A Multiple Imputation Approach to Indicators

Luis Huergo

luis.huergo@uni-tuebingen.de

University of Tuebingen, Germany

5^{th} September 2006

Introduction

KEI-Datasets have missing data ...

Eberhard Karls
Universität
Tübingen

Introduction

KEI-Datasets have missing data... lots of them.

Introduction

In order to build Composite Indicators, those missing values must be imputed.

Introduction

How should imputations be? (Little and Rubin, 2002) I

- Conditional on observed variables, to reduce bias due to nonresponse, improve precision and preserve association between missing and observed variables.
- Multivariate, to preserve associations between missing variables.

Eberhard Karls
Universität
Tübingen

Introduction

How should imputations be? (Little and Rubin, 2002) II

- Draws from the predictive distribution rather than means, to provide valid estimates of a wide range of estimands.
- Multiple, in order to account for imputation uncertainty.

Eberhard Karls
Universität
Tübingen

Introduction

However, datasets for Composite Indicators have special, stable features which must be taken into account

- The number of variables in the dataset is usually bigger than the sample size.
- Rows represent countries.
- Variables (columns) are subject to political decision.
- Panel structure: small number of time periods.
- Non normal data with nonlinear relationships between variables.
- As a rule, one should expect missing values to occur nonmonotonically, on all variables.

Eberhard Karls
Universität
Tübingen

Introduction

With following consequences

- Imputation models must restrict themselves to estimable ones (in spite of loss of generality)$\longrightarrow$ Sample sizes are almost invariably too small for parameter estimation of large multivariate models.
- Possibilities to enlarge the effective sample size must be considered $\longrightarrow$ time dimension.
- No row deletion allowed.
- No column deletion allowed $\longrightarrow$ In general, it is not possible to reduce the number of variables in the data set.

Eberhard Karls
Universität
Tübingen

Introduction

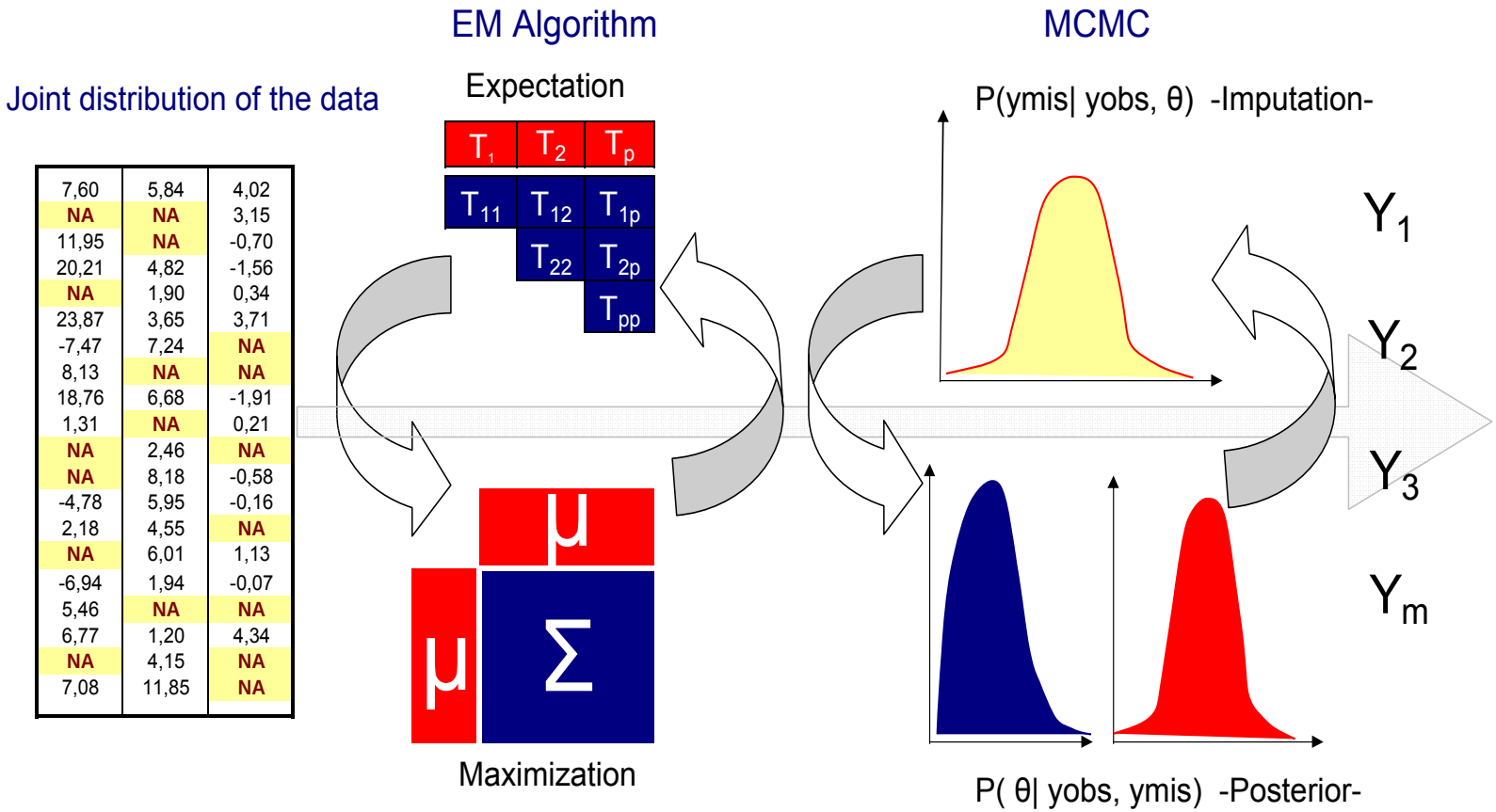
Which in turn determine the applicability of imputation methods

- Hot deck methods are not applicable.
- It is not possible (as a general rule) to estimate monotone missing patterns by factorizing the (log-)likelihood function.
- It is not possible to estimate autoregressive relationships.

Eberhard Karls
Universität
Tübingen

Methodological section

The general strategy: EM Algorithm + MCMC



The EM Algorithm

The *Expectation Maximization* Algorithm (Dempster, Laird and Rubin; 1977) is an iterative method for calculation of *Maximum Likelihood* estimates in *incomplete data* settings.

The EM Algorithm

Some definitions

In a slight abuse of notation, let $f_y(y|\theta)$ where $y = (y_{obs}, y_{mis})$ denote the joint distribution of the complete data, $f_{y_{obs}}(y_{obs}|\theta)$ the joint distribution of the observed data and $f_{y_{mis}}(y_{mis}|\theta)$ the joint distribution of the missing data.

Furthermore, let $L(\theta|y)$ denote the likelihood of the complete data and $L(\theta|y_{obs})$ the likelihood of the observed data.

Analogously, let $l(\theta|y)$ denote the log-likelihood of the complete data and $l(\theta|y_{obs})$ the log-likelihood of the observed data.

The EM Algorithm

The main problem

In many cases $l(\theta|y_{obs})$ has a complex structure which makes its maximization very difficult or even impossible. On the other hand, $l(\theta|y)$ has usually a much easier structure.

However, the information about $l(\theta|y)$ is incomplete, since a part of y is missing. The stochastics proposes as the best estimation of $l(\theta|y)$ its conditional expectation given the informaton contained in y_{obs} (and some parameter vector $\theta^{(i)}$).

How do $l(\theta|y_{obs})$ and $\mathrm{E}\left[l(\theta|y)|y_{obs}\right]$ relate to each other?

