LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

LONGITUDINAL DATA WITH MISSING DATA AND MEASUREMENT ERROR

MASTER THESIS

Handed in by Lisa Baganz
Supervised by Prof. Dr. Christian Heumann
Munich, 08.09.2015

Abstract

In this thesis we examine the performance of a method to correct for missing response, missing covariates and mismeasured covariates in longitudinal datasets. The method is called Multiple Overimpuatation and was proposed by Blackwell et al. (2015a,b). To our knowledge it was not tested before for longitudinal missing and mismeasured data. Many researchers do not correct for measurement error but only for missing data, probably due to the fact that there exist easy-to-use methods in many statistical programs as multiple overimputation for missing data but no easy-to-use methods for mismeasured data. We conduct a simulation study to show that additional correction for measurement error using Multiple Overimputation improves the results a lot. Even if the exact amount of measurement error is unknown, Multiple Overimputation performs better than normal Multiple Imputation or a Complete Case Analysis. Multiple Overimputation was included in the R-package Amelia II, so researchers do not have to implement anything by themselves and can easily use this new approach to improve their results.

Contents

1 Introduction

Empirical studies are very important in many fields of research like medical, sociological and economic research. For example in medical research often new therapies are tested against older therapies. A common procedure for testing this is to conduct a longitudinal study where one half of the propositi get the old therapy (placebo group) and the other half the new one (treatment group). Over a given time period the propositi are medically examined multiple times, leading to a longitudinal datasets containing by definition several measurements over time per subject. This study type often occurs in clinical, pharmaceutical, sociological studies. For example, if a new medicine that should lower the blood pressure is tested, one half of the propositi ingest this new medicine and the other half the old or none over a time period and the blood pressure and other variables of interest are measured at several time points before, during and/or after this period. So one can test whether the new medicine actually lowers the blood pressure over time. For longitudinal datsets there are several analysis methods, e.g. different types of mixed models which can control for individual specific effects. But what if a propositi moves away so she cannot take part in the study any longer or, the extreme, if a propositi dies? This phenomenon is called drop-out since the propositi do not rejoin the study after missing one or several appointments and we obtain missing values in the data. Missing values can be obtained also for single values for example if a propositi misses a checkup. Variables which are not time-dependent (baseline variables) can be missing, too. The methods which can be used for missing data depends on the types of missing data. Three types are distinguished: missing completely at random, missing at random and not missing at random. If data are missing completely at random, standard methods can be used for the complete cases (the rows of the data without missings), since with missing completely at random the missingness probability does not depend on other values of the dataset. If the data are missing at random, the missingness probability depends on some values of the observed data only. For this scenario there exist several methods, like imputation methods where the missing values are replaced, likelihood methods and bayesian methods. The fewest assumptions have to be made if the missing data are not missing at random since in this case the missingness probability of a value is allowed to depend on all other values in the data, regardless if missing or observed and even on the missing value itself. The analysis methods for this type are much more complicated since the missing data mechanism has to be specified which is mostly not known.

Besides the problem of missing data, in empirical studies it may happen that single values or even whole variables are not measured correctly. In other terms: The variable or value is subject to measurement error. For example, a mismeasured variable can result if the measurement device is not precise enough and so the measurements vary around the true value. There are also specific analysis methods for data measured with error even for longitudinal data. But most methods are complicated to use.

But why should these problems not occur together? Usually researchers correct only for one of these two problems, missingness or measurement error, if the data are longitudinal. In our opinion the reason is that there are no easy-to-use methods to correct for both. There is already some literature concerning longitudinal data with measurement error and missing data but on the one hand most methods are complicated since they are not implemented in statistical programs and on the other hand to our knowledge they only address mismeasured covariates and missing response. We think that many longitudinal studies exist where covariates are missing, too and that the results could be improved by correcting for it. Therefore we want to examine the performance of an easy-to-use method [1], called Multiple Overimputation (Blackwell et al. (2015a, 2015b)), if we face longitudinal data with measurement error and missing data. Multiple Overimputation is an extension of Multiple Imputation (Honaker and King (2010)) that also corrects for measurement error. Multiple Imputation replaces (imputes) the missings multiply by plausible values using the information from the observed values. The result are several imputed datasets without missings which can be analysed by standard statisical methods for complete data and the results for the several datatsets can be combined to one result. It is better to impute multiply because it ensures that the variability due to the imputation is included (we do not know the exact value but replace the missing by a plausible random value). The same holds for Multiple Overimpuatation which overwrites the mismeasured values,but uses in addition information from the observed values and the information from the mismeasured value itself. The main idea of Overimputation is that missing data can be interpreted as enormous measurement errors. If the measurement error goes to infinity, the mismeasured variable contains no information on the true value and therefore this is the same as if the value was missing. The ideas of Blackwell et al. (2015a, 2015b) and Honaker and King (2010) are implemented in the R-package Amelia II. We will use this package to analyse a simulated longitudinal dataset with missing covariate values, a mismeasured covariate and drop-out in the response (for simplicity all covariates are baseline covariates) and examine if the method is appropriate for this phenomenon. Therefore we calculate the bias and MSE of the coefficients of a linear mixed model. We compare the results of a linear mixed model with random intercept for the true data without measurement error or missings and three analyses methods for the observed data: Multiple Overimputation, Multiple Imputation (measurement error ignored) and a complete case analysis (both problems ignored). To support our results we also conduct a sensitivity analysis to examine how the results

[1] It is easy to use since it is implemented in the R-package Amelia II and therefore can be used without further programming.

are influenced by different values of measurement error or missingness and by the number of measurements per subject. As a part of the sensitivity analysis we relax the assumption that we know the amount of measurement error, assume that it is over-/underestimated and examine the influence on the resulting coefficients. Additional to the comparison of the coefficients we also compare the results for the variance of the random intercept for the different methods and different amounts of measurement error.

For a better understanding of the underlying problems we want to explain in the following section the different missing data types and the most important measurement error types in general. Moreover we describe the consequences and how we can correct for measurement error or missing data in general. We also explain how a measurement error variance can be estimated and we outline some methods for longitudinal data with missing response and mismeasured covariates. In the third section we explain the Multiple Impuatation and Multiple Overimputation approach. Then we describe the model which we use in the simulation study. The simulation study is presented in section five where we describe the data generation, the imputation, the analysis, the results and at the end the sensitivity analysis.

2 Missing Data and Measurement Error

Missing Data is a problem that often occurs in empirical studies. Imagine that propositi test a new medicine that should lower the blood pressure and have to visit the doctor several times for the duration of the study. It is possible that for some propositi measurements of the blood pressure or other covariates are missing. In which way one can deal with missing data depends strongly on the missing-process. It is possible that people just forget a visit or that their blood pressure is so high that they cannot keep the appointment. The different types of missing data will be described in the following part. In the second part of this section we will outline some existing methods to handle missing data.
Some of these methods can be modified to deal with another problem that could occur - measurement error. In the blood pressure example a measurement error could be that the blood pressure is not measured exactly since the haemodynamometer has a non-negligible variance in its measurements. We will describe some types of measurement error and methods for handling this problem in the third and fourth part of this section. It is not unusual that both problems, missing data and measurement error, occur together. So we outline some methods for handling missing data and measurement error for longitudinal data in section 2.7.

2.1 Notation

To describe the scnearios in the next sections, we first have to define several variables:
The response vector $\boldsymbol{y}_i$ $(T \times 1)$ for subject $i \in 1, ..., I$ is denoted by $y_i = (y_{i1}, y_{i2}, ..., y_{iT})'$ where T is the number of measurements per subject. The associated covariate matrix $(T \times P)$ is given by $\boldsymbol{X}_i$ where P denotes the numbers of covariates. The $(I \cdot T) \times (P + 1)$ dataset is denoted by $\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X})$. To locate the missing values, we define an $(I \cdot T) \times (P + 1)$ matrix as missingness indicator, $\boldsymbol{M}$ with $m_{itp} = 1$ if the corresponding value z_{itp} is missing and $m_{itp} = 0$ if it is observed. The distribution of $\boldsymbol{M}$ is given by $P(\boldsymbol{M}|\boldsymbol{Z}) = P(\boldsymbol{M}|\boldsymbol{y}, \boldsymbol{X})$. If $\boldsymbol{Z}$ is subject to missingness, we can split this up into two subsets, one containing the observed values denoted by $\boldsymbol{Z}^{obs}$ and one containing the missing values denoted by $\boldsymbol{Z}^{mis}$.

2.2 Types of Missing Data

In general, there are three types of missing data (Little and Rubin (2002)):

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)

3. Not Missing at Random (NMAR)

It is important to decide between these types to conduct reliable analysis. Therefore we want to describe the types and their consequences for analysis.
1. MCAR: Missing completely at random means that the probability of missingness of a variable does not depend on values of other variables or the variable itself:

$$P(\boldsymbol{M}|\boldsymbol{Z}) = P(\boldsymbol{M}).$$

If a participant just forgets her visit at the doctor, this missing value could be regarded as MAR. The observed subset of the dataset $\boldsymbol{Z}^{obs}$ is a random sample of the whole dataset $\boldsymbol{Z}$ because the missing values are fully random and do not depend on values of any variable. This implies that the joint distribution and the moments are the same for the complete data and the observed subset. So we do not have to correct for missing data since we will obtain the same results if we just use the observed data as if we would have used the whole dataset. This "method" of just using the observed data is called complete case analysis and will be shortly described in section 2.3. But the observed data are often not a random sample of the complete data, hence the missing mechanism is either MAR or NMAR.
2. MAR: The assumptions for MCAR are often not met in reality. With Missing at random, the probability of missingness is allowed to depend on the observed values but not on the missing values:

$$P(\boldsymbol{M}|\boldsymbol{Z}) = P(\boldsymbol{M}|\boldsymbol{Z}^{obs}).$$

Imagine that the blood pressure is missing and another variable is the weather which is measured to control for the variations in blood pressure. The weather is observed for a corresponding missing blood pressure value. Then the Blood pressure would be missing at random if the person was not showing up because the weather was bad. Here the distribution of the observed data and the complete data are not the same ($\boldsymbol{Z}^{obs}$ is no random sample of $\boldsymbol{Z}$). Therefore we cannot use the observed data for analysis, only. If we would use a complete case analysis although the data are MAR, we obtain biased estimates of the moments for the complete data. This missing mechanism (MAR) is still restrictive because the missing values are not allowed to depend on other missings or the missing value itself. So a third missing data mechanism, called not missing at random, is defined which covers this situation.
3. NMAR: If the data are not missing at random, the probability of a missing value is

allowed to depend on this missing value itself or other missing values:

$$P(\boldsymbol{M}|\boldsymbol{Z}) = P(\boldsymbol{M}|\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}).$$

An example for NMAR is that the value of the blood pressure is missing because the persons blood pressure was so high that she could not take the appointment. Consequently valid inference is only possible if we specify the model for $P(\boldsymbol{M}|\boldsymbol{Z})$. The choice of the missing data model influences the results of the analysis. It is difficult to specify the missing data model since the observed data contain no information on this mechanism. Hence we cannot support or refuse a hypothesis concerning a specific mechanism and therefore sensitivity analyses with different mechanisms are useful.

In addition to dividing between these types of missing data, Ibrahim and Molenberghs (2009) draw a distinction between ignorable and non-ignorable missing data. If a data mechanism is ignorable, we can use only the observed data for an analysis and still get valid results. It can be deduced that MCAR data are ignorable since we already described that we can use only the complete cases for valid inference. MAR data are also ignorable if some light conditions are fulfilled - but unfortunately only in the likelihood and Bayes framework and not in frequentist. In a frequentist franework MAR is non-ignorable and NMAR is always non-ignorable. This means that the missing data mechanism cannot be ignored if we want to do inference for the complete data but the missing data mechanism needs to be specified.

