimator is then

$$L(\widehat{\phi}|y) = \sup_{\phi \in \theta} L(\theta|y) \ .$$

A complete summary of the ML method is given in GREENE (2003, Ch. 17).

One drawback of this method is that it requires the derivatives of the log likelihood function to be computed, which often turns out to be difficult. Thus, numerical or iterative methods have to be used. One example of an iterative method which does not require the calculation of derivatives is the EM algorithm.

2.4 The EM Algorithm

2.4.1 Overview

Consider a data matrix Y_{obs}, which was generated out of a specific probability distribution. The parameters are however unknown and the aim is precisely to find the parameters that describe the distribution of the data as best as possible. The EM algorithm (DEMPSTER et al., 1977) iteratively yields Maximum Likelihood estimates (MLEs), even if there are incomplete data. This algorithm can be split up in two steps:

(a) Expectation-step (e-step)

The **e-step** finds the expectations of the log-likelihood of θ, $l(\theta \mid Y)$, where the expectation regarding Y_{mis} conditioning Y_{obs} and $\theta^{(t)}$, has to be calculated:

$$Q(\theta \mid \theta^{(t)}) = \int l(\theta \mid Y) P(Y_{mis} \mid Y_{obs}, \theta^{(t)}) dY_{mis} \; .$$

(b) Maximization-step (m-step)

The **m-step** then finds the updated value

$$\theta^{(t+1)} = \arg\max_{\theta} \; Q(\theta \mid \theta^{(t)}) \; . \tag{2.4}$$

The updated $\theta^{(t+1)}$ replaces the $\theta^{(t)}$ in the e-step and $\theta^{(t+2)}$ maximizes $Q(\theta \mid \theta^{(t+1)})$. This routine is repeated until the sequence $\left\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\right\}$ converges.

2.4.2 Theory

This section discusses the functioning of the EM algorithm based on the book of Schafer (1997, pp. 38).

As long as the probability function in the *incomplete-data* problem of Y, i.e. $P(Y \mid \theta) = P(Y_{obs}, Y_{mis} \mid \theta)$, can be split in two factorizable terms, then following holds

$$P(Y \mid \theta) = P(Y_{obs} \mid \theta) \, P(Y_{mis} \mid Y_{obs}, \theta) \; . \tag{2.5}$$

Making use of Equation (2.3), the likelihood functions of θ given Y or Y_{obs} can be stated as

$$\begin{aligned} L(\theta \mid Y) &= P(Y \mid \theta), \\ L(\theta \mid Y_{obs}) &= P(Y_{obs} \mid \theta) \; , \end{aligned}$$

and inserted into Equation (2.5)

$$L(\theta \mid Y) = L(\theta \mid Y_{obs}) \, P(Y_{mis} \mid Y_{obs}, \theta) \; . \tag{2.6}$$

Taking the logarithm of each side in Equation (2.6)

$$\log L(\theta \mid Y) = \log L(\theta \mid Y_{obs}) + \log P(Y_{mis} \mid Y_{obs}, \theta) + c \; ,$$

it follows that

$$l(\theta \mid Y) = l(\theta \mid Y_{obs}) + \log P(Y_{mis} \mid Y_{obs}, \theta) + c \; , \tag{2.7}$$

where $l(\theta \mid Y)$ is the *complete-data* loglikelihood, $l(\theta \mid Y_{obs})$ the *observed-data* loglikelihood and c an arbitrary constant. Of high importance is the expression $P(Y_{mis} \mid Y_{obs}, \theta)$, which is described as the conditional predictive distribution of the missing data given

parameter θ. It describes the interdependence between Y_{mis} and θ: Considered as a probability distribution, it imparts knowledge about Y_{mis} for any assumed value of θ. Whereas considered as a function of θ, it provides information about θ, which is contained in Y_{mis}.

Now one has to find the MLE of θ, that is, the value of θ which maximizes $l(\theta \mid Y_{obs})$. Equation (2.7) shows, that this is equivalent to searching for the the value for θ that maximizes $l(\theta \mid Y) - \log P(Y_{mis} \mid Y_{obs}, \theta)$, as long as

$$l(\theta \mid Y_{obs}) = l(\theta \mid Y) - \log P(Y_{mis} \mid Y_{obs}, \theta) \; . \tag{2.8}$$

Given that Y_{mis} is not observable, the above mentioned term cannot be computed. Therefore, one takes the average of Equation (2.8) over the predictive distribution $P(Y_{mis} \mid Y_{obs}, \theta^{(t)})$, where $\theta^{(t)}$ is a preliminary estimate of the unknown parameter. Hence it follows

$$\begin{aligned} & \mathrm{E}\,[\, l(\theta \mid Y_{obs})\,] &=& \mathrm{E}\,[\, l(\theta|Y)\,] - \mathrm{E}\,[\log P(Y_{mis} \mid Y_{obs}, \theta)\,] \\ \Leftrightarrow \quad & l(\theta \mid Y_{obs}) &=& Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}) \; , \end{aligned} \tag{2.9}$$

where

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \int l(\theta \mid Y)\, P(Y_{mis} \mid Y_{obs}, \theta^{(t)})\, dY_{mis} \\ H(\theta \mid \theta^{(t)}) &= \int \log P(Y_{mis} \mid Y_{obs}, \theta)\, P(Y_{mis} \mid Y_{obs}, \theta^{(t)})\, dY_{mis} \; . \end{aligned}$$

If the loglikelihood is linear in the data, then the expectations in Equation (2.9) are computable by imputing the missing data with their conditional expectation given the the observed data and some parameters. However, in general it is not the case.

2.4.3 Properties

DEMPSTER et al. (1977) show that when the likelihood is bounded the EM algorithm always converges to a stationary point. The algorithm constructs a sequence $\left\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\right\}$ so that the loglikelihood is a non-decreasing function

$$l(\theta^{(t+1)} \mid Y_{obs}) \geq l(\theta^{(t)} \mid Y_{obs}) \; , \tag{2.10}$$

which can also be written as follows

$$\begin{aligned} l(\theta^{(t+1)} \mid Y_{obs}) - l(\theta^{(t)} \mid Y_{obs}) &\geq 0 \\ Q(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)}) + H(\theta^{(t)} \mid \theta^{(t)}) &\geq 0 \\ \underbrace{Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})}_{A} + \underbrace{H(\theta^{(t)} \mid \theta^{(t)}) - H(\theta^{(t+1)} \mid \theta^{(t)})}_{B} &\geq 0 \; . \end{aligned}$$

This holds, because it can be shown that both quantities A and B are non-negative:

- **Quantity A:** This is non-negative, because $\theta^{(t+1)}$ in Equation (2.4) has been chosen to satisfy $Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)})$ for all θ.
- **Quantity B:** This is non-negative as well, as can be shown by Jensen's inequality (Sydsaeter et al., 2005, p. 67) and the concavity of the logarithmic function.

$$\begin{aligned} B &= -\mathrm{E}\left[\log \tfrac{P(Y_{mis}|Y_{obs},\theta^{(t+1)})}{P(Y_{mis}|Y_{obs},\theta^{(t)})} \,\middle|\, Y_{obs}, \theta^{(t)} \right] \\ &\geq -\log \mathrm{E}\left[\tfrac{P(Y_{mis}|Y_{obs},\theta^{(t+1)})}{P(Y_{mis}|Y_{obs},\theta^{(t)})} \,\middle|\, Y_{obs}, \theta^{(t)} \right] \\ &= -\log \int \tfrac{P(Y_{mis}|Y_{obs},\theta^{(t+1)})}{P(Y_{mis}|Y_{obs},\theta^{(t)})} \quad P(Y_{mis} \mid Y_{obs}, \theta^{(t)}) dY_{mis} \\ &= 0 \,. \end{aligned}$$

If the EM algorithm constructs a sequence $(\theta_t)_{t\in\mathbb{N}}$, then the sequence $(L(\theta_t))_{t\in\mathbb{N}}$ is monotone increasing. Is there a $M \in \mathbb{R}$ with $(L(\theta_t))_{t\in\mathbb{N}} < M$, i.e. L is bounded, so is the sequence $(L(\theta_t))_{t\in\mathbb{N}}$ already necessarily convergent (Königsberger, 2000, p. 46). More precisely, the sequence of the $\theta^{(t)}$'s leads to an increase of θ with each iteration step. But it can also be shown that, under relatively weak assumptions, a sequence $l(\theta)$ converges to $l(\widehat{\theta})$, where $\widehat{\theta}$ is a stationary point.

It cannot in general be guaranteed that the EM algorithm will converge to a maximum. That requires an unimodal and concave loglikelihood function over the whole parameter space θ.

To the advantages of the EM algorithm one can list its straightforwardness, its stability and its simple applicability, because the restrictions on the parameters are mostly fulfilled automatically and one obtains an iteratively increasing likelihood. But convergence can be very slow and depends particularly on the proportion of missing information. The linear rate of convergence is an often noted disadvantage, which cannot compete for example with the quadratic convergence of the Newton-Raphson method (see Deuflhard, 2004). Furthermore, standard errors are computable indirectly, though in a rather technical way. As with all iterative methods, the selection of starting values is an important issue.

However, it should be mentioned that the arising problems regarding saddle points, extrema, boundary estimation, etc. are not exclusive of the EM algorithm, but intrinsic to the maximum likelihood method. For an extensive description of the properties Schafer (1997, Ch. 3.3, pp. 51) and McLachlan and Krishnan (1997, Ch. 3, pp. 82) give a wider overview.

2.5 Data Augmentation

2.5.1 Bayes Statistics

Since the EM algorithm is a method for Maximum Likelihood estimation rather than an imputation method, it can only provide one estimate for each missing value and hence

cannot fill in missing data multiple times to account for imputation uncertainty. To derive a multiple imputation method, such as data augmentation, one must throw a glance at the Bayesian framework. Ideally, one draws from the predictive posterior distribution, $P(Y_{mis} \mid Y_{obs})$, which can be obtained by taking the average of the conditional predictive distribution $P(Y_{mis} \mid Y_{obs}, \theta)$ over the *observed-data* posterior distribution of the unknown parameters, θ. That is, the posterior distribution contains evidence of the observed-data posterior distribution of θ.

To obtain $P(Y_{mis} \mid Y_{obs})$, one has to gain knowledge about the posterior distribution, $P(\theta \mid Y)$. This distribution is a combination of the likelihood-function, $L(\theta \mid Y)$, and the prior distribution $P(\theta)$:

$$P(\theta \mid Y) \propto L(\theta \mid Y)\, P(\theta) \ ,$$

which can be obtained from Bayes' theorem (BAYES, 1958).

The posterior knowledge of θ can be regarded as prior knowledge of a θ which is modified by the likelihood-function. In order for the posterior distribution to be tractable and the resultant simulated parameter values to be obtained, a suitable prior (SCHAFER, 1997, pp. 154) distribution has to be chosen. The calculation of the posterior distribution can be done analytically or by using a Monte Carlo method.

2.5.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are employed to get random draws from a probability distribution, called the *target* distribution $P(Z)$. In cases where the direct simulation of this distribution turns out to be difficult and cannot be done directly, the idea is to construct a *Markov Chain* which is designed to have $P(Z)$ as its stationary distribution. Such a chain is a sequence of random variables $\{Z^{(1)}, Z^{(2)}, \ldots, Z^{(t)}, \ldots\}$, where each value only depends in some way on the immediately previous ones, which converges under certain conditions to the target distribution. In practice, to eliminate the dependence on starting values, a sufficiently large t or *burn-in* period K has to be chosen. The *burn-in* period is an initial number of iterations which are normally discarded, in order for the chain to lose its dependence on the starting values

A *Gibbs sampling* (GEMAN and GEMAN, 1984) is the best known and most implemented sampling method for MCMC. Consider a problem with a random vector partitioned into two subvectors $Z = (z_1, z_2)$ and suppose the joint distribution of Z, say $P(Z)$, exists. Then starting at some initial point, the sequence

$$\left\{(z_1^{(1)}, z_2^{(1)}), (z_1^{(2)}, z_2^{(2)}), \ldots, (z_1^{(t)}, z_2^{(t)}), \ldots\right\}$$

can be obtained by successively drawing from the distributions

$$\begin{aligned} z_1^{(s)} &\sim P(z_1^{(s)} \mid z_2^{(s-1)}) \\ z_2^{(s)} &\sim P(z_2^{(s)} \mid z_1^{(s)}) \ , \end{aligned}$$

where $(s = 1, 2, \ldots, t, \ldots)$. Thus the values at the $(s+1)$-*th* step depend entirely on the values at the sth step and, given those values, are independent of the previous history. Under mild regularity conditions, the Markov chain converges to a stationary distribution, which is the target distribution, that is $Z^{(t)} \xrightarrow{d} Z$ as $t \to \infty$. The method generalizes to any number of subvectors (see Schafer, 1997, Eq. (3.31)).

Example: The Gibbs Sampler for random number generation from a bivariate distribution with known parameters (taken from Huergo, 2008)

The only purpose of this example is to show the mode of operation of the Gibbs Sampler, mostly in a graphical way. Indeed, drawing i.i.d. numbers from a normal distribution represents no problem whatsoever and there are lots of algorithms which can accomplish this in an efficient way. However, it is for didactic purposes interesting to deal with a multivariate distribution, whose marginal and conditional distributions belong to the same distribution family.

Setup: Let Y be a bivariate normally distributed random variable, whose parameters Θ are assumed to be known. This parameter vector be made up of $\mu_1 = 1, \mu_2 = 0{,}5$, $\sigma_1^2 = 1, \sigma_2^2 = 1$ and $\rho = 0.5$. The aim of this example is to obtain drawings from this bivariate distribution without resorting to the joint density to accomplish it.

Additionally, let both conditional distributions

$$f_{Y_1|Y_2=y_2;\Theta} \sim N(\mu_1 + \rho\tfrac{\sigma_1}{\sigma_2}(y_2 - \mu_2);\ \sigma_1^2(1-\rho^2))$$

and

$$f_{Y_2|Y_1=y_1;\Theta} \sim N(\mu_2 + \rho\tfrac{\sigma_2}{\sigma_1}(y_1 - \mu_1);\ \sigma_2^2(1-\rho^2))$$

be available as well.

The Gibbs Sampler has to be initialized with a starting value y_1^0 for Y_1. Conditioning on y_1^0 the distribution $f_{Y_2|Y_1=y_1^0;\Theta}$ is completely characterized and it is possible to draw a value y_2^1 from it. Conditioning on this value, a value y_1^1 from $f_{Y_1|Y_2=y_2^1;\Theta}$ can in turn be drawn, which completes one cycle. This iteration scheme can be repeated until there are enough observations to characterize the joint distribution.

Figure 2.2 depicts the joint distribution and its marginals. Previous to the start of the simulation there are no conditionals drawn. The process in an advanced status, which can easily be recognized on the green dots underneath the marginal distribution, is illustrated in Figure 2.3. Both panels depict two consecutive steps of the simulation. The conditional distribution in the right panel is located exactly over the last value drawn, which is plotted in the left panel.

After 500 Iterations a comparison is drawn between the target distribution and the kernel density estimation of the drawn sample. Additionally, the marginals are compared to its empirical counterparts. Figures 2.4, 2.5 and 2.6 show the results of this comparison. Despite the small deviations between the bivariate target distribution an the kernel density estimation, it is already clear that the algorithm is able to capture the dependence structure of the joint distribution. Because of the existence of the bivariate distribution and the fact that the drawings are from the whole set of full conditionals, this distribution can be approximated to any degree of accuracy.

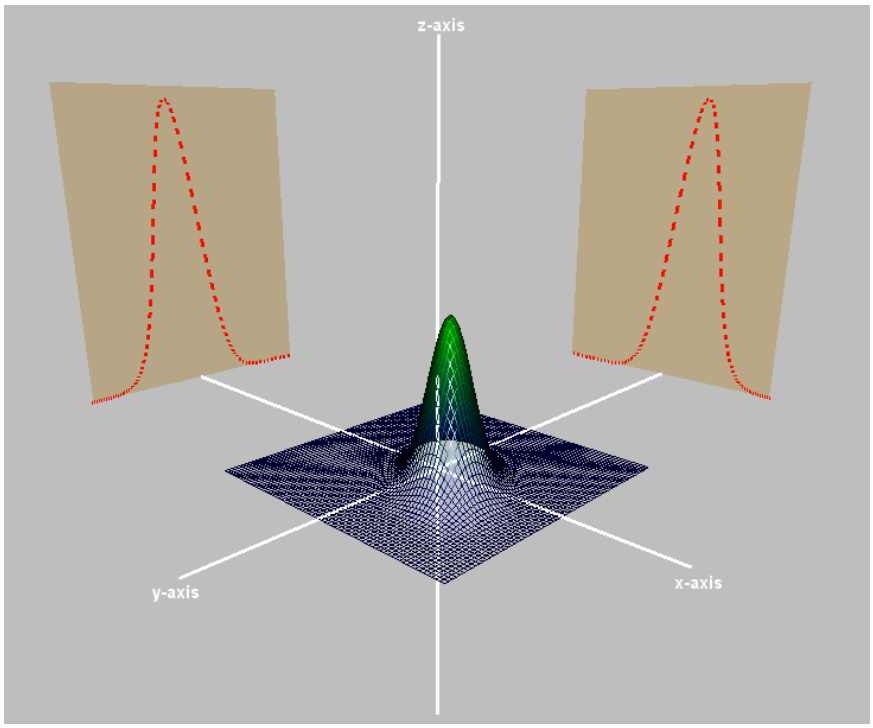

Figure 2.2: Begin of the Simulation.

The objective is to draw from the joint distribution in the middle of the graphic. As an additional check, both marginals are included (red, dashed lines). It must be pointed out, that neither the joint distribution nor the marginals are used for the sample generation. (Source: HUERGO, 2008.)

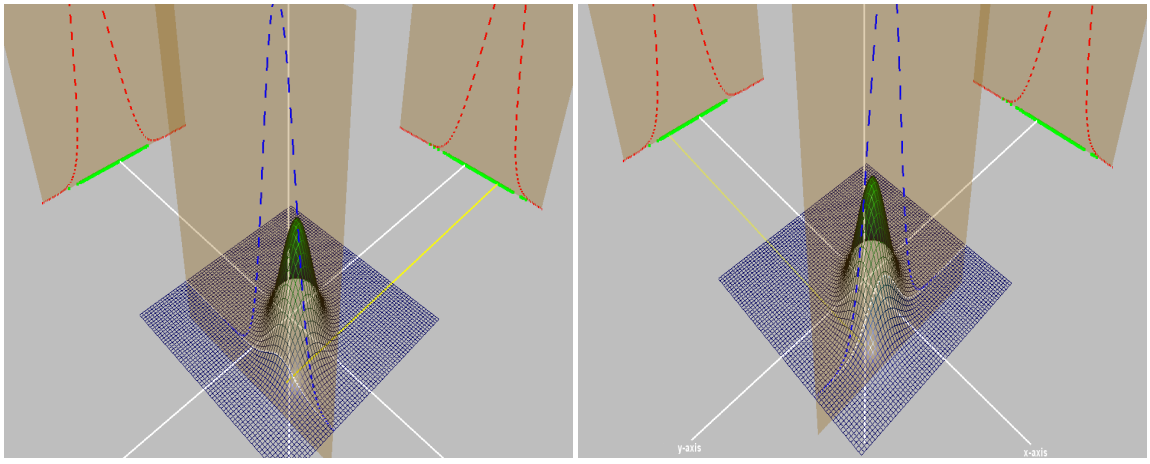

Figure 2.3: Draws from the conditional distributions.

It can be seen that the conditional distribution on the right panel is positioned exactly over the last value drawn (left panel, yellow line). (Source: HUERGO, 2008.)

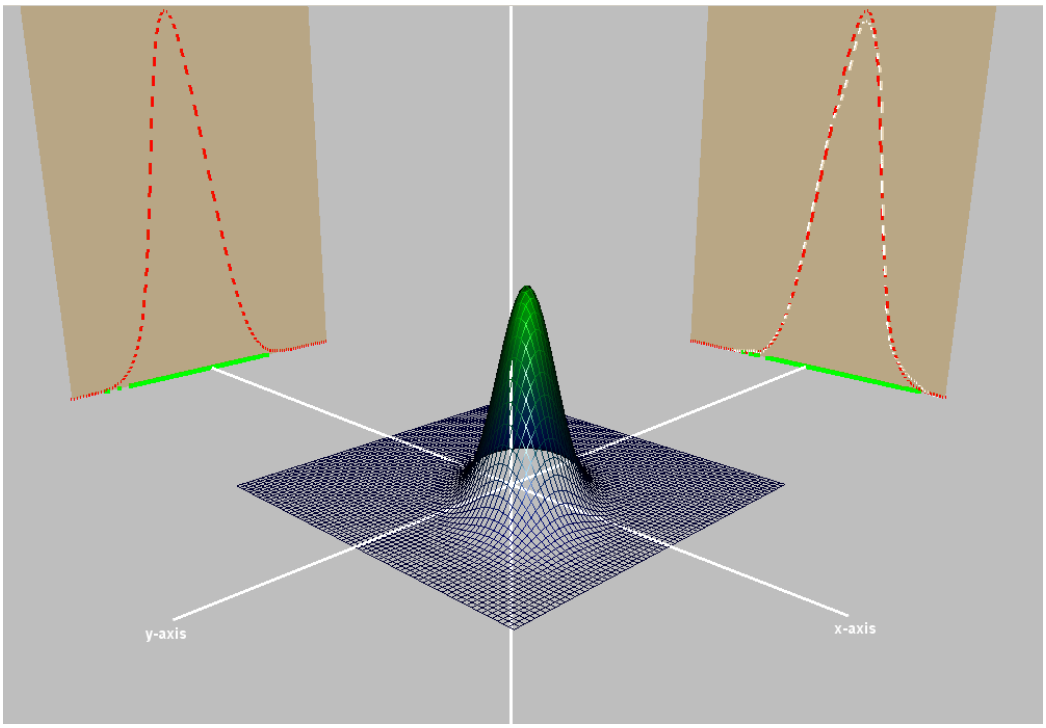

Figure 2.4: First marginal distribution.

The dashed white line shows the kernel density estimation of the first 500 draws. (Source: Huergo, 2008.)

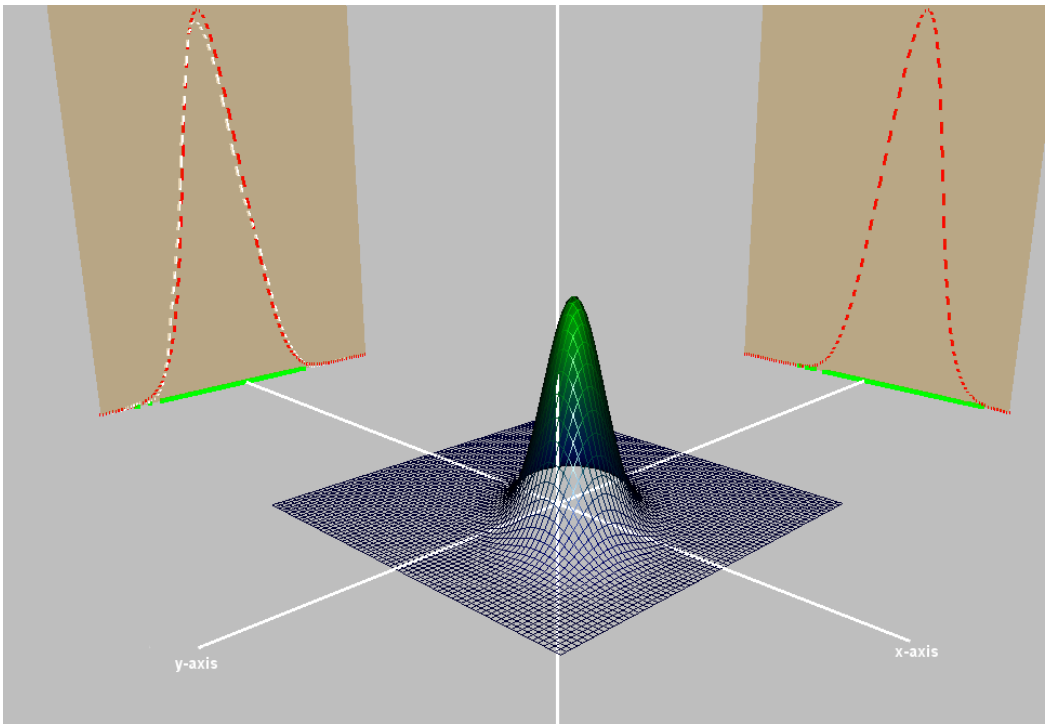

Figure 2.5: Second marginal distribution

After 500 iterations is the approximation fairly accurate. (Source: Huergo, 2008.)