The EM Algorithm

From elementary probability theory

$$f_{y_{mis}}(y_{mis}|y_{obs},\theta) = \frac{f_y(y|\theta)}{f_{y_{obs}}(y_{obs}|\theta)} \quad \text{and hence}$$

$$f_{y_{obs}}(y_{obs}|\theta) = \frac{f_y(y|\theta)}{f_{y_{mis}}(y_{mis}|y_{obs},\theta)}$$

Taking logarithms

$$\log f_{y_{obs}}(y_{obs}|\theta) = \log f_y(y|\theta) - \log f_{y_{mis}}(y_{mis}|y_{obs},\theta)$$

$$l(\theta|y_{obs}) = l(\theta|y) - \log f_{y_{mis}}(y_{mis}|y_{obs},\theta)$$

The EM Algorithm

and building an expectation over the predictive distribution of y_{mis} given y_{obs} and $\theta^{(i)}$

$$\mathrm{E}[l(\theta|y_{obs})] = \mathrm{E}[l(\theta|y)] - \mathrm{E}[\log f_{y_{mis}}(y_{mis}|y_{obs},\theta)]$$
$$l(\theta|y_{obs}) = \underbrace{\mathrm{E}[l(\theta|y)]}_{\mathrm{Q}} - \underbrace{\mathrm{E}[\log f_{y_{mis}}(y_{mis}|y_{obs},\theta)]}_{\mathrm{H}}$$

If the log-likelihood is linear in the data, then its expectation can be computed by imputing the missing data with their conditional expectation given the observed data (and some parameter vector $\theta^{(i)}$)

The EM Algorithm

If the data belong to the *Exponential family of distributions* then the loglikelihood $l(\theta|y)$ is not linear in the data, but rather is linear in a set of sufficient statistics (T)
Example(bivariate case)

$$T_1 = \sum_{i=1}^{n} y_{i1},\ T_2 = \sum_{i=1}^{n} y_{i2},\ T_{11} = \sum_{i=1}^{n} y_{i1}^2,\ T_{22} = \sum_{i=1}^{n} y_{i2}^2,\ T_{12} = \sum_{i=1}^{n} y_{i1}y_{i2}$$

Filling in missing values in the E-step does not work. The expectations of the sufficient statistics must be computed.

The EM Algorithm

The computation of expected values already suggests the iterative nature of the algorithm: In order to build an expected value, distribution parameters are needed, which in turn represent the objective of the estimation.

The typical steps are:

1. Choose a set of starting values.
2. **Expectation:** Compute the expected value of the *Likelihood* given the current values of the parameters.
3. **Maximization:** Compute new parameter values which maximize this *Likelihood*.
4. Iterate to convergence.

The core of the EM Algorithm for normal data: the *Sweep Operator* (Schafer, 1997)

Alternative parameterizations of the normal distribution

Suppose that z is a $p\,x1$ random vector distributed as $N(\mu, \Sigma)$, which we partition as $z' = (z_1', z_2')$ where z_1 and z_2 are subvectors of length p_1 and $p_2 = p - p1$ respectively.

It is well known that the marginal distributions of z_1 and z_2 are $N(\mu_1, \Sigma_{11})$ and $N(\mu_2, \Sigma_{22})$, where $\mu' = (\mu_1', \mu_2')$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

are the partitions of μ and Σ corresponding to $z' = (z_1', z_2')$.

The Sweep Operator (Schafer, 1997)

Moreover, the conditional distributions are also normal; in particular, the distribution of z_2 given z_1 is normal with mean

$$\mathrm{E}(z_2|z_1) = \alpha_{2\cdot1} + \beta_{2\cdot1} z_1$$

and covariance matrix $\Sigma_{22\cdot1}$, where

$$\begin{aligned}
\alpha_{2\cdot1} &= \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \\
\beta_{2\cdot1} &= \Sigma_{21}\Sigma_{11}^{-1}, \\
\Sigma_{22\cdot1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.
\end{aligned}$$

are the vector of intercepts, matrix of slopes and matrix of residual covariances, respectively, from the (multivariate) regression of z_2 on z_1.

The Sweep Operator (Schafer, 1997)

Because specifying the joint distribution of z_1 and z_2 is equivalent to specifying the marginal distribution of z_1 and the conditional distribution of z_2 given z_1, it is possible to characterize the parameters of the distribution of z either by $\theta = (\mu, \Sigma)$ or by $\phi = (\phi_1, \phi_2)$ where

$$\phi_1 = (\mu_1, \Sigma_{11}) \quad \text{and}$$
$$\phi_2 = (\alpha_{2\cdot 1}, \beta_{2\cdot 1}, \Sigma_{22\cdot 1})$$

The Sweep Operator (Little and Rubin, 2002)

Sweeping

The sweep operator is defined for symmetric matrices as follows. A $p \times p$ symmetric matrix G is said to be *swept on row and column k* if it is replaced by another $p \times p$ symmetric matrix H with elements defined as follows

$$h_{kk} = \frac{-1}{g_{kk}}$$

$$h_{jk} = h_{kj} = \frac{g_{jk}}{g_{kk}}, \; j \neq k,$$

$$h_{jl} = g_{jl} - \frac{g_{jk} g_{kl}}{g_{kk}}, \; j \neq k, \; l \neq k.$$

The Sweep Operator

The operation of sweeping on a variable turns that variable from a dependent variable into a predictor or independent variable.

The swept matrix contains the corresponding regression coefficients along with the variance-covariance matrix of the residuals.

By means of the Sweep Operator, it is possible to compute all possible uni- and multivariate regressions among the variables of the joint distribution of the data without having to use sets of completely observed values to do it.

Eberhard Karls
Universität
Tübingen

The Data Augmentation Algorithm (Schafer, 1997)

Consider the following iterative sampling scheme: given a current guess $\theta^{(t)}$ of the parameter, first draw a value of the missing data from the conditional predictive distribution of Y_{mis}

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)}) \quad (1)$$

Then, conditioning on $Y_{mis}^{(t+1)}$, draw a new value of θ from its complete-data posterior,

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}) \quad (2)$$

The Data Augmentation Algorithm (Schafer, 1997)

Repeating this scheme from a starting value $\theta^{(0)}$ yields a stochastic sequence $\{\theta^{(t)}, Y_{mis}^{(t)} : t = 1, 2, \ldots\}$ whose stationary distribution is $P(\theta, Y_{mis}|Y_{obs})$.

The subsequences $\{\theta^{(t)}, : t = 1, 2, \ldots\}$ and $\{Y_{mis}^{(t)} : t = 1, 2, \ldots\}$ have $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$ as their respective stationary distributions.

The Data Augmentation Algorithm (Schafer, 1997)

Tanner and Wong (1987) refer to (1) as the Imputation or I-step and (2) as the Posterior or P-step, because (1) corresponds to imputing a value of the missing data Y_{mis} and (2) corresponds to drawing a value of θ from a complete-data posterior.

For a value of t that is suitably large, we car regard $\theta^{(t)}$ as an approximate draw from $P(\theta|Y_{obs})$.

Alternatively, we can regard $Y_{mis}^{(t)}$ as an approximate draw from $P(Y_{mis}|Y_{obs})$.

Empirical section

EM vs. LS

A simple example

Without NA's	
-0.2317790	1.5645425
1.3977843	3.4177385
1.0965123	1.5624080
4.2057202	5.0685650
5.1601624	6.3115680
-0.6176621	0.3965697
2.1914055	2.5698609
0.6822623	0.8115055
0.8397043	1.0817316
2.4754820	3.5060047
0.2276409	0.2446433
-1.1512896	1.8198668
0.8433541	3.0730297
1.8972929	2.8986880
3.5244921	4.0894709
2.4650780	5.0531604
1.9660875	3.2031609
1.0907652	3.0170962
-0.4410336	3.7961236
-1.1765625	-1.7791738