In principle, we have to distinguish between ignorable and nonignorable missingness to choose an appropriate analysis method but there is a third classification for missing data which is also important for the choice of the analysis method - missing data patterns. To explain what missing data patterns are, we want to describe drop-outs first. Drop-out means that after one missing value all others are missing, too. For example, propositi of a clinical longitudinal study could move away during the study duration so that they cannot take part in the study any more or the extreme case would be that a person dies. Drop-outs are one missing data pattern which is also called monotone missing data pattern. If the missingness of a value does not imply that the following values of the variable for the corresponding subject are also missing and some values are observed after one missing value, the missing data pattern is non-monotone.

As already mentioned before, the handling of missing data depends on the missing data patterns and the missing data mechanisms. We want to describe some methods for missing data in the following section. We only distinguish between the missing data mechanism and not the missing data patterns since in general all these methods can be used with both patterns after some adjustments.

2.3 How to handle Missing Data

There is not a "best" method for a missing data type but there are different ways to deal with missing data. It depends on the type which methods are appropriate.

The easiest way is to use only the complete cases (CC). In a complete case analysis all rows are dropped in which a value is missing. If the data would be non-ignorably missing there would be a structure in missingness and for example rows from patients with very high blood pressure would be dropped. Since the CC analysis is a frequentist methods, NMAR and MAR are non-ignoroble. If you would use the CC analysis for MAr or NMAR data nevertheless, the results are biased and you cannot generalize the results of such an analysis to the population which is basically the meaning of statistical inference. Similarly, with the Available Case Analysis (AC), all cells with missing values are deleted. So for each analysis model, all observed cases are used but the models can not be compared anymore since they may contain different observations. Both methods, CC and AC, can be used if the data are missing completely at random (MCAR) because then the observed data are a random sample of the complete data and therefore the results of the analysis can be generalized and are not biased.

But if the data is non-ignorable missing, we can still use the data for complete [2] data inference. There are several methods which can deal with MAR data and with some (strict) assumptions also with NMAR data. These methods can be divided in the following sections:

1. Maximum Likelihood
2. Fully Bayesian
3. Weighted Estimating Equations (WEE)
4. Single Imputation methods
5. Multiple Imputation

We want to describe the mentioned possibilities for handling missing data briefly based on Molenberghs et al. (2014).

1. Maximum Likelihood:

Maximum Likelihood is an often used method for inference in statistics not only if data are missing. It is based on the idea of searching the parameter which maximizes the probability of obtaining the observed random sample of the total population. Therefore

[2] The term complete is a bit misleading since it is used for two different things in the context of missing data: complete data are the whole true unobserved dataset which by definition do not contain any missings; complete cases describe the rows of the observed data which do not contain missing values.

the likelihood of this parameter is constructed, interpreted which is a function of the unknown parameter, only. To estimate this parameter, the likelihood is maximized. The standard ML-method can be modified so that it can be used with missing data.
There is lots of literature about ML-estimation combined with missing data in general and also a lot for longitudinal studies. ML-methods can be used for data which are MCAR or MAR, but with some modifications it is also possible to use them if the data are not missing at random. To use ML methods for partially missing data a parametric model for the complete data have to be specified and if the data are NMAR a parametric model for the missing data need to be specified, also. This can be difficult because we rarely know the missing data model and therefore we have to make assumptions about it. But here we want to concentrate primarily on MAR and MCAR data. With the help of the complete data model we can construct a likelihood for the parameters of interest. Since the complete data are not available, the observed data likelihood is used to obtain the ML-estimates. For simplicity we assume that only values of the response $\boldsymbol{y}_i$ are missing and $\boldsymbol{X}_i$ is fully observed. Hence, $\boldsymbol{M}_i = (\boldsymbol{M}_{i1}, ..., \boldsymbol{M}_{iT})'$ is now a $T \times 1$ vector. The observed-data likelihood is given by

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}|\boldsymbol{y}_i^{obs}, \boldsymbol{M}_i, \boldsymbol{X}_i) = c \cdot \prod_{i=1}^{I} \int f(\boldsymbol{y}_i, \boldsymbol{M}_i|\boldsymbol{X}_i, \boldsymbol{\gamma}, \boldsymbol{\phi}) d\boldsymbol{y}_i^{miss},$$

where γ are the parameters corresponding to $\boldsymbol{y}$, the parameters ϕ correspond to the missingness indicators $\boldsymbol{M}$ and c is a constant value. Here the missing values are integrated out which means that the missing data mechanism is ignored. The likelihood for MAR data is based on the fact that the likelihood contribution of subject i under MAR can be factorized as

$$f(\boldsymbol{y}_i^{obs}, \boldsymbol{M}_i|\boldsymbol{X}_i, \boldsymbol{\gamma}, \boldsymbol{\phi}) = f(\boldsymbol{M}_i|\boldsymbol{y}_i^{obs}, \boldsymbol{X}_i, \boldsymbol{\phi}) \cdot \int f(\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{miss}|\boldsymbol{X}_i, \boldsymbol{\gamma}) d\boldsymbol{y}_i^{miss}.$$

We can factorize the likelihood under MAR in this way since the missingness probability does not depend on the missing data and therefore $f(\boldsymbol{M}_i|\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{miss}, \boldsymbol{X}_i, \boldsymbol{\phi}) = f(\boldsymbol{M}_i|\boldsymbol{y}_i^{obs}, \boldsymbol{X}_i, \boldsymbol{\phi})$. This first factor represents the missing data mechanism MAR. So the mechanism is ignorable since this method is a likelihood method. Therefore the first factor can be left out in the likelihood. The likelihood for MAR data is simplified to

$$\mathcal{L}(\boldsymbol{\gamma}|\boldsymbol{y}_i^{obs}, \boldsymbol{X}_i) = c \cdot \prod_{i=1}^{I} \int f(\boldsymbol{y}_i^{obs}, \boldsymbol{y}_i^{miss}, |\boldsymbol{X}_i, \boldsymbol{\gamma}) d\boldsymbol{y}_i^{miss}.$$

To obtain the ML-estimates for γ, the Likelihood needs to be maximized using normal ML-methods.

The ML-approach can be used also for NMAR data, for example, look at Ibrahim and Molenberghs (2009) who describe this method for non-ignorable missing data. The estimation is rather complex because they have to specify the missing data mechanism and the joint distribution of the data. In the ML-approach there are two possible methods for missing data which are NMAR: selection models and pattern mixture models. For detailed explanation of these please consult Ibrahim and Molenberghs (2009) and the references in there.

Regardless of whether we face MAR or NMAR data, likelihood methods can be seen as imputation methods. For example, if we assume that the missing data mechanism is ignorable, the missing values are imputed[3] using the parameter estimates resulting from the above described observed data likelihood for MAR data. Mostly the Expectation Maximization (EM) algorithm is used for imputation of the missing values. This algorithm for maximum likelihood with incomplete data was proposed by Dempster et al. (1977). It is an iterative process with the aim to find maximum likelihood estimates. We want to describe this algorithm shortly:

Parenthesis Expectation Maximization algorithm: The EM algorithm consists of two steps, one Expectation step and one Maximization step which are basically given by:

1. E-step: filling in missing data with their conditional expectations given the current estimates of parameters and the observed data:

$$Q(\boldsymbol{\theta}) = E_{y^{miss}|y^{obs}}[l(\gamma; \boldsymbol{y}^{obs}, \boldsymbol{y}^{miss}|\boldsymbol{y}, \gamma^{(0)})]$$

2. M-step: estimate complete-data parameters by maximizing the resulting complete data likelihood:

$$\boldsymbol{\gamma}^{(1)} = \underset{\gamma}{\text{argmax}} Q(\boldsymbol{\gamma}).$$

These steps are repeated until convergence of the parameters at step k. So the resulting parameter vector is $\boldsymbol{\gamma}^{(k)}$. If the complete response data $\boldsymbol{y}_i$ are normally distributed these parameters are given by the mean vector and the covariance matrix, $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, of the multivariate normal distribution.

According to Molenberghs et al. (2014) the conditional mean of the missing values given

[3] Imputation means that the missing values are replaced so that the result is a complete dataset without missings.

the observed values and parameters is

$$E(\boldsymbol{y}_i^{miss}|\boldsymbol{y}_i^{obs},\gamma) = \boldsymbol{\mu}_i^{miss} + \Sigma_i^{omiss}\Sigma_i^{o-1}(\boldsymbol{y}_i^{obs} - \mu_i^{obs}),$$

where $\boldsymbol{\mu}_i^{miss}$ corresponds to $\boldsymbol{y}_i^{miss}$, $\boldsymbol{\mu}_i^{obs}, \Sigma_i^{obs}$ to $\boldsymbol{y}_i^{obs}$ and $\Sigma_i^{obsmiss}$ contains the covariances between y_i^{miss} and y_i^{obs}. This expectation would be used to fill in the missing data. Here is still assumed that the data are MAR and therefore we only have to make assumptions on the distribution of the complete response and the model for the complete data, $f(\boldsymbol{y}_i|\boldsymbol{X}_i,\gamma)$, has to be specified correctly.

2. Fully Bayesian:

Another approach for inference in statistics are Fully Bayesian methods. In the Bayesian framework the parameters are no longer fixed values but random variables itself. Moreover if we have beliefs about parameters these beliefs can be consistently updated in the Bayesian framework. Ibrahim and Molenberghs (2009) describe a fully Bayesian method for missing data. For using this methods you have to specify priors for all parameters and the distributions for missing variables. By multiplying these priors and the observed data likelihood, as described in the Maximum Likelihood part above, we obtain the posterior distribution for the parameters. The point estimates can be calculated for example as the posterior mean or mode. Here the missing values will be imputed (Ibrahim et al. 2005), also and therefore normal inference methods can be used after imputing. The values are imputed by sampling from the conditional distribution of the missing data using a Gibbs Sampler [4]. According to Ibrahim et al. (2012) Bayesian methods are the best related to generality and power. In comparison they are easy to implement because only some steps has to be added to the existing Gibbs sampler.

3. Weighted Estimating Equations:

The Weighted Estimating Equations (WEE) approach is based on the Generalized Estimating Equations (GEE) approach which is a semi-parametric method where no full distribution of the joint distribution of a subjects observation is necessary, and an extension of the quasi-likelihood method for generalized linear models to longitudinal data. The covariance does not have to be specified correctly but a working covariance for the repeated measurements vector need to be specified. The WEE corresponds in fact to the GEE method for missing data and has been proposed by Robins et al. (1995). According to Ibrahim et al. (2012) WEE require an explicit missing data mechanism or the specification of the estimating equations for the missing data given the observed data but fewer assumptions on the data model are needed. This method is called weighted EE because the contribution to the estimating equations from a complete observation is weighted by

[4] A Gibbs Sampler can be used if direct sampling is difficult. It is a MCMC algorithm which samples from the conditional distribution.

the inverse probability that the covariate which contains missings is observed (Lipsitz et al. (1999)). This method can be used for MCAR data leading to valid inference, is not hard to compute and the joint distribution does not have to be specified. The resulting estimators are consistent and asymptotically normal under some assumptions (Robins et al. (1995)). But for MAR and NMAR the results are generally biased.