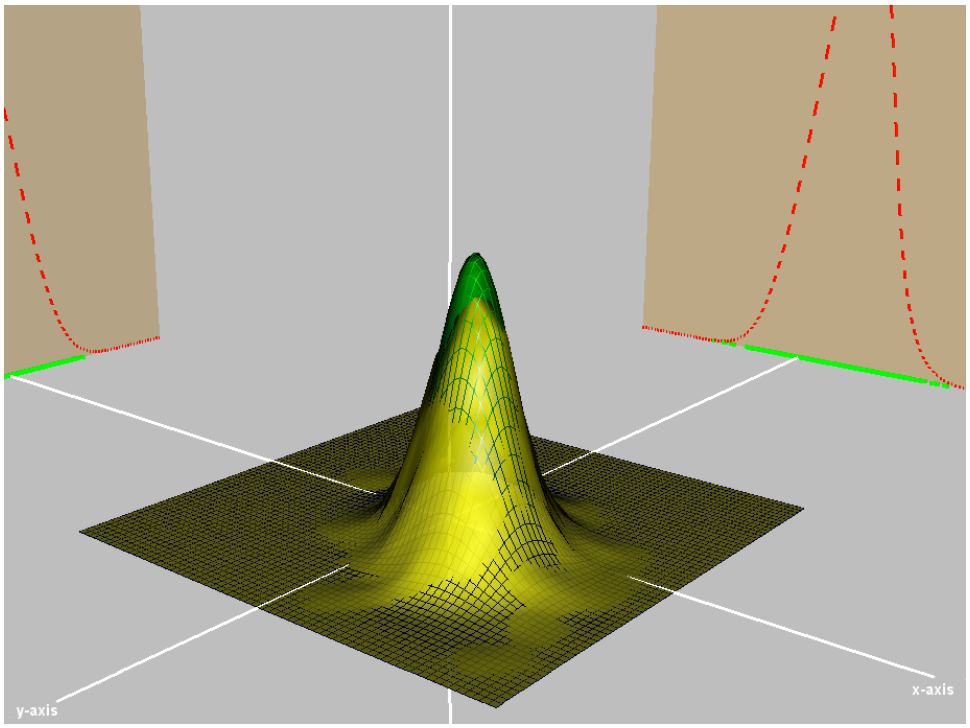

Figure 2.6: Approximation of the joint distribution

The yellow surface curve represents 500 Pairs drawn from the conditional distributions. It can clearly be seen, that the algorithm is able to reproduce the correlation structure of the target distribution. (Source: HUERGO, 2008.)

Another Markov Chain Monte Carlo method which is closely related to the Gibbs sampler is the Data Augmentation algorithm, which goes back to the paper of TANNER and WONG (1987). In fact, under certain conditions, the former can be shown to be a special case of the latter (GELFAND and SMITH, 1990).

Beginning with the starting value $\theta^{(0)}$, data augmentation yields the stochastic sequence $\left\{(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \ldots\right\}$, whose stationary distribution is the distribution of interest, namely $P(Y_{mis}, \theta \mid Y_{obs})$, or respectively the stationary distributions $P(\theta \mid Y_{obs})$ and $P(Y_{mis} \mid Y_{obs})$ of the subsequences $\left\{\theta^{(t)} : t = 1, 2, \ldots\right\}$ and $\left\{Y_{mis}^{(t)} : t = 1, 2, \ldots\right\}$. Such a sequence can be obtained by iterative draws of $Y_{mis}^{(t+1)}$ and $\theta^{(t+1)}$ from $P(Y_{mis} \mid Y_{obs}, \theta^{(t)})$ and $P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$ in a two step approach which turns out to be simpler than drawing directly from the posterior distributions $P(Y_{mis} \mid Y_{obs})$ and $P(\theta \mid Y_{obs})$. For large t this sequence converges to a draw from the joint posterior distribution $P(\theta, Y_{mis} \mid Y_{obs})$ (LITTLE and RUBIN, 2002, p. 201).

2.5.3 Structure of the Data Augmentation algorithm

The DA algorithm is an iterative method which bears a striking resemblance to the EM algorithm. Indeed, it combines the properties of the EM algorithm and multiple imputation to simulate the posterior distribution of θ. The main difference with the EM algorithm is that the deterministic e- and m-steps get replaced by the stochastic i- and p-steps:

(a) Imputation-Step (i-step)

To impute missing values, a current estimation of θ is needed. The parameter estimate θ already calculated by the EM algorithm can serve as a starting value. In contrast to the EM algorithm, which makes use of conditional expectations, the DA algorithm applies draws from the conditional predictive distribution of the missing data given the observed values and the current parameters. The imputation step (**i-step**) can be stated as follows:

$$Y_{mis}^{(t+1)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)}) \; .$$

(b) Posterior-Step (p-step)

New estimators for the parameters can be calculated with the augmented dataset $Y^{(t+1)} = (Y_{obs}, Y_{mis}^{(t+1)})$ from the i-step. These parameters represent what is known about the parameter values which are contained in the data. Thus, the new distribution will be combined with the already known prior distribution. In doing so, one obtains a new posterior distribution from which the new parameters can be drawn. This is called the posterior step (**p-step**)

$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)}) \; .$$

2.6 Extensions to EM and DA

2.6.1 The ECM Algorithm

In cases where the m-step turns out to be complicated, a possible approach is to let Q increase rather than maximize it. This is the idea behind the GEM algorithm (Dempster et al., 1977). A special case of it is the ECM algorithm (Meng and Rubin, 1993), which shares its convergence properties with the EM algorithm, such as monotone convergence. However, the ECM algorithm takes advantage of the simplicity of complete-data conditional maximization. A m-step of the EM algorithm will be replaced with $S > 1$ *conditional maximization* (CM) steps. In doing so, each CM step maximizes the Q-function with respect to one subvector of θ, $(\theta_1, \theta_2, \ldots, \theta_s)$, holding the other $S - 1$ fixed, which turns out to be computationally simpler than a maximization over the whole parameter space of θ as is done in a m-step.

2.6.2 The ECME Algorithm

The *expectation-conditional maximization either* (ECME) algorithm (Liu and Rubin, 1994) is a generalization of the ECM algorithm that maximizes *either* the constrained expected loglikelihood, this is the Q-function, or the correspondingly constrained actual loglikelihood function. Moreover, ECME shares the same convergence properties of EM or ECM but has a faster convergence because it maximizes the actual likelihood (conditionally) and not an approximation of it.

2.6.3 The PX-EM Algorithm

A further extension to EM is the so-called *parameter-expanded* EM (PX-EM) algorithm which can be used to speed up the convergence. LIU et al. (1998) provide the basic theory of PX-EM. However, in this work only a simple adjustment proposed by KENT et al. (1994) will be implemented in the robust models in Chapter 5.

Chapter 3

Multivariate Normal Model

3.1 The Sweep Operator

3.1.1 Purpose

The *Sweep Operator* is an important tool when applying the EM algorithm on missing data. The sweep operator itself is also an algorithm that involves a finite sequence of relatively easy steps. It delivers a conditional distribution of a dataset with multivariate normal distributed random variables given another dataset. In doing so, this approach is a powerful and simple tool to impute missing values and to compute conditional expectations. Its functionality is presented in Appendix B.1. A detailed introduction on the sweep operator is given in Goodnights' tutorial (Goodnight, 1979). Further information and properties in case of missing-data problems are presented in Little and Rubin (1987) and Schafer (1997) at full length.

3.1.2 Alternative Parameterizations of the Normal Distribution

The sweep operator is a very useful tool, which can convert the *response* variables of a multivariate normal distribution into *predictors*. Assume that $z \sim MVN(\mu, \Sigma)$ is a p-dimensional random vector. If $1 \leq p_1 < p$, one can partion z into two random vectors $z' = (z_1', z_2')$, where z_1 contains the first p_1 elements and z_2 the last $p - p_1$ elements. Then μ and Σ can be stated as follows

$$\begin{aligned} \mu' &= (\mu_1', \mu_2') \,, \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \end{aligned}$$

where $E[z_i] = \mu_i$, $Cov[z_i] = \Sigma_{ii}$ and $Cov[z_1, z_2] = \Sigma_{12} = \Sigma_{21}'$. Now one has to know, that in case of a multivariate normal distribution the conditional distribution of z_2 given z_1 is also normally distributed with

$$E[z_2 \mid z_1 = x] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x - \mu_1) = \underbrace{(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1)}_{\alpha_{2\cdot1}} + \underbrace{(\Sigma_{21}\Sigma_{11}^{-1})}_{\beta_{2\cdot1}} x \tag{3.1}$$

and

$$Var[z_2 \mid z_1 = x] = \underbrace{\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}_{\Sigma_{2\cdot 1}} \ , \tag{3.2}$$

where $\alpha_{2\cdot 1}$ is the vector of intercepts, $\beta_{2\cdot 1}$ the matrix of slope coefficients and $\Sigma_{2\cdot 1}$ the variance-covariance matrix of the residuals when realizing a multivariate regression of z_2 on z_1.

With a correct sweep operation of the parameters of the multivariate normal distribution into an alternate form, the results of Equations (3.1) and (3.2) can be derived. Next, one arranges for θ:

$$\theta = \begin{bmatrix} -1 & \mu^{'} \\ \mu & \Sigma \end{bmatrix} = \begin{bmatrix} -1 & \mu_1^{'} & \mu_2^{'} \\ \mu_1 & \Sigma_{11} & \Sigma_{12} \\ \mu_2 & \Sigma_{21} & \Sigma_{22} \end{bmatrix} \ ,$$

where the first position of θ has to be labeled with 0. Then, a sweep operation of θ at positions $1, \ldots, p_1$ results in

$$SWP[1, \ldots, p_1]\,\theta = \begin{bmatrix} -1 - \mu_1^{'}\Sigma_{11}^{-1}\mu_1 & \mu_1^{'}\Sigma_{11}^{-1} & \mu_2^{'} - \mu_1^{'}\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} & \Sigma_{11}^{-1}\Sigma_{12} \\ \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1 & \Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{bmatrix} .$$

This swept matrix contains the parameters of the conditional distribution of z_2 given z_1, as described above, whereas the upper left submatrix, with dimensions $(p_1 + 1) \times (p_1 + 1)$

$$\begin{bmatrix} -1 & \mu_1^{'} \\ \mu_1 & \Sigma_{11} \end{bmatrix} \ \Rightarrow \ \begin{bmatrix} -1 - \mu_1^{'}\Sigma_{11}^{-1}\mu_1 & \mu_1^{'}\Sigma_{11}^{-1} \\ \Sigma_{11}^{-1}\mu_1 & -\Sigma_{11}^{-1} \end{bmatrix}$$

includes the marginal distribution of z_1 in swept form. The reason why the upper left cell of θ contains -1 can be explained by the fact, that the parameter matrix θ is assumed to be already swept on position 0. A reverse sweep operation of θ on position 0,

$$RSWP[0]\,\theta = \begin{bmatrix} 1 & \mu^{'} \\ \mu & \Sigma + \mu\mu^{'} \end{bmatrix} \ , \tag{3.3}$$

yields the parameters of the unconditional multivariate normal distribution, expressed in terms of the first two moments. This unswept form of the matrix can be used to compute the MLEs.

In the following sections the EM and DA algorithms will be implemented to multivariate normally distributed data with missing values.

3.2 Parametrization

To do an imputation in a multivariate dataset, one has to compute the conditional distribution of the missing data given the observed data. To perform this task the sweep operator turns the response variables into predictors. Suppose, that Y is a matrix of

independent observations of the multivariate normal distribution $MVN(\mu, \Sigma)$ which belongs to the exponential family of distributions. Therefore, the loglikelihood in Equation (2.9) is not linear in the data but rather linear in a set of sufficient statistics whose expectations have to be computed. The sufficient statistics for this model can be stated as follows: $T_1 = Y'\mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)'$, and $T_2 = Y'Y$, which can be arranged in a $(p+1) \times (p+1)$ matrix

$$T = \begin{bmatrix} n & T_1' \\ T_1 & T_2 \end{bmatrix} .$$

To make the sweep operator feasible, one arranges the parameters μ and Σ into the following matrix

$$\theta = \begin{bmatrix} -1 & \mu' \\ \mu & \Sigma \end{bmatrix} .$$

The moment equations of a ML estimation in this model can be expressed in terms of a reverse sweep operator, similar to Equation (3.3) . The ML estimators of μ and Σ have to solve

$$RSWP[0]\,\theta = n^{-1}T .$$

Hence, this can be done by following sweep operation:

$$\widehat{\theta} = SWP[0]\,n^{-1}T . \tag{3.4}$$

The result can be used to provide a concise description of the EM algorithm, as long as the two steps of the algorithm, as described in Section 2.4.1, can be summarized in the following equation

$$\theta^{(t+1)} = SWP[0]\,n^{-1}\,E\left[T \mid Y_{obs}, \theta^{(t)}\right] , \tag{3.5}$$

where $\theta^{(t)}$ and $\theta^{(t+1)}$ denote the successive parameter estimates. The term in Equation (3.5) is basically a linear regression of Y_{mis} on Y_{obs}, whose computation needs evidence of the parameters of the conditional distribution $Y_{mis} \mid Y_{obs}$. These parameters can be estimated by means of the sweep operator.

There are miscellaneous methods for the use of the EM algorithm. So, Schafer describes an implementation that sorts the rows in Y by patterns of missingness as a first step. In doing so, the number of sweep operations can be kept to a minimum. But the implementation is only meaningful when the rows in Y are i.i.d. This method works excellently for datasets that aren't too large and is technically simple and realizable as long as it provides an explicit and easily implementable design of the EM operator.

3.3 Implementation of the EM Algorithm

3.3.1 Preliminary Preparation of the Data

As already mentioned, in a first step the data must be manipulated so as to minimize the number of sweep operations. Therefore, the rows in the data matrix Y have to be

sorted into S groups by all S occurring patterns of missingness. The groups are indexed by $s = 1, \ldots, S$. Observations that only have missing values must be excluded, because they make no contribution to the observed data likelihood (see Equation (2.3)). Furthermore, they just slow down the convergence of the EM algorithm because the proportion of missing information increases. However, the models to be imputed within the KEI project were chosen such as to avoid exclusions.

To arrange the data suppose the initially introduced indicator matrix R and simply change its dimensions to $S \times p$, so that

$$r_{sj} = \begin{cases} 1, & \text{if } Y_j \text{ in group } S \text{ is observed} \\ 0, & \text{if } Y_j \text{ in group } S \text{ is missing.} \end{cases}$$

Next, one introduces $\mathcal{O}(s)$ and $\mathcal{M}(s)$ to label the observed and missing values of the variables for each of the S patterns. This can be done using the subset $\{1, \ldots, p\}$:

$$\mathcal{O}(s) = \{j : r_{sj} = 1\},$$
$$\mathcal{M}(s) = \{j : r_{sj} = 0\}.$$

Then determine $\mathcal{I}(s)$ with the subset $\{1, \ldots, n\}$, which gives the corresponding rows of pattern s of the data matrix Y.

3.3.2 The E-step

With that information the expected value of the sufficient statistics for an assumed value of θ can be computed. Because of the met assumption of independent rows one can write

$$P(Y_{mis} \mid Y_{obs}, \theta) = \prod_{i=1}^{n} P(y_{i(mis)} \mid y_{i(obs)}, \theta) \ , \tag{3.6}$$

where $y_{i(mis)}$ and $y_{i(obs)}$ are the subvectors with the corresponding missing and observed values. To apply a multivariate normal linear regression of Y_{mis} on Y_{obs}, which is equivalent to Equation (3.6), one has to sweep the parameter matrix θ at the corresponding positions. This happens successively for each of the S groups. One then obtains the parameters for $P(y_{i(mis)} \mid y_{i(obs)}, \theta)$, if row i belongs to the corresponding pattern s, by a sweep operation in the respective rows and columns, which are labeled by $\mathcal{M}(s)$. Then, the swept matrix of this regression reads as follows:

$$A = SWP[\mathcal{O}(s)]\theta \ , \tag{3.7}$$

where a_{jk} represents the (j,k)th element, $(j, k = 0, \ldots, p)$.

As long as it is about observed values - that is $j \in \mathcal{O}(s)$ - one can make clear for the first moment:

$$E(y_{i(obs)} \mid Y_{obs}, \theta) = y_{i(obs)}$$

and for the second moment (the covariance is built with a $y_{i(obs)}$, which is assumed to be fixed):

$$Cov(y_{i(obs)}, y_{ik} \mid Y_{obs}, \theta) = 0$$

(with an arbitrary k). For unobserved values it has to be picked from the swept matrix:

$$E(y_{i(mis)} \mid Y_{obs}, \theta) = a_{0(mis)} + \sum_k a_{(obs)(mis)} y_{i(obs)}$$

and

$$Cov(y_{i(mis1)}, y_{i(mis2)} \mid Y_{obs}, \theta) = a_{(mis1)(mis2)}.$$

Somewhat more generally:

$$E(y_{ij}, y_{ik} \mid Y_{obs}, \theta) = Cov(y_{ij}, y_{ik} \mid Y_{obs}, \theta) + E(y_{ij} \mid Y_{obs}, \theta) E(y_{ik} \mid Y_{obs}, \theta) \ ,$$

where

$$Cov(y_{ij}, y_{ik} \mid Y_{obs}, \theta) = \begin{cases} 0, & \text{if } j \in \mathcal{O}(s) \\ a_{jk}, & \text{if } j,k \in \mathcal{M}(s) \ . \end{cases}$$

Thus, for the expected values of y_{ij} and $y_{ij}y_{ik}$

$$E(y_{ij} \mid Y_{obs}, \theta) = \begin{cases} y_{ij}, & \text{if } j \in \mathcal{O}(s) \\ y^*_{ij}, & \text{if } j \in \mathcal{M}(s) \end{cases}$$

and

$$E(y_{ij}y_{ik} \mid Y_{obs}, \theta) = \begin{cases} y_{ij}y_{ik}, & \text{if } j,k \in \mathcal{O}(s) \\ y^*_{ij}y_{ik}, & \text{if } j \in \mathcal{M}(s), k \in \mathcal{O}(s) \\ a_{jk} + y^*_{ij}y^*_{ik}, & \text{if } j,k \in \mathcal{M}(s) \ , \end{cases}$$

where $y^*_{ij} = a_{0j} + \sum_k a_{kj} y_{ik}$ represents the value of the sufficient matrices, completed by the correction factor. In the e-step this expectation will be summarized over all rows for each j and k. Practically, one at first adds up the expectations of all rows within a pattern. The S patterns then have to be added up and one obtains the output of the e-step:

$$E\left[\, T \mid Y_{obs}, \theta \,\right] \ ,$$

where T represents the sufficient statistics.

3.3.3 The M-step

To obtain the parameter matrix θ from the result above, one simply needs to apply Equation (3.4), a sweep operation on position 0. Thus, one leaves an iteration step behind. The maximization-step is a simple repetition of the e-step above. But instead of the initializing matrix $\theta^{(0)}$ one puts the newly computed $\theta^{(t)}$ into the sweep operator and obtains $\theta^{(t+1)}$ as already shown in Equation (3.5). This happens until an iteration criterion stops the algorithm.

3.3.4 Iteration Criterion

The means and covariance matrix from the observed values turn out to be the most appropriate starting value for the EM algorithm:

$$\theta^{(0)} = \begin{bmatrix} -1 & \mu_{obs}^{(0)'} \\ \mu_{obs}^{(0)} & \Sigma_{obs}^{(0)} \end{bmatrix} .$$

A simpler, alternative approach is to assume that mean and covariances are 0 and variances 1. LITTLE and RUBIN (1987) discuss further possibilities for configurating starting values.

The algorithm proceeds until the estimation converges. A tolerance can be defined as abort criterion (e.g., if a further iteration step yields no further improvement in the parameters, the algorithm can be stopped).

3.4 Implementation of the DA Algorithm

3.4.1 The I-step

Because it is assumed, that the rows in Y are conditionally independent given θ, each missing $y_{i(mis)}$ can be drawn independently:

$$y_{i(mis)}^{(t+1)} \sim P(y_{i(mis)} \mid y_{i(obs)}, \theta^{(t)}) .$$

Hence, the i-step is an independent simulation of random normal vectors for each row of the data matrix Y, with means and covariances according to $E(y_{ij} \mid Y_{obs}, \theta) = y_{ij}^{*}$ and $Cov(y_{ij}, y_{ik} \mid Y_{obs}, \theta) = a_{jk}$, where $j, k \in \mathcal{M}(s)$. The calculation is mostly similar to that of the EM algorithm.

A Cholesky factorization enables simulating random normal vectors. To draw from the distribution of $y_{i(mis)}$ given $y_{i(obs)}$ and θ, one just computes the Cholesky factor of the submatrix in Equation (3.7) corresponding to the rows and columns, which are labeled by $\mathcal{M}(s)$. The remaining elements of A remain unaffected

$$A := Chol_{\mathcal{S}} A .$$

With the effected Cholesky factorization, the i-step is just a straight routine computation through all occurring patterns $s = 1, \ldots, S$

$$Chol_{\mathcal{M}(s)} SWP[\mathcal{O}(s)]\theta ,$$

with a following simulation of the missing values $y_{i(mis)}$ for each $i \in \mathcal{I}(s)$.

3.4.2 The P-step

As described in Schafer (1997, Ch. 5.2.2 and 5.2.3), the complete data posterior distribution $P(\theta \mid Y_{obs}, Y_{mis})$ is a *normal inverted* Wishart distribution. Therefore, the p-step is merely meant as a simulation of the *normal inverted* Wishart distribution:

$$\begin{aligned} \mu \mid \Sigma &\sim N(\mu_0, \tau^{-1}\Sigma) \,, \\ \Sigma &\sim W^{-1}(m, \Lambda) \,, \end{aligned}$$

where $(\tau, m, \mu_0, \Lambda)$ are derived from the prior distribution and the missing data $Y_{mis}^{(t)}$ from the last i-step.

To obtain a normal inverted Wishart distribution, one arranges a matrix B in such a way that the elements of its minor diagonal are χ^2-distributed and the elements above that diagonal are standard normal distributed. The cross product has the property $B'B \sim W(m, I)$, which implies that

$$M = (B')^{-1}C \,,$$

where C is the Cholesky factor of $\Lambda^{-1} = C'C$. In calculating the cross product of M, one maintains *normal inverted* Wishart distributed matrices, because

$$(M'M)^{-1} = C^{-1}B'B(C')^{-1} \sim W(m, \Lambda) \,. \tag{3.8}$$

This method of obtaining a normal inverted Wishart distribution is known as a *Bartlett Decomposition*. Next, the expectations have to be determined

$$\mu = \mu_0 + \tau^{-1/2}M'z \mid \Sigma \sim N(\mu_0, \tau^{-1}\Sigma), \tag{3.9}$$

where $z \sim N(0, I)$ is a $p \times 1$ vector with independent standard normal distributed variables.

Hence, the p-step runs as follows: In a first step Σ will be drawn from Equation (3.8) and then μ conditioned on Σ from Equation (3.9).

Example: Predictive distribution of a missing values (NA) in the case of a multivariate normal distribution (taken from Huergo, 2008).

Setup: Starting point of the simulation is a sample of size 400 from a 5-Variate normal distribution with the following Parameters

$$\mu = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 2 \\ 1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 4{,}000 & 1{,}340 & 1{,}200 & 0{,}304 & 0{,}144 \\ 1{,}340 & 1{,}000 & 0{,}750 & 0{,}208 & 0{,}120 \\ 1{,}200 & 0{,}750 & 2{,}250 & 0{,}744 & 0{,}531 \\ 0{,}304 & 0{,}208 & 0{,}744 & 0{,}640 & 0{,}264 \\ 0{,}144 & 0{,}120 & 0{,}531 & 0{,}264 & 0{,}360 \end{pmatrix} \,, \tag{3.10}$$

which are assumed unknown and allow a comparison of the simulation results.