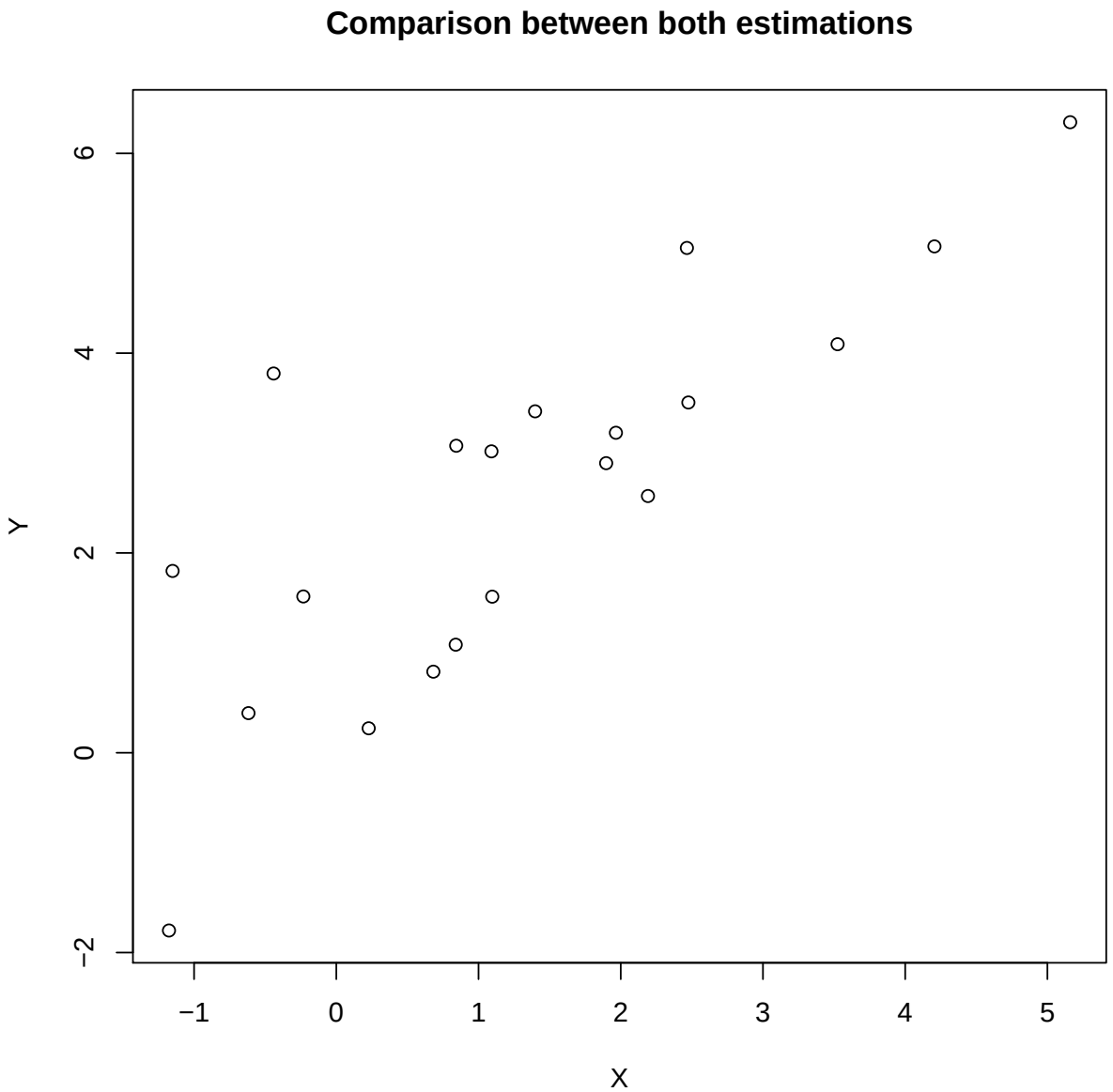
A simple example

Without NA's		With NA's	
-0.2317790	1.5645425	-0.2317790	NA
1.3977843	3.4177385	NA	3.4177385
1.0965123	1.5624080	NA	1.5624080
4.2057202	5.0685650	4.2057202	NA
5.1601624	6.3115680	NA	6.3115680
-0.6176621	0.3965697	-0.6176621	0.3965697
2.1914055	2.5698609	NA	2.5698609
0.6822623	0.8115055	NA	0.8115055
0.8397043	1.0817316	0.8397043	1.0817316
2.4754820	3.5060047	NA	3.5060047
0.2276409	0.2446433	0.2276409	NA
-1.1512896	1.8198668	-1.1512896	NA
0.8433541	3.0730297	NA	3.0730297
1.8972929	2.8986880	1.8972929	2.8986880
3.5244921	4.0894709	NA	4.0894709
2.4650780	5.0531604	2.4650780	NA
1.9660875	3.2031609	1.9660875	NA
1.0907652	3.0170962	1.0907652	NA
-0.4410336	3.7961236	-0.4410336	3.7961236
-1.1765625	-1.7791738	-1.1765625	NA

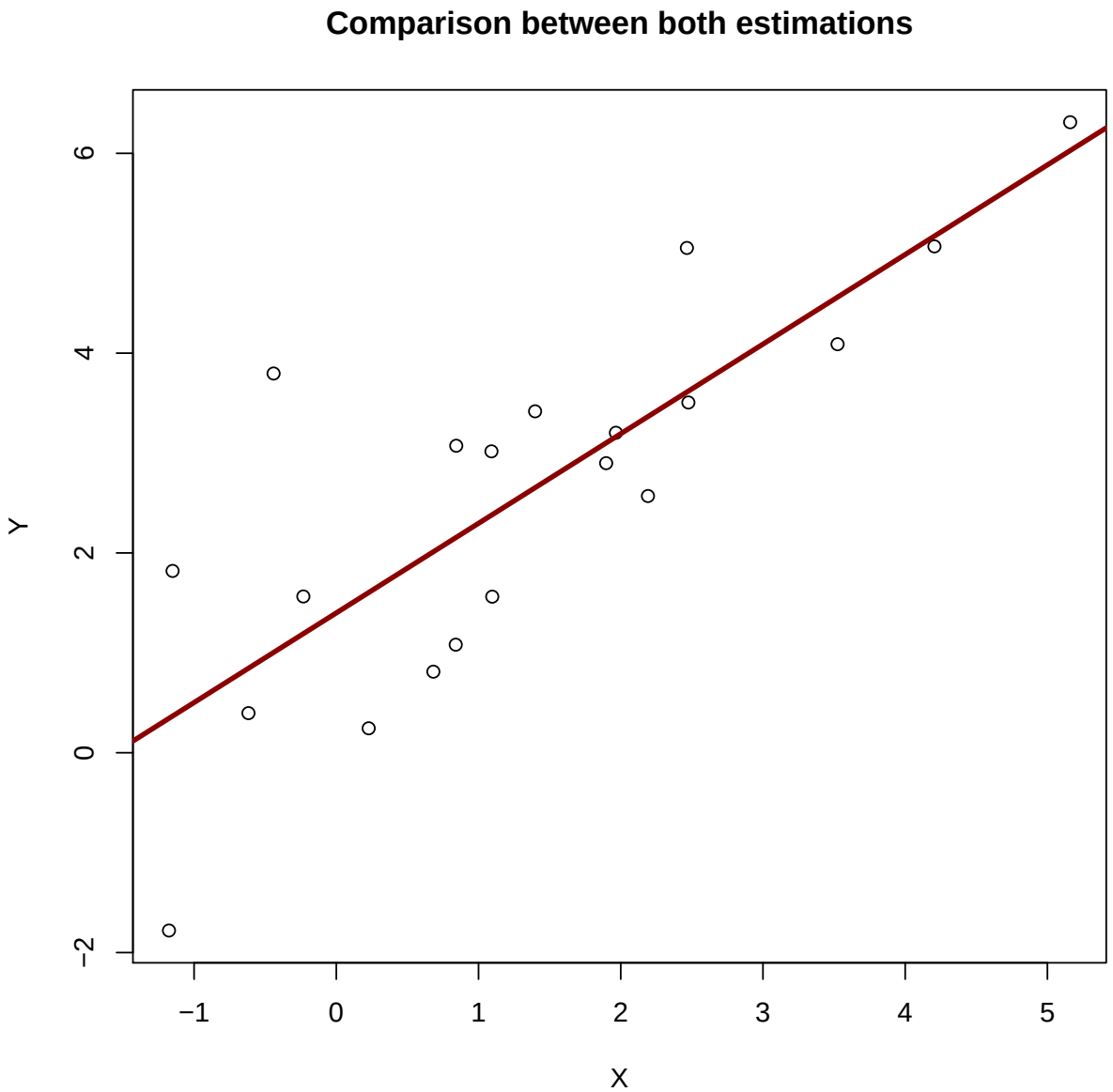
A simple example

Without NA's		With NA's	
-0.2317790	1.5645425	-0.2317790	NA
1.3977843	3.4177385	NA	3.4177385
1.0965123	1.5624080	NA	1.5624080
4.2057202	5.0685650	4.2057202	NA
5.1601624	6.3115680	NA	6.3115680
-0.6176621	0.3965697	-0.6176621	0.3965697
2.1914055	2.5698609	NA	2.5698609
0.6822623	0.8115055	NA	0.8115055
0.8397043	1.0817316	0.8397043	1.0817316
2.4754820	3.5060047	NA	3.5060047
0.2276409	0.2446433	0.2276409	NA
-1.1512896	1.8198668	-1.1512896	NA
0.8433541	3.0730297	NA	3.0730297
1.8972929	2.8986880	1.8972929	2.8986880
3.5244921	4.0894709	NA	4.0894709
2.4650780	5.0531604	2.4650780	NA
1.9660875	3.2031609	1.9660875	NA
1.0907652	3.0170962	1.0907652	NA
-0.4410336	3.7961236	-0.4410336	3.7961236
-1.1765625	-1.7791738	-1.1765625	NA