4. Simple Imputation:

In general, imputation means that the missing values are replaced by other values. Imputation methods can be divided in simple/single and multiple imputation. With single imputation all missing value are replaced at once, yielding one complete dataset which can be analysed with standard statistical methods. But with multiple imputation the missings are replaced k times so that we obtain k different complete datasets. First we want to shortly describe some single imputation methods and in the next part we outline multiple imputation methods.

One simple imputation method is the Mean Value Imputation (MV). With MV missing values are replaced by the mean of the observed values of the corresponding variable for a subject. Another possibility is to use a method called "last observation carried forward" (LOCF) where the missing values are replaced by the last observed value of the variable for the respective subject. There are other simple imputation methods as baseline value imputation where the values of the baseline (first observation) are carried forward or worst value imputation where the missing values are replaced by the worst value of the variable for the corresponding subject.

5. Multiple Imputation:

Single imputation does not take the uncertainty of the missing data into account and therefore may produce misleading covariate effects. To avoid this we can produce multiple imputations. The missings are replaced k times yielding to k imputed datasets. These datasets does not contain missing values and standard analysis methods can be used. Each of the k datasets is analysed separately and the results are combined in the end. This method was proposed by Rubin (1987) for missing data in surveys and Blackwell et al. (2011) proposed an algorithm for computing multiple imputed datasets which is part of the R-package Amelia. We will use MI in our simulation so we will describe it in more detail in section 2.7.

2.4 Types of Measurement Error

Another problem in empirical studies may be that variables are measured with errors. Imagine that in the blood pressure example a covariate is the persons weight. This could be measured with error for example since the weighing-machine was not measuring accurately enough. Or it could also be that the response, the blood pressure, was measured with error but we restrict our discriptions here to mismeasured covariates.

Even with this limitation there are still different types of measurement errors. Some we want to describe in this part based on Buonaccorsi (2010). They distinguish between the classical measurement error model and the Berkson error model. The difference is that in the classical model we assume a specific distribution for the mismeasured observed values given the unobserved true values. Whereas in the Berkson model we make an assumption regarding the distribution of the unobserved true values given the observed values. So the assumption is the reverse.

The classical additive measurement error model is given by:

$$\boldsymbol{x}^e = \boldsymbol{x} + \boldsymbol{u} \quad \text{(classical model)}, \tag{1}$$

where $\boldsymbol{x}$ is the real unobserved variable, $\boldsymbol{u}$ is the measurement error and $\boldsymbol{x}^e$ is the observed covariate measured with error. If some observations of $\boldsymbol{x}^e$ are measured correctly, the corresponding values of $\boldsymbol{u}$ are zero. $\boldsymbol{u}$ is independent from $\boldsymbol{x}$ so the measurement error is random. Furthermore there are following conditions for the normal classical additive measurement error:

$$\begin{aligned}
\mathbb{E}(\boldsymbol{x}^e|\boldsymbol{x}) &= x \\
\mathbb{E}(\boldsymbol{u}|\boldsymbol{x}) &= 0 \\
\text{Var}(\boldsymbol{u}|\boldsymbol{x}) &= \sigma^2_{\boldsymbol{u}} \\
\boldsymbol{u} &\sim \mathcal{N}(0, \sigma^2_{\boldsymbol{u}}).
\end{aligned}$$

The first of these conditions describes that the mismeasured covariate is unbiased given the true covariate and the second means that the measurement error is additive because the expected value of the measurement error $\boldsymbol{u}$ given the true covariates is zero. Moreover we assume that the variance of the measurement error does not depend on the subject but is the same for all observations.

By contrast the additive Berkson error model is given by:

$$\boldsymbol{x} = \boldsymbol{x}^B + \boldsymbol{u}_B \quad \text{(Berkson model)},$$

where the unobserved true covariate $\boldsymbol{x}$ is random and the observed mismeasured covariate $\boldsymbol{x}^B$ is fixed. The error $\boldsymbol{u}_B$ has to fulfill similar conditions as the classical error:

$$\begin{aligned}\mathbb{E}(\boldsymbol{u}^B) &= 0\\ \text{Var}(\boldsymbol{u}^B) &= \sigma^2_{\boldsymbol{u}^B}.\end{aligned}$$

We want to give a short example for the Berkson error: A variable in a dataset is the air pollution a person is exposed to but the air pollution is measured at a central place in the residential area or the city. So the observed value is not the exact value for the person. Here the mismeasured value at the measuring station is a fixed values and the true unobserved value is random. According to Buonaccorsi (2010) this type of measurement error can usually be ignored. Since we do not use the Berkson error in our model but the classical additive measurement error we do not want to explain the Berkson error any further.

There are also nonadditive measurement errors which are all measurement errors shich fulfil the following condition: $\mathbb{E}(\boldsymbol{x}^e, \boldsymbol{x}) = g(\boldsymbol{\theta}, \boldsymbol{x}) \neq x$. So with a nonadditive measurement error the mismeasuered covariate is biased. Some examples for nonadditive measurement errors are:

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}^e, \boldsymbol{x}) &= g(\boldsymbol{\theta}, \boldsymbol{x}) = \theta + x \text{ (constant bias)},\\ \mathbb{E}(\boldsymbol{x}^e, \boldsymbol{x}) &= g(\boldsymbol{\theta}, \boldsymbol{x}) = \theta_0 + \theta_1 x \text{ (linear measurement error)},\\ \mathbb{E}(\boldsymbol{x}^e, \boldsymbol{x}) &= g(\boldsymbol{\theta}, \boldsymbol{x}) = \underbrace{g(\boldsymbol{\theta}, x)}_{\text{nonlinear in } \theta} \text{ (nonlinear measurement error)}.\end{aligned}$$

For more complex types of measurement errors we are referring to Buonaccorsi (2010) chapter 6.4. All these measurement error types can influence the analysis of a dataset so correction for measurement errors may be useful to obtain valid results. We want to describe the consequences and some correction methods in the following section.

2.5 Consequences of Measurement Errors and Methods to correct for it

The easiest way of handling a measurement error would be to ignore it. But this can have consequences for the estimates. In this section we want to describe these consequences and methods to correct for measurement error whereby we restrict our explanations to classical additive measurement errors.

2.5.1 Consequences

At first we want to describe the influence of a measurement error on analysis results. Therefore we assume a simple linear model with just one covariate and then substitute the true covariate $\boldsymbol{x}$ which is normally distributed, $\boldsymbol{x} \sim \mathcal{N}(\mu, \sigma)$, by $\boldsymbol{x} = \boldsymbol{x}^e - \boldsymbol{u}$:

$$\begin{aligned}
&\boldsymbol{y} = \beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{\varepsilon} \\
\Leftrightarrow\ &\boldsymbol{y} = \beta_0 + (\boldsymbol{x}^e - \boldsymbol{u})\beta_1 + \boldsymbol{\varepsilon} \\
\Leftrightarrow\ &\boldsymbol{y} = \beta_0 + \boldsymbol{x}^e\beta_1 + \underbrace{(\boldsymbol{\varepsilon} - \beta_1\boldsymbol{u})}_{\boldsymbol{\varepsilon}^*}.
\end{aligned}$$

Here $\boldsymbol{\varepsilon}^*$ is correlated with $\boldsymbol{x}^e$, hence the regression error is correlated. Therefore the estimators will be biased if we ignore the measurement error. Since we assume that only the covariates are measured with error (the response is measured correctly) and that measurement errors are uncorrelated, the naive estimator for the slope is given by

$$\beta_1^{\mathrm{e}} = \left[\frac{\sigma^2}{\sigma^2 + \sigma_{\boldsymbol{u}}^2}\right] \cdot \beta_1 = \kappa\beta_1,$$

where $\kappa = \frac{\sigma^2}{\sigma^2+\sigma_{\boldsymbol{u}}^2}$ is the reliability ratio and β_1^{e} is the biased estimator for β_1. The reliability ratio describes the part of the variance of the mismeasured variable which is not explained by the measurement error but which is explained by the variance of the true unobserved variable. The reliability ratio is lower than one if $\sigma_{\boldsymbol{u}^2} > 0$. Then $|\beta_1| > |\beta_1^{\mathrm{e}}|$ and therefore the naive estimator β^{e} is biased towards zero. This effect is called attenuation.

Consequence with Measurement Error: Attenuation

The naive estimator for the slope is biased towards zero.

Although the naive estimator is biased, the naive t-test for $\beta_1 = 0$ is correct in our scenario since $\beta_1^{\mathrm{e}} = 0$ is only possible if $\beta_1 = 0$. Under this conditions the t-test is not correct if the error is nonadditive or if there is also a measurement error in $\boldsymbol{y}$ which is correlated with the measurement error in $\boldsymbol{x}$. Often the interest lies also in the estimator for expected value of $\boldsymbol{y}$ for a specific value $\boldsymbol{x}^*$. The naive estimator for this expected value is biased also. Without measurement error in the response the bias of the estimated response is $\beta_1(\kappa - 1)(\boldsymbol{x}^* - \mu)$. This bias is zero if $\beta_1 = 0$ or if the covariate value for which the response was calculated equals the mean of the covariate ($\boldsymbol{x}^* = \mu$) (under the assumption that $\boldsymbol{x}$ is mismeasured, so that $\kappa < 1$). The expected value is the same if there is a measurement error in the response which is not correlated with the covariates measurement error. Otherwise, if the measurement error of the response and covariate are correlated, the bias is no longer given by the above expression and much more complicated.

If we want to obtain unbiased results we have to correct for the measurement error. There are several methods to do so which require different assumptions or additional data. For example we need to know the reliability ratio (or make an assumption about it), the measurement error itself or the ratio of measurement variances (Fuller, 1987).

In the previous part of this consequences section we only considered simple regression models but in reality most regression models contain more than one covariate. So now we want to describe the consequences for such multiple regression models, still relying on Buanoccorsi (2010). We also still assume that only covariates are measured with error and the response is measured correctly. The mismeasured variables are now given by $\boldsymbol{x}^e_{it} = \boldsymbol{x}_{it} + \boldsymbol{u}_{it}$ but there are also other correctly measured variables in the model. The naive estimators for the coefficients of the covariates $\hat{\boldsymbol{\beta}}_{c,naive}$ and the intercept $\hat{\boldsymbol{\beta}}_{0naive}$ are given by

$$\hat{\boldsymbol{\beta}}_{c,naive} = \boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e}{}^{-1}\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}}, \hat{\beta}_{0naive} = \bar{y} - \hat{\boldsymbol{\beta}}_{c,naive}{}'\bar{\boldsymbol{X}}^e$$

where

$$\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e} = \frac{\sum_{it}(\boldsymbol{x}^e_{it} - \bar{\boldsymbol{x}}^e)(\boldsymbol{x}^e_{it} - \bar{\boldsymbol{x}}^e)'}{n-1}, \boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}} = \frac{\sum_{it}(\boldsymbol{x}^e_{it} - \bar{\boldsymbol{x}}^e)(y_{it} - \bar{y})'}{n-1},$$

with $\boldsymbol{x}^e_{it} = (x^e_{1it}, ..., x^e_{(p-1)it})'$, $\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e}$ a $P \times P$ matrix and $\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}}$ a $P \times 1$ vector. An important consequence of measurement errors in multiple regression models is:

If one variable is mismeasured, not only its estimator of the coefficient is biased but often the estimators of the coefficients of all other covariates are biased, too.

We want to explain this phenomenon with the help of an example from Buonaccorsi (2010). We only adjust the notation.