For the purposes of the simulation, 400 values from the sample were set as NA. These values were appropriately chosen to make sure that no row of the data set was completely

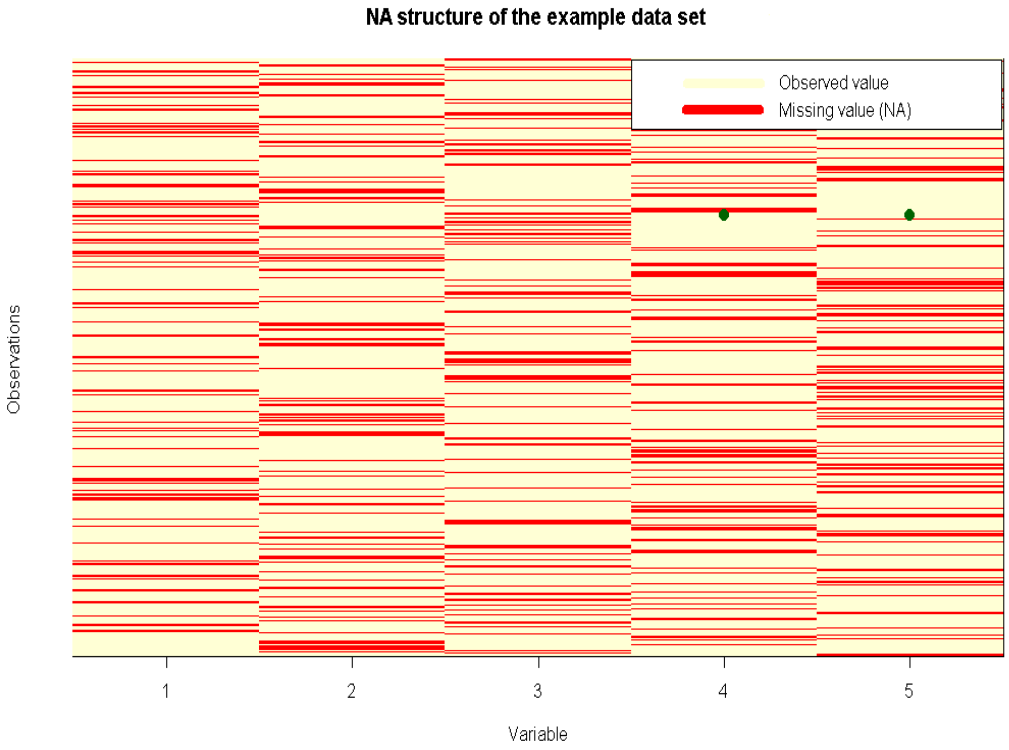

Figure 3.1: NA structure of the data.

The green dots flag the values, whose joint distribution has to be estimated. (Source: HUERGO, 2008.)

observed. The resulting structure of the data set can be seen in Figure 3.1. The goal of the exercise is the simulation of the predictive distribution of an arbitrary pair of missing values, given the available data, in order to proceed to its imputation.

In order to simulate the predictive distribution of these missing values, a *Data Augmentation* Algorithm will be used.

Let θ denote the unknown parameters μ und Σ, with resulting joint distribution $f_\Sigma(\cdot) \cdot f_{\mu|\Sigma}$. Further let X denote the underlying data set. Because of the presence of missing data let X be composed of two parts, the observed data Y and the missing data Z. Thus it holds $X := (Y, Z)$.

The density function $f(\theta|X) = f(\mu, \Sigma|Y, Z)$ denotes thus the posterior distribution of the parameters μ und Σ given the information contained in the partially observed sample X.

In the case of a fully observed data set, Bayes' theorem would suggest the following structure for this posterior distribution:

$$f(\theta|X) \propto f(\theta) f(X|\theta) \ , \tag{3.11}$$

where $f(\theta)$ denotes the prior distribution of μ and Σ, and the Symbol $\propto$ makes clear, that this distribution is uniquely characterized up to a proportionality constant $\int f(\theta) f(X|\theta) \delta\theta$.

In the considered case of a multivariate normal distribution it is straightforward to calculate $f(\theta|X)$ in closed form.

The calculation of the posterior distribution of the parameters under both conjugate and noninformative priors is extensively discussed in Gelman et al. (2004, pp. 87-88).

In the presence of missing data the situation changes considerably: The expression $f(\theta|X)$ cannot be calculated analytically and one has to resort to simulative methods.

Let $f(Z|Y)$ denote the predictive distribution of the missing data given the observed ones, whereby the dependence to the parameters θ was eliminated by means of integration. It holds thus

$$f(Z|Y) = \int f(Z|\theta, Y) f(\theta|Y) \delta\theta \; . \tag{3.12}$$

The integrator in 3.12 is not available in closed form and thus the predictive distribution $f(Z|Y)$ must be simulated as well.

The Data Augmentation algorithm proceeds as follows:

1. Starting values $\theta_0 = \{\Sigma_0, \mu_0\}$ for $\theta = \{\Sigma, \mu\}$ must be chosen.

2. Conditional on the chosen values θ_0 and the observed data Y, values Z_0 for the missing data are drawn from $f(Z|\theta_0, Y)$, which is a normal distribution with vector of expected values ϕ resulting from the multivariate linear regression of Z on Y and variance-covariance matrix Ψ, equal the variance-covariance matrix of the residuals of this regression. The *imputation* step is now complete.

3. With the completed data it is now possible to estimate the parameters of the joint distribution:

   (a) Given the new set of complete data X_0 is the new variance-covariance matrix Wishart[1] distributed with scaling parameter $\widehat{\Sigma}_0$, where

   $$\widehat{\Sigma}_0 = \tfrac{1}{n} \left(X_0' X_0 - \overline{x}_0 \overline{x}_0' \right) \quad \text{and} \tag{3.13}$$

   $$\overline{x}_0^{\cdot, j} = \tfrac{1}{n} \sum_{i=1}^{n} x_0^{i,j} \quad \text{for } j \in \{1 : k\}, \quad k = \text{Number of dimensions.} \tag{3.14}$$

   $\overline{x} = (\overline{x}^{\cdot, 1}, \ldots, \overline{x}^{\cdot, k})$ is thus the vector of the columnwise computed mean values of the completed data.

   (b) Conditional on Σ_0 and X_0, μ_1 is a draw from a normal distribution with parameters $(\overline{x}, \Sigma_0/n)$, where n represents the number of rows of the completed data matrix X_0 bezeichnet. The *Posterior*-Step is now complete.

4. The algorithm continues iterating between *Imputation* and *Posterior* steps until a konvergence criterion is reached.

[1] The Wishart distribution is a multivariate generalization of the Chi squared distribution.

Simulation of the joint predictive distribution of a pair of missing values: The values of the 106-*th* row, columns four an five of the data set were set to NA for the purposes of the simulation. From now on these values will be referred to as $z_{106,4}$ and $z_{106,5}$. The choice of the row is arbitrary and is not expected to affect the accuracy of the simulation.

The aim of the example is the simulation of the joint predictive distribution of both values, given the observed data, $f(z_{106,4}, z_{106,5}|Y)$.

After a burn in period of 2000 cycles, 1000 iterations of the Data Augmentation algorithm will be used to estimate the joint distribution of the missing data. The results will be compared to the following distributions:

1. The distribution of $z_{106,4}$ and $z_{106,5}$ given the observed values of the 106-*th* row and under the assumption of known parameters. The resulting distribution is normal with the following parameters:

   - Vector of expected values: Results from the linear regression of $z_{106,4}$ and $z_{106,5}$ on the observed values of the 106-*th* row.
   - Variance-covariance matrix: Equals the variance-covariance matrix of the Residals of this Regression.

   The regression parameters result from the factorization of the parameters of the multivariate normal distribution. This factorization can easily be computed for all combinations of observed an missing values by means of the *Sweep* Operator.

   The conditional distribution of the missing values given the parameters and the observed values of the 106-*th* row $f(z_{106,4}, z_{106,5}|\phi, x_{106,\{1,2,3\}})^2$, represent the the maximal achievable knowledge of the missing values. This distribution will be used as a benchmark to test the accuracy of the results from the Data Augmentation algorithm.

2. The predictive distribution of the values to be imputed, given a completely observed data set.

   In this case the imputation has the properties of a forecast: a fully observed data set allows the estimation of the parameters of the underlying distribution. An additional row with two NAs enlarges the data set. The estimated parameters and the additional data can be used for the estimation of the conditional distribution of the missing data $f(z_{106,4}, z_{106,5}|\phi = \widehat{\phi}, X)$. In order to eliminate the dependence on estimated parameters it is usual in a bayesian context to construct the predictive distribution of the missing data given the observed data:

   $$f(z_{106,4}, z_{106,5}|X) = \int_{\phi} f(z_{106,4}, z_{106,5}|\phi, X) f(\phi|X) \delta\phi \ .$$

   This distribution will be simulated as well.

   Because of the availability of a completely observed data set, the distribution can be simulated with conventional Monte Carlo methods and there is no need to resort to iterative algorithms.

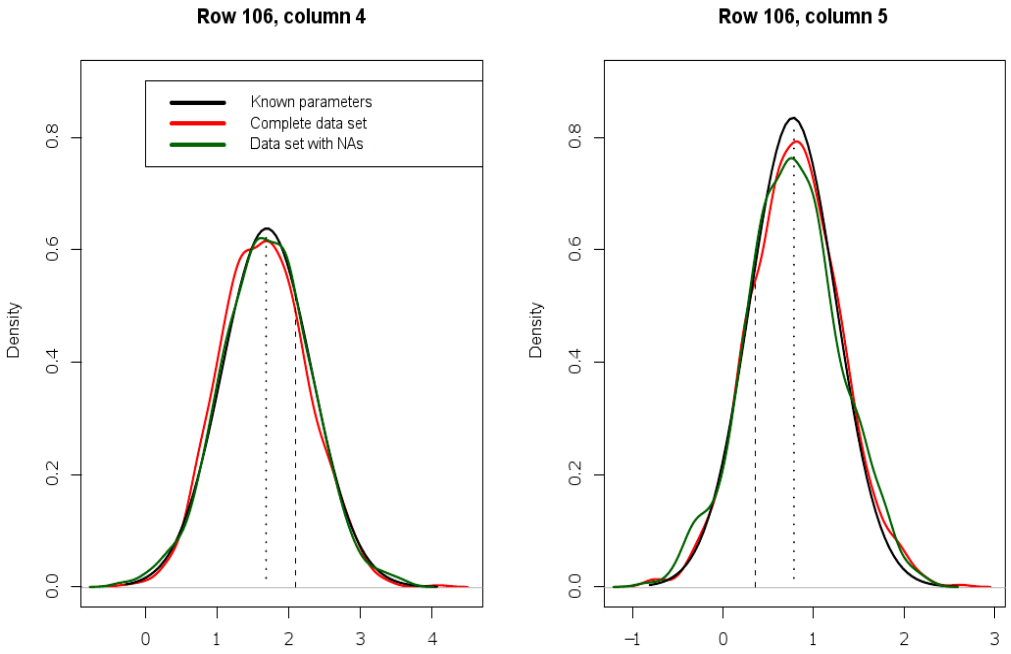

Figure 3.2: Comparison of the marginal distributions.

The figure shows the conditional distribution of the missing values in the case of known parameters (black solid line), along with the kernel density estimation of the predictive distribution with: -completely observed data (red solid line) and -data with missing values (green solid line). The dashed lines flag the original data, which were set to NA for the simulation. The dotted line flags the estimate from the EM-Algorithm. (Source: Huergo, 2008.)

Figure 3.2 shows the results of the simulation. In spite of the high number of missing values are the by means of the Data Augmentation algorithm simulated marginals a close approximation to the theoretical conditional distribution (black solid line). The fact that the predictive distributions are a bit wider and lower is a natural consequence of the fact that the parameters had to be integrated out of the distribution.

Additionally, the EM algorithm was applied to the data. The results are flagged with a dotted black line in Figure 3.2. After convergence, the estimates of the EM-Algorithm are $\widehat{z}_{106,4}$: 1,686 and $\widehat{z}_{106,5}$: 0,785. Both values are quite close to the modal values of the respective theoretical distributions.

The theoretical correlation between $z_{106,4}$ and $z_{106,5}$ amounts to 0,281. Because of the available correlation structure of the observed data it is not possible to completely reconstruct the original correlation of 0,55 between the fifth and sixth columns. The simulation in the case of completely observed data and unknown parameters yields a correlation coefficient of 0,246. The simulated sample in the presence of missing values yields a correlation of 0,309. Figure 3.3 shows simulated data from the three settings. All of them show a similar correlation structure.

[2]In the case of known parameters the data from rows other than the 106-*th* are not involved in the estimation of the predictive distribution and can be safely ignored.

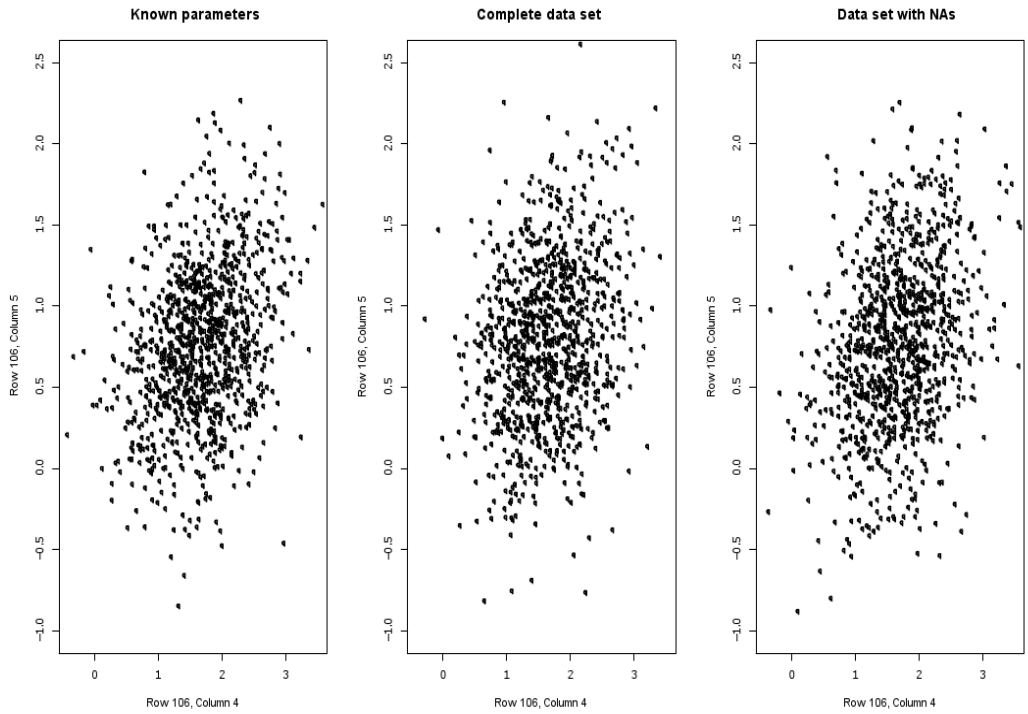

Figure 3.3: Comparison of the scatterplots.

The three scatterplots show the bivariate distributions in the three cases compared. The first panel depicts the expected correlation structure. It is evident that the Data Augmentation algorithm was able to reconstruct this correlation structure to a high amount, in spite of the missing data. (Source: HUERGO, 2008.)

Chapter 4

Power Transformation to induce approximate Normality

4.1 Justification of correcting the shape of the Data

Since most of the KEI indicators deviate from normality, the data must be transformed. Figure 4.1 presents an illustrative example of why there is a need for such a transformation. The proposed power transformation introduced in this chapter goes back to Huergo (2008).

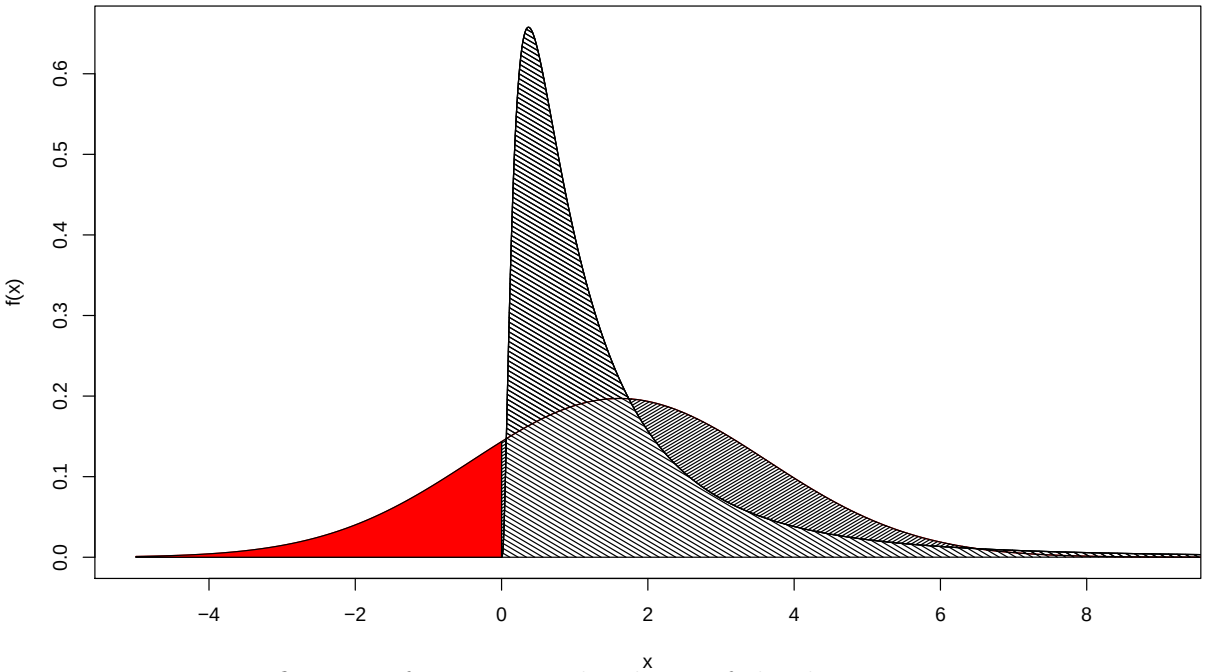

Figure 4.1: Justification of correcting the shape of the data. (Source: Huergo, 2008.)

The diagonally hatched density function in the front corresponds to the original data-set. The EM algorithm computes mean and variance of these data as these parameters characterize a normal distribution completely. Then it constructs a normal distribution to impute the missing data using these parameters (note that both densities have an

identical mean and variance). The assumed distribution is the one in the background. The red area represents values that are not valid under the original distribution since the data are defined on $\mathbb{R}^+$. The normal distribution is, unformally speaking, what the (classical) EM algorithm for normally distributed data *sees* when it receives data from a right-skewed density function. Figure 4.2 compares imputed values of the EM algorithm with and without transformation. While the former imputes negative values, the proposed transformation provides a valid imputation. Also Schafer confirms to this suggestion:

> *Datasets encountered in the real world often deviate from multivariate normality, but in many cases the normal model will be useful even when the actual data are nonnormal. [...] Sometimes the normality assumption may be made more plausible by applying suitable transformations to one or more variables* (Schafer, 1997, pp. 29 and 147)

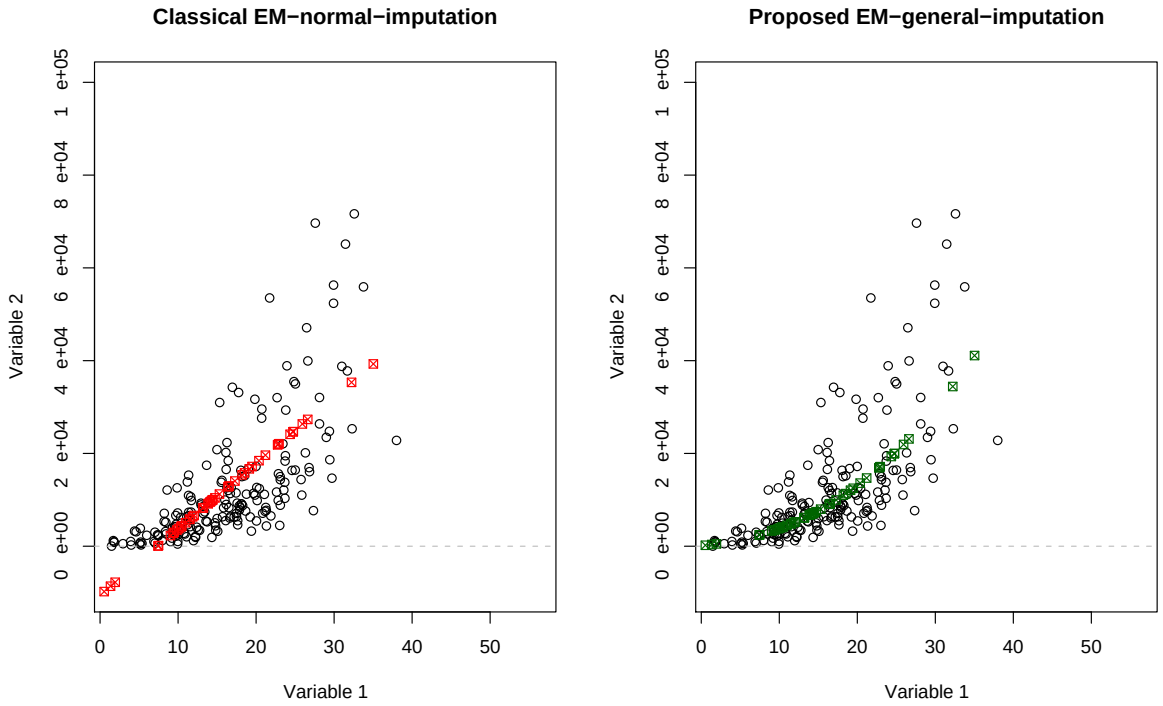

Figure 4.2: Comparison of the imputations.

A sample is drawn from a bivariate normal distribution and then raised to powers in such a way that the resulting marginal distributions are skewed to the right but with different skewnesses. Due to the different skewnesses, the cloud of points has a curved form. Both variables are defined on $\mathbb{R}^+$. Only variable two has missing values, which have to be imputed. In order to simplify matters, the comparison is performed by the EM-algorithm and not the Data Augmentation algorithm. The classical EM-algorithm for normal data imputes linearly (and so does the DA algorithm) thus yielding invalid (i.e. negative) values (left panel). To avoid that, the objective is to transform the dataset by a still to be presented power transformation before imputing the missing values. After the imputation, the augmented dataset can be transformed back. The better fit of the imputed values in the transformed dataset is evident (right panel). (Source: Huergo, 2008.)

4.2 Univariate Transformation

4.2.1 Purpose

As already mentioned, it is necessary to transform the KEI dataset to adjust the empirical distributions of the indicators to approximate normality. Beside location and dispersion, a further characterization of the data includes skewness and kurtosis. They are measures of the lack of symmetry and of the peakedness of a distribution. Skewness and kurtosis for a normal distribution are respectively zero and three. Hence, the main task of the transformation is to find a transformation parameter such that these values can be approximated as best as possible. A justification of why the proposed optimization routine only makes use of the third and fourth central moments is discussed in Appendix B.2.3.

Since the minimization cannot be carried out analytically, a numerical optimization routine is needed. The proposed algorithm uses a power transformation and has a structure which resembles a GMM estimation procedure.

4.2.2 The Transformation Algorithm

The proposed algorithm to search for the power transformation parameter works as follows:

1. **Sample Moments:** In a first step, the ML estimators μ and σ^2 have to be calculated for the transformed dataset $y^\star = y_{obs}^\theta$:

$$\begin{aligned} \overline{y} &= \tfrac{1}{n}\sum y_i^\star \,, \\ s &= \sqrt{\tfrac{1}{n}\sum (y_i^\star - \overline{y})^2} \,. \end{aligned}$$

2. **Z-Score** To prevent a collapse of the distributions, the transformed dataset has to be standardized. Therefore each value must be subjected to a z transformation:

$$z_i = \tfrac{y_i^\star - \overline{y}}{s} \,.$$

   Then z has a mean value of zero and a variance of one.

3. **Moment conditions:** The moment conditions of the third and fourth central moments of these z-variables have to be stated:

$$\overline{m}(\widehat{\theta}) = \begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \,,$$

   with

$$\overline{m}_1 = \tfrac{1}{n}\sum_{i=1}^{n}(z_i - \mu_z)^3 = \tfrac{1}{n}\sum_{i=1}^{n} z_i^3 \,,$$

$$\overline{m}_2 \;=\; \tfrac{1}{n}\sum_{i=1}^{n}\left((z_i-\mu_z)^4-3\sigma^4\right) = \tfrac{1}{n}\sum_{i=1}^{n} z_i^4 - 3\;.$$

The arranged moment conditions can be seen as the mean estimation errors, i.e. the average deviation between the standardized variables and their expected values.

There are more moment restrictions ($\overline{m}_1$ and $\overline{m}_2$) than parameters to estimate (θ). Therefore, there is no estimator $\widehat{\theta}$ which solves the sample moment conditions uniquely. The GMM-like idea is to look for a θ such that $\overline{m}$ is as close as possible to zero.

4. **Minimization:** The search can be done by using a quadratical form of $\overline{m}$ and minimizing it. In analogy to GMM a more general quadratic form using a weighting matrix has to be employed. Without loss of generality an identity matrix I_n can be used, so that both moment conditions receive an equal weight. But also other choices of weights are possible and discussed in literature. Formally, the transformation parameter is then defined to be the estimator that solves the following minimization problem:

   $$\arg\min Q(\theta) = \overline{m}'\,\mathrm{I}_n\,\overline{m}.$$

   To solve this optimization problem a numerical procedure must be used that searches the interval from a lower to an upper endpoint for a minimum of the moment conditions with respect to θ.

   As already mentioned, because of the use of higher moments of the normal distribution, this transformation would be more suitable for large sample sizes.