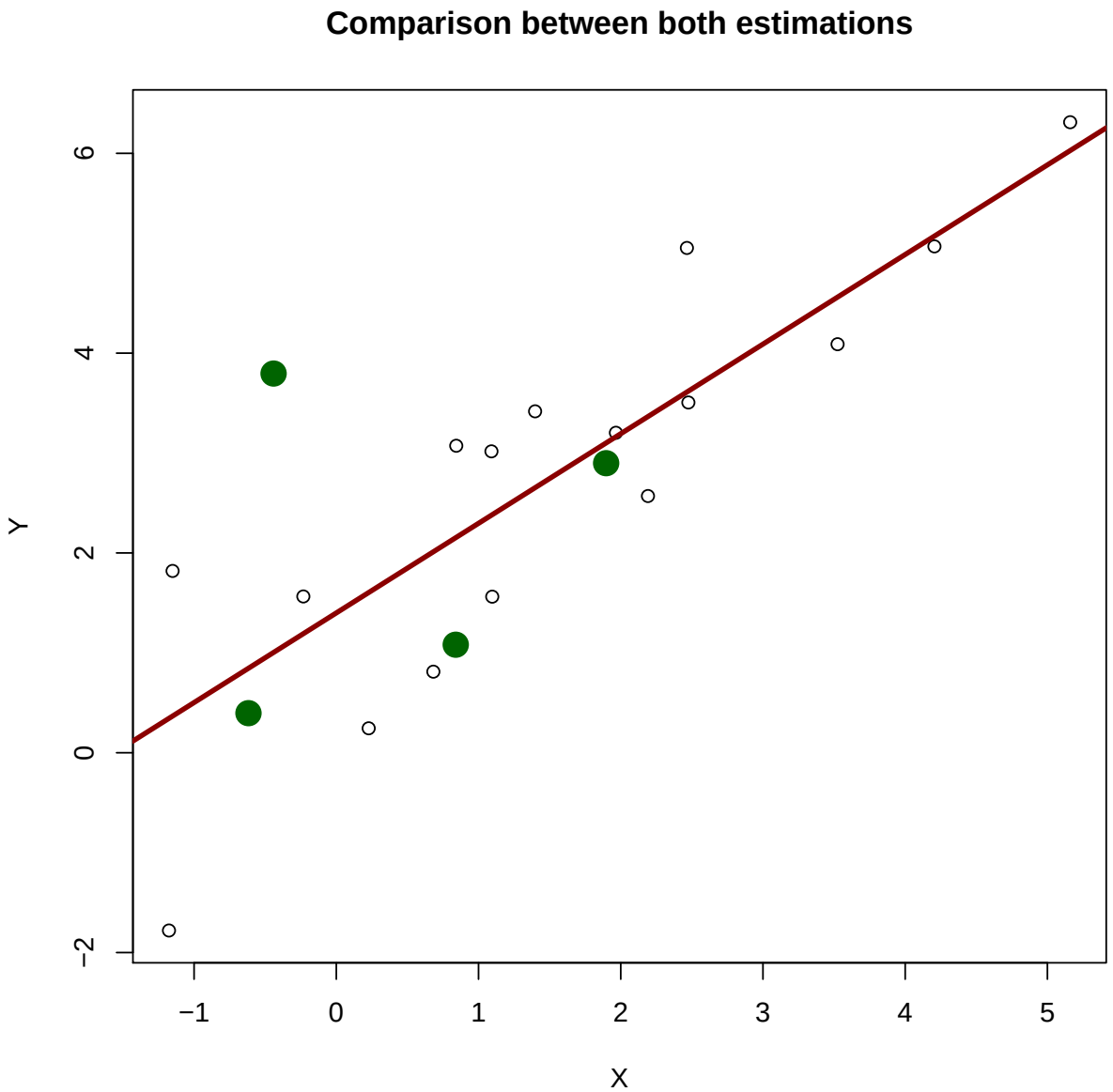
Scatterplot of the real data



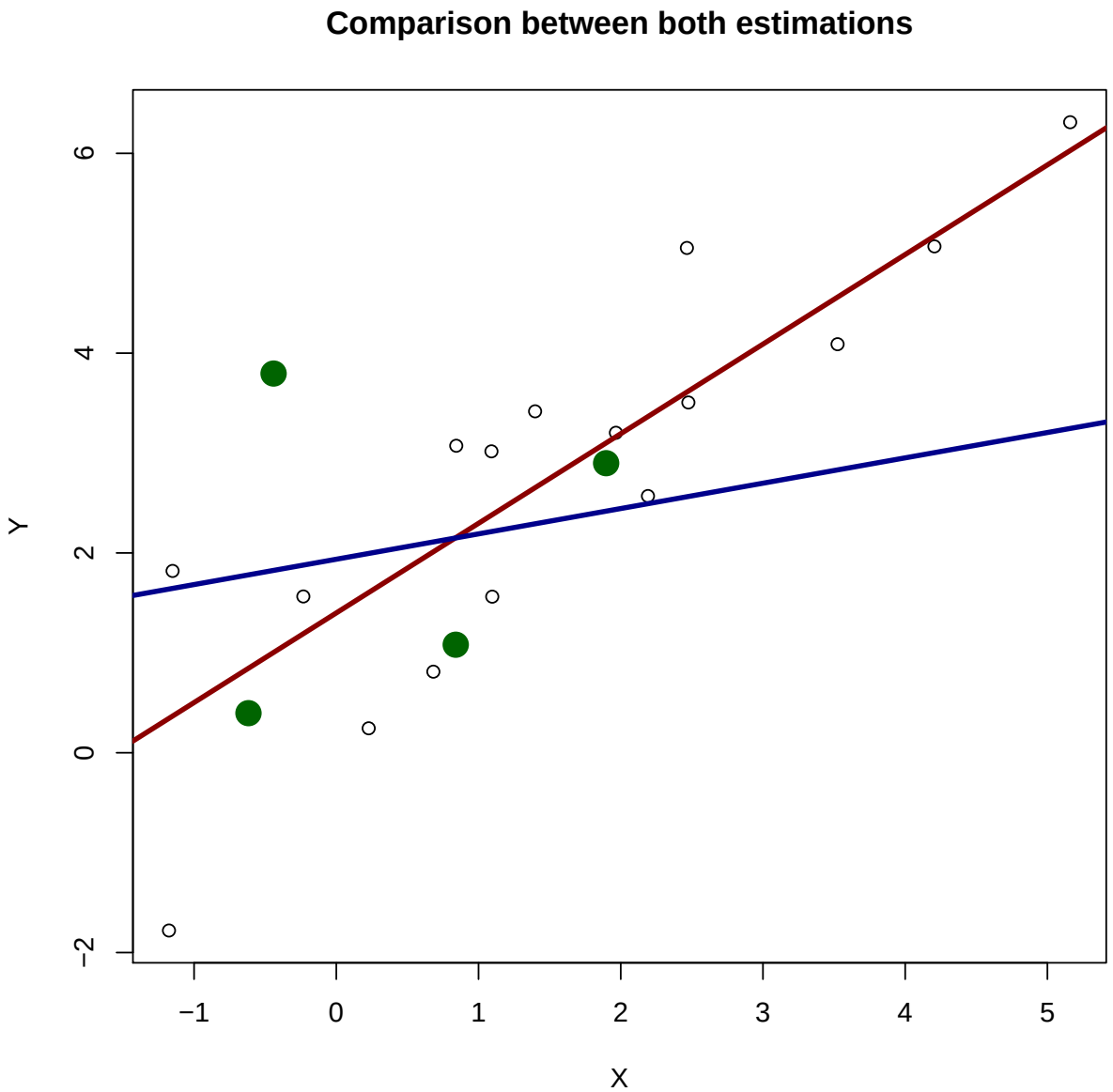
True (sample) regression line



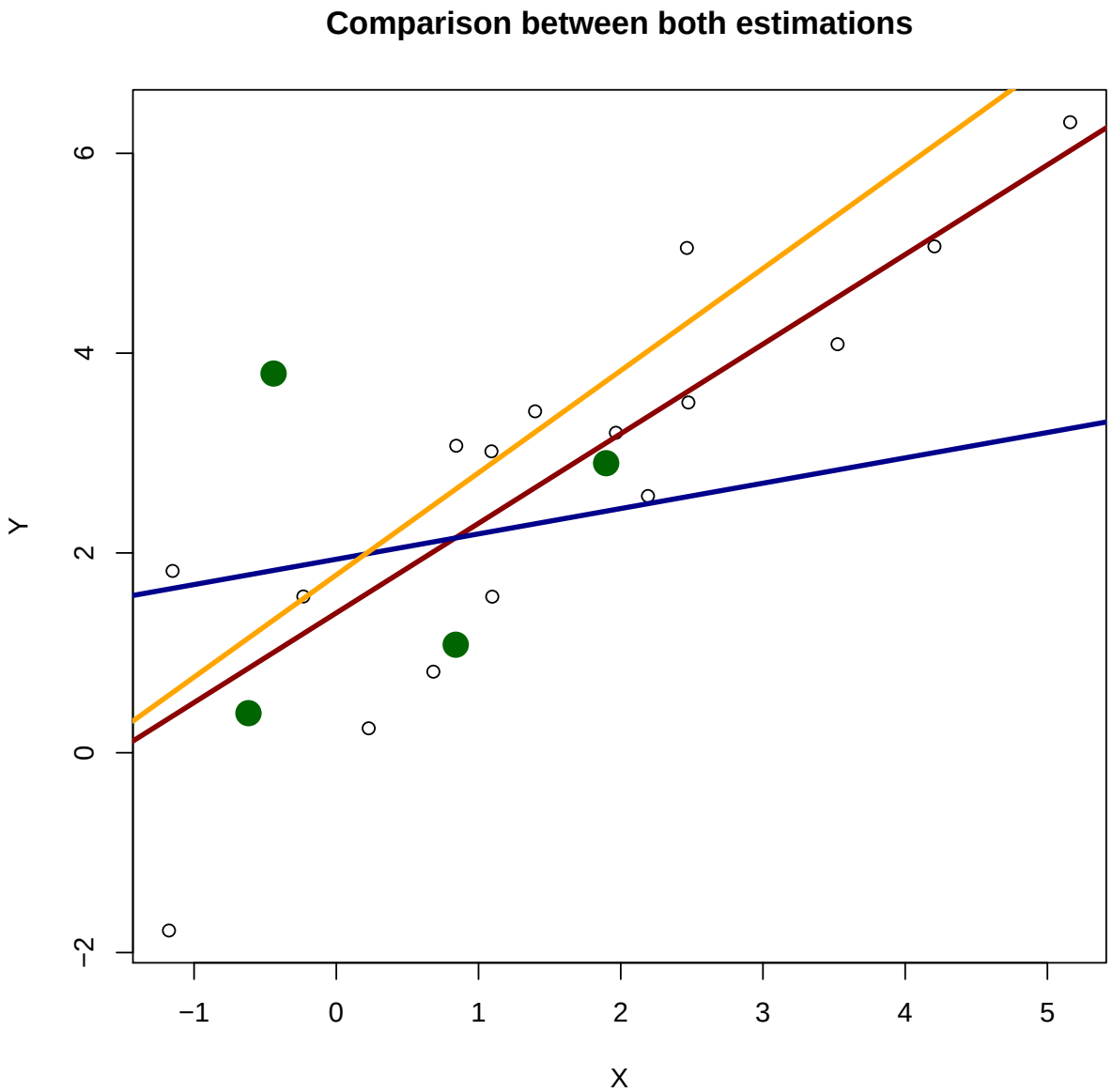
Pairs of fully observed data



LS-regression