We have a model with two covariates $\boldsymbol{x}_1, \boldsymbol{x}_2$. $\boldsymbol{x}_1$ is not observed directly but only in a mismeasured form $\boldsymbol{x}^e_1$. The second covariate (and by assumption the response) are measured correctly and the measurement error variance for $\boldsymbol{x}^e_1$ is denoted by σ^2_u since we assume that the measurement error variance is the same for all observations. The measurement error covariance matrix is given by

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} \sigma^2_u & 0 \\ 0 & 0 \end{pmatrix},$$

since $\boldsymbol{x}_2$ is measured correctly which induces that the measurement error variance of $\boldsymbol{x}_2$ and the covariances of measurement errors are zero. The covariance matrix of the true observed covariate vector $\boldsymbol{X}$ is

$$\boldsymbol{\Sigma}_{XX} = \begin{pmatrix} \sigma^2_1 & \sigma^2_{12} \\ \sigma^2_{12} & \sigma^2_2 \end{pmatrix}$$

Then the expected value of the naive extimator for the mismeasured covariate is given by

$$E(\hat{\beta}_{1naive}) \approx \left(\frac{\sigma_2^2 \sigma_1^2 - \sigma_{12}^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right) \beta_1$$

and the biases in the naive estimators of β_1 and β_2 are

$$\begin{aligned} Bias(\hat{\beta}_{1naive}) = E(\hat{\beta}_{1naive}) - \beta_1 &= \left(\frac{-\sigma_2^2 \sigma_u^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right) \beta_1 \\ Bias(\hat{\beta}_{2naive}) &= \left(\frac{\sigma_{12}^2 \sigma_u^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right) \beta_1 . \end{aligned}$$

The bias in the naive estimator of β_2 does not depend on β_2 itself but on β_1 and it is zero if the covariates are not correlated (σ_{12}). In the case that the covariates are uncorrelated, the naive estimator for β_1 is the same as in the simple linear model ($\hat{\beta}_{naive} = [\sigma_1^2/(\sigma_1^2 + \sigma_u^2)]$). With a considerable measurement error the consequences should not be ignored because the biases can be very large. So it is necessary to correct for measurement error if we want to have good estimators in terms of bias. Buonaccorsi (2010) presents some methods for correcting which we want to summarize here but we restrict this to measurement errors in the covariates and the same measurement error covariances for each observation. They assume that these covariances can be estimated and the estimates are given by $\hat{\mathbf{\Sigma}}_u$. Estimation methods for the measurement error variance are describe in section 2.6. But at first we summarize five methods to correct for measurement error:

2.5.2 Moment Corrected Estimators

The first approach which can correct for measurement errors are moment corrected estimators. This is a very simple method since it is based on the expected value of the sample covariance $E(\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e}) = \hat{\mathbf{\Sigma}}_{XX} + \hat{\mathbf{\Sigma}}_u$. For the correction the naive estimator of the covariate coefficients, $\hat{\beta}_{c,naive} = \boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e}{}^{-1}\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}}$, is modified by subtracting the estimated measurement error covariance from the sample covariance:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_c &= (\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e} - \hat{\mathbf{\Sigma}}_u)^{-1}\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}} = \hat{\mathbf{\Sigma}}_{XX}^{-1}\hat{\mathbf{\Sigma}}_{XY} \\ &= (\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e} - \hat{\mathbf{\Sigma}}_u)^{-1}\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{X}^e}\boldsymbol{\beta}_{c,naive} = \hat{\boldsymbol{\kappa}}^{-1}\boldsymbol{\beta}_{c,naive} \\ \hat{\beta}_0 &= \bar{y} - \hat{\boldsymbol{\beta}}_c{}'\bar{\boldsymbol{X}}^e, \end{aligned}$$

where $\hat{\boldsymbol{\kappa}} = (\hat{\mathbf{\Sigma}}_{XX} + \hat{\mathbf{\Sigma}}_u)^{-1}\hat{\mathbf{\Sigma}}_{XX}$ is the estimated reliability ratio. We assume that there is no error in the response. Hence the sample covariance $\boldsymbol{S}_{\boldsymbol{X}^e\boldsymbol{Y}}$ equals the estimated covariance between the true covariates and the response $\hat{\mathbf{\Sigma}}_{XY}$ because there cannot be a correlation between the measurement error variance.

2.5.3 Regression calibration

Another method is regression calibration as described by Carroll et al. (1995). Here the mismeasured values are replaced by estimates from the regression of the unobserved part of $\boldsymbol{X}$ (the true values of the mismeasured covariates) on X^e and the covariates which are measured correctly. Then standard analysis methods can be used and the resulting standard errors have to be adjusted because otherwise they would ignore that for the regression of the unobserved part of $\boldsymbol{X}$ parameters had to be estimated, too.

2.5.4 SIMEX

Buonaccorsi (2010) also explain the SIMEX (simulation extraploation) approach. This method consists of two steps: simulation and extrapolation. In general, the first step simulates what happens if the measurement error would be larger and in the second step it extrapolates back to no measurement error. Buonaccorsi defines [5]:

$$\beta_p(\lambda) = E(\beta_{p,naive}) \text{ if } Cov(\boldsymbol{u}) = (1+\lambda)\boldsymbol{\Sigma}_u.$$

With the help of $\lambda > 0 \in (\lambda_1, ..., \lambda_M)$ we can control for the amount of measurement error covariance. In the SIMEX algorithm, the first step is to calculate the actualized value of the mismeasured covariates $\boldsymbol{X}^e_b(\lambda_m)$ so that the covariance of the new measurement error $\boldsymbol{u}$ is $(1+\lambda)\boldsymbol{\Sigma}_u = \boldsymbol{\Sigma}_u + \lambda\boldsymbol{\Sigma}_u$ with $b = 1, ..., B$ denoting the number of the step and B is a large number. By assumption $\boldsymbol{u}_b$ is i.i.d. with mean zero and covariance $\hat{\boldsymbol{\Sigma}}_u$. So it follows that

$$\boldsymbol{X}^e_b(\lambda_m) = \boldsymbol{X} + \underbrace{\boldsymbol{u}_b + \lambda^{1/2}\boldsymbol{u}_b}_{\boldsymbol{u}} = \boldsymbol{X}^e + \lambda^{1/2}\boldsymbol{u}_b.$$

So $cov(\boldsymbol{u}) = cov(\boldsymbol{X}^e) + cov(\lambda^{1/2}\boldsymbol{u}_b) = (1+\lambda)\boldsymbol{\Sigma}_u$ if $\boldsymbol{\Sigma}_u = \hat{\boldsymbol{\Sigma}}_u$. This was the Simulation step. Then the coefficients are naively estimated for each b and the mean over all B coefficient vectors is calculated and denoted by $\bar{\boldsymbol{\lambda}}_m$. Next for each covariate p a model $g_p(\lambda)$ is fitted as a function of λ_m for the p-th component of $\bar{\boldsymbol{\lambda}}_m$. In the last step the SIMEX estimator for each p is calculated:

$$\hat{\beta}_{p,SIMEX} = g_p(-1),$$

[5] We modified the definition a little bit because we assume that the measurement error is the same for all observations and therefore it does not depend on i.

because so $cov(\boldsymbol{u}) = (1+(-1))\boldsymbol{\Sigma}_u = 0$ and hence there is no longer a measurement error in the model. According to Blackwell et al. (2015a) SIMEX is not so good if there are multiple mismeasured variables because the computation is harder and the results depend more on the extrapolation model.

2.5.5 Likelihood Methods

Likelihood methods where already describe as a methods which can be used with missing data but likelihood methods can correct for measurement error, too. To compute the likelihoods for data with measurement error the model for $\boldsymbol{y}$ given $\boldsymbol{x}$, the model for $\boldsymbol{X}$ in the structural setting [6] and the measurement error model or some of these models need to be specified. The likelihood with no error in the response for the case that the true $\boldsymbol{X}$ is a random variable with parameters $\boldsymbol{\theta}$ is given by

$$\begin{aligned}\mathcal{L}_{YX^e}(\boldsymbol{\theta}_y, \boldsymbol{\omega}, \boldsymbol{\theta}) &= \prod_{it} f(y_{it}, x_{it}^e; \boldsymbol{\theta}_y, \boldsymbol{\omega}, \boldsymbol{\theta}) \\ &= \prod_{it} \int_x f(y_{it}|\boldsymbol{x}; \boldsymbol{\theta}_{y,it}) f(\boldsymbol{x}_{it}^e|\boldsymbol{x}; \boldsymbol{\omega}) f_{X_{it}}(\boldsymbol{x}; \boldsymbol{\theta}) dx,\end{aligned}$$

where $\boldsymbol{\omega}$ are the parameters of the measurement error model and $\boldsymbol{\theta}_y = (\boldsymbol{\beta}, \boldsymbol{\sigma}_y)$ are the parameters of the model for $\boldsymbol{Y}|\boldsymbol{x}$. With this likelihood we can calculate ML-estimates but therefore we often need to know some parameters or make assumptions about them. Sometimes all parameters are identifiable without assumptions. To solve the problem of unknown parameters, additional data could be used, for example external validation data, internal validation data or replicate data.

External validation data describe a dataset containing true values for the mismeasured variables. This dataset is independent from the main dataset. With the help of external validation data, pseudo-maximum likelihood estimators can be calculated. At first, the measurement error parameters from the external data are estimated. These estimated measurement error parameters are used to compute the pseudo-MLEs for the observed data likelihood.

There is also the possibility to use internal validation data. If some values of a mismeasured variable are measured correctly and therefore represent the true values, a subset containing these true observed values can be created and this is called internal validation data. We assume here that we do not have longitudinal data so there is no time indices necessary and we have I observations. If there is no error in the response , with internal

[6] In the structural setting the $\boldsymbol{X}_i$ are random.

validation data the full likelihood is given by

$$\mathcal{L}_{Int} = \prod_{i=1}^{I_V} f(\boldsymbol{y}_i, \boldsymbol{x}_i^e, \boldsymbol{x}_i) \prod_{I_V+1}^{I} f(\boldsymbol{y}_i, \boldsymbol{x}_i^e),$$

where $i \in (1, ..., I_V)$ are the observations for which the true values for the mismeasured variable are observed. So for $I - I_V$ observations the true values are not observed or in other words the true values are missing (MAR according to Buanaccorsi et al. (2010)). Missing data methods could be used (see 2.3) .

Another way would be to use likelihood methods with replicate data. Replicate data means that for every observation there are several additional measurements of the mismeasured variable. These replicate values are subject to measurement error, too. The measurement error can differ between the different replicate datasets. In the likelihood the density of the mismeasured variable is replaced by the density of the replicates:

$$\mathcal{L}_{Rep}(\boldsymbol{\theta}_y, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \prod_{it} \int_x f(y_{it}|\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\sigma}_y) f(\boldsymbol{x}_{it,rep}^e|\boldsymbol{x}; \boldsymbol{\omega}) f_{X_{it}}(\boldsymbol{x}; \boldsymbol{\theta}) dx.$$

According to Buonaccorsi (2010) a problem of all likelihood methods is that the implementation is often complicated but they yield good results when the assumptions are fulfilled.

2.5.6 Modified Estimating Equations

Estimates which are corrected for measurement error can also be obtained using modified estimating equations. The resulting estimators are consistent if some regulatory conditions are fulfilled. Buonaccorsi (2010) describes the modified estimating equations (MEE) approach. At first the measurement error parameters are estimated. Second the naive estimating equations are modified by using the estimates from the first step. Then the MEE estimates have to be computed iteratively using standard methods for data without measurement error.

Often we need to know the measurement error variance to use specific methods. Since we do not know this variance in reality we need to make assumptions about it or estimate it. In the next section we want to describe a possible method to estimate the measurement error variance.