4.2.3 Properties

In order to explore the properties of the transformation parameter, a series of simulations was carried out. This section shows some of the results of these simulations.

(a) Convergence of the transformation parameter

One important feature of a transformation method is its ability to recognize whether a transformation is necessary or not. In other words, it should not transform data which already have the right shape. In order to test the behavior of the power transformation a simulation was carried out in which samples of increasing size from a normal distribution were drawn and subject to the transformation method. Except for sample effects the optimal transformation parameter in such cases should be one, since the data are already normally distributed. It is also to be expected that asystematic sample biases cancel out and that the amplitude of the deviations decreases with increasing sample sizes. Figure 4.3 shows the results of the simulation.

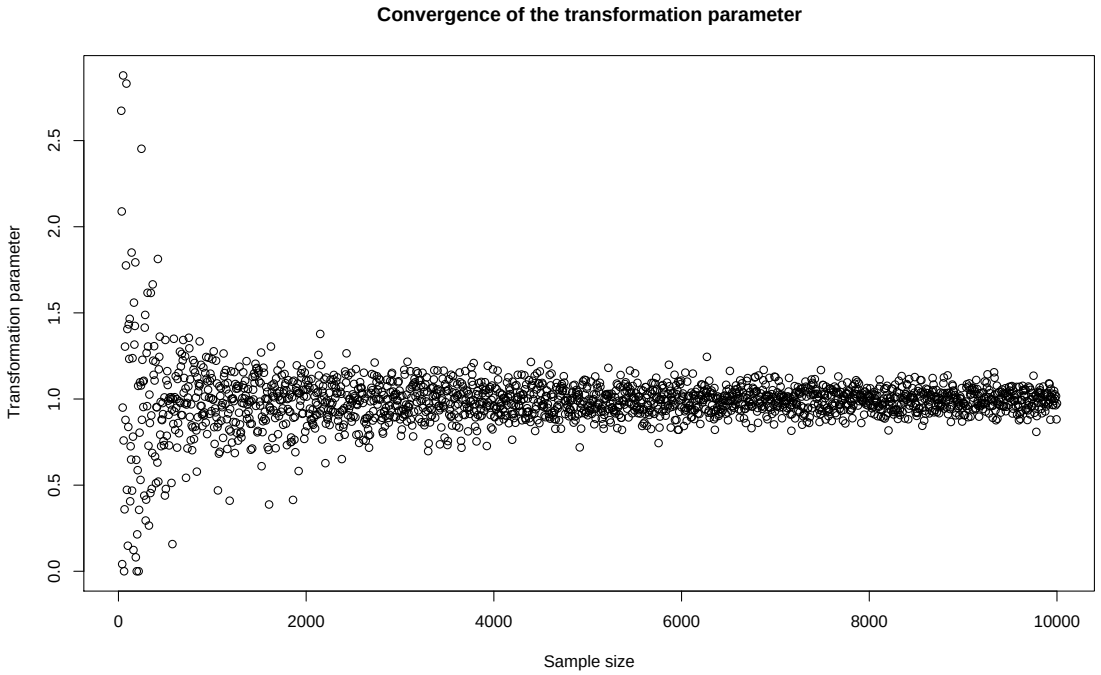

Figure 4.3: Convergence of the transformation parameter.

The normally distributed data - $X \sim N(8, 1.44)$ - used for the simulation are drawn with the `rnorm()` function of the R program. (Source: Huergo, 2008.)

(b) Consistency[3] of the transformation parameter

Three sets of Monte-Carlo simulations yield means and variances of transformation parameters (of 500 repetitions) which transform the given samples to approximate normality. The data are drawn with increasing sample sizes

b1) from a normal distribution where $X \sim N(8, 1.44)$. The transformation parameter must be on average one, because, except for stochastic effects, there is no need to transform a normal distribution. As it can be seen in Figure 4.4, the mean of the transformation parameters asymptotically converges to 1 and the variance decreases monotonically.

[3]It must be emphatically pointed out, that the word *consistency* as used in this section is in the sense of a Monte Carlo simulation. A further investigation of the theoretical properties of the transformation parameter is needed to establish its consistency. The same is valid for expressions such as *unbiasedness*.

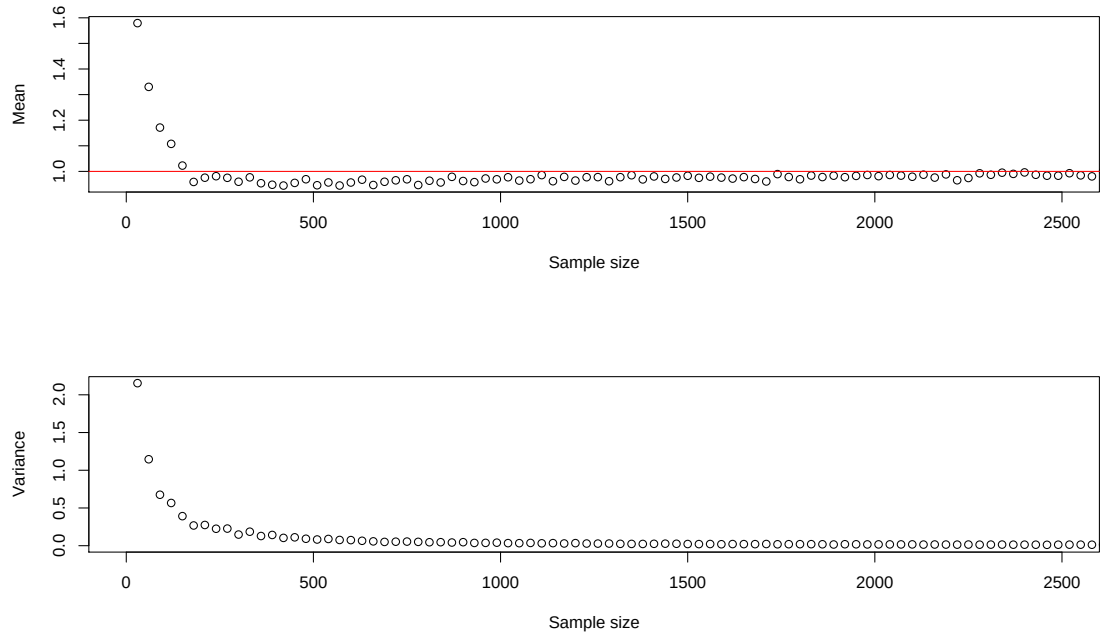

Figure 4.4: Evolution of the mean and the variance of the transformation parameter (1).

Evolution of the mean and the variance of the transformation parameter with increasing sample size. Samples from a $N(8, 1.44)$. The data are drawn with the `rnorm()` function of the R program. (Source: Huergo, 2008.)

b2) from a transformed normal distribution $f(X) = X^3$ where $X \sim N(8, 1.44)$. Thus, the transformation parameter must be $\frac{1}{3}$ to transform the data to normality. It can be seen in Figure 4.5 that the mean of the transformation parameters converges asymptotically to $\frac{1}{3}$ and the variance decreases monotonically.

b3) from a Gamma distribution where $X \sim \Gamma(4, 1)$. In contrast to the other cases it is not clear what the optimal transformation value should be. However it is interesting to see whether the transformation routine converges to a plausible value and to observe the behavior of the variance. Figure 4.6 shows that the mean of the transformation parameters converges asymptotically to $\frac{1}{3}$ and the variance decreases monotonically.

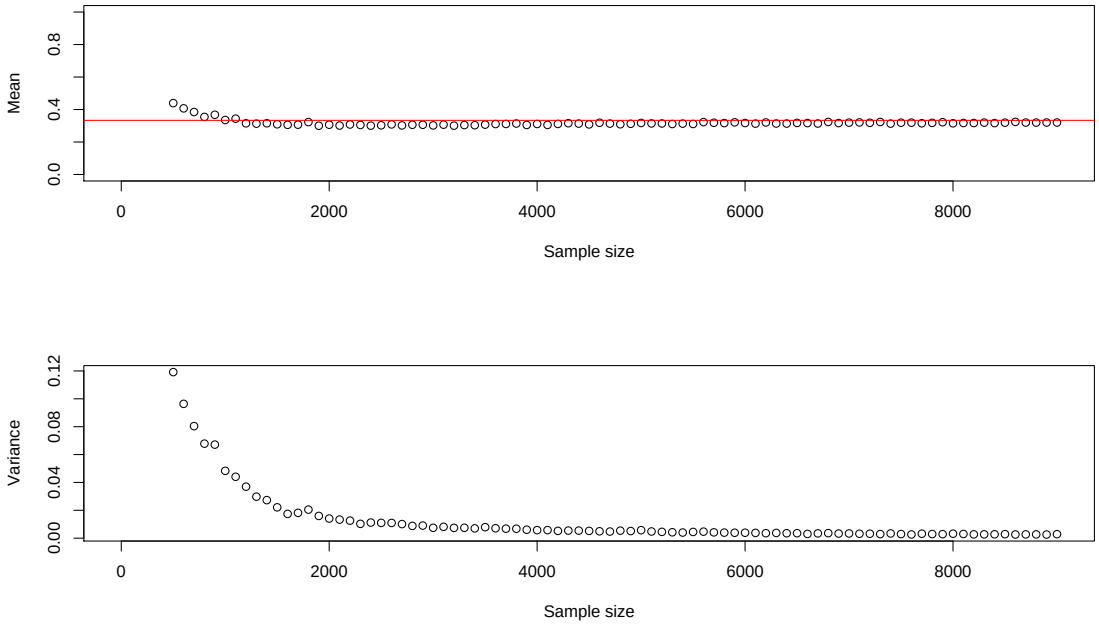

Figure 4.5: Evolution of the mean and the variance of the transformation parameter (2).

Evolution of the mean and the variance of the transformation parameter with increasing sample size. Samples from a transformed $N(8, 1.44)$. The transformed data - $f(X) = X^3$ - are drawn with a *Metropolis-Hastings* algorithm. (Source: Huergo, 2008.)

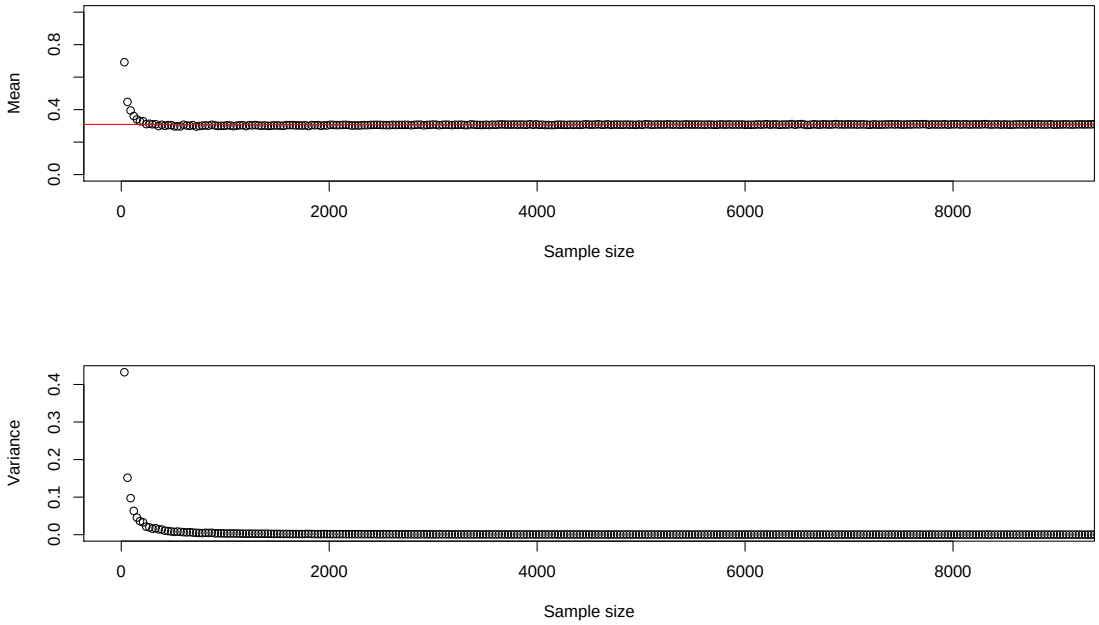

Figure 4.6: Evolution of the mean and the variance of the transformation parameter (3).

Evolution of the mean and the variance of the transformation parameter with increasing sample size. Samples from a $\Gamma(4, 1)$. The data are drawn with the `rgamma()` function of the R program. (Source: Huergo, 2008.)

(c) Asymptotic distribution of the transformation parameter

The empirical distributions of the resulting transformation parameters from b1) and b3) have been subjected to a Shapiro-Wilk normality test (where the significance level α is 0.05). Figure 4.7 shows the behavior of the acceptance rate of the test for different sample sizes. The acceptance rate increases approximately monotonically with increasing sample sizes. Huergo (2008) offers a more detailed explanation of the steps followed for this analysis.

It must be pointed out, however, that the simulative examination of the distribution of the transformation parameter can only be taken as a preliminary result. In order to draw definitive conclusions, this must be proven analytically by an appropriate *central limit theorem* (CLT).

(d) Comparison to the logarithmic transformation

Another frequently used non-linear transformation is the logarithm(ic) transformation. This transformation is often able to correct right-skewed data to a roughly symmetric and often normal-looking shape. Further advantages are its simplicity and its reversibility, i.e. $e^{\log(Y)} = Y$.

Because the proposed power transformation requires an numerical optimization algorithm and is in general of a more detailed structure, it is necessary to test whether the extra complexity is justified. In the next simulation, the two approaches for transforming data were compared.

Frequently occurring shapes of data (in the KEI dataset) skewed to the right and multimodal distributions:

d.1) The data in Figure 4.8 are drawn from an exponential distribution (i.e. skewed to the right). It is evident, that the proposed power transformation yields a better approximation to normality than the logarithmic transformation. Furthermore, the quality of the approximation improves with increasing sample sizes.

d.2) A random variable is said to have a lognormal distribution if its logarithm is normally distributed. That is, if Y is a random variable with a lognormal distribution, then $\log(Y)$ is normally distributed. The data for the test in Figure 4.9 are drawn from a lognormal distribution (i.e. skewed to the right) and thus represent the best possible case for the logarithmic transformation. The objective is to test the performance of the proposed power- against the logarithmic transformation in a setting favorable to the latter. To test the accuracy of the transformations, the p-values of the Shapiro-Wilk normality test are computed for the transformed data yielded by both approaches. The null hypothesis, that the data come from a normal distribution, must be rejected with a p-value smaller than or equal to a significance level α (set at 0.05). Figure 4.9 shows the discrepancy between both approaches by counting the number of non-identical results (whether rejected or not). On average, the two approaches differ only in approximatively one case out of 100 repetitions (independent from the sample size).

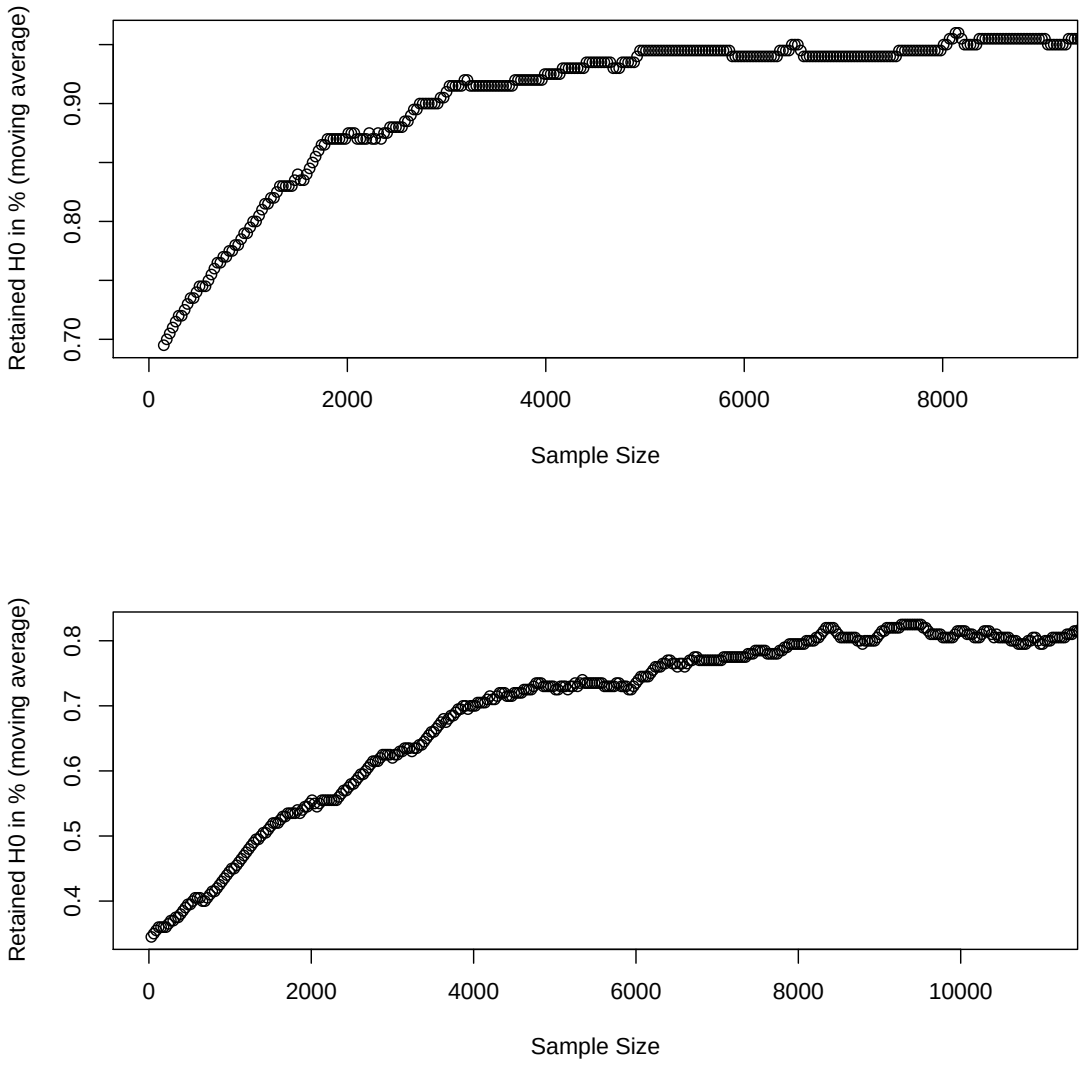

Figure 4.7: Empirical distributions of the transformation parameter by means of Shapiro-Wilk normality tests.

Upper figure: The normally distributed data - $X \sim N(8, 1.44)$ - used for the simulation are drawn with the `rnorm()` function of the R program. Lower figure: The gamma distributed data - $X \sim \Gamma(4, 1)$ - used for the simulation are drawn with the `rgamma()` function of the R program. (Source: Huergo, 2008.)

d.3) For the seek of completeness a counterexample is offered, in which the power transformation is not able to correct the shape of the data. The data for the following example were drawn from a uniform distribution. Figure 4.10 shows that neither the power transformation nor the logarithm transformation are able to correct the data in such a situation. Since uniformly distributed data can be brought to almost any distribution shape by the well known *Inverse Transform Method* (see Fishman, 2006, p. 77), this weakness of the power transformation is more of theoretical than of practical relevance.

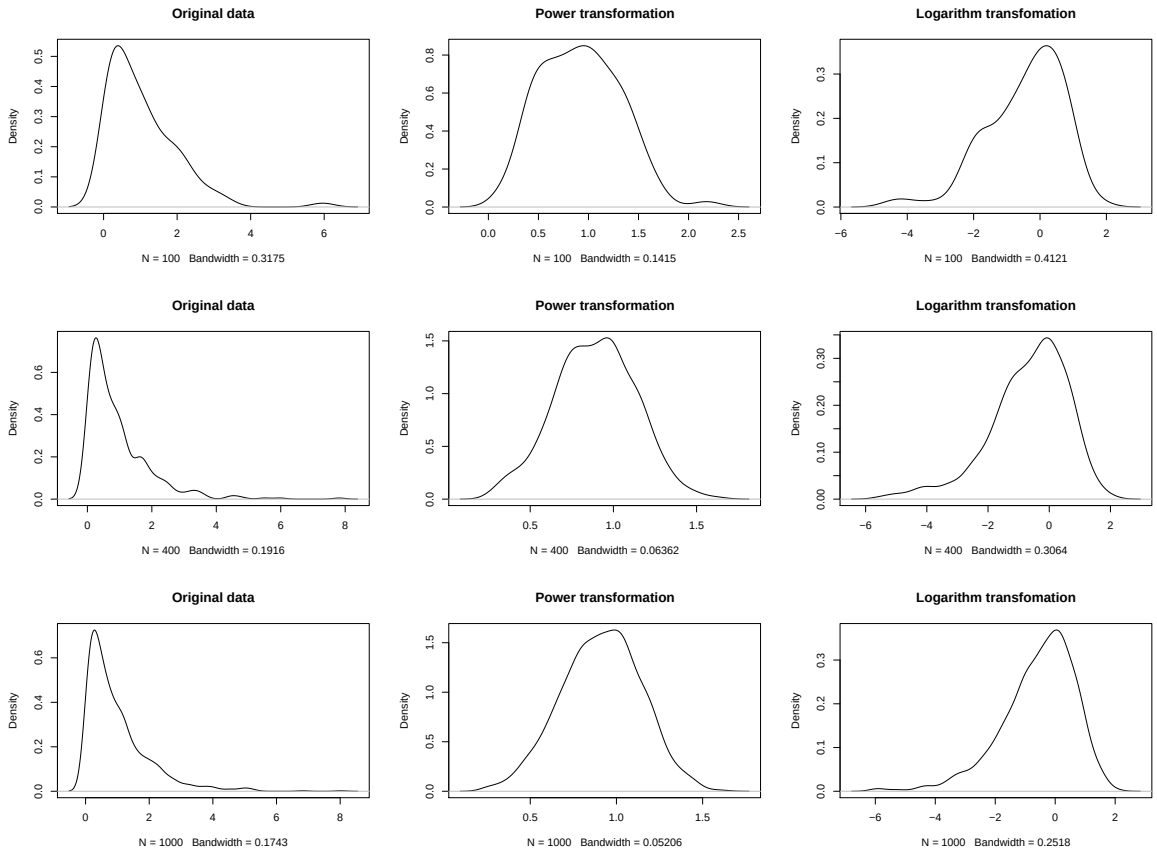

Figure 4.8: Power transformation and right-skewed data.

The exponentially distributed data - $X \sim \text{Exponential}(\lambda)$ where $\lambda = 1$ - used for the simulation are drawn with the `rexp()` function of the R program. (Source: Huergo, 2008.)

(e) Reversibility

The proposed power transformation is invertible in the sense that the dataset can be transformed back to its originally empirical distribution after the imputation step. Invertibility is a very important property for a transformation method. The following example illustrates the invertibility property on indicator A2a3. While Figure 4.11 presents the density of indicator A2a3 that has been transformed to normality and back to its original shape, Table 4.1 shows the computed values. The obtained values are rounded off to the 8th decimal place. A proof (in the real case) of the invertibility property of the power transformation is given in Appendix B.2.1.

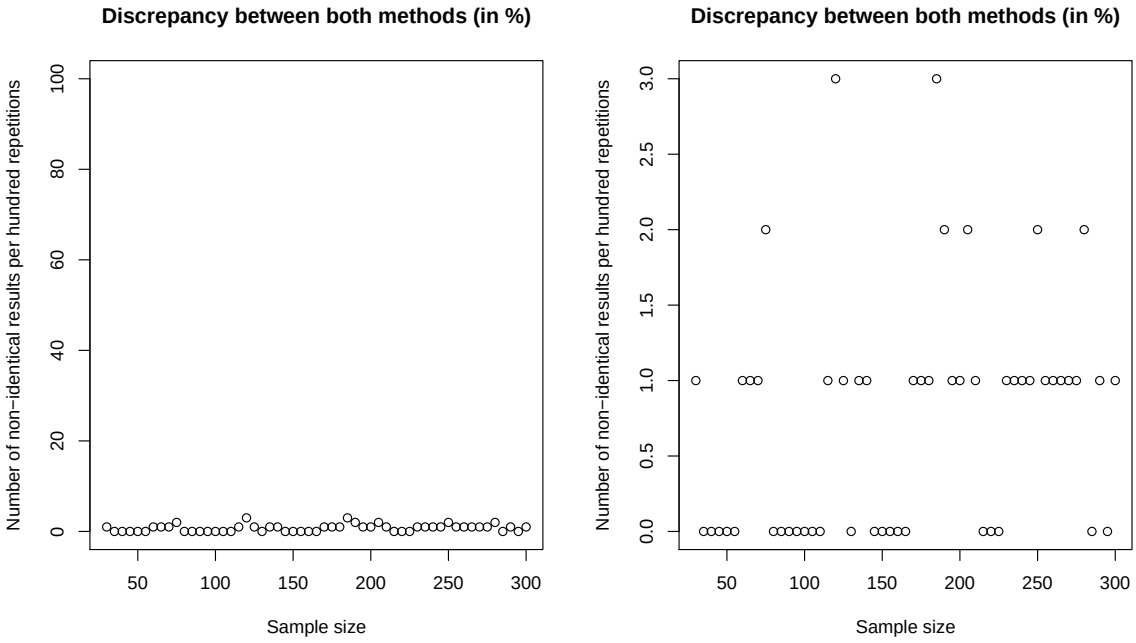

Figure 4.9: Comparison to logarithmization in case of lognormal data.