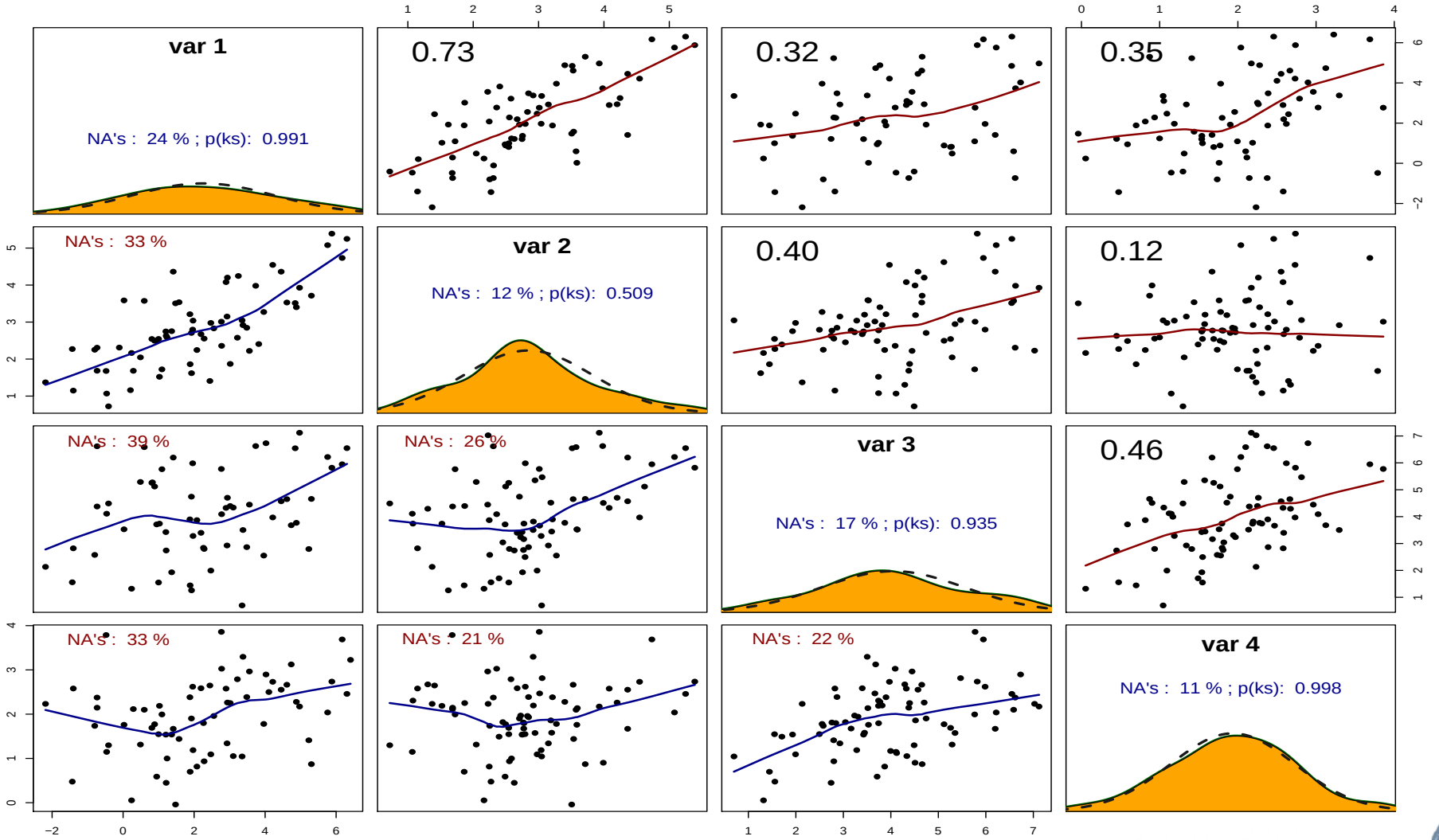


EM-regression



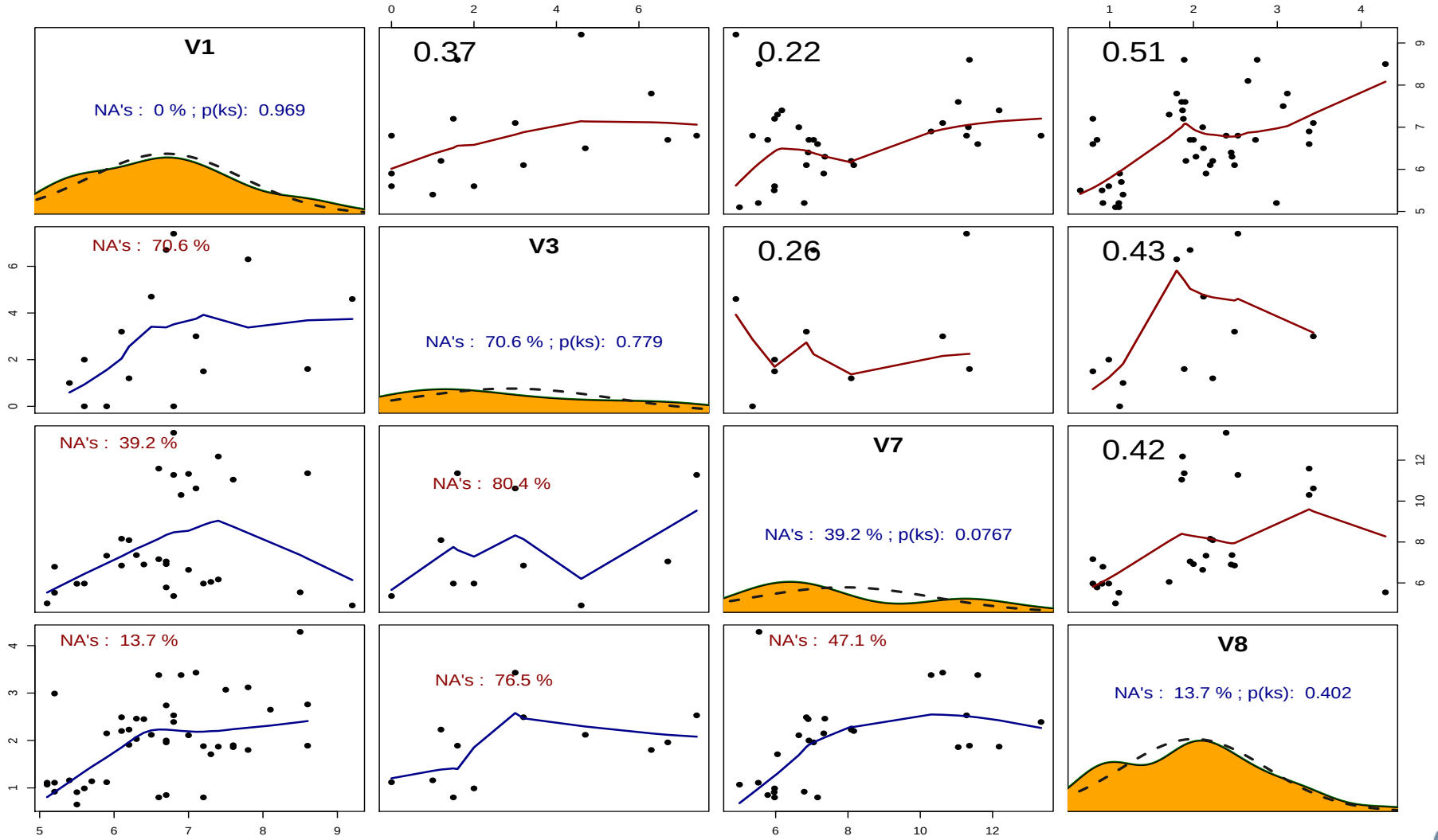
Visualizing multivariate data

Normally distributed multivariate data with missing values



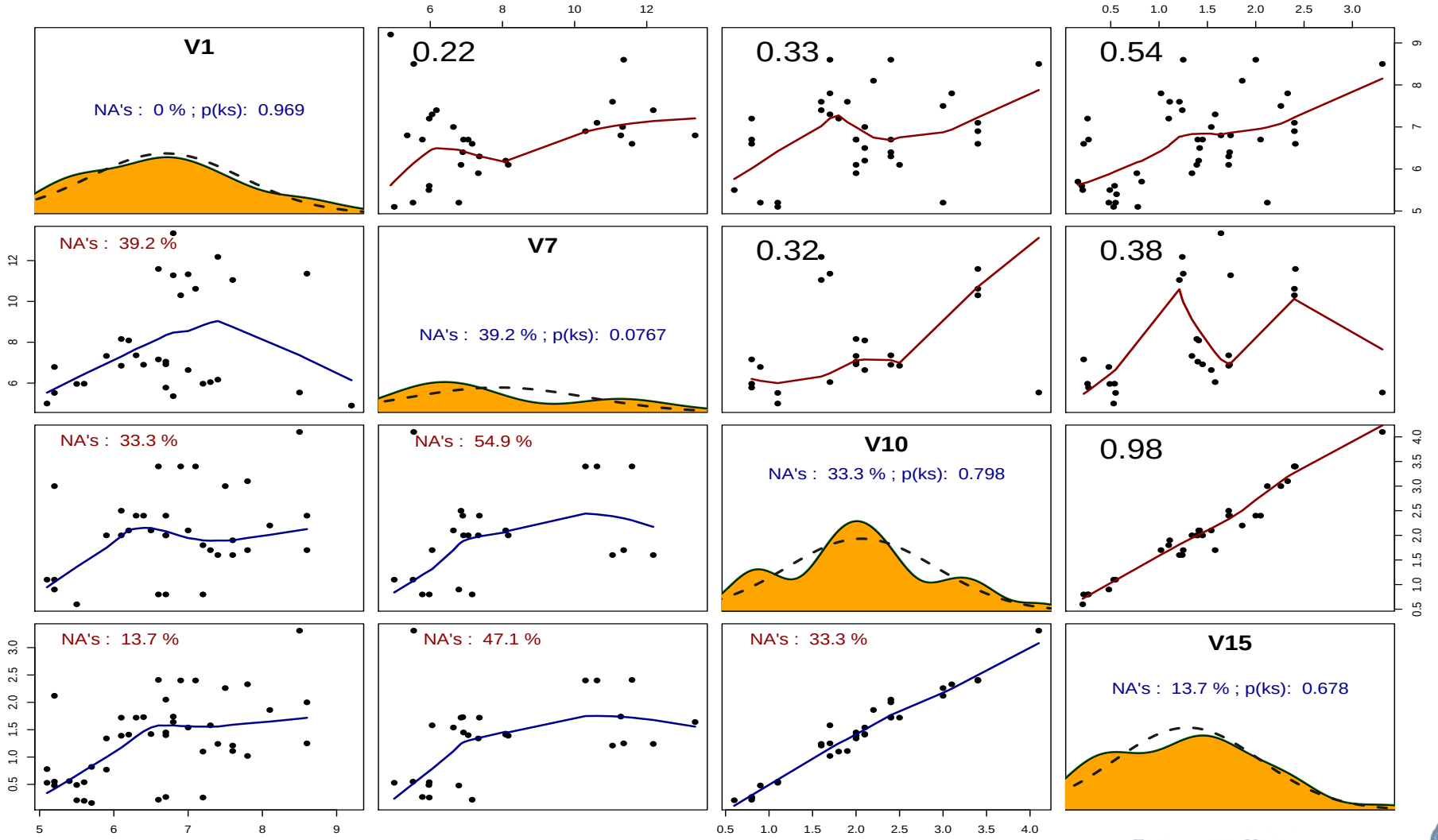