2.6 Estimating the Measurement Error Variance

For many methods that correct for measurement error we need to specify the variance of the measurement error but mostly this is not known in reality. If it is misspecified the estimators could be biased as we see in the sensitivity analysis of this work. So it is desirable to obtain best possible estimators for this variance. Buonaccorsi (2010) describes a way to estimate the measurement error variance using replicate data (shortly described in the previous section). It is assumed that these replicates are taken on units in the main study and that we do not have longitudinal data so we only have one observation per subject and no time indices is needed. Here we assume that different observations can have different measurement error variances. The R replicate values for covariate p and observation i are given by $X^e_{i1j}, ..., X^e_{iR_{ij}p}$ where R_{ij} is the number of replicate values for covariate p and observation i and

$$X^e_{irp}|x_{ip} = x_{ip} + u_{irp}.$$

Analogously to 1 the true values of observation i and covariate p are denoted by x_{ip} and the measurement error u_{irp} has mean zero and the variance is given by $\sigma^2_{uip(1)}$ (per-replicate variance). We assume here that the measurement errors of two variables are not correlated (interested readers can consult Buonaccorsi (2010)). Assuming that there is more than one replicate per observation and variable ($R_{ip} > 1$), the per-replicate variance can be estimated using

$$\hat{\sigma}^2_{uip(1)} = S^2_{ip} = \frac{1}{R_{ip} - 1} \sum_r (X^e_{irp} - \bar{X}^e_{ip})^2,$$

where $\bar{X}^e_{ip} = \sum_r X^e_{irp}/R_{ip}$. The estimated variance of u_{ip} is obtained by $\hat{\sigma}^2_{uip} = \hat{\sigma}^2_{uip(1)}/R_{ip}$. This estimated measurement error variance can be used to correct for measurement error with different methods. For example we could use this estimated measurement error variance to compute $\hat{\boldsymbol{\kappa}} = (\hat{\boldsymbol{\Sigma}}_{XX} + \hat{\boldsymbol{\Sigma}}_u)^{-1}\hat{\boldsymbol{\Sigma}}_{XX}$ for a moment corrected estimator.
In the previous sections we described missing data methods and methods to correct for measurement errors separately. But in reality we often face data which contain missing data and mismeasured variables. In the following section we want to outline some methods which can be used for this combined problems.

2.7 Longitudinal Data with Missing Data and Measurement Error

After this explanation of missing data and measurement errors and how we can deal with it if they occur separately, we now look at the combined problems of missing data and measurement error but only for longitudinal data. There can be different combinations of missing data and measurement error. For example only covariates or only the response could be mismeasured and missing or the response is mismeasured and some covariates contain missing values or both are subject to missingness and measurement error. To our knowledge the most examined combination of the problems is a measurement error in the covariates and a missing response or drop-out. There are different methods used in the literature. Some methods are outlined in this chapter. One approach uses SIMEX, one a joint model with MCEM for inference, another one uses approximate likelihoods and MCEM, one uses estimating equations based on moments and the last described method is multiple overimputation. We start with the SIMEX based approach.

Yi (2008) takes a look at response data subject to dropout which are missing at random and mismeasured covariates in a logistic regression model. They use two inference methods. The first method, inverse probability weighted GEE, is used often if data are incomplete. With measurement error inverse probability weighted GEE yields biased estimates. The other one is SIMEX (Cook and Stefanski (1995)) which contains a simulation step, an estimation step and an extrapolation step. Yi (2008) explains the idea of SIMEX as follows: Instead of replacing the true $\boldsymbol{x}$ by $\boldsymbol{x} + \boldsymbol{u}$ (as in the normal SIMEX approach, see 2.5.4), it is replaced by $\boldsymbol{x} + \boldsymbol{u} + \lambda\sigma_{\boldsymbol{u}}\boldsymbol{u}_s$ where $\boldsymbol{u}_s \sim \mathcal{N}(0,1)$. The result is that the estimator for β_1 converges in probability to the true value of β_1 if $\lambda = -1$. Comparing the SIMEX approach in a simulation study to the naive method of ignoring the measurement error shows that the SIMEX approach performs much better in terms of bias. Since this approach is functional, it is not necessary to model the distribution of the mismeasured covariates.

Yi et al. (2011) examine longitudinal data with covariate measurement error and missing response, too. But they use a generalized linear mixed model (GLMM). For simultaneous inference they use a joint model method with MCEM (Monte Carlo EM-Algorithm) because the complete data log-likelihood is very complex. They conduct a simulation study where they compare a naive method (Complete Case with ignoring measurements errors) and their proposed method. The latter performs very good with different degree of measurement error (small bias). Whereas the naive approach leads to large biases if there is a measurement error and missing data. Yi et al. also compared different missing data mechanisms and describe the results for their methods with and without measurement error. They also conducted a bias analysis which we do not want to explain further.

Liu and Wu (2007) use a similar approach: They use two approximate likelihood methods but for sempiparametric NLME- (nonlinear mixed effect) models with mismeasured time-varying covariates and missing response. Both are implemented by a MCEM- algorithm combined with a Gibbs sampler. They assume that the missing data process is nonignorable. The first method uses a joint model with MCEM and Gibbs sampler to compute the approximate maximum likelihood estimates simultaneously. The second method is an approximate method which uses Taylor expansion. The first method is more accurate than the second one but computationally very intensive. So they propose to use the second method if the first one is too complex or to calculate starting values for the first method. Both have performed much better than the naive approach (ignoring missings and measurement errors) in a simulation study.

Another method for dealing with missing response and time-varying covariates measured with error is proposed by Yi et al. (2012). They are using estimating equations based on moments. Benefits of this method are that only a few assumptions are necessary. No assumptions about the distribution of mismeasured covariates are necessary because Yi et al. (2012) use a functional measurement error method. Moreover they do not need to specify the distribution of the complete data since the model is marginal. Their approach consists of three steps. In the first step the marginal moments for the error process are used to correct for the measurement errors. Then inverse probability weighting is used to adjust for the missing response. The results of the first two steps are time-specific estimating equations for each subject which are corrected for measurement error and missing response. These estimating equations are combined to one efficient estimating equation by the generalized method of moments (Hansen (1982)). Yi et al. (2012) compare their proposed approach to three naive ones (one that ignores covariate measurement error and missing responses, one that ignores only measurement error and one that ignores only missing responses) with the result that the naive methods not correcting for measurement error and/or missing response perform badly regarding the bias. In contrast, the method of Yi et al. (2012) performs better in terms of bias and the proposed method exhibits a low bias and a high coverage rate (close to the nominal 95%) [7].

To our knowledge, none of the above described methods was modified in a way so that it can be used for missing covariates, a response subject to drop-out and covariate measurement error. But this combination is a realistic scenario. A method which is not restricted to specific combination of missing data and measurement error is multiple overimputation which is an extension of multiple imputation. With multiple overimputation, the missing values are filled in (as with multiple imputation, see last part of 2.3) and the mismeasured values are overwritten (overimputed). To our knowledge, it was not examined yet if this method can be used for the combined problems of missing covariates, response drop-out and covariate measurement error if the data are longitudinal. We want to examine if multiple overimputation can be used for this problem of incomplete longitudinal data. We conduct a simulation study given in section 5 and the corresponding model is described in section 4. But at first we want to explain multiple impuatation and its extension multiple overimputation.

[7] This means that the corresponding confidence interval covers the true value in 95 out of 100 cases.

3 Multiple (Over-)Imputation

Multiple Imputation (Rubin (1987)) is a method to correct for missing data where missing values are filled in multiple times. The results are several datasets without missings so that standard statistical methods can be used. We want to describe the multiple imputation algorithm used in the R-package Amelia II (Honaker et al. (2011), Honaker and King (2010)). With a later version of Amelia II (relying on Blackwell et al. (2015 a,b)) it is also possible to correct for measurement error using multiple overimputation which we want to present in the second part of this section.

3.1 Multiple Imputation (MI)

For a better understanding of the Amelia package for multiple imputation (MI), we want to outline what multiple imputation is in general. Therefore we rely on King et al. (2001). As already described in the last part of section 2.3, multiple imputation means that a missing value is filled in (imputed) k-times. Hence we obtain k datasets without missings in which the observed values are the same for each dataset. The filled in values can differ from dataset to dataset. If the prediction is good, the variance across these filled in values between the datasets is small and the variance is large if the prediction is not good. An explanation is that if the information used for imputation is more clearly, it leads to similar filled-in values in the k imputations. The resulting k datasets can be analysed by normal statistical procedures. And the k analysis results can be combined by simple methods. Imagine you estimate a regression coefficient β_l for each dataset $l \in \{1, 2, ..., k\}$, the overall point estimate is given by

$$\bar{\beta} = \frac{1}{k}\sum_{l=1}^{k}\beta_l. \tag{2}$$

Similarly we can combine estimates for other quantities of interest, such as univariate mean, predicted probability or first differences in the datasets. The variance of the MI point estimate of a coefficient can be calculated by:

$$Var(\beta) = \frac{1}{k}\sum_{l=1}^{k}Var(\beta_l) + \underbrace{\frac{1}{k-1}\sum_{l=1}^{k}(\beta_l - \bar{\beta})}_{\text{sample variance across the } k \text{ regression coefficients}} .$$

Now we want to explain the MI algorithm used in the Amelia package. Therefore we rely on the explanations in King et al. (2001), Honaker and King (2010) and Honaker et al. (2011). The first important assumption to use the `amelia()` function for imputation is that the missing data are MAR. The MAR assumption can be encouraged by including more informative variables - even variables which are not part of the model of interest. Because so there is more information which can be used to draw the imputations. This additional informative variables could be variables measured after the treatment, variables which are endogenously determined or measures of the same quantity as others. This inclusion of additional variables is possible since imputation models are not used for causal explanation or parameter estimates but for predictions of the distribution of each of the missing values.

The next assumption is that the variables are multivariate normally distributed. Transformations can be used to make this assumption more reliable. Moreover for example Schafer (1997) and Schafer and Olsen (1998) have shown that this approximated model (normal distribution) often works as well as more complicated models for categorical or mixed data.

To explain the used algorithm in the amelia package, we need do describe an imputation model (King et al. (2001)). We define a $(1 \times (P+1))$ vector $\boldsymbol{Z}_{it}$ which contains the values of the covariates and the response for subject $i \in 1, 2, ..., I$ at time $t \in 1, 2, ..., T$. If there are no missing values, the $((I \cdot T) \times (P+1))$ matrix $\boldsymbol{Z}$ is multivariate normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The covariances (off-diagonal elements of Σ) allow the variables to be correlated. As an example, if one covariate is a persons weight at the baseline and another is a persons height, both covariates are positively correlated since it may be plausible that on average the weight increases with the height. The complete data Likelihood is given by

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma | \boldsymbol{Z}) \propto \prod_{i,t} p(\boldsymbol{Z}_{it} | \boldsymbol{\mu}, \Sigma) \tag{3}$$

Then the observed data likelihood if data is given by

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma | \boldsymbol{Z}^{obs}) \propto \prod_{i=1}^{I} \prod_{t=1}^{T} p(\boldsymbol{Z}_{it}^{obs} | \boldsymbol{\mu}_{it}^{obs}, \boldsymbol{\Sigma}_{it}^{obs}), \tag{4}$$

where $\boldsymbol{Z}_{it}^{obs}$ are the observed elements of $\boldsymbol{Z}_{it}$. $\boldsymbol{\mu}_{it}^{obs}$ is the corresponding subvector of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{it}^{obs}$ is the corresponding submatrix of Σ. The subvector and submatrix do not vary over i and t since we assume that the mean and covariance are the same for all observations. An imputation for subject i at time point t for a variable (covariates and response) $p_y \in \{1, ..., (P+1)\}$ can be calculated by

$$\tilde{Z}_{itp_y} = \boldsymbol{Z}_{it,-p_y} \tilde{\beta} + \tilde{\epsilon}_{it}, \tag{5}$$

where $\tilde{Z}_{itp_y}$ is a simulated value for variable p_y and subject i at time point t. β is the coefficient from a regression of $\boldsymbol{Z}_{p_y}$ on all other variables $\boldsymbol{Z}_{it,-p_y}$ in the $(I \cdot T \times (P+1))$ dataset $\boldsymbol{Z}$. ϵ_{it} denotes the fundamental uncertainty. Additional uncertainty is obtained by estimating the coefficient because we do not know the exact value β. The $\sim$-sign means that the corresponding value is a random draw from an appropriate posteriori. The generated imputations are continuous values. If the imputed values are categorical there are two possibilities: The first is to round the value to the nearest categorical value (Schafer 1997). The better method would be to draw from a suitable discrete distribution with mean equal to that of the normal distribution. In the `amelia()` function values are imputed k times. Therefore it is a multiple imputation.