The lognormally distributed data - $X \sim \text{Log-N}(0, 1)$ - used for the simulation are drawn with the `rlnorm()` function of the R program. (Source: Huergo, 2008.)

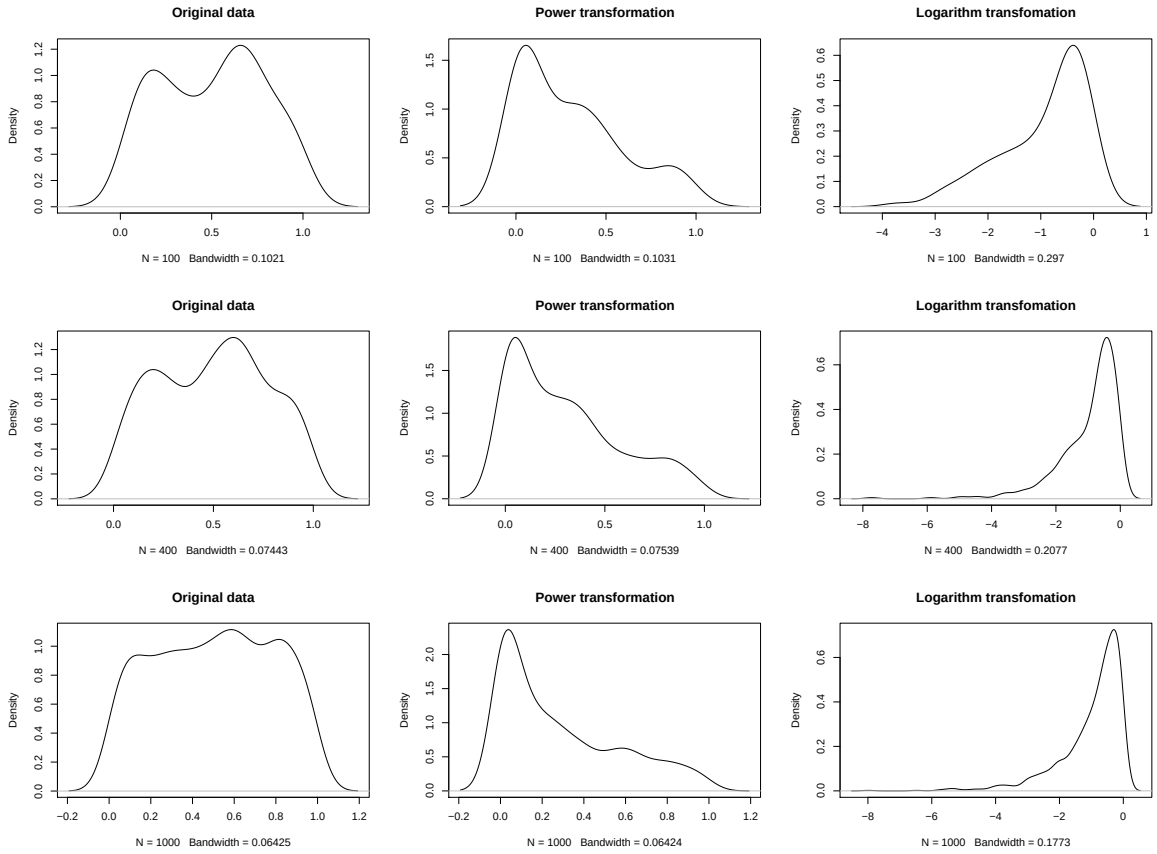

Figure 4.10: Power transformation and uniformly distributed data.

The uniformly distributed data - $X \sim \mathrm{U}(0,1)$ - used for the simulation are drawn with the `runif()` function of the R program. (Source: Huergo, 2008.)

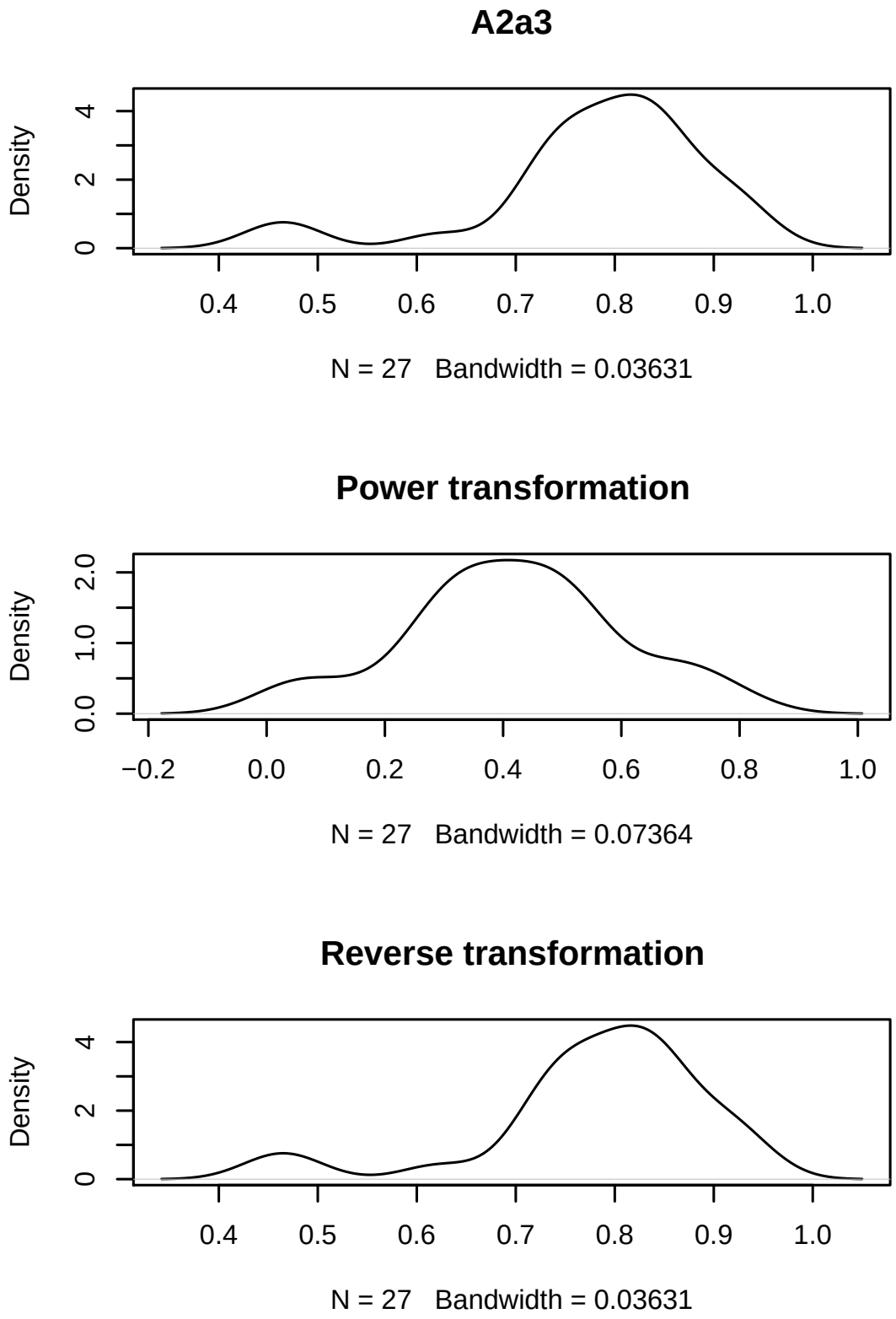

Figure 4.11: Reverse transformation of indicator A2a3.

The density of indicator A2a3 (in 2004) has been transformed to normality and back to its original shape. The transformation parameter θ is 3.868025. (A Shapiro-Wilk normality test yields a p-value of 0 for indicator A2a3 and 0.79 for the transformed data.) (Source: Enderle, 2008.)

Country	Original data	Power transformation	Reverse transformation
at	0.858	0.54602946	0.858
be	0.818	0.45216666	0.818
cy	0.776	0.36716103	0.776
cz	0.914	0.70097576	0.914
de	0.728	0.28529703	0.728
dk	0.762	0.34167872	0.762
ee	0.803	0.42028346	0.803
es	0.612	0.14370765	0.612
eu15	0.743	0.30923734	0.743
eu25	0.772	0.35974033	0.772
fi	0.845	0.51406666	0.845
fr	0.814	0.44349375	0.814
gr	0.830	0.47894653	0.830
hu	0.835	0.49044741	0.835
ie	0.853	0.53356548	0.853
it	0.734	0.29470051	0.734
jp	NA	NA	NA
lt	0.850	0.52618983	0.850
lu	0.725	0.28068024	0.725
lv	0.795	0.40398218	0.795
mt	0.510	0.06992687	0.510
nl	0.750	0.32090886	0.750
pl	0.909	0.68594732	0.909
pt	0.496	0.06264463	0.496
se	0.860	0.55107545	0.860
si	0.905	0.67409890	0.905
sk	0.917	0.71011005	0.917
uk	0.770	0.35607228	0.770
us	NA	NA	NA

Table 4.1: Reverse transformation of indicator A2a3. (Source: Enderle, 2008.)

(f) Simulative examination of the variance of the transformation parameter as a function of the sample size

To test its behavior, the variance of the transformation parameter will be estimated by a *constant elasticity model* where the logarithm of the variance is regressed on the logarithm of the sample size:

$$\log\left(var(\widehat{\theta})\right) = \beta_0 + \beta_1 \log(n) + \text{error} \ .$$

Thus, the variance can be expressed as a function of the sample size. Figure 4.12 shows the variance of the simulation in b1) which is used to estimate the order of $var(\widehat{\theta})$. Under the assumption of valid Monte Carlo results it holds that

$$\begin{aligned}
&\text{var}(\widehat{\theta}) \approx \tfrac{100}{n^{1,13}} \\
&\text{var}(\widehat{\theta})\, n^{1,13} \approx 100 \\
&\text{var}(\widehat{\theta})\, n^{1,13} \approx \text{c} \\
&\text{var}(\widehat{\theta}) \in \mathcal{O}(n^{-1,13}) \ .
\end{aligned}$$

Hence, the variance of the transformation parameter is of the order $n^{-1.13}$ (for more details see Appendix B.2.2).

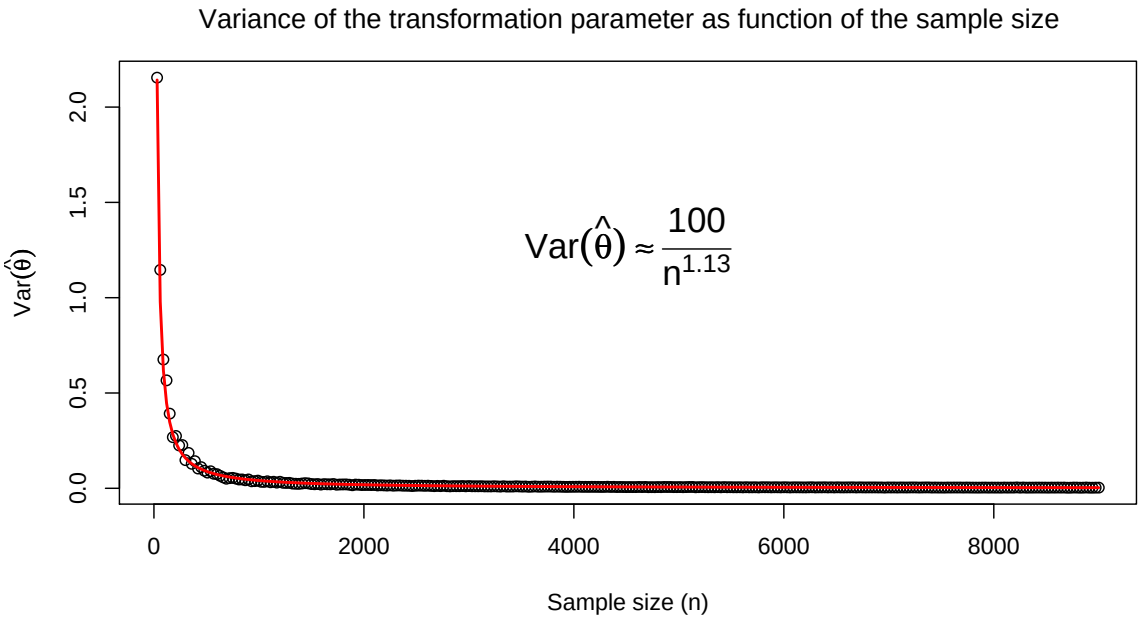

Figure 4.12: Variance of the transformation parameter.

Figure 4.4 continued. The red line is an estimation of the variance (from a log-log or constant elasticity model): $\log\left(var(\widehat{\theta})\right) = 4,57 - 1,13 \log(n)$. (Source: Huergo, 2008.)

4.3 Multivariate Transformation

4.3.1 Complete Dataset

Since the primary algorithm is merely thought to univariately transform the indicators, a proposed extension to deal with multivariate datasets is now discussed. Suppose $v = (v_1, v_2, \ldots, v_p)$ is such a dataset with p indicators which have to be transformed to approximate multivariate normality. The multivariate algorithm works as follows:

1. First of all, the marginal distribution of the first indicator[4], v_1, must be transformed with the univariate algorithm. This univariate transformation yields parameter θ_1.

2. The second step takes into account the fact that indicator v_1 has alread been transformed. To account for that, the univariate algorithm is replaced by a multivariate one. This new algorithm has two sets of moment conditions:

   (a) **y-conditions:** Two moment conditions for the marginal distributions, in analogy to the univariate transformation algorithm (i.e. $\overline{m}_1$ and $\overline{m}_2$).

   (b) **e-conditions:** Two moment conditions for the distribution of the residuals, when a linear regression of the new variables on the already transformed variable(s) is performed.

   As a result, the second indicator v_2 is transformed in such a way that both the marginal distribution and the distribution of the residuals of a linear regression on the already transfomed indicator v_1 are corrected to an approximate normal shape.

3. At each additional iteration step a further variable of the remaining $p-2$ indicators $(v_3, \ldots, v_p)$ gets transformed. The number of regressors increases with each step. The number of moment conditions remains constant.

Comment on the e-conditions: The idea behind the e-conditions is simple, since all conditional distributions of a joint normal distribution are also normal, the residuals of a regression of two normally distributed variables are normally distributed as well. For this reason, when regressing some indicators on already transformed (or normally distributed) indicators, one can assume normally distributed residuals. Thus, since the e-conditions control the normality of the residuals, the new indicators become by and by approximatively normal. The formal derivation of the e-conditions proceeds as follows:

Run a regression of y on X to obtain the residuals

$$e = y^* - X\left(X'X\right)^{-1} X'y^* \ ,$$

which must be standardized by a z transformation:

$$z_e = \frac{e-\bar{e}}{\sigma_e} = \frac{e}{\sigma_e} \ ,$$

[4]Without loss of generality the natural order was chosen.

where $\sigma_e = \sqrt{\frac{1}{n}\sum e^2}$ and the residuals add to zero by definition. Once again, z_e is assumed to be $N(\mu_z, \sigma^2_{z_e}) = N(0,1)$. Then, the moment conditions of the third and fourth central moments of these z_e-variables can be stated:

$$\begin{aligned}
\overline{m}_3 &= \tfrac{1}{n}\sum_{i=1}^{n}(z_{e,i} - \mu_{z_e})^3 = \tfrac{1}{n}\sum_{i=1}^{n}(z_{e,i})^3 \,, \\
\overline{m}_4 &= \tfrac{1}{n}\sum_{i=1}^{n}((z_{e,i} - \mu_{z_e})^4 - 3\sigma^4_{z_e}) = \tfrac{1}{n}\sum_{i=1}^{n}(z_{e,i})^4 - 3 \,.
\end{aligned}$$

Apart from two more moment conditions being included (i.e. all in all four), the minimization task of the second step turns out to be identical to the univariate approach. Note, that the y- and e-conditions are necessary but not sufficient conditions for the multivariate normality.

To obtain all p transformation parameters, such a minimization must be achieved $p-1$ times. At each iteration step, a further indicator of the dataset will be included as dependent variable whereas the newly transformed indicator (i.e., the dependent variable from the iteration step before) will then be added to the regressor(s). In doing so, the number of moment conditions doesn't increase with the number of variables.

4.3.2 Incomplete Dataset

So far, the transformation algorithms have dealt with completely observed data. However, the presence of missing values is the reason for the imputation procedures and hence for the present algorithms.

One possibility when dealing with datasets with missing values is to univariately transform the observed data. In analogy to the general imputation problem, and depending on the mechanism of missingness, there may occur a rough bias when the univariate transformation is applied to the observed data. Indeed, this effect increases with the amount of missing data. For example, a high amount of missing data on a sample from a skewed distribution can cause a shift in the skewness, which causes the transformation algorithm to deliver an incorrect transformation parameter. This is more likely to occur under MAR or NMAR mechanisms.

The described multivariate algorithm is based upon regressions and therefore designed for complete datasets. It is then natural to look for a way to suitably *complete* the data before the transformation.

The proposed idea is to develop an iterative tandem approach between the transformation- and the EM-algorithm:

1. In a first step, the EM algorithm must be applied to the still untransformed data in order to get a *complete* dataset.

2. All variables (including the imputed data) undertake an univariate transformation. The resulting transformation parameters, θ_0 are saved.

3. The EM algorithm is applied to the θ_0-transformed variables.

4. The expanded variables are multivariately transformed and a new vector of parameters, θ_1, results.

5. The EM algorithm is applied to the $(\theta_0 \cdot \theta_1)$-transformed variables.

6. The iteration runs until the product of the parameters does not change with additional iterations.

Extensive simulations have shown that the proposed algorithm performs fairly well in practice, and that the resulting imputations in general settings are more accurate than the imputations of the uncorrected data. However, further work is necessary, especially concerning the moment conditions and the tandem with the EM-algorithm, in order to identify the conditions under which the algorithm is likely to perform well, as well as possible pitfalls. Until then it is advisable to consider it a prototype.

Chapter 5

Robust Models

5.1 General Mixture Model

5.1.1 Parametrization

In addition to the classical normal model, some positive unobserved scalars q_i ($i = 1, 2, \ldots, n$) which are an i.i.d. sample from the density $h(q)$ will be taken into account, so that the new models' supposed multivariate distribution of Y will read as follows

$$(y_i \mid \theta, q_i) \overset{ind}{\sim} MVN_k(\mu, \Psi/q_i) \;, \tag{5.1}$$

where k is the number of variables. Because this multivariate normal distribution belongs to the exponential family of distributions, the loglikelihood in Equation (2.9) is not linear in the data but rather linear in a set of sufficient statistics. For the extended model the complete-data sufficient statistics are: $T_0 = \sum_{i=1}^{n} q_i$, $T_1 = \sum_{i=1}^{n} q_i y_i$ and $T_2 = \sum_{i=1}^{n} q_i \, y_i y_i'$, which can be arranged in a $(p+1) \times (p+1)$ matrix

$$T = \begin{bmatrix} T_0 & T_1' \\ T_1 & T_2 \end{bmatrix} .$$

For a complete setting, this is if q and Y are completely observed, the ML estimates of $\theta = (\mu, \Psi)$ could be found by weighted least squares:

$$\widehat{\mu} = \tfrac{T_1}{T_0} \;, \tag{5.2}$$

$$\widehat{\Psi} = \tfrac{1}{n}\left(T_2 - \tfrac{T_1 T_1'}{T_0}\right) \;. \tag{5.3}$$

But in case of missing data and unknown weights, some extensions to the EM algorithm have to be accommodated:

5.1.2 Implementation

In the e-step, the complete-data sufficient statistics will be estimated by their conditional expectations:

$$E\left[T_0 \mid Y_{obs}, \theta\right] = E\left(\sum_{i=1}^{n} q_i \mid Y_{obs}, \theta\right) = \sum_{i=1}^{n} w_i^{(t)}$$

with the estimated weights $w_i^{(t)} = E(q_i \mid Y_{obs}, \theta^{(t)})$.

The jth component of $E\left[T_1 \mid Y_{obs}\right]$ is

$$\begin{aligned} E\left(\sum_{i=1}^{n} q_i y_{ij} \mid Y_{obs}, \theta\right) &= \sum_{i=1}^{n} E\big\{q_i\, E(y_{ij} \mid Y_{obs}, q_i, \theta) \mid Y_{obs}, \theta\big\} \\ &= \sum_{i=1}^{n} w_i^{(t)} E(y_{ij} \mid Y_{obs}, \theta) \\ \text{or} \quad E(q_i y_{ij} \mid Y_{obs}, \theta) &= \begin{cases} w_i^{(t)} y_{ij}, & \text{if } j \in \mathcal{O}(s) \\ w_i^{(t)} y_{ij}^*, & \text{if } j \in \mathcal{M}(s) \end{cases} \end{aligned}$$

and the (j,k)th element of $E\left[T_2 \mid Y_{obs}\right]$

$$\begin{aligned} E\left(\sum_{i=1}^{n} q_i y_{ij} y_{ik} \mid Y_{obs}, \theta\right) &= \sum_{i=1}^{n} E\{q_i E(y_{ij} y_{ik} \mid Y_{obs}, q_i, \theta) \mid Y_{obs}, \theta\} \\ &= \sum_{i=1}^{n} E\Big\{q_i \left[E(y_{ij} \mid Y_{obs}, \theta) E(y_{ik} \mid Y_{obs}, \theta)\right. \\ &\qquad \left. + \operatorname{cov}(y_{ij} y_{ik} \mid Y_{obs}, q_i, \theta)\right] \Big| Y_{obs}, \theta\Big\} \\ &= \sum_{i=1}^{n} w_i\, E(y_{ij} \mid Y_{obs}, \theta)\, E(y_{ik} \mid Y_{obs}, \theta) + \Psi_{ik,obs,i} \\ \text{or} \quad E(q_i y_{ij} y_{ik} \mid Y_{obs}, \theta) &= \begin{cases} w_i\, y_{ij} y_{ik}, & \text{if } j,k \in \mathcal{O}(s) \\ w_i\, y_{ij}^* y_{ik}, & \text{if } j \in \mathcal{M}(s), k \in \mathcal{O}(s) \\ w_i\, y_{ij}^* y_{ik}^* + a_{jk}, & \text{if } j,k \in \mathcal{M}(s) \end{cases} . \end{aligned}$$

The m-step turns out to calculate the new estimates $(\mu^{(t+1)}, \Psi^{(t+1)})$ from Equations (5.2) and (5.3) as done in the classical model but with T_0, T_1 and T_2 replaced by their estimates from the e-step. The PX-EM speeds the convergence by replacing the denominator n in Equation (5.3) by the sum of the current weights, $\sum_{i=1}^{n} w_i^{(t)}$.

Before stating different models, i.e. defining a distribution for the weights, a measure for detecting outliers has to be defined.

5.1.3 Mahalanobis Distance

The size and shape of the distribution of a multivariate dataset are quantified by its covariance matrix. A very useful distance measure which takes into account the covariance among variables is the Mahalanobis distance (Mahalanobis, 1936). It measures the distance of a case from the centroid (multivariate mean) of a distribution, given the covariance (multivariate variance) of the distribution. Therefore it is often used as a multivariate outlier detection method, which turns out to be useful for weighting observations.

The squared distance from the mean for observed variables in case i reads as follows

$$d_i^{(t)} = \sqrt{(y_{obs,i} - \mu_{obs,i}^{(t)})' \Psi_{obs,i}^{(t)-1} (y_{obs,i} - \mu_{obs,i}^{(t)})} \,,$$

where $\mu_{obs,i}$ is the a vector of means and $\Psi_{obs,i}$ is the variance covariance matrix. Large squared distances d_i^2 denote outliers and involve downweighting of cases depending on the purposed model.