Eberhard Karls
Universität
Tübingen

Exploration of the data I



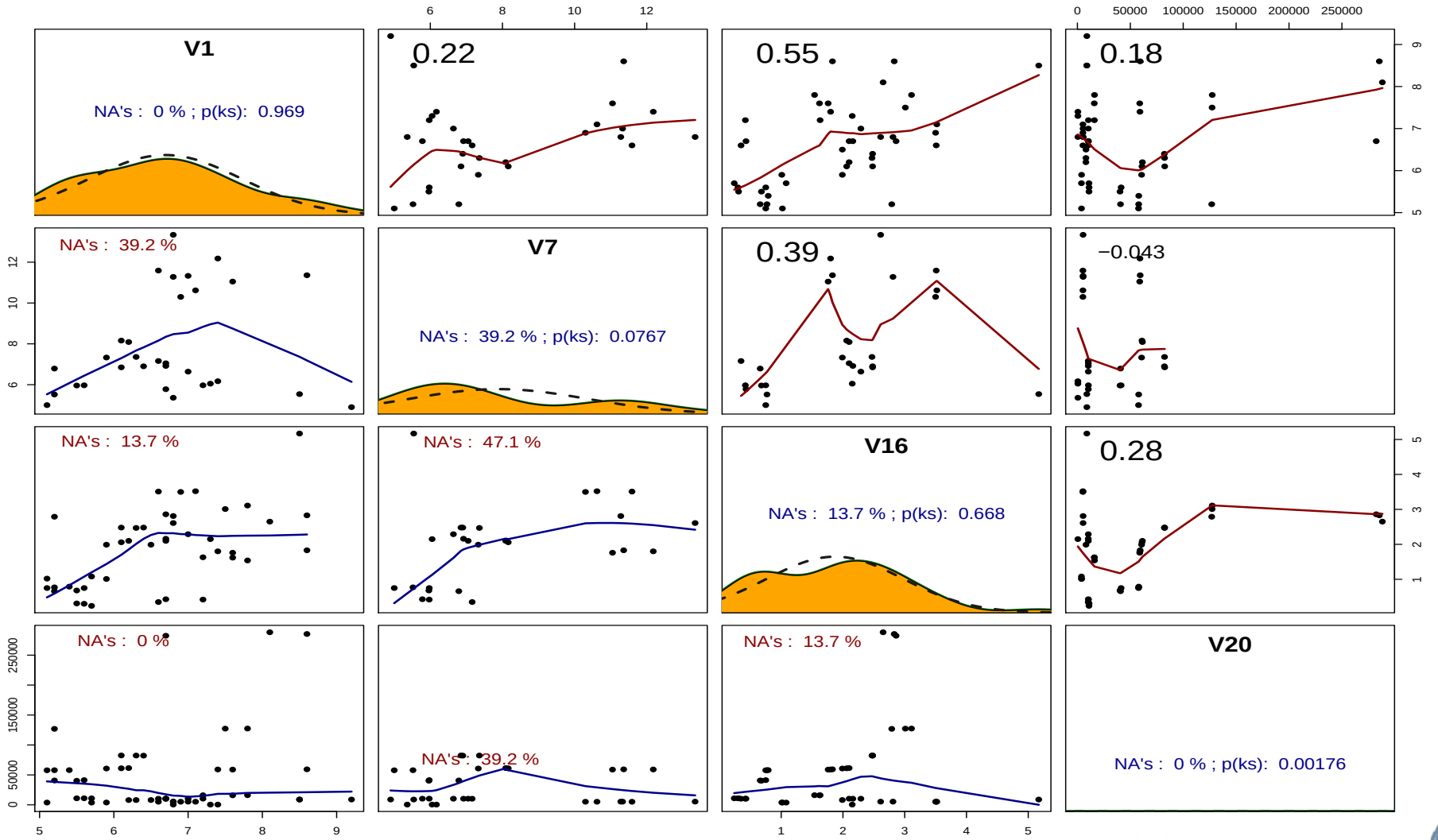
Eberhard Karls
Universität
Tübingen

Exploration of the data II



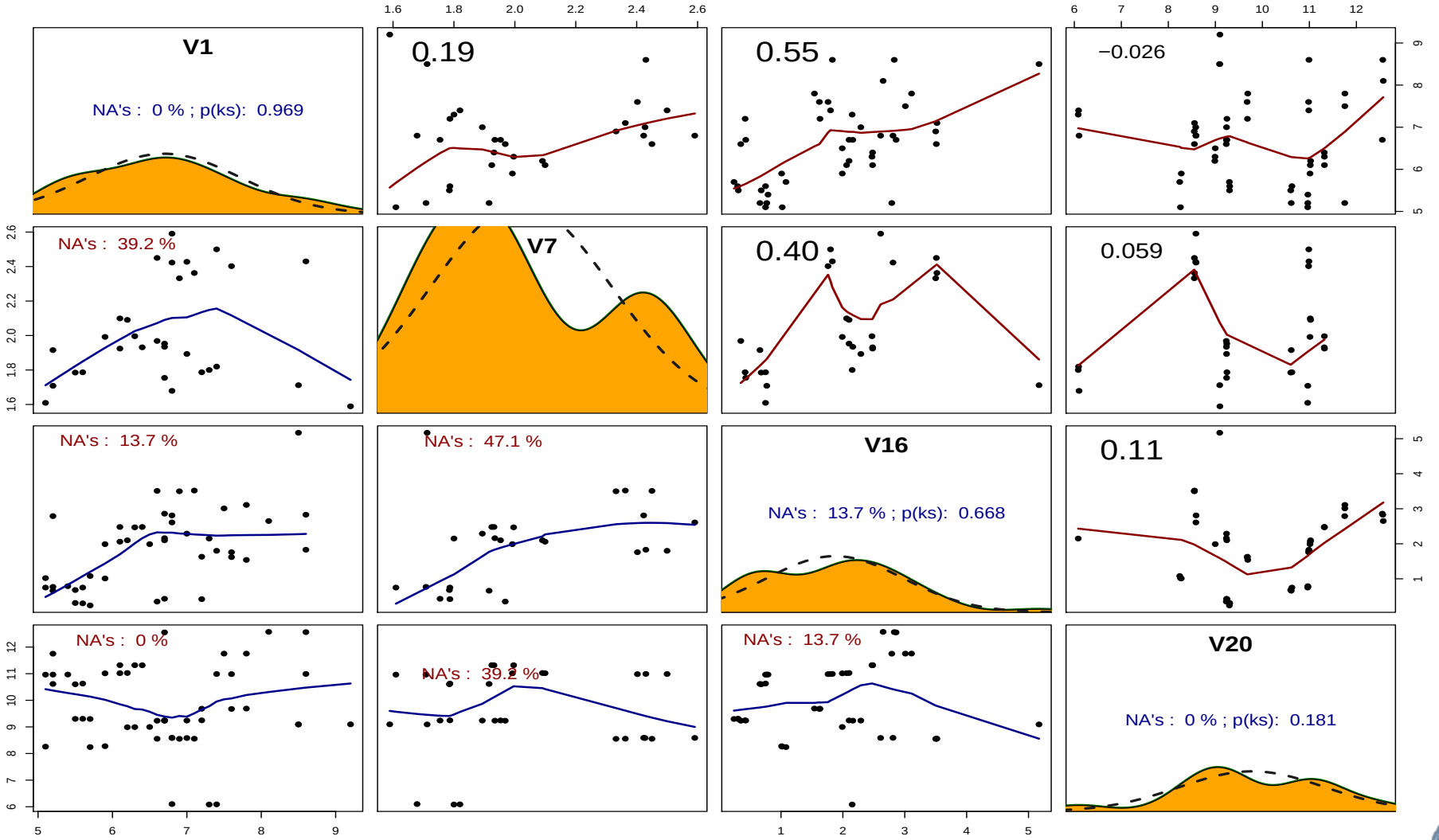
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Exploration of the data III



Eberhard Karls
Universität
Tübingen

Exploration of the (transformed) data



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Experiences from the first imputation round