But why do we need to impute k-times? MI has three major advantages over single imputation (Rubin (1987)). The first is that the efficiency of the estimation is higher when using MI if the imputations are random draws which shall represent the distribution of the data. Due to the missing values there is additional variability in the model which should not be ignored. The second advantage of MI is that this variability is displayed by combining the results of the k imputations. Moreover if we have several models for which repeated imputations are drawn randomly, we can use complete data methods to do sensitivity inference (third advantage). The disadvantages are actually not worth mentioning. Obviously MI is more costly than single imputation regarding the imputation itself and the analysis. Moreover, more memory is needed since we have to store several datasets. But if k is not to large this is not a problem especially not nowadays where we have huge servers with much more memory than in the 1980s. Besides that, Rubin (1987) outlines that $k = 5$ or $k = 10$ would be sufficient. Considering that, the needed memory is not that much. For more details please consult Rubin (1987). A nice description of the disadvantage of single imputation is given by Honaker and King (2010). They said that single imputation "cause[s] statistical analysis software to think the data have more observations than were actually observed and to exaggerate the confidence you have in your results by biasing standard errors and confidence intervals"(Honaker and King (2010), p. 563).

A problem regarding implementing the described imputation model is that generating random draws from the posterior of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is computationally very difficult. There are two possible methods to solve this problem: IP (Imputation-Posterior) and EM (Expectation Maximization) algorithm (see 2.3) whereas King at al. (2001) recommend the EMis algorithm (EM with importance sampling). EMis is faster than IP and it draws from the same posterior $P(Z_{mis}|Z_{obs})$ (multivariate normal observed data posterior). Please consult King et al. (2001) for detailed explanation of these algorithms.

Honaker and King (2010) propose to combine EM with Bootstrapping (EMB). This algorithm is used in the Amelia R-package. Since we use this package in our simlation study, we want to describe EMB and the MI algorithm in Amelia more accurately. Here we rely our explanations on Honaker and King (2011). Two assumptions are needed for using Amelia:

$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \text{ (multivariate normal distribution)} \tag{6}$$

$$p(\boldsymbol{M}|\boldsymbol{Z}) = p(\boldsymbol{M}|\boldsymbol{Z}_{obs}) \text{ (MAR assumption)}, \tag{7}$$

where $\boldsymbol{M}$ is a $(I \cdot T \times (P+1))$ missingness matrix. A cellvalue of $\boldsymbol{M}$ is one if the corresponding cellvalue of $\boldsymbol{Z}$ is missing and zero if it is observed. The first assumption describes that the dataset $\boldsymbol{Z}$ is multivariate normally distributed. The second assumption means that the data have to be at least MAR. Since MCAR is a more restrictive assumption, it would fulfil this assumption, too. To take draws from the posterior of the complete data parameters $\boldsymbol{\theta}$ using EMB, we need to specify this posterior. Therefore we first need the observed data likelihood $p(\boldsymbol{Z}^{obs}, \boldsymbol{M}|\boldsymbol{\theta})$. To specify the observed data likelihood we use the MAR assumption and that $\boldsymbol{M}$ is not dependent on the complete data parameters (Honaker et al. (2011)). The complete data likelihood is given by:

$$p(\boldsymbol{Z}^{obs}, \boldsymbol{M}|\boldsymbol{\theta}) = p(\boldsymbol{M}|\boldsymbol{Z}^{obs})p(\boldsymbol{Z}^{obs}|\boldsymbol{\theta}). \tag{8}$$

This leads to the likelihood in equation 4. Since we are only interested in the inference on complete data parameters, we cross the other part out:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{Z}^{obs}) \propto p(\boldsymbol{Z}^{obs}|\boldsymbol{\theta}) \tag{9}$$

By using the law of iterated expectations the likelihood can also be expressed by

$$p(\boldsymbol{Z}^{obs}|\boldsymbol{\theta}) = \int p(\boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z}^{mis}. \tag{10}$$

So we integrate out the missing values. Now the posterior is given by

$$p(\boldsymbol{\theta}|\boldsymbol{Z}_{obs}) \propto p(\boldsymbol{Z}^{obs}|\boldsymbol{\theta}) = \int p(\boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z}^{mis}. \tag{11}$$

Hence the posterior is proportional to the observed data likelihood (9). Draws from this posterior can be created using the EMB algorithm:

1. create $(I \cdot T \times (P+1))$ bootstrap sample of the data with replacement
2. use EM to create point estimates of $\boldsymbol{\mu}$ and Σ
3. impute missing values using the estimates and the original sample units

We already described the EM algorithm in the Maximum Likelihood part of section 2.3. The EMB algorithm is repeated k times. So we obtain k imputed datasets without missing values. Practically this is what `amelia()` does. The output when using `amelia()` contains several amelia-objects whereas the most important for analysis are the imputations. Moreover the output contains the number of imputed datasets k, the missing Matrix M, a $(P+1)+1 \times (P+1)+1 \times k$ array which contains the converged parameters for each of the k EM algorithms (see 2. in the above algorithm). The posterior modes for the k EM chains $\boldsymbol{\mu}$ $((P+1) \times k)$ and $\boldsymbol{\Sigma}$ $((P+1) \times (P+1) \times k)$ are given, too and some other objects (see the document "Package Amelia").

Now we can use every standard statistical method to analyse the datasets and combine the k results for example as described in 2. Another possibility is to use the `zelig()` function from the R-package Zelig (Imai et al. (2007,2008)). We can pass the amelia output directly to this function just like a normal dataset to a regression function. We also have to specify the desired model and formula. There are many different options for models which can be find in the vignettes "Zelig Core Model Reference Manual". To name just a few: least squares, gamma, negative binomial, different gee models and several bayesian regression models. Models for multilevel data are included in the R-package ZeligMultilevel (**reference?!**), for example mixed effects linear (LME) regression which we will use in our simulation. The `zelig()` function automatically combines the results of k multiple imputed datasets. The normal `summary()` function does not print the combined results if a multiply imputed dataset with a LME model is used. Some modifications are necessary to make it work. Alternatively we can use the normal function but this leads to seperate regression outputs for each imputed dataset so we have to combine the results ourselves by using equation 2 for the coefficients. Other parameter can be calculated analogously.

Multiple Imputation only corrects for missing data but not for measurement error. Therefore Blackwell et al.(2015 a,b) extend their MI algorithm and call the resulting approach Multiple Overimputation (MO) which can correct for measurement error, too. We want to describe this approach in the next section.

3.2 Multiple Overimputation (MO)

Multiple overimputation is a method which can correct for missing and mismeasured data simultaneously. It was proposed by Blackwell et al. (2015a,b) as an extension of multiple imputation to mismeasured data. In the MO approach missing data are imputed as with MI and mismeasured values are overwritten, or in terms of Blackwell at al., overimputed. Other methods for longitudinal missing data and mismeasured variables are very complicated because we cannot use standard statistical methods. But with this approach it is possible to use statistical analysis methods for compelete data which makes it really easy to use. In this section we rely on explanations in Blackwell et al. (2015a,b). Easier than using MO would be to treat mismeasured values as unobserved values. Then the mismeasured values can be deleted from the dataset and MI can be used to replace the missing values. But Blackwell et al. (2015a,b) think that even if the values are mismeasured, they still are informative so we should not handle them as if they were missing. This additional information from the mismeasured variables can be used to improve the imputation of the missings and mismeasured values and therefore to increase the efficiency of estimates. MO takes this additional information into account. To overimpute a mismeasured value two sources of information can be used. The first is similar to MI: The mismeasured value is treated as missing and the information of all other observation is used. Additionally the information from the mismeasured value itself can be used.

Blackwell et al. (2015a) describe that missing data can be interpreted as special, enormous measurement errors. We want to illustrate this by using an example from Blackwell et al. (2015a). Assume that a mismeasured variable $\boldsymbol{x}^e$ is given by equation 1, so the variable is subject to a classical additive measurement error. Furthermore it is assumed that the measurement error $\boldsymbol{u}$ given the latent variable $\boldsymbol{x}$ is normally distributed with mean zero and variance σ_u^2. σ_u^2 denotes the measurement error variance which can take values

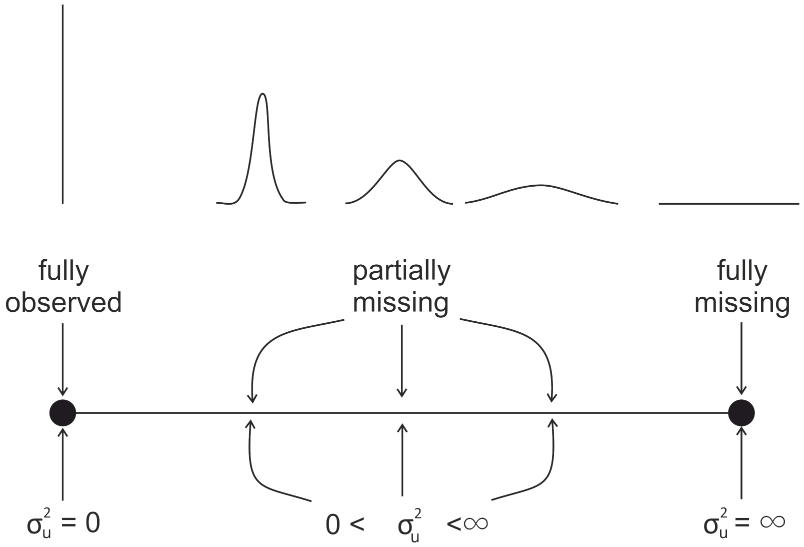


Figure 1: The continuum of measurement error with observation-level priors illustrated in top row.[9]

between zero and infinity. In figure 1 these values between zero and infinity are pictured as a line since all values in between are possible. On the left end the measurement error variance σ_u^2 is zero and on the right side it is infinite so the measurement error variance is

[9] This graphic was copied from Blackwell et al. (2015a).

increasing from left to right. If this variance is zero it means that the variable is measured correctly so there is no uncertainty about the location of the true value. Above the line, which represents the measurement error variance, the distributions of the unobserved true values are drawn. If there is no measurement error this distribution is a vertical line since we know the exact location of the true value. If the measurement error is greater than zero but still small, the distribution of the unobserved true values is a tight normal distribution. With increasing measurement error variance the distribution gets heavier tails so that it is a flat line if the measurement error variance goes to infinity. In the case of an infinite measurement error variance we know nothing about the location of the true value since the observed value is entirely uninformative. So if the measurement error variance goes to infinity, the value can be interpreted as missing. The values between the two extremes are partially missing since we have some information on the location of the true value but do not know it exactly. This additional information on the location of the true value is used to form overimputed values for the mismeasured data by incorporating it in cell-level priors. With the help of these cell-level priors MI is extended to MO.