5.2 Contaminated Normal Model

To derive the contaminated (multivariate) normal model, following distribution for $q_i = w_i$ must be assumed:

$$h(w_i) = \begin{cases} 1-\delta & \text{if } w_i = 1 \\ \delta & \text{if } w_i = \lambda \\ 0 & \text{otherwise,} \end{cases} \tag{5.4}$$

where $0 < \delta < 1$, $\lambda > 0$ with known probability of contamination δ and known λ. Then the marginal distribution for y_i is a mixture of the two distributions

$$N(\mu, \Psi) \text{ and } N(\mu, \Psi/\lambda) \,.$$

For the contaminated normal model, one sets $\lambda \ll 1$ (say 0.1). Little and Rubin (2002) show that the weight can be derived by a simple application of Bayes' theorem. For case i the distribution in Equation (5.4) yields the weight

$$w_i^{(t)} = \frac{1-\delta+\delta\lambda^{1+\frac{k_i}{2}} \exp\left\{(1-\lambda)\frac{d_i^{(t)2}}{2}\right\}}{1-\delta+\delta\lambda^{\frac{k_i}{2}} \exp\left\{(1-\lambda)\frac{d_i^{(t)2}}{2}\right\}} \,.$$

Whereas the contaminated normal model is designed for especially downweighting outliers, the following t-model produces smoothly declining weights with increasing d_i^2.

5.3 Multivariate t-Model

5.3.1 t-Model (with known ν)

Another choice of deriving a form of the weights is to suppose the weights w_i are such that $w_i\nu$ is chi-squared distributed with degrees of freedom ν, that is $w_i \sim_{ind} \chi_\nu^2/\nu$. Therefore,

by mixing Equation (5.1) with the scaling variable $q = w$, the marginal distribution for y_i is defined by

$$y_i \sim_{ind} t_k(\mu, \Psi, \nu),$$

where t_k denotes a k-variate Student's t-distribution with the probability function

$$P(Y \mid \mu, \Psi, \nu) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)|\Psi|^{-\frac{1}{2}}}{\Gamma(\frac{\nu}{2})\left\{\Gamma\left(\frac{1}{2}\right)\right\}^k \nu^{\frac{k}{2}}} \times \left(1 + \frac{(Y-\mu)\Psi^{-1}(Y-\mu)'}{\nu}\right)^{-(\frac{\nu+k}{2})} .$$

The weights can be yielded by a application of Bayes' theorem. Hence, for case i results

$$w_i^{(t)} = E(q_i \mid Y_{obs}, \theta^{(t)}) = \frac{(\nu+k_i)}{(\nu+d_i^{(t)2})} . \tag{5.5}$$

Whereas both the contaminated normal model as well as the t-model assume fixed parameters to calculate the weights, they are said to not be very flexible.

5.3.2 Adaptive t-Model (with unknown ν)

Therefore, a further extension to the t-model relaxes the assumption of a fixed parameter ν such that it becomes more flexible. Thus, the degrees of freedom ν in Equation (5.5) must be replaced by a current estimate $\nu^{(t)}$. Then, the m-step which calculates new estimates $(\mu^{(t+1)}, \Psi^{(t+1)})$ has to be extended by also computing a new $\nu^{(t+1)}$. To make this feasible, one applies the ECME algorithm, which splits up the m-step in two CM steps:

CM1: To find new parameters $(\mu^{(t+1)}, \Psi^{(t+1)})$, maximize the Q-function with respect to $\theta = (\mu, \Psi)$.

CM2: Maximization of the observed, true likelihood with respect to ν with fixed parameters $(\mu^{(t+1)}, \Psi^{(t+1)})$ to find new parameter $\nu^{(t+1)}$. This can be done by a one-dimensional maximization of the observed likelihood

$$\begin{aligned} \ell(\nu, \mu, \Psi \mid Y_{obs}, \mu^{(t+1)}, \Psi^{(t+1)}) \;=\; & -\tfrac{n}{2}\log|\Psi| + n\log\left(\Gamma(\tfrac{\nu+k}{2})\right) \\ & -\tfrac{nk}{2}\log(\nu) - n\log\left(\Gamma(\tfrac{\nu}{2})\right) \\ & -\tfrac{\nu+k}{2}\sum_{i=1}^{n}\left(\log(1 + \tfrac{(y_i-\mu)\Psi^{-1}(y_i-\mu)'}{\nu})\right) , \end{aligned}$$

this is the sum of the logarithm of the density of a multivariate Student's t distribution over all i.

5.4 Draws from the Posterior Distribution

Since the aforementioned accommodations are presented for the EM algorithm, this section touches on how to implement draws from the posterior distribution of the parameters. Whereas the modifications to the models in Sections 5.2 and 5.3.1 are straightforward without any larger obstacles, the adaptive multivariate t-model must be expanded by draws of ν. But these cannot be drawn directly due to numerical underflow which means that numerical values can't be displayed by a computer due to their length. Since probabilities range in the interval 0 to 1, where infinitely many values can be assumed, one needs a method that can exploit the range of the computer and reduce rounding errors.

At first, the p-step has to be split up in two sub-steps P1 and P2 as follows:

P1: In analogy to the p-step for the normal model, new parameters $\Psi^{(t+1)}$ and $\mu^{(t+1)}$ will be drawn from the distributions in Equations (3.8) and (3.9).

P2: Given the computed parameters $(\Psi^{(t+1)}, \mu^{(t+1)})$ and a constant prior for ν, a new estimation of the degrees of freedom must be drawn from the posterior distribution.

Since it cannot be drawn directly from the posterior distribution, an approximation to the inverse cumulative distribution function (cdf) must be formed. This can be done by the simple and intuitive *Griddy Gibbs* sampler (Tanner, 1991, pp. 101) which is based on the empirical distribution method. Following steps are required to create such a Griddy Gibbs sampler:

1. Compute a grid of points $v_1, v_2, \ldots$, such that most of them are in the neighborhood of high mass and fewer points of low mass. The proposed approach is that the grid is based on the slope of the first derivatives rather than spaced uniformly. So, it will result in a finer grid near to the maximum and computer time can be reduced.

2. Evaluate $P(\nu^{(t+1)} \mid Y, \Psi^{(t+1)}, \mu^{(t+1)})$ at these points to form an approximation of the inverse cdf. Generally, the inverse cdf is used to obtain the original values that were used for calculating the cdf. If the distribution is continuous, the result of the inverse cdf is the original value.

3. Sample a random variate p from a continuous uniform distribution on the interval $[0, 1]$ and draw ν from the inverse cdf via an approximation of the p-th quantile of the distribution.

Chapter 6

Imputation Round

6.1 Purpose

The main objective of this work was to implement and develop suitable methods which allow for an imputation of the challenging KEI dataset. Therefore, it is important to show how the results in the underlying work have been implemented. The imputation of missing data within the KEI project is, in contrast to the examination and improvement of imputation models, predominantly a practical task. Its results can have an impact on the whole project. Hence, the imputation of such a voluminous dataset is not at all capable of being automated but rather must be undertaken manually with care.

In the course of the last imputation rounds some graphical tools were used and to some extent specially developed for this task. Thus, next section will throw a glance at these tools.

6.2 Graphical Tools

6.2.1 The Correlation Map

The *correlation map* is a matrix graphic, which converts correlation matrices into colored matrices by assigning a color to each cell of the matrix, where the scale conforms to the absolute value of the correlations. Whereas in recent imputation rounds one had to examine the correlation matrix with its huge dimension ($p \times p$), this graphic enables to get a quick overview of the the underlying data situation. In doing so, it is much easier to construct models on the basis of correlations. Figure 6.1 shows the correlation structure of the whole dataset. The graphic used for the imputation is a modification of the add-on library spatstat (Baddeley, 2008).

6.2.2 The Exploration Graphic

The purpose-built exploration graphic for the project is also a matrix graphic. It provides different information for a (small) dataset, which is very useful when constructing and

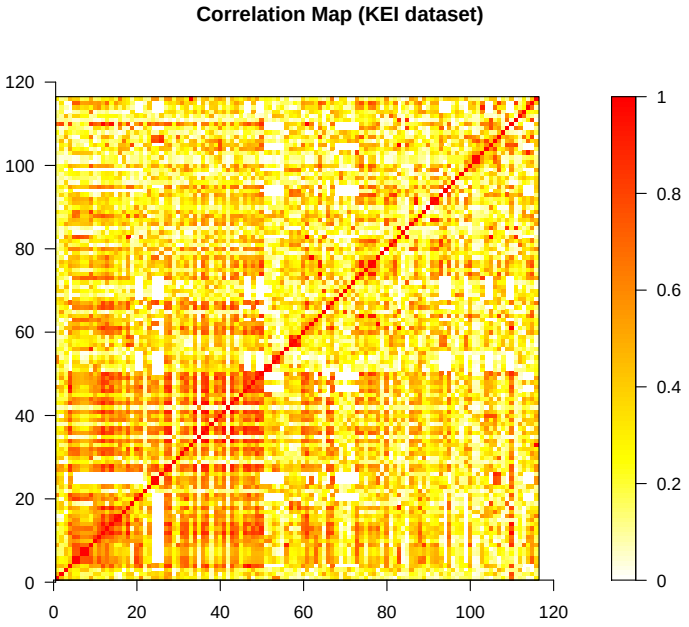

Figure 6.1: Correlation map of the whole KEI dataset. (Source: Enderle, 2008.)

verifying models for the imputation. Figure 6.2 shows an example of this tool. The main diagonal contains:

- a kernel density estimation of the empirical (marginal) distributions and probability functions of normal distribution which are parameterized by the empirical arithmetic mean and variance,
- the proportion of missing data (NAs) of the marginal distributions and
- the p-values of the Kolmogorov-Smirnov tests for normality of the marginal distributions. The null hypothesis of normality has to be rejected with a p-value smaller than or equal to a significance level α (often set at 0.05).

Furthermore, the remaining cells of the graphic provide:

- the proportions of pairwise completely observed data (below the main diagonal),
- the correlations of all observed pairs of variables (above the main diagonal) and
- the scatter plots of all available pairs of variables, featured with a smooth regression curve (i.e. a non-parametric *Lowess* (Cleveland, 1981, 1979) regression to correct for extreme outliers and check the linearity of the relation between indicators).

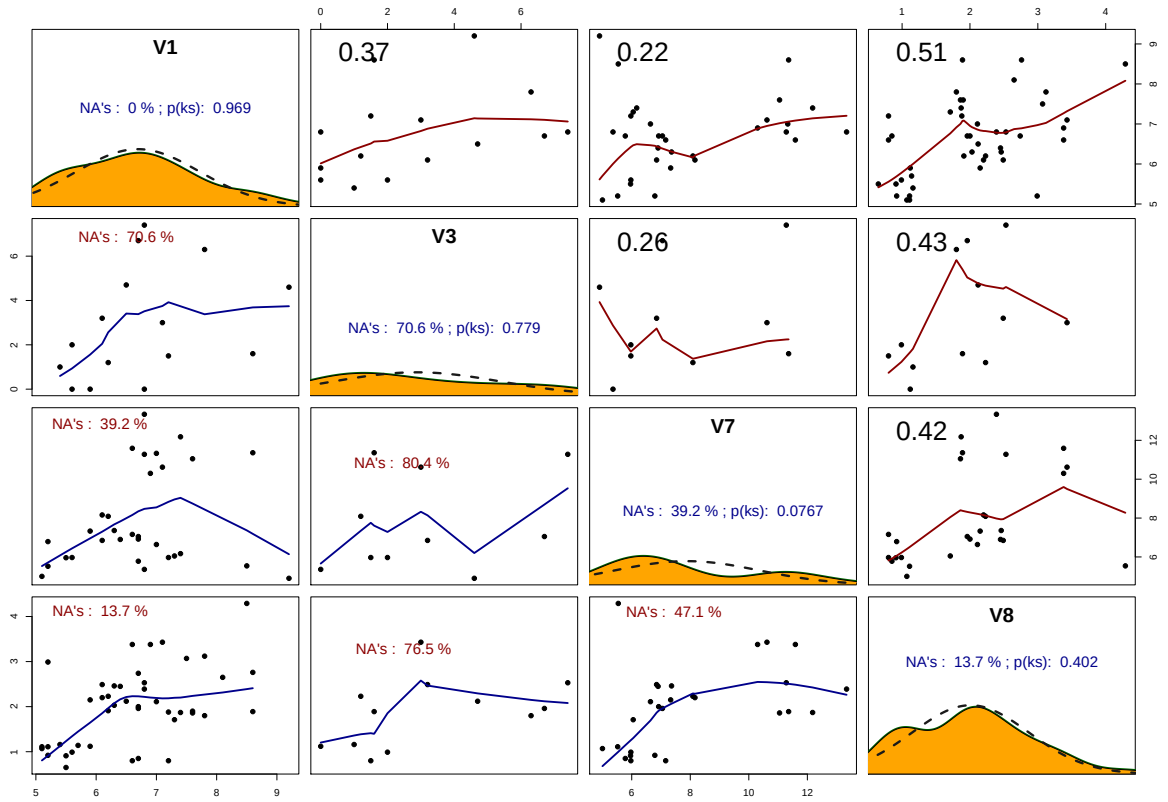

Figure 6.2: Exploration Graphic of a dataset.

6.2.3 The Cluster Dendrogram

The cluster dendrogram is a graphical output of the *Cluster Analysis* which strives to bundle objects. Its aim is to cluster these objects into groups such that the objects within groups are as similar as possible (high intra-correlation) and that the groups among themselves are as dissimilar as possible (low inter-correlation). The typical approach of a cluster analysis is to measure the differences between objects (i.e. indicators). For the KEI project, a simple correction was applied: To obtain differences, correlations between indicators (within the v-th model) have been used *as distance* as follows

$$D_v = 1 - |\rho_v| \ ,$$

where ρ_v presents the correlation matrix of model v.

The distance matrix was computed by using the Euclidean distance measure to weight larger differences more strongly. Then, the cluster analysis must be carried out by trying to build homogenous clusters, i.e. such that the variance within groups is small. This can be achieved by the *Ward* method, which computes a heterogeneity measure and is conservative in the sense that it builds equally large groups. A dendrogram of the indicators of the KEI dataset is given in Figure A.2 (p. 74). Only 94 indicators were used because the others had to few complete pairwise observations.

6.2.4 The Multiway Dot Plot

Relating to the KEI project, the Multiway Dot Plot (Cleveland, 1994) is a graphical approach that compares the imputed values of both implicit and multiple imputation methods. An example of the graphic is given in Figure 6.3.

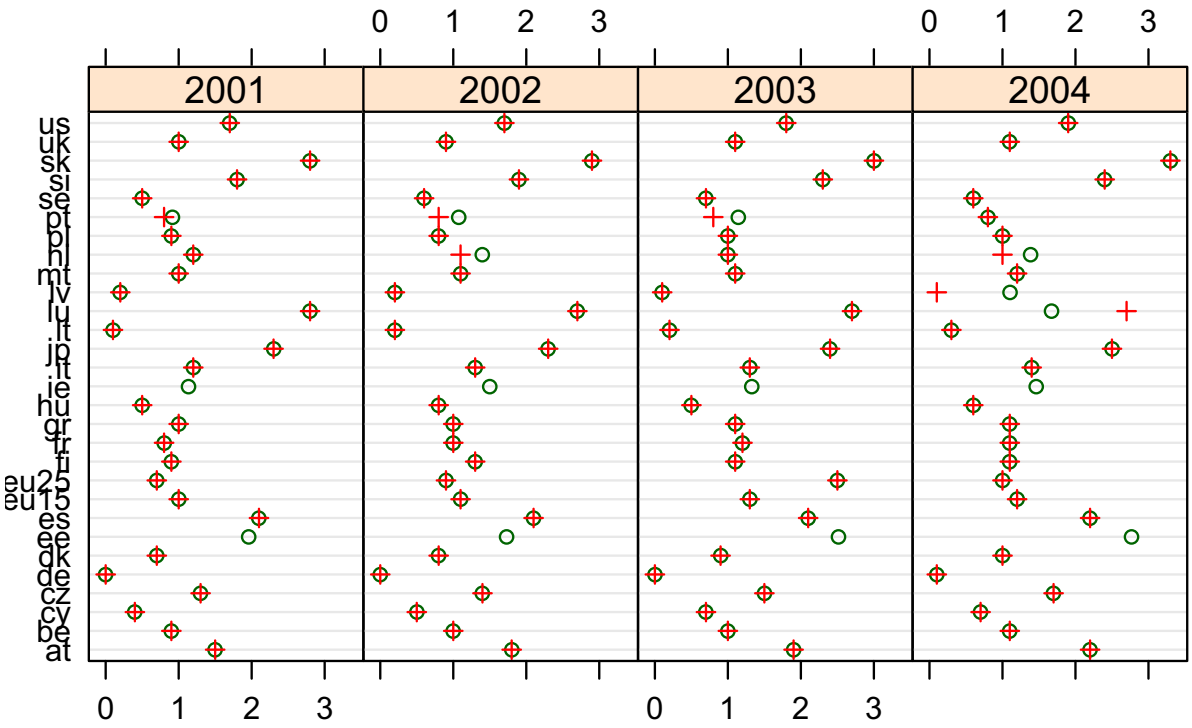

Figure 6.3: Example of the Multiway Dot Plot.

The graphic compares the values computed by two different methods (represented by circles and crosses) for all observations (i.e. countries).

6.3 The Sequence of the Imputation

As already mentioned the imputation within the KEI project is not an automated routine and consists of several steps. Each of them has been carried out with care. The steps for the actual imputation round are documented as follows:

1. **Transformation of the dataset**

   At first, the KEI dataset was subjected to an univariate power transformation as described in Chapter 4. The transformation parameters have been saved for a later reverse transformation. No multivariate power transformations were carried out for lack of time (these would actually go between steps 3 and 4).

2. **Graphical Analysis and Preliminary Grouping of the Data**

   Since the KEI dataset consists of 125 indicators, the exploratory data analysis began with a graphical overview of the correlations. The correlation map for all indicators

is given in Figure 6.1. Then those indicators that had too few observations to compute correlations with other indicators were excluded from the examination. These indicators are distinguishable by the white (i.e. low correlation) horizontals, verticals and rectangles on the correlation matrix map. The remaining 94 indicators were hierarchically clustered to obtain groups with high intra-correlations. For that purpose, the correlation matrix of the data (with pairwise complete observations) was used. The dendrogram in Figure A.2 (p. 74) presents the proposed clusters with high intra-correlations. However, it turned out that the models couldn't been created that simply. The example in Appendix A.4 uses an earlier dataset (which was updated for the actual imputation round) in order to clearly exhibit several problems.

Hence, the transformed dataset was analyzed and classified into preliminary groups. However, these shortcomings show the need for a thorough control of the imputation process and the time consuming handwork involved in it.

3. **Building of the Models**

   When the preliminary grouping was done, the models were finally created by adding dummy variables. Altogether, 46 models v were built, which can be assigned to one of three types:

   (a) **Contaminated Dummy Model:** Only when the data situation guaranteed enough observed values for all four years, was it able to build contaminated normal models which account for dummy variables for all years (i.e. 3 dummies). Examples are models `m01` to `m20`.

   (b) **Contaminated Dummy Model + LVCF:** Since the countries had collected indicators for only two or three years, these indicators were used to construct a contaminated dummy model with dummies corresponding to these years (i.e. 1 or 2 dummies). Afterwards, the missing values for the remaining year(s) were replaced by the LVCF routine. An example is model `m21`, where the years 2003 and 2004 were imputed using the contaminated dummy model and where 2001 and 2002 were augmented by 2003.

   (c) **Contaminated Model + LVCF:** The contaminated normal model without dummy variables was built when indicators were only available for one year. Afterwards the missing values for the remaining three years were replaced using the LVCF routine. So for example in model `m31` the years 2001 to 2003 were augmented by the values of 2004 which were imputed using the contaminated model.

4. **Documentation of the models**

   To allow conclusions to be drawn in subsequent processing with the imputed KEI dataset, the quality of the built models has been recorded and transferred to Tables A.1 and A.2 (pp. 70 and 71):

   - **indicators:** These two columns identify which indicators were finally used to build the models. Whereas *direct* marks the indicators that actually were imputed, *auxiliary* indicators only permitted imputations (because of high correlations to the direct indicators or to improve the multivariate setting).

- **how the imputation was done:** These columns highlight according to the chosen type (from step 3) how the imputation was done with respect to the years concerned. It can be noted that in 18 models all four years, in 5 models three years and in 7 models two years were involved in the imputation. The remaining years have been augmented by LVCF (the year given in parentheses). Then, 16 models without a dummy structure were built. Out of these 7 occurred in 2004 .
- **observed correlations:** The pairwise computed correlations between the indicators within the v-th model have been averaged, $\overline{r}_v$ and evaluated using the following criterion:

$$\text{(correlation quality)}_v \quad = \quad \begin{cases} \textit{very bad} & \text{if } \overline{r}_v < 0.1 \\ \textit{bad} & \text{if } 0.1 \leqslant \overline{r}_v < 0.3 \\ \textit{middle} & \text{if } 0.3 \leqslant \overline{r}_v < 0.5 \\ \textit{good} & \text{if } 0.5 \leqslant \overline{r}_v < 0.7 \\ \textit{very good} & \text{if } 0.7 \leqslant \overline{r}_v \ . \end{cases}$$

- **quality:** A model score has been computed for the potential model quality with respect to the proportion of missing values and the observed correlations:

$$\kappa_v \ \widehat{=} \ \overline{r}_v \ \cdot \ \frac{\text{\# of observed values in model } v}{\text{\# of indicators in model } v \cdot \text{\# of years} \cdot \text{\# of countries}} \ .$$

Then the v-th model was evaluated using the following criterion:

$$\text{(model quality)}_v \quad = \quad \begin{cases} \textit{very bad} & \text{if } \frac{\kappa}{100} < 0.1 \\ \textit{bad} & \text{if } 0.1 \leqslant \frac{\kappa}{100} < 0.3 \\ \textit{middle} & \text{if } 0.3 \leqslant \frac{\kappa}{100} < 0.5 \\ \textit{good} & \text{if } 0.5 \leqslant \frac{\kappa}{100} < 0.7 \\ \textit{very good} & \text{if } 0.7 \leqslant \frac{\kappa}{100} \ . \end{cases}$$

The following table summarizes the evaluation results of the 46 models given in Table A.2 (p. 71):

Criterion	very bad	bad	middle	good	very good
Correlation quality	-	2	20	17	7
Potential model quality	-	5	22	19	-

Table 6.1: Summary of the models' quality.

Remarks:

- Several completely observed indicators were used as auxiliary covariates multiple times because of their completeness and (good) correlations to many other indicators (i.e. B1a2, B1c2, B2a1 and B2b1).
- A very high correlation (almost equal to 1.00) between certain indicators prevented the construction of models because of multicollinearity. Thus, a particular model has been built for each of the affected indicators .
- Some very adverse indicators were standardized in addition to the power transformation to prevent their distributions from collapsing (here as well these

transformations have been arranged such that the indicators could be transformed back after the imputation).

- Because of the bad data situation, in only one model (i.e. `m42`) were the indicators able to impute themselves without the help of auxiliary indicators.

5. **Calculation of the Starting Values**

   According to model chosen in step 3, a contaminated EM algorithm yielded the starting values for the DA algorithm. The obtained imputation was saved as well as the starting values.

6. **Generation of $k = 5$ Multiple Imputes**

   After a *burn-in* phase of 1000 iterations, a robust DA algorithm according to the chosen model yielded 5 multiple imputes for each missing value. Thus, 5 augmented datasets including observed and imputed values were stored in addition to those from the previous step have been stored.

7. **Reverse transformation and plausibility check**

   When the imputation was done, the data was transformed back with the saved parameters. Although the reversibility of the power transformation and the standardization is given, a subsequent plausibility check of the new datasets was carried out to guarantee a correct proceeding in the former steps and to exclude human failure. This was done by a straight routine comparison of the observed values of each new generated dataset with the original one.

8. **Implicit Method: Spline Imputation + LVCF**

   The implicit method, a combined approach of the described spline imputation and LVCF method (see Section 1.4 on page 6), was applied to the original KEI dataset. Then, the output was stored.

9. **Weighting function**

   The weighting function is a simple routine that asks, for a given indicator, how many years are available for a given country. Depending on the number of years, it yields

$$\alpha_j = \begin{cases} 1 & \text{if u=4} \\ 0.85 & \text{if u=3} \\ 0.7 & \text{if u=2} \\ 0.25 & \text{if u=1} \\ 0 & \text{if u=0} \end{cases}, \tag{6.1}$$

   where u is the number of years and α_j has the dimensions (29×1). The weights have been saved for the final convex combination.

10. **Sensitivity Analysis**

    As described in Section 6.2.4 both approaches the implicit and the imputation method were compared by a Multiway Dot Plot. In doing so, the quality of the methods and approaches used was recorded. Figure 6.4 gives an example.