Setting

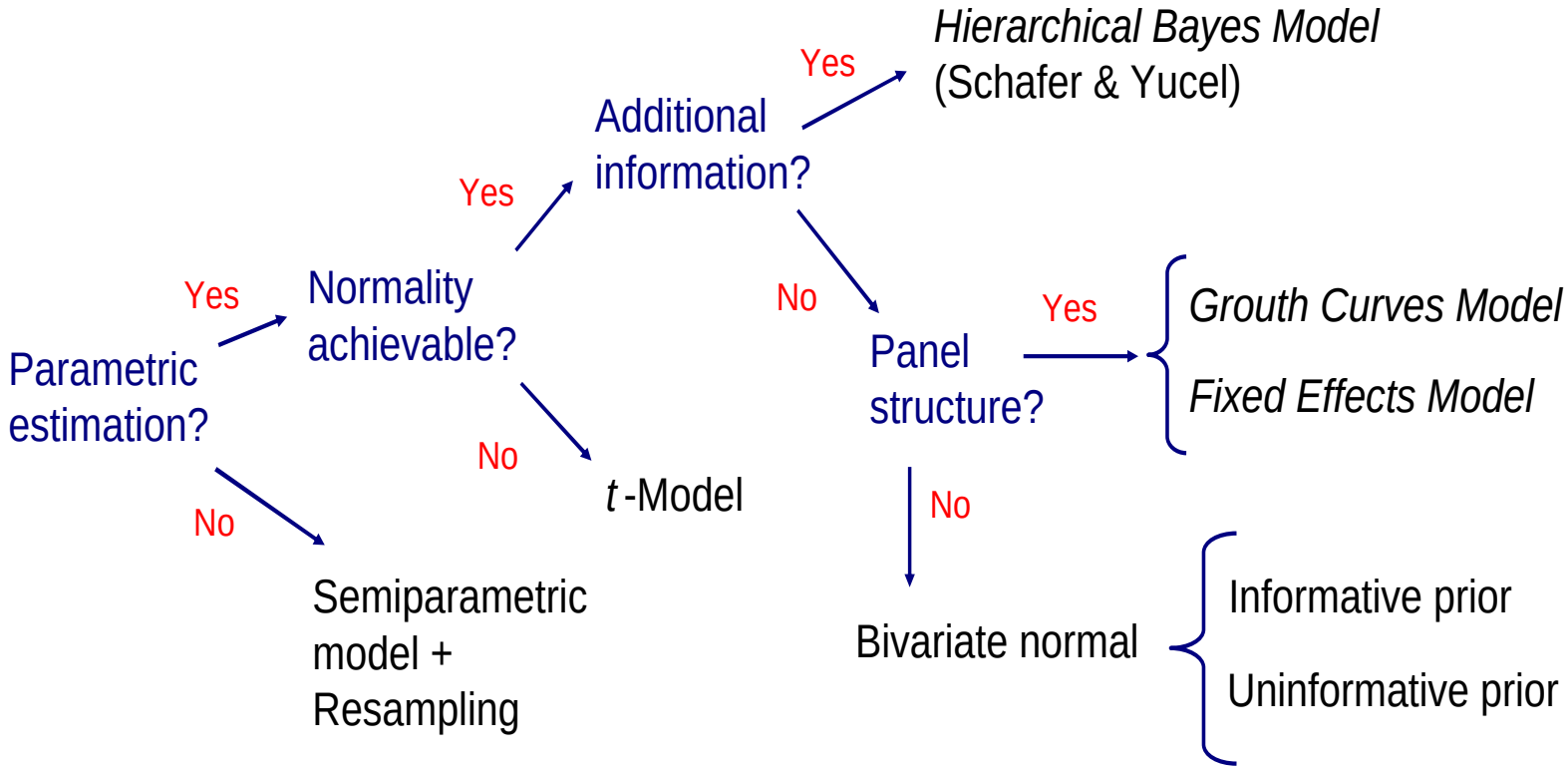
- The presented methods for normal data were tested on a real KEI-Dataset.
- Several Models of increasing complexity were used.
- An attempt was made to apply traditional imputation methods to the data in order to compare its performance.

Experiences from the first imputation round

Results

- The nonnormality of the variables slowed down the convergence of the Algorithm. The parameters were also poorly estimated.
- The nonlinearity of the relationships diminished the quality of the imputation.
- Simple models seemed to perform better than bigger ones: even when the EM Algorithm was able to estimate the parameters of the joint distribution, the resulting parameters were probably just to poorly estimated and provided imprecise imputed values.
- The high proportion of missing data rendered methods not basing on the EM Algorithm completely useless.

Although no model can utterly cope with all the problems found, the following models give partial answers to them:



Further reading

- Casella, G. and George, E. I. Explaining the Gibbs Sampler. *The American Statistician*, **46**, 167-174. 1992.
- Little, R. J. A. and Rubin, D. B. Statistical Analysis with Missing Data. Second edition. Wiley. 2002
- Schafer, J.L. Analysis of Incomplete Multivariate Data. Chapman and Hall. 1997
- Tanner, M.A. Tools for statistical inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions. Second Edition. Springer-Verlag. 1993.
- Tanner, M.A. and Wong, W. H. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528-550.

Procedures based on Completely Recorded Units:

When some variables are not recorded for some of the units, a simple expedient mentioned in Section 1.1 is simply to discard incompletely recorded units and to analyse only the units with complete data (Nie et. al., 1975). [...]. It is generally easy to carry out and may be satisfactory with small amounts of missing data. **It can lead to serious biases, however, and it is not usually very efficient, especially when drawing inferences for subpopulations.**

Little and Rubin, 2002; p. 19.

Ignorability

The missing data mechanism is ignorable for likelihood inference if:

MAR: the missing data are missing at random.
Distinctness: the parameters θ and ψ are distinct, in the sense that the joint parameter space of (θ, ψ) is the product of the parameter space of θ and the parameter space of ψ.

Little and Rubin, 2002; p. 120.

Ignorability

The missing data mechanism is ignorable for bayesian inference if:

MAR: the missing data are missing at random.
The parameters θ and ψ are *a priori* indepedent, that is, the prior distribution has the form

$$p(\theta, \psi) = p(\theta) \cdot p(\psi)$$

Little and Rubin, 2002; p. 120.

Ignorability

Filling in missing values in the E-step does not work because the loglikelihood $l(\theta|y)$ is not linear in the data, but rather is linear in the following sufficient statistics (bivariate case)

$$T_1 = \sum_{i=1}^{n} y_{i1},\ T_2 = \sum_{i=1}^{n} y_{i2},\ T_{11} = \sum_{i=1}^{n} y_{i1}^2,\ T_{22} = \sum_{i=1}^{n} y_{i2}^2,\ T_{12} = \sum_{i=1}^{n} y_{i1}y_{i2}$$

Little and Rubin, 2002; p. 171.

General Ignorable Procedures

We shall call a missing-data procedure a *general ignorable procedure* if it is based upon either an observed-data likelihood or an observed-data posterior. The EM algorithm, for example, will be seen to be a general ignorable procedure because it maximizes the observed-data likelihood.

Schafer, 1997; p. 23

Limited practical experience with real data also suggests that general ignorable procedures may tend to perform well even when the ignorability assumption is suspect, especially in multivariate settings.

Schafer, 1997; p. 27.

General Ignorable Procedures

David *et al.*(1986) found little evidence of bias in ignorable procedures that imputed missing values of income on the basis of other demographic and questionnaire items that were observed.

This evidence came from knowledge of the missing values obtained from an external source, the actual wages and salary reported to the Internal Revenue Service on the individuals' tax returns.

David *et al.*(1986) also concluded that further improvements in the missing-data procedures would probably come from better modeling of the multivariate structure of the data, not from nonignorable modeling.

Schafer, 1997; p. 27.

Eberhard Karls
Universität
Tübingen

About proper imputation models

From a practical standpoint, knowing whether an imputation method is technically proper for a particular analysis is less important than knowing whether it actually behaves well or poorly over repeated samples. The latter question can be addressed through simulation studies with realistic complete-data populations and realistic response mechanisms.

Schafer, 1997; p. 145.

Eberhard Karls
Universität
Tübingen