So when using MO for missing data and measurement error we have two distributions describing the available information: On the one hand there are the cell-level priors which represent the information from the observed mismeasured value and on the other hand we obtain a distribution based on all the rest of observed data in the dataset. These two sources of information are combined. We can now correct for all amounts of measurement errors not only for the extreme case with infinite measurement error variance. But if the measurement error is infinite and therefore the mismeasured value does not contain any information about the true latent value, only information from the observed values is used for imputation. On the other hand if the measurement error variance is neither zero nor infinite, we draw several times from the posterior distribution of the latent true value to overwrite the mismeasured value.

Now we want to describe more precisely how MO works. Therefore we use the explanations in Blackwell et al. (2012b). We modify the notation a little bit since we want to use MO for longitudinal data. We add an indices for the time of the measurement, $t = 1, ..., T$. For the case with only missing values we defined a missingness matrix M where the values were either zero (for not missing) and one (for missing). Now a measurement mechanism is defined which takes into account that values can also be mismeasured (partially missing). For the data $\boldsymbol{x}_{itp}$, with $i = 1, ..., I$ (subjects), $t = 1, ..., T$ (time/number of measurements per subject) and $p = 1, ..., P$ (covariates), the measurement mechanism is given by:

$$\tilde{m}_{itp} = \begin{cases} 0 \text{ if } x_{itp} \text{ is observed} \\ 1 \text{ if } x_{itp} \text{ is missing and an unbiased proxy } x^e_{itp} \\ 0 \text{ if } x_{itp} \text{ is missing} \end{cases} . \qquad (12)$$

Blackwell et al. (2015b) consider here only covariates and not the response. With the help of this mechanism we can define two subsets, one containing the (partially) missing true values and one containing the observed true values:

$$\boldsymbol{x}^{\text{obs}} = \{x_{itp} | (\forall i, j) \wedge (\tilde{m}_{ijt} = 0)\}$$
$$\boldsymbol{x}^{\text{mis}} = \{x_{itp} | (\forall i, j) \wedge (\tilde{m}_{ijt} > 0)\}.$$

These two subsets can be combined to $\boldsymbol{x} = (\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}})$. The measurement mechanism and the data subsets can be used to construct the following observed data probability density function:

$$p(\tilde{\boldsymbol{m}}, \boldsymbol{x}^e, \boldsymbol{x}^{\text{obs}} | \boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}}) = \int p(\tilde{\boldsymbol{m}} | \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) p(\boldsymbol{x}^e | \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{\eta}) p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}} | \boldsymbol{\theta}) d\boldsymbol{x}^{\text{mis}}, \qquad (13)$$

where $\tilde{\boldsymbol{\phi}}$ are the parameters corresponding to the distribution of measurement mechanism, $\tilde{\boldsymbol{m}}$ and $\boldsymbol{\eta}$ are the parameters corresponding to the mismeasured data $\boldsymbol{x}^e$ [10]. This is the first step to determine the distribution of the true data $\boldsymbol{x}$. Moreover two assumptions are made, one concerning the measurement mechanism and one concerning the measurement error. The first assumption ensures that the distribution of the measurement indicator $\boldsymbol{m}$ does not depend on the missing data. Blackwell at al. (2015b) call this property Ignorable Measurement Mechanism Assignment (IMMA). IMMA is the extension of the missing at random assumption necessary for MI. Two things have to be fulfilled for IMMA (and for the above described probability density function):

1. For any values of $\boldsymbol{m}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^e$ and two possible ralizations of missing data ,$\boldsymbol{x}^{\text{mis}}$ and $\boldsymbol{x}^{\text{mis}\prime}$,

$$p(\tilde{\boldsymbol{m}} | \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) = p(\tilde{\boldsymbol{m}} | \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}\prime}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) \qquad (14)$$

[10] Blackwell et al. (2015b) write that $\boldsymbol{\eta}$ (they call it γ) are the parameters of the distribution of measurement meachanism but the measurement mechanism is denoted by $\boldsymbol{m}$ (see page 2 in Blackwell et al. (2015b)) and the parameters for the distribution of $\boldsymbol{m}$ are given by $\tilde{\boldsymbol{\phi}}$. We conclude that $\boldsymbol{\eta}$ are the parameters corresponding to the distribution of the mismeasured variable since it is part of the distribtuion in equation 13

2. the parameters $\tilde{\boldsymbol{\phi}}$ corresponding to the distribution of $\boldsymbol{m}$ are distinct from the parameters corresponding to $\boldsymbol{x}^e$ and $\boldsymbol{x}$, which are denoted by $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$.

The second assumption concerns the measurement error distribution:

> Except for the parameters $\boldsymbol{\eta}$ the measurement error data generating process or measurement error distribution $p(\boldsymbol{x}^e|\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{\eta})$ is known. For the parameters $\boldsymbol{\eta}$ exists an consistent estimator if they are unknown.

To describe MO Blackwell et al. (2015b) define three likelihoods: joint likelihood, profile likelihood and ignorance likelihood. They show that the profile likelihood for $(\boldsymbol{\theta}, \boldsymbol{\eta})$ is proportional to the ignorance likelihood if the IMMA assumption is fulfilled. The joint likelihood of the parameters corresponding to the true data $\boldsymbol{x}$, the distribution of the missing data mechanism and the missingness indicator $\tilde{\boldsymbol{m}}$ is given by:

$$\mathcal{L}_j(\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}}) = p(\tilde{\boldsymbol{m}}, \boldsymbol{x}^e, \boldsymbol{x}^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}}). \tag{15}$$

The profile likelihood of the parameters corresponding to the true data $\boldsymbol{x}$ and to the distribution of the missing data mechanism with respect to $\boldsymbol{\phi}$ is constructed as

$$\mathcal{L}_p(\boldsymbol{\theta}, \boldsymbol{\eta}) = \max_{\tilde{\boldsymbol{\phi}}} \left[\mathcal{L}_j(\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}})\right]. \tag{16}$$

So the profile likelihood describes the maximum of the joint likelihood for the parameters of the distribution of $\tilde{\boldsymbol{m}}$, $\tilde{\boldsymbol{\phi}}$. The third specified likelihood is the ignorance likelihood which ignores the measurement indicator $\tilde{\boldsymbol{m}}$:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{x}^e, \boldsymbol{x}^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{17}$$

Since the measurement indicator is ignored in the ignorance likelihood, it does not depend on the corresponding parameters $\tilde{\boldsymbol{\phi}}$. Now it can be shown that under IMMA the above described profile likelihood is proportional to the ignorance likelihood. If IMMA holds, the joint likelihood can be factored as

$$\begin{aligned}
\mathcal{L}_j(\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}}) &= p(\boldsymbol{m}, \boldsymbol{x}^e, \boldsymbol{x}^{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{\phi}}) \\
&\overset{\text{def. joint Lik.}}{=} \int p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}|\boldsymbol{\theta}) p(\boldsymbol{x}^e|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{\eta}) p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) d\boldsymbol{x}^{\text{mis}} \\
&\overset{\text{IMMA}}{=} \int p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}|\boldsymbol{\theta}) p(\boldsymbol{x}^e|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{\eta}) p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) d\boldsymbol{x}^{\text{mis}} \\
&= p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) \int p(\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}|\boldsymbol{\theta}) p(\boldsymbol{x}^e|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}}, \boldsymbol{\eta}) d\boldsymbol{x}^{\text{mis}} \\
&= p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^e, \tilde{\boldsymbol{\phi}}) \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}).
\end{aligned}$$

The fourth equality holds since the measurement mechanism does not depend on the missing values and therefore we do not have to integrate out the missings and can write it in front of the integral. The resulting integral describes the ignorance likelihood 17 since the distribution to which $\tilde{\boldsymbol{\phi}}$ corresponds, is ignored. So we obtain the last term. The profile likelihood is a function of the joint likelihood. The factored likelihood can be inserted in the profile likelihood:

$$\begin{aligned}
\mathcal{L}_p(\boldsymbol{\theta},\boldsymbol{\eta}) &= \max_{\tilde{\phi}} \left[p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^e,\tilde{\boldsymbol{\phi}})\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta})\right] \\
&= \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta})\max_{\tilde{\phi}} \left[p(\tilde{\boldsymbol{m}}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^e,\tilde{\boldsymbol{\phi}})\right]^{11} \\
&\propto \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta}).
\end{aligned}$$

Here the second equation holds since the second factor in the first line (the ignorance likelihood) does not depend on $\tilde{\boldsymbol{\phi}}$ and therefore the maximum of this likelihood with respect to $\tilde{\boldsymbol{\phi}}$ is simply the likelihood itself. The last equality follows because the second factor in the second line is independent of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. So it is shown that the profile likelihood 16 is proportional to the ignorance likelihood 17 if IMMA holds.

□

Since the profile and ignorance likelihood are proportional, the inferences based on the ignorance likelihood and inferences about $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ from the profile likelihood are the same. In the following parts of this work $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta})$ denotes as observed data likelihood.

In the maximum likelihood part of section 2.3 we described the EM algorithm in general to find maximum likelihood estimates if the likelihood depends on data which are subject to missingness. This algorithm can be used also if there are missing data and mismeasured variables. Blackwell et al. (2015b) assume that the parameters of the measurement error distribution, $\boldsymbol{\eta}$ are known. In the first step of the EM algorithm the conditional expectations given the current parameter estimates are calculated. With missing data and measurement error, this expectation averaged across the missing data is given by:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h)}) = \int \log\left[p(\boldsymbol{x}^{\text{mis}},\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^e|\boldsymbol{\theta},\boldsymbol{\eta})\right] p(\boldsymbol{x}^{\text{mis}}|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^e,\boldsymbol{\theta}^{(h)},\boldsymbol{\eta})d\boldsymbol{x}^{\text{mis}}, \tag{18}$$

where $\boldsymbol{\theta}^{(h)}$ are the parameters estimated in the previous M-step. So the second factor, the posterior distribution of the missing data, is fixed in $\boldsymbol{\theta}$. The first factor is the log likelihood

[11] Blackwell et al. wrote $\mathcal{L}_p(\boldsymbol{\theta},\boldsymbol{\eta})$ instead of $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\eta})$. We think this is a typing error since in the first equation the likelihood does not depend on $\tilde{\boldsymbol{\phi}}$ and therefore the maximization over $\tilde{\boldsymbol{\phi}}$ does not yield the profile likelihood but the likelihood itself.

of the complete data which is given by

$$\log\left[p(\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{e}|\boldsymbol{\theta}, \boldsymbol{\eta})\right] = \log\left[p(\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}|\boldsymbol{\theta})\right] + \log\left[p(\boldsymbol{x}^{e}|\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{\eta})\right].$$

Given this log likelihood, the expectation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h)})$ can be simplified since the second term of the complete data likelihood does only depend on the measurement error process which is assumed to be known. So this term can be crossed out in the expectation since the parameters $\boldsymbol{\eta}$ of the measurement error process are distinct from $\boldsymbol{\theta}$. The other part of the expectation, the posterior distribution of the missing data, is given by

$$p(\boldsymbol{x}^{\text{mis}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{e}, \boldsymbol{\theta}^{(h)}, \boldsymbol{\eta}) \propto p(\boldsymbol{x}^{\text{mis}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{\theta}^{(h)})p(\boldsymbol{x}^{e}|\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{\eta}).$$

The simplified expectation combined with this expression of the posterior distribution of the missing data yields the following objective function for the E-step:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h)}) = \int \log\left[p(\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}|\boldsymbol{\theta})\right] p(\boldsymbol{x}^{\text{mis}}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{\theta}^{(h)})p(\boldsymbol{x}^{e}|\boldsymbol{x}^{\text{mis}}, \boldsymbol{x}^{\text{obs}}, \boldsymbol{\eta})d\boldsymbol{x}^{\text{mis}}. \tag{19}$$

So the information from the mismeasured values enter the objective function via the posterior distribution of the missing data. After calculating the expectation in the E-step, updated estimates of $\boldsymbol{\theta}$ are created in the M-step by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(h)})$ with respect to $\boldsymbol{\theta}$. The resulting estimate is denoted by $\boldsymbol{\theta}^{(h+1)}$ and used to compute the expectation in the next step. Both steps are repeated until convergence. The results are the maximum likelihood estimates of $\boldsymbol{\theta}$ under the observed data likelihood. In reality the assumption that the parameters $\boldsymbol{\eta}$ of the distribution of the mismeasured values are known, is not fulfilled often. So the parameters have to be estimated. If consistent estimates are used instead of the true $\boldsymbol{\eta}$, the algorithm still converges to a consistent estimator for $\boldsymbol{\theta}$.