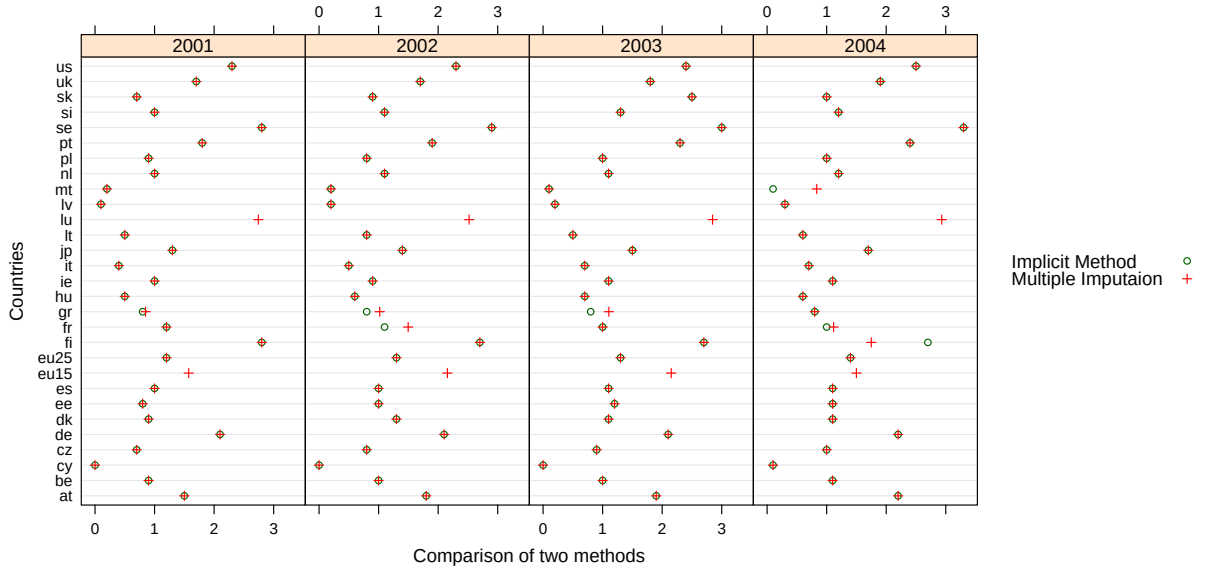

Figure 6.4: Sensitivity analysis of indicator A2a4.

The Multiway Dot Plot compares the average of the $k = 5$ imputes with the result from the implicit method. As can be seen for Malta, there are different imputation results in 2004. Therefore, it seems that one needs more robust results, which can be achieved by the convex combination as described in step 11. (Source: Enderle, 2008.)

11. **Convex Combination**

In the last step, final convex combinations of the augmented data were carried out using the implicit method and multiple imputation, weighted by the calculated weights from the penultimate step. The convex combination reads as follows

$$\mathrm{CoCo}_{j,k} = \alpha_j \, \mathrm{IM}_j + (1 - \alpha_j) \, \mathrm{MI}_{j,k} \ ,$$

where IM and MI stand respectively for the dataset generated by the implicit method and the k-th (single or multiple) imputation.

Chapter 7

Conclusion

Because of the high rate of missing information, the KEI dataset was completed by multiple imputations, which are

> *a device for representing missing-data uncertainty.* (Schafer, 1999, p. 8)

As already described, the KEI project had to make several investigations in order to cope with the adverse data situation. However, it can be assumed that the project will make a contribution to the improvement of the situation for the following reasons:

- The results of the project will cause a sensitization such that the statistical departments of the different countries see the need for well-timed surveys and supply timely estimates for the indicators.
- Because even the particular indicators are a result of aggregation processes, the sensitization can cause an improvement in the collection of data on the micro level. The imputation of microdata would produce better estimations especially because of the improvement of the quality regarding two problems listed in Section 1.3:
  1. The i.i.d. assumptions is more realistic on the micro level.
  2. The sample sizes are larger, which improves the whole imputation process: On the one hand, the proposed transformation is more suitable. On the other hand, larger models can be constructed and thus more attention can be paid to the correlations. In doing so, the quality of the imputation increases.
- Fewer outliers will occur because of the unification or at least the convergence of the indicators' construction methods.

Thus, it is expected that the results of the project can contribute to an increase in the quality of the indicators, a harmonization of their construction and measurement and a reduction of the number of missing values. When this aim is achieved, the whole process can be automated and higher developed imputation methods can be used.

Appendix A

The KEI Dataset

A.1 Description of the KEI Dataset

The treated KEI dataset was composed of 29 countries or country groups (i.e. 25 European countries, Japan, US, EU_{15}, EU_{25}) over a time period of 4 years (hence the panel structure). Originally, it was intended to collect 125 indicators but 9 indicators have not been collected so far. These have been excluded from the imputation. This results in an array with dimensions $29 \times 116 \times 4$ or 116×116 (when stringing the 4 years together vertically).

So each country can notice 4 observations per indicator at most, what will act as reference number.

A.1.1 Indicators

16 indicators (A1a1, A1a2, A3e1, A4b3, A4b5, B1a1, B1a2, B1b3, B1c2, B2a1, B2a2, B2b1, B2b2, B2b6, B2b7, C1a1) are observed completely.

67 indicators have an average number of observations, which is smaller than 3. This means that on average there are less than 3 observations per country.

The average number of observations per indicator is 2.35.

28 indicators (A1b5, A4b7, C1b7, A3c2, A3d3, A3d4, A3d5, A4b4, A4d1, A4d2, A4d4, A4d3, A4d5, C1d3, C1d4, A2a1, A2a2, A2a5, A2a6, A3d1, A4e1, C1c1, A3d2, B2c7, C1d5, A2a7, C1b3, B2c2) have on average less than one observed value per country.

These data suggest that there is a high heterogeneity between the indicators. The variation coefficient of the averaged values is 0.52.

A.1.2 Countries

Not even one country has collected all values completely.

The highest average number of observations per indicator is 2.89 (Finland). The US, Japan and Malta bring up the rear with 1.66, 1.6 and 1.5 values per indicator respectively. In total, there are 7 countries which have collected less than 2 observations per indicator on average.

The countries' dispersion about the averaged value of 2.35 observations per indicator is much smaller than the dispersion indikatorweise. The Variations coefficient for the countries is 0.17.

In total, there are 42 % missing values but with big differences between the indicators which exacerbates the imputation.

A.2 Summary

Model number	Indicators		How the imputation is done in ...			
	Direct	Auxiliary	2001	2002	2003	2004
m01	A2a4, A2b5, B1b2, C1c2	A4b5	cont. dummy model			
m02	A2c3, A2d2, A4b1, B1b1	B1c2	cont. dummy model			
m03	A2b2, A3a1, A3a6, A3a7	B2b1	cont. dummy model			
m04	A2b3, A2e1, A3a10	B2b1	cont. dummy model			
m05	A3a2, A3a9	A4b3	cont. dummy model			
m06	A3a3, A4c1	B1c2, B2b1	cont. dummy model			
m07	A3a5, A3a8	B1b3	cont. dummy model			
m08	A3a4, A3a12, A3b5	B1b3	cont. dummy model			
m09	A4b6, C1b6	B1a2	cont. dummy model			
m10	B2a5, C1a3	B2a2	cont. dummy model			
m12	B2c4	B1c1, B2b2	cont. dummy model			
m13	C1b1, C1b4	A1a2, A4b5	cont. dummy model			
m14	B2b4	B1c2, B2b1	cont. dummy model			
m15	A4a1, A4a2, A4a3	B2a1	cont. dummy model			
m16	A3a11, B2b3 C1b5	A4b5, C1a1	cont. dummy model			
m17	A2a3, B2b5	B2a2	cont. dummy model			
m18	B2a4	A2a4, C1b2	cont. dummy model			
m19	C1a2	B1a1	cont. dummy model			
m20	A2b4, B1c1 C1d2	B2a1	cont. dummy model			
m30	A1b1, A1b3	A3e1, B1c2	LVCF (2003)	LVCF (2003)	cont. dummy model	
m31	A1b2, A1b4, A1b5	A3e1, B1c2	LVCF (2003)	LVCF (2003)	cont. dummy model	
m32	A1b6, A1b7, A4b2	B1c2	LVCF (2003)	LVCF (2003)	cont. dummy model	
m33	A1c1, A1c3, A1d3	B1c2, B2b1	LVCF (2003)	LVCF (2003)	cont. dummy model	
m40	A1a4, A3b1	A5b5	cont. dummy model			LVCF (2003)
m41	B2c1	A4b5	cont. dummy model			LVCF (2003)
m42	A3b2, A3c1, C1b2		cont. dummy model			LVCF (2003)

Continued on next page ...

Model number	Indicators		How the imputation is done in ...			
	Direct	Auxiliary	2001	2002	2003	2004
m45	A1c2, A1c4, A1d1, A1d2	A4b3, B1c2	LVCF (2002)	cont. dummy model		
m50	A2a1	A2a2, A3c2, B1c2	LVCF (2003)	LVCF (2003)	cont. model	LVCF (2003)
m51	A3c2, C1b3, C1b7	A4b5	LVCF (2003)	LVCF (2003)	cont. model	LVCF (2003)
m52	A3d3, A3d5	A3d4, B2a1	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m53	A4b7, A4d1, A4d3	A1a2	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m54	A4e1, C1d4, C1d5	A4b5	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m70	B2c5, B2c6	A3e1, B2b2	cont. dummy model			
m71	B2a3	A1a1, C1a1	cont. dummy model			LVCF (2003)
m72	A1a3	B1c2, B2b1	LVCF (2003)	LVCF (2003)	cont. dummy model	
m73	B2c3	B1a1, B2b2	LVCF (2002)	cont. dummy model		LVCF (2002)
m74	B2b2, B2c7	A3e1, B2a1	cont. model	LVCF (2001)	LVCF (2001)	LVCF (2001)
m75	A2a7	B2a1, C1b2	cont. model	LVCF (2001)	LVCF (2001)	LVCF (2001)
m77	A3d4, A4d2, C1d3	B2a1, B2b6	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m78	A4b4, A4d5	B1a2, C1a1	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m79	A4d4	A3a4, B2b4	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m81	A2a5, A2a6	A4b3, A4c1	cont. model	LVCF (2001)	LVCF (2001)	LVCF (2001)
m82	A3d1, A3d2	A3a12, B2c6, C1b4	LVCF (2004)	LVCF (2004)	LVCF (2004)	cont. model
m83	A2a2	A1a1, A4b3	LVCF (2003)	LVCF (2003)	cont. model	LVCF (2003)
m84	C1c1	A1a1, A1a2	cont. dummy model			LVCF (2003)
m86	A1c5	B1c2, A1c4, A3b5	LVCF (2003)	LVCF (2003)	cont. dummy model	

Table A.1: How the imputation is done. (Source: ENDERLE, 2008.)

Model	Correlation quality	Potential model quality
m01	middle	middle
m02	good	middle
m03	good	good
m04	good	middle
m05	middle	bad
m06	good	good
m07	middle	middle
m08	good	middle
m09	middle	middle
m10	middle	bad
m12	bad	bad
m13	middle	bad
m14	good	middle
m15	very good	middle
m16	good	middle
m17	middle	bad
m18	middle	middle
m19	middle	middle
m20	middle	middle
m30	good	middle
m31	good	middle
m32	good	middle
m33	good	middle
m40	very good	good
m41	very good	good
m42	very good	middle
m45	middle	middle
m50	good	bad
m51	very good	bad
m52	good	bad
m53	middle	bad
m54	good	bad
m70	middle	bad
m71	bad	bad
m72	good	middle
m73	very good	good
m74	middle	bad
m75	middle	bad
m77	middle	bad
m78	middle	bad
m79	middle	bad
m81	very good	middle
m82	middle	bad
m83	middle	bad
m84	good	middle
m86	good	middle

Table A.2: Evaluation of the models. (Source: Enderle, 2008.)

Indicator	Country (plus year)
A1b3	several
A1b4	several
A1b4	ee
A1b5	se in 2004
A1b6	hu, lv and sk
A1b7	hu, lv and sk
A1c1	lv
A1c2	several
A1c3	lv
A1c4	several
A1c5	several
A1d1	several
A1d2	several
A1d3	several
A2a4	mt and fi in 2004
A2b2	lu in 2001 and 2002
A2b3	lu in 2001 and 2002
A2b4	dk and nl
A2b5	gr, fr, mt and fi
A2c3	ie and cz
A2e1	se
A3a10	se and lu
A3a12	pt and lu
A3a3	lu
A3a4	se in 2001 and 2004
A3a5	at
A3a8	se in 2002
A4d1	jp
B2a3	se, it and be
B2a5	eu15
B2b4	several
B2c1	several
B2c3	several in 2003
B2c6	ie, es and cz
C1a2	lu
C1a3	lu
C1b1	ie and fi in 2004
C1b4	pl and lu
C1b6	mt and gr
C1b7	lu
C1c1	too few data
C1c2	gr, dk, lu and be
C1d4	us and lu
C1d5	lu

Table A.3: Sensitivity analysis. (Source: ENDERLE, 2008.)

A.3 Dendrograms

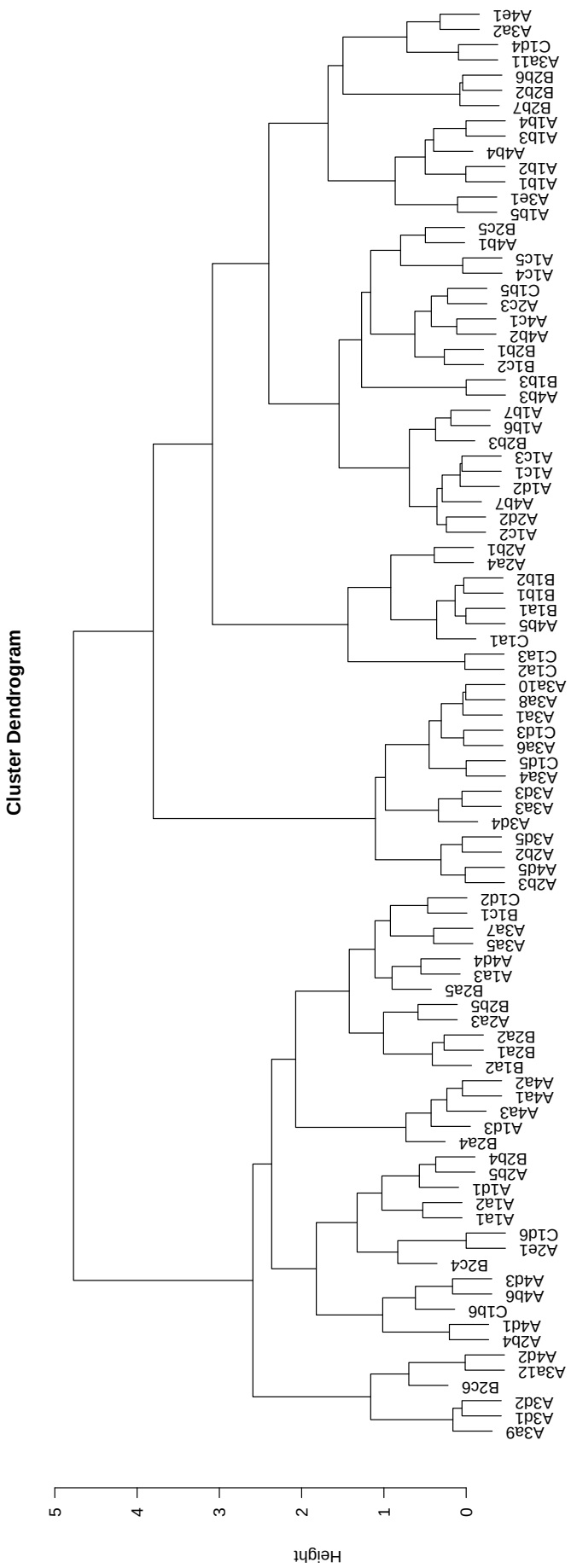

Figure A.1: Dendrogram of an earlier dataset. (Source: Enderle, 2008.)

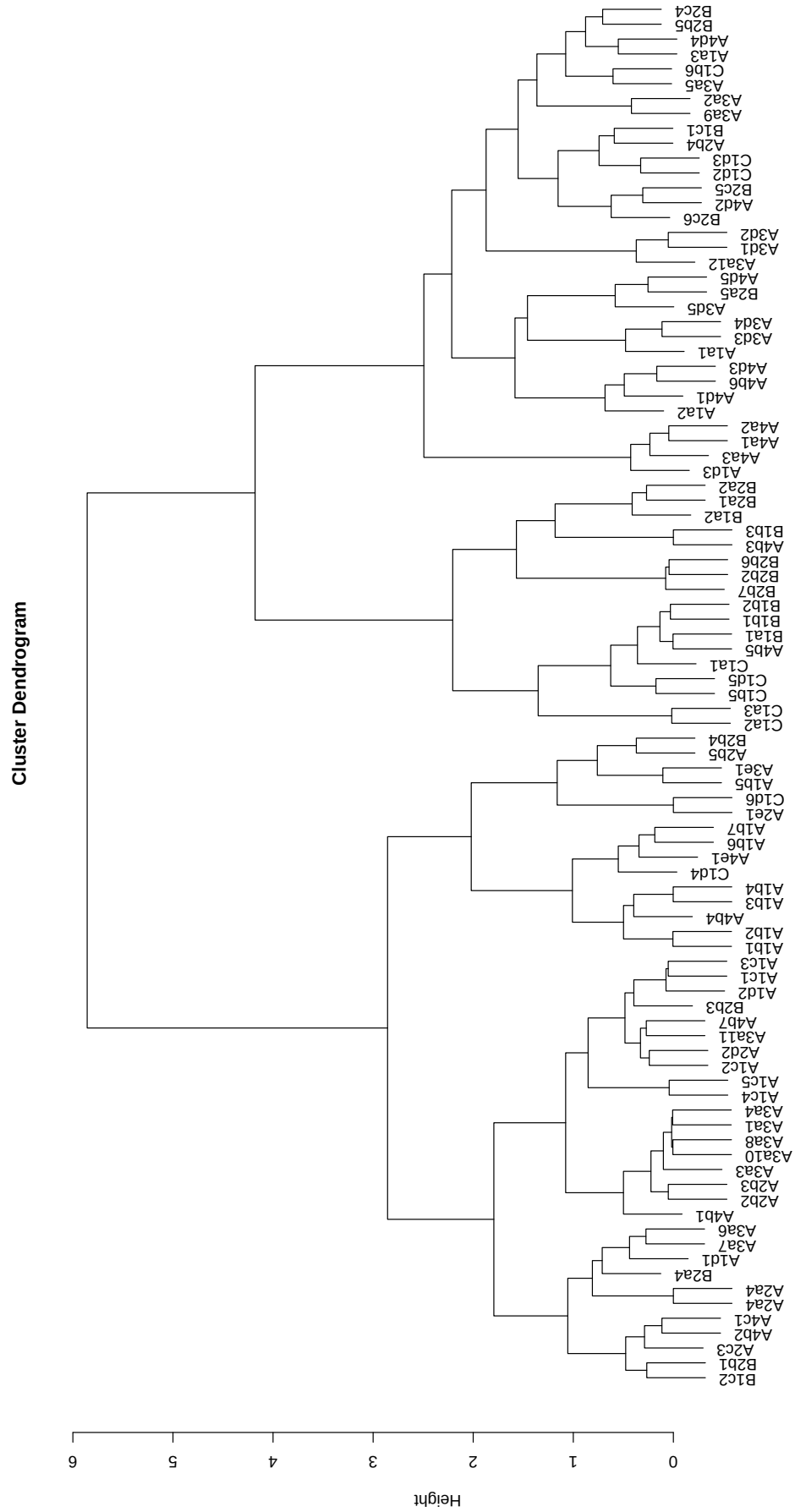

Figure A.2: Dendrogram of the actual dataset. (Source: ENDERLE, 2008.)

A.4 An Exemplification

An example is given by the first cluster (from the left hand side) in Figure A.1, which combines the following indicators into a group: A3a9, A3d1, A3d2, B2c6, A3a12 and A4d2. Then the group was analyzed by its correlation map, given in Figure A.3. The graph shows that two inner-groups (A3a9, A3d1, A3d2) and (A3a12, A4d2) with very high intra-correlations result. Going back to the dendrogram, one can see that the result goes hand in hand with the graph: (1) first these two inner-groups were clustered and then (2) indicators A3a12 and (A3d1, A3d2) ensured that the two inner-groups were clustered together.

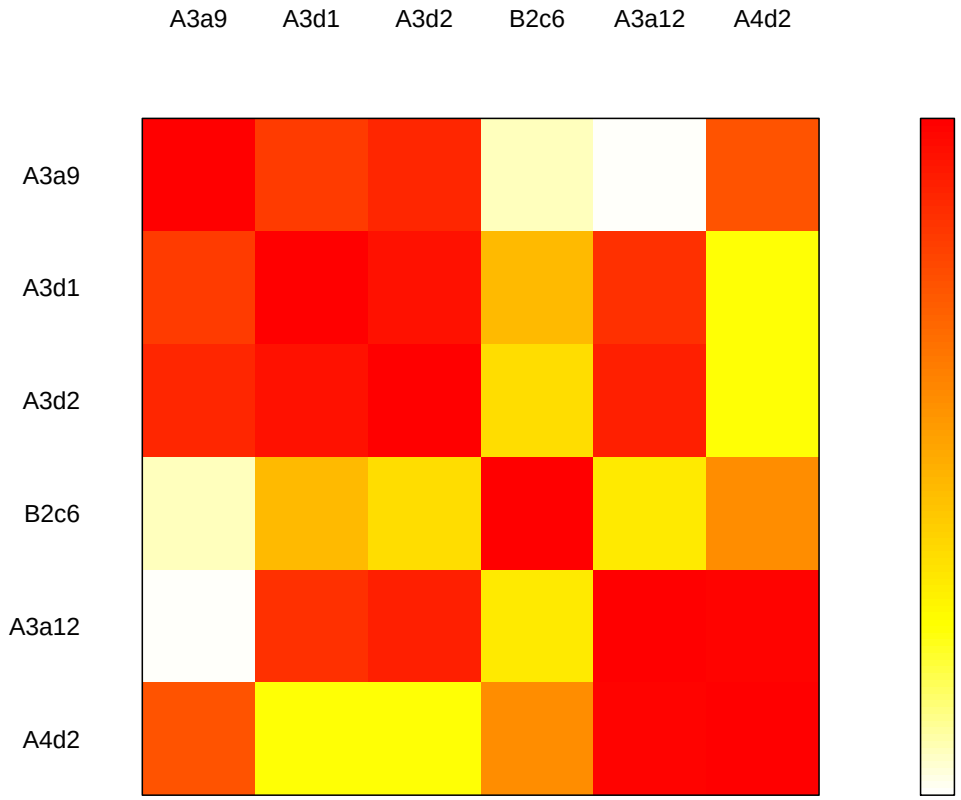

Figure A.3: Correlation map of the 1st cluster. (Source: Enderle, 2008.)

But due to the bad data situation and different patterns of missingness between indicators, the models often couldn't be created as suggested by these approaches. The cluster must be explored more precisely to reveal some of its shortcomings:

- The exploration graphic for the cluster is given in Figure A.4. Since the missing values have been deleted pairwise for the computed correlations, some correlations are based upon very few observations. So, for example, the correlation between A3a12 and A4d2 is based upon just 4 observations (i.e. 96.6 % missing values). Thus, the high correlation of 0.99 must be handled with care.
- A further problem comes up when the correlation of 0.95 between A3d1 and A3d2 is analyzed more precisely in Table A.4 (a). The value is based upon 19 pairs, all in the year 2004. Indicator A3d1 has a total of 9 missing values. All of them are present in the indicator A3d2 as well. As a consequence of that, these indicators cannot impute themselves.

- In Table A.4 (b) indicator B2c6 shows the already mentioned problem for 4 countries, that there are no values at all. This is an undesirable situation as well and the panel structure has a reduced impact.

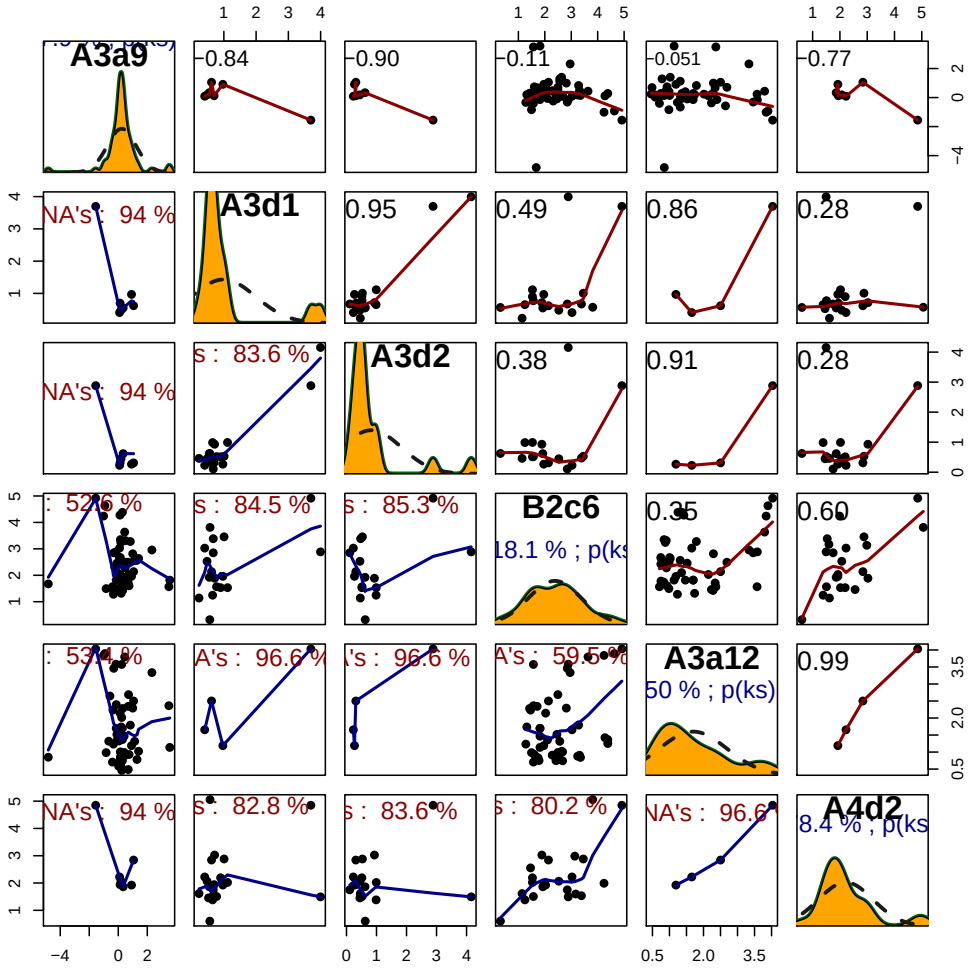

Figure A.4: Exploration graphic of the 1st cluster. (Source: ENDERLE, 2008.)

Country	A3d1	A3d2
at	NA	NA
be	0.0377	0.0232
cy	0.0275	0.0279
cz	0.0301	0.0139
de	0.0340	0.0141
dk	0.0327	0.0048
ee	0.0327	0.0211
es	0.0317	0.0441
eu15	NA	NA
eu25	NA	NA
fi	0.0490	0.0237
fr	0.0233	0.0200
gr	0.0438	0.0229
hu	0.0472	0.0119
ie	0.0267	0.0276
it	0.0198	0.0102
jp	NA	NA
lt	0.0112	0.0206
lu	0.0542	0.0443
lv	NA	NA
mt	0.0278	NA
nl	0.0256	0.0200
pl	0.0347	0.0416
pt	0.1945	0.1857
se	NA	NA
si	NA	NA
sk	0.1800	0.1289
uk	NA	NA
us	NA	NA

(a)

Country	2001	2002	2003	2004
at	0.099	0.092	0.094	0.089
be	0.045	0.042	0.042	0.044
cy	0.015	0.012	0.011	0.009
cz	NA	0.040	0.048	0.060
de	NA	NA	NA	NA
dk	0.079	0.070	0.070	0.080
ee	0.073	0.079	0.088	0.095
es	0.037	NA	NA	0.035
eu15	0.070	0.075	0.075	0.077
eu25	0.066	0.070	0.072	0.074
fi	0.094	0.091	0.093	0.097
fr	0.049	0.050	0.065	0.071
gr	0.044	0.042	0.041	0.043
hu	0.072	0.069	0.059	0.055
ie	0.057	NA	0.057	0.054
it	0.051	0.051	0.051	0.085
jp	NA	NA	NA	NA
lt	0.021	0.025	0.027	0.032
lu	0.030	0.034	0.044	0.043
lv	NA	0.081	0.079	0.057
mt	0.097	0.090	0.100	0.107
nl	NA	NA	NA	NA
pl	0.036	0.047	0.051	0.053
pt	0.083	0.080	0.081	0.081
se	0.072	0.075	0.044	0.041
si	0.074	0.064	0.068	0.088
sk	0.102	0.119	0.130	0.138
uk	0.123	0.123	0.120	0.119
us	NA	NA	NA	NA

(b)

Table A.4: An exemplification. (Source: Enderle, 2008.)

Appendix B

Some Mathematics

B.1 The Sweep Operator

The sweep operator is defined for symmetric matrices: Assume a symmetric $p \times p$ matrix G, whose (i, j)th element is given by g_{ij}. Sweeping this matrix G on position k (for any $k \in 1, \ldots, p$), $SWP[k]$, the sweep operator creates another $p \times p$ matrix H

$$SWP[k]\, G = H \ ,$$

whose elements are given by:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} = h_{kj} &= g_{jk}/g_{kk} \qquad \text{for} \quad j \neq k \\ h_{jl} = h_{lj} &= g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad \text{for} \quad j \neq k \quad \text{and} \quad l \neq k \ . \end{aligned}$$

As long as a sweep operation of matrix G in all positions contains no division by 0, we obtain the negative inverse of matrix G

$$SWP[1, \ldots, p]\, G = SWP[1] \ldots SWP[p]\, G = -G^{-1} \ . \tag{B.1}$$

A mentionable property of the Sweep Operator is its commutativity

$$SWP[k_1]\, SWP[k_2]\, G = SWP[k_2]\, SWP[k_1]\, G$$

for any $k_1 \neq k_2$ with $k_1, k_2 \in 1, \ldots, p$. Thus, the order of the sweeps in Equation (B.1) is not of interest.

Furthermore, we can define the *reverse sweep operator*. In sweeping matrix G on position k we receive a new matrix H

$$RSWP[k]\, G = H \ ,$$

with its elements

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} = h_{kj} &= -g_{jk}/g_{kk} \qquad \text{for} \quad j \neq k \\ h_{jl} = h_{lj} &= g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad \text{for} \quad j \neq k \quad \text{and} \quad l \neq k \, . \end{aligned}$$

Also the reverse sweep operator is commutative, wherewith it can be shown that it returns a swept matrix to its original form

$$RSWP[k]\, SWP[k]\, G = G \, ,$$

for any $k \in 1, \ldots, p$.

B.2 The Power Transformation

B.2.1 Reversibility of the Power Transformation

Real-valued version (Königsberger, 2000, p. 31):

Let $f : X \to \mathbb{R}$ be injective, where $X \subset \mathbb{R}$. *Injective* signifies that there exists for every function value $y \in f(X)$ exactly one $x \in X$ where $y = f(x)$. The function g, which specifies a so-called *preimage* to every $y \in f(x)$, is called the *inverse function* of f:

$$g : f(X) \to \mathbb{R}, \quad g(f(x)) = x \, .$$

Injective are for example all strictly monotonic functions. *Consequently any strictly monotonic function $f : X \to \mathbb{R}$ has an inverse function : $f(X) \to \mathbb{R}$, which is monotonic in the same sense.*

the power function $x^r, x > 0, r \in \mathbb{R}$ is strictly monotonic increasing for $r > 0$ and strictly monotonic decreasing for $r < 0$. Moreover, $g(x) = x^{\frac{1}{r}}$ is the inverse funtion of $f(x) = x^r$.

B.2.2 Asymptotical Properties of Sequences and Functions

Deterministic version (Mittelhammer, 1996, p. 231)

Let $\{x_n\}, n \in \mathbb{N}$ be a real number sequence, which is said to be *at most of order n^k*,denoted by $\mathcal{O}(n^k)$, if there exists a finite real number c such that $|n^{-k}x_n| \leq c \, \forall \, n$.

Stochastic version (Mittelhammer, 1996, p. 248)

Let $\{x_n\}$ be a sequence of random scalars, $X \in \mathbb{R}\, and\, n \in \mathbb{N}$. This sequence is said to be *at most of order n^k in probability*, denoted by $\mathcal{O}_p(n^k)$, if for every $\epsilon > 0$ there exists a positive constant $c(\epsilon) < \infty$ with the property that $\mathbb{P}(n^{-k}|X_n| \leq c(\epsilon)) \geq 1 - \epsilon, \forall \, n$.

B.2.3 Justification of the Plug-in

In a first step it will be shown that both methods ML (GREENE, 2003, p. 472, Ex. 17.2) and MM (GREENE, 2003, pp. 527f., Ex. 18.2) yield the same estimates for a normal distribution with mean μ and variance σ^2. Furthermore it will be shown that the GMM method (with two more moment conditions) achieves the Cramér-Rao bound, which is a small sample property of an estimator, as well as the ML method. The Crámer-Rao bound provides a lower bound for the variance of an estimator obtained by any estimation method. An unbiased estimator which achieves this bound is said to be efficient. Thus, additional moment conditions have no further effect on the efficiency of the estimator. Therefore, instead of stating the conditions for the first two moments, the MLEs $\widehat{\mu}$ and $\widehat{\sigma}^2$ can serve as plug-in, for the third and fourth central moment conditions as proposed in Chapter 4.

ML and MM estimators

(i) Maximum Likelihood estimator

The loglikelihood for a normal distribution is

$$l(\mu, \sigma^2) = -\tfrac{n}{2}\log(2\pi) - \tfrac{n}{2}\log(\sigma^2) - \tfrac{1}{2}\sum_{i=1}^{n}\left[\tfrac{(y_i-\mu)^2}{\sigma^2}\right] \ .$$

To compute the MLEs are then computed from the equation system based on the first derivatives of the loglikelihood with respect to the parameters respectively:

$$\tfrac{\partial l}{\partial \mu} = \tfrac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0$$

and

$$\tfrac{\partial l}{\partial \sigma^2} = -\tfrac{n}{2\sigma^2} + \tfrac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0 \ .$$

Then, to obtain the MLEs for a normal distribution both likelihood equations must be solved for the parameters

$$\widehat{\mu}_{ML} = \overline{y}$$

and

$$\widehat{\sigma}^2_{ML} = \sum_{i=1}^{n}(y_i - \overline{y})^2/n \ .$$

These empirical estimators are efficient and can be used to state the third and fourth central moment conditions without further moment conditions.

(ii) Method of Moments Estimator

In random sampling from $\mathrm{N}(\mu, \sigma^2)$,

$$\operatorname{plim} \tfrac{1}{n} \sum_{i=1}^{n} y_i = \operatorname{plim} \overline{m}_1' = E(y_i) = \mu$$

and

$$\operatorname{plim} \tfrac{1}{n} \sum_{i=1}^{n} y_i^2 = \operatorname{plim} \overline{m}_2' = \operatorname{var}(y_i) + \mu^2 = \sigma^2 + \mu^2 \; .$$

Equating both sides of the probability limits gives the two moment estimators

$$\widehat{\mu}_{MM} = \overline{m}_1' = \tfrac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$$

and

$$\widehat{\sigma}^2_{MM} = \overline{m}_2' - \overline{m}_1'^2 = \left(\tfrac{1}{n} \sum_{i=1}^{n} y_i^2 \right) - \left(\tfrac{1}{n} \sum_{i=1}^{n} y_i \right)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 / n \; .$$

Hence, it has been shown, that both methods ML and MM yield the identical estimators such that holds

$$\widehat{\mu}_{ML} = \widehat{\mu}_{MM} = \overline{y}$$

and

$$\widehat{\sigma}^2_{ML} = \widehat{\sigma}^2_{MM} = \sum (y_i - \overline{y})^2 / n \; .$$

Cramér-Rao bounds of ML and MM

(i) Maximum Likelihood

At first the second derivatives $\frac{\partial^2 l}{\partial \mu \partial \mu}$ and $\frac{\partial^2 l}{\partial \sigma^2 \partial \sigma^2}$ that are collected in the Hessian matrix must be computed

$$H(\mu, \sigma^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum (y_i - \mu) \\ -\frac{1}{\sigma^4} \sum (y_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum (y_i - \mu)^2 \end{bmatrix} .$$

The negative value of the expected Hessian matrix is called the information matrix

$$I \equiv -E[H(\mu, \sigma^2)] = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} .$$

Under valid regularity conditions the estimator of a parameter vector will always be as large as the inverse of the information matrix, i.e. the Cramér-Rao bound

$$I[(\mu, \sigma^2)] = \tfrac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} .$$

(ii) (Generalized) Methods of Moments

HANSEN (1982) showed that the inverse of the variance covariance matrix of the moment conditions $\mathbf{S}^{-1}$ is a optimal weighting matrix to achieve a minimal variance covariance matrix $\mathbf{V}$.

So first, the variance covariance matrix of the moment conditions must be derived:

$$
\begin{aligned}
\mathbf{S} &= \mathrm{E}[m_t(\widehat{\Theta})\, m_t(\widehat{\Theta})'] \\
&= \mathrm{E}\left(\begin{array}{c} (x_t-\mu) \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{array} \right) \left(\begin{array}{c} (x_t-\mu) \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{array} \right)' \\
&= \mathrm{E}\left(\begin{array}{cccc} (x_t-\mu)^2 & \underbrace{(x_t-\mu)(x_t-\mu)^2-(x_t-\mu)\sigma^2}_{=0} & \cdots & \\ & \cdots & (x_t-\mu)^4 & \underbrace{(x_t-\mu)^5-(x_t-\mu)3\sigma^4}_{=0} \\ \vdots & & \vdots & \end{array} \right) \\
&= \left[\begin{array}{cccc} \sigma^2 & 0 & 3\,\sigma^4 & 0 \\ 0 & 2\,\sigma^4 & 0 & 12\,\sigma^6 \\ 3\,\sigma^4 & 0 & 15\,\sigma^6 & 0 \\ 0 & 12\,\sigma^6 & 0 & 96\,\sigma^8 \end{array} \right],
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathrm{E}(x_t-\mu) = 0 \\
&\mathrm{E}(x_t-\mu)^2 = \sigma^2 \\
&\mathrm{E}(x_t-\mu)^{3+2k} = 0 \qquad \text{for } k = 0, 1 \ldots \\
&\mathrm{E}(x_t-\mu)^4 = 3\sigma^4\,.
\end{aligned}
$$

The partial first derivatives of the moment conditions are

$$
\mathfrak{D} = \mathrm{E}\left[\begin{array}{cc} \frac{\partial m_1}{\partial \mu} & \frac{\partial m_1}{\partial \sigma^2} \\ \frac{\partial m_2}{\partial \mu} & \frac{\partial m_2}{\partial \sigma^2} \\ \frac{\partial m_3}{\partial \mu} & \frac{\partial m_3}{\partial \sigma^2} \\ \frac{\partial m_4}{\partial \mu} & \frac{\partial m_4}{\partial \sigma^2} \end{array} \right] = \mathrm{E}\left[\begin{array}{cc} -1 & 0 \\ -2(x_t-\mu) & -1 \\ -3(x_t-\mu)^2 & 0 \\ -4(x_t-\mu)^3 & -6\sigma^2 \end{array} \right] = \left[\begin{array}{cc} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{array} \right].
$$

With $\mathbf{S}^{-1}$ as weighting matrix, the optimal variance covariance matrix for the GMM estimator is then found to be

$$
\begin{aligned}
&\text{Asy.Cov}(\Theta) = \mathbf{V} = \tfrac{1}{n}(\mathfrak{D}'\,\mathbf{S}^{-1}\,\mathfrak{D})^{-1} \\
&= \tfrac{1}{n}\left(\left[\begin{array}{cccc} -1 & 0 & -3\sigma^2 & 0 \\ 0 & -1 & 0 & -6\sigma^2 \end{array} \right] \left[\begin{array}{cccc} \frac{5}{2\sigma^2} & 0 & \frac{-1}{2\sigma^4} & 0 \\ 0 & \frac{2}{\sigma^4} & 0 & \frac{-1}{4\sigma^6} \\ \frac{-1}{2\sigma^4} & 0 & \frac{-1}{6\sigma^6} & 0 \\ 0 & \frac{-1}{4\sigma^6} & 0 & \frac{1}{24\sigma^8} \end{array} \right] \left[\begin{array}{cc} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{array} \right] \right)^{-1}
\end{aligned}
$$

$$= \frac{1}{n\frac{1}{\sigma^2}\frac{1}{2\sigma^4}} \begin{bmatrix} \frac{1}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix}$$

$$= \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \quad \text{Cramer-Rao bound for } \widehat{\mu} \text{ and } \widehat{\sigma}^2.$$

B.2.4 Transformed Normal Distribution

Let X be a normally distributed random variable with parameters over a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$

$$\mathrm{X} \sim \mathrm{N}(\mu, \sigma^2) \ .$$

Furthermore, define the following bijective image:

$$\psi : \mathbb{R} \to \mathbb{R};\ X \mapsto \psi(X) = X^3 \ .$$

The density function of the transformed random variable can be obtained via the density transformation theorem.

For that purpose, the following terms are necessary:

$$\begin{aligned} \psi(x) &= x^3 =: y \\ \psi^{-1}(y) &= y^{\frac{1}{3}} \\ (\psi^{-1}(y))' &= \tfrac{1}{3} y^{\frac{-2}{3}} \end{aligned}$$

then it holds

$$\begin{aligned} g_Y(y) &= f_X(y^{\frac{1}{3}}) \left| \tfrac{1}{3} y^{\frac{-2}{3}} \right| && \text{for } y \in \mathbb{R} \\ &= \tfrac{1}{\sqrt{2\pi\sigma^2}} \left| \tfrac{1}{3} y^{\frac{-2}{3}} \right| \mathrm{e}^{\frac{-1}{2\sigma^2}(y^{1/3}-\mu)^2} \ . \end{aligned}$$

One can draw from this distribution via a Metropolis-Hasting algorithm (e.g. TANNER and WONG, 1987).

Bibliography

Albert, J. (**2007**): Bayesian Computation with R. Springer.

Baddeley, A. (**2008**): *Analysing spatial point patterns in R*. Technical report, CSIRO Mathematical and Information Sciences.

Bayes, T. (**1958**): *Studies in the History of Probability and Statistics: IX. Thomas Bayes' Essay Towards Solving a Problem in the Doctrine of Chances*. Biometrika, 45, pp. 296–315, (Bayes' essay in modernized notation).

Bollen, K. A. (**1989**): Structural Equations with Latent Variables. Wiley-Interscience.

Buck, S. F. (**1960**): *A method of estimation of missing values in multivariate data suitable for use with an electronic computer*. J. Roy. Statist. Soc., B 22, pp. 302–306.

Cleveland, W. S. (**1979**): *Robust Locally Weighted Regression and Smoothing Scatterplots*. Journal of the American Statistical Association, 74, pp. 829–836.

Cleveland, W. S. (**1981**): *LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression*. The American Statistician, 35 (1), p. 54.

Cleveland, W. S. (**1994**): The elements of graphing data. AT&T Bell Laboratories.

David, M., **Little, R. J. A.**, **Samuhel, M. E.** and **Triest, R. K.** (**1986**): *Alternative methods for CPS income imputation*. Journal of the American Statistical Association, 81, pp. 29–41.

Dempster, A. P., **Laird, N. M.** and **Rubin, D. B.** (**1977**): *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 39 (1), pp. 1–38.

Deuflhard, P. (**2004**): Newton Methods for Nonlinear Problems. Springer.

Enderle, T. M. (**2008**): Multiple Imputation of Macro Data for Knowledge Economy Indicators. Diploma thesis, University of Tübingen.

Fishman, G. S. (**2006**): A first Course in Monte Carlo. Thomson.

Gelfand, A. E. and **Smith, A. F. M.** (**1990**): *Sampling Based Approaches to Calculating Marginal Densities*. Journal of the American Statistical Association, 85 (410), pp. 398–409.

Gelman, A., **Carlin, J. B.**, **Stern, H. S.** and **Rubin, D. B.** (**2004**): Bayesian Data Analysis. Chapman & Hall, second ed.

Geman, S. and **Geman, D.** (**1984**): *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, pp. 721–741.

Goodnight, J. H. (**1979**): *A Tutorial on the SWEEP Operator*. The American Statistician, 33 (3), pp. 149–158.

Greene, W. H. (**2003**): Econometric Analysis. Prince & Hall.

Hansen, L. P. (**1982**): *Large Sample Properties of Generalized Method of Moments*. Econometrica, 50 (4), pp. 1029–1054.

Hayashi, F. (**2000**): Econometrics. Princeton University Press.

Huergo, L. (**2008**): Markov Chain Monte Carlo Methoden zur multiplen Imputation für die Knowledge Economy Indicators: Anwendung und Verbesserungsvorschläge. Ph.D. thesis, University of Tübingen.

Kent, J. T., **Tyler, D. E.** and **Vardi, Y.** (**1994**): *A curious likelihood identity for the multivariate T-distribution*. Commun. Statist. B - Simulation and Computation, 23, pp. 441–453.

Kim, J.-O. and **Curry, J.** (**1977**): *The Treatment of Missing Data in Multivariate Analysis*. Sociological Methods & Research, 6 (2), pp. 215–240.

Königsberger, K. (**2000**): Analysis 1. Springer.

Kofman, P. and **Sharpe, I. G.** (**2003**): *Using Multiple Imputation in the Analysis of Incomplete Observations in Finance*. Journal of Financial Econometrics, 1, pp. 216–249.

Krosnick, J. A., **Weisberg, H. F.** and **Bowen, B. D.** (**2000**): An introduction to survey research, polling, and data analysis. SAGE Publications Inc.

Lange, K. L., **Little, R. J. A.** and **Taylor, J. M. G.** (**1989**): *Robust Statistical Modeling Using the* t *Distribution*. Journal of the American Statistical Association, 84 (408), pp. 881–896.

Ligges, U. (**2006**): Programmieren mit R (Statistik und ihre Anwendungen), vol. 2. Springer Berlin.

Little, R. J. A. (**1988**): *Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values*. Applied Statistics, 3 (1), pp. 23–38.

Little, R. J. A. and **Rubin, D. B.** (**1987**): Statistical Analysis with Missing Data. Wiley Interscience.

Little, R. J. A. and **Rubin, D. B.** (**2002**): Statistical Analysis with Missing Data. Wiley Interscience, 2 ed.

Liu, C. (**1995**): *Missing Data Imputation Using the Multivariate* t *Distribution*. Journal of Multivariate Analysis, 53, pp. 139–158.

Liu, C. and **Rubin, D. B.** (**1994**): *The ECME algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence*. Biometrika, 81 (4), pp. 633–648.

Liu, C. and **Rubin, D. B.** (**1995**): *ML Estimation of the* t *Distribution using EM and its Extensions, ECM and ECME*. Statistica Sinica, 5, pp. 19–39.

Liu, C., **Rubin, D. B.** and **Wu, Y. N.** (**1998**): *Parameter Expansion to Accelerate EM: The PX-EM Algorithm*. Biometrika, 85 (4), pp. 755–770.

Mahalanobis, P. C. (**1936**): *On the generalized distance in statistics*. Proceedings of the National Institute of Science of India, 2, pp. 49–55.

McLachlan, G. J. and **Krishnan, T.** (**1997**): The EM Algorithm and Extensions. Wiley Interscience.

Meng, X.-L. and **Rubin, D. B.** (**1993**): *Maximum likelihood estimation via the ECM algorithm: a general framework*. Biometrika, 80, pp. 267–278.

Mittelhammer, R. C. (**1996**): Mathematical Statistics for Economics and Business. Springer.

Murrell, P. (**2005**): R Graphics, vol. 2. Chapman & Hall.

Nijman, T. and **Verbeek, M.** (**1992**): *Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function*. Journal of Applied Econometrics, 7 (3), p. 243:257.

Rizzo, M. L. (**2008**): Statistical Computing with R. Chapman & Hall/ CRC.

Roth, P. L. (**1994**): *Missing data: A conceptual review for applied psychologists*. Personnel Psychology, 47, pp. 537–560.

Rubin, D. B. (**1976**): *Inference and Missing Data*. Biometrika, 63, pp. 581–592.

Rubin, D. B. (**1987**): Multiple Imputations for Nonresponse in Surveys. Wiley.

Saisana, M. and **Munda, G.** (**2008**): *Final report on simulation results for indicators*. KEI deliverable D5.4, 5.6, 5.8, 7.2, and 7.3, `http://kei.publicstatistics.net`.

Schafer, J. L. (**1997**): Analysis of Incomplete Multivariate Data. Chapman & Hall.

Schafer, J. L. (**1999**): *Multiple imputation: a primer*. Statistical Methods in Medical Research, 8, pp. 3–15.

Schafer, J. L. and **Yucel, R. M.** (**2002**): *Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values*. Journal of Computational and Graphical Statistics, 11 (2), pp. 437–457.

Schaich, E. and **Münnich, R.** (**2001**): Mathematische Statistik für Ökonomen. Vahlen.

Sydsaeter, K., **Hammond, P. J.**, **Seierstad, A.** and **Strom, A.** (**2005**): Further Mathematics for Economic Analysis. Prentice Hall.

Tanner, M. A. (**1991**): Tools for Statistical Inference. Observed Data and Data Augmentation Methods. Springer.

Tanner, M. A. and **Wong, W. H.** (**1987**): *The Calculation of Posterior Distributions by Data Augmentation*. Journal of the American Statistical Association, 82 (398), pp. 528–540.