This EM algorithm is used to compute the multiple overimputations. So the result of MO are several datasets without missing values and corrected for measurement error which are denoted by $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(k)}$ with $\boldsymbol{x}^{(k)} = (\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{\text{mis}(k)})$. So only the missing and mismeasured data can differ from imputed datasets to imputed dataset, the correctly observed data are the same. The mismeasured values are overwritten with draws from the predictive posterior distribution of the unobserved data:

$$(\boldsymbol{x}^{\text{mis}(k)}) \sim p(\boldsymbol{x}^{\text{mis}(k)}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{z}^{e}) = \int p(\boldsymbol{x}^{\text{mis}(k)}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{z}^{e}, \boldsymbol{\theta}, \boldsymbol{\eta})p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{e})d\boldsymbol{\theta}d\boldsymbol{\eta}. \tag{20}$$

Given this predictive posterior, a possible way to create multiple overimputations can be deduced. Therefore $\boldsymbol{\theta^{(k)}}$ is drawn from its posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{x}^{e}, \boldsymbol{\eta})$ and then $(\boldsymbol{x}^{\text{mis}(k)})$ is drawn from $p(\boldsymbol{x}^{\text{mis}(k)}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{z}^{e}) = \int p(\boldsymbol{x}^{\text{mis}(k)}|\boldsymbol{x}^{\text{obs}}, \boldsymbol{z}^{e}, \boldsymbol{\theta}^{(b)}, \boldsymbol{\eta})$. To generate these draws the EM algorithm can be used in combination with an additional sampling step. This EM approach is used in Blackwell et al. (2015a,b). For additional sampling they use bootstrapping and in sample k $\hat{\boldsymbol{\theta}}_k$ is used as draw from the posterior of $\boldsymbol{\theta}$. If the EM algorithm as described above is used, we obtain a maximum likelihood estimate or a maximum a posteriori estimate $\hat{\boldsymbol{\theta}}$. This is used to draw from the predictive posteriori. Another possibility is to use for example Gibbs sampling or other Markov Chain Monte Carlo approaches.

To specify the distributions more precisely, we have to make an assumption on the distribution of the true data. Often it is assumed that the complete data $\boldsymbol{x}$ are multivariate normally distributed with parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The normal distribution is statistically dual. Imagine a function of two variables (a, b). The distribution of a with parameter b and the distribution of b with parameter a are statistically dual if they both describe the function of (a, b) (Bityukov et al. (2008)). Or for the normal distribution in other words: Any conditional distribution of the ideal data is also normal if the ideal data is normal. Additional to the distribution of the complete data, assumptions concerning the distribution of the mismeasured values x_{itp} are necessary. Blackwell et al. (2015b) assume that the mismeasured values are normally distributed with mean z_{itp} (unobserved true value) and measurement error variance $\sigma^2_{u,p}$ [12]. The distribution of the mismeasured values can also be written as multivariate normal distribution by changing the notation. Since only values contained in $\boldsymbol{x}^{\text{mis}}_{it}$ [13] can be mismeasured, this is the mean of the multivariate normal distribution. The covariance matrix is given by $\boldsymbol{\Sigma}_u = \boldsymbol{\sigma}^2_u \boldsymbol{I}$ with $\boldsymbol{\sigma}^2_u = \{\sigma^2_{u,p}; m_{itp} = 1 \wedge \sigma^2_{u,p} = \sigma^2_{u,itp} \forall i, t\}$ and $\boldsymbol{I}$ is the identity matrix with dimension $\sum_p \mathbb{I}\{m_{itp} = 1\}$. The measurement error variance is assumed to be known or estimable and the measurement error is unbiased. If otherwise the measurement error is biased or depends on other variables, this information can be included in the cell-level means. For example assume that we want to measure the weight of people as variable $\boldsymbol{x}_1$ and we know that the weighing machine always measures three kilograms less than would be right. This error can be incorporated in the cell-level means of the distribution by subtracting three from the cell-level mean. So the resulting distribution would be $\mathcal{N}(x_{it1} - 3, \sigma^2_{u,j})$ and if the weighing machine measured three kilograms less for all observations and times this distribution is the same for all i and t. But Blackwell et al. (2015b) assume that the measurement error is unbiased. They describe the E-step for normal complete data and normal mismeasured values but assume that fully missing values ($m_{itp} = 2$) do not exist.

[12] Blackwell et al. (2015b) assume that this variance differs for different observation i. For longitudinal data this would mean that the variance have to depend on it but we assume that the measurement error variance is the same for all observations. It is only allowed to differ for different covariates p

[13] $\boldsymbol{x}^{\text{mis}}_{it}$ are the true unobserved values for the missing and mismeasured values.

Then the E-step is given by

$$\begin{aligned}\mathbb{E}\left[T(\boldsymbol{x})|\boldsymbol{x}^{\text{obs}},\boldsymbol{x}^{e},\boldsymbol{\theta}^{(h)}\right] &= \int T(\boldsymbol{x})p(\boldsymbol{x}^{\text{mis}}|\boldsymbol{x}^{\text{obs}},\boldsymbol{\theta}^{(h)})p(\boldsymbol{x}^{e}|\boldsymbol{x}^{\text{mis}},\boldsymbol{\eta})d\boldsymbol{x}^{\text{mis}}\\ &= \int T(\boldsymbol{x})\prod_{i,t}\underbrace{p(\boldsymbol{x}_{it}^{\text{mis}}|\boldsymbol{x}_{it}^{\text{obs}},\boldsymbol{\theta}^{(h)})p(\boldsymbol{x}_{it}^{e}|\boldsymbol{x}_{it}^{\text{mis}},\boldsymbol{\Sigma}_u)}_{\text{full conditional distribution}}d\boldsymbol{x}_{it}^{\text{mis}},\end{aligned}$$

where $T(\boldsymbol{x})$ are the sufficient statistics for the multivariate normal and the parameter $\boldsymbol{\eta}$ of the mismeasured variables is the measurement error variance $\boldsymbol{\Sigma}_u$. Instead of calculating the above E-step, we could calculate the objective function 19. Both yield the same results. To calculate the E-step the full conditional have to be determined. Therefore the distributions of $\left[\boldsymbol{x}_{it}^{\text{mis}}|\boldsymbol{x}_{it}^{\text{obs}},\boldsymbol{\theta}\right]$ and $\left[\boldsymbol{x}_{it}^{e}|\boldsymbol{x}_{it}^{\text{mis}},\boldsymbol{\Sigma}_u\right]$ are necessary. Both distributions are (multivariate) normal:

$$\begin{aligned}\left[\boldsymbol{x}_{it}^{\text{mis}}|\boldsymbol{x}_{it}^{\text{obs}},\boldsymbol{\theta}\right] &\sim \mathcal{N}(\boldsymbol{\mu}_{e|o},\boldsymbol{\Sigma}_{e|o})\\ \left[\boldsymbol{x}_{it}^{e}|\boldsymbol{x}_{it}^{\text{mis}},\boldsymbol{\Sigma}_u\right] &\sim \mathcal{N}(\boldsymbol{x}_{it}^{\text{mis}},\boldsymbol{\Sigma}_u),\end{aligned}$$

where $(\boldsymbol{\mu}_{e|o},\boldsymbol{\Sigma}_{e|o})$ are deterministic functions of the complete data parameters $\boldsymbol{\theta}$ and $\boldsymbol{x}_{it}^{\text{obs}}$. The first factor of the full conditional contains information from a regression of the missing data on the observed data and the second factor represents information from the mismeasured values. The distribution of the full conditional can be calculated as follows

$$\begin{aligned}p(\boldsymbol{x}_{it}^{\text{mis}}|\boldsymbol{x}_{it}^{\text{obs}},\boldsymbol{x}_{it}^{e},\boldsymbol{\theta}^{(h)},\boldsymbol{\Sigma}_u) &\propto p(\boldsymbol{x}_{it}^{\text{mis}}|\boldsymbol{x}_{it}^{\text{obs}},\boldsymbol{\theta}^{(h)})p(\boldsymbol{x}_{it}^{e}|\boldsymbol{x}_{it}^{\text{mis}},\boldsymbol{\Sigma}_u)\\ &\propto \exp(-\frac{1}{2}(\boldsymbol{x}_{it}^{\text{mis}}-\boldsymbol{\mu}_{e|o})^T\boldsymbol{\Sigma}_{e|o}^{-1}(\boldsymbol{x}_{it}^{\text{mis}}-\boldsymbol{\mu}_{e|o}))\cdot\exp(-\frac{1}{2}(\boldsymbol{x}_{it}^{\text{e}}-\boldsymbol{x}_{it}^{\text{mis}})^T\boldsymbol{\Sigma}_u^{-1}(\boldsymbol{x}_{it}^{\text{e}}-\boldsymbol{x}_{it}^{\text{mis}}))\\ &= \exp(-\frac{1}{2}\left[(\boldsymbol{x}_{it}^{\text{mis}}-\boldsymbol{\mu}_{e|o})^T\boldsymbol{\Sigma}_{e|o}^{-1}(\boldsymbol{x}_{it}^{\text{mis}}-\boldsymbol{\mu}_{e|o})\cdot(\boldsymbol{x}_{it}^{\text{e}}-\boldsymbol{x}_{it}^{\text{mis}})^T\boldsymbol{\Sigma}_u^{-1}(\boldsymbol{x}_{it}^{\text{e}}-\boldsymbol{x}_{it}^{\text{mis}})\right])\\ &= \exp(-\frac{1}{2}\left[\boldsymbol{x}_{it}^{\text{mis}T}\boldsymbol{\Sigma}_{e|o}^{-1}\boldsymbol{x}_{it}^{\text{mis}}-2\boldsymbol{x}_{it}^{\text{mis}T}\boldsymbol{\Sigma}_{e|o}^{-1}\boldsymbol{\mu}_{e|o}+\boldsymbol{\mu}_{e|o}{}^T\boldsymbol{\Sigma}_{e|o}^{-1}\boldsymbol{\mu}_{e|o}\right.\\ &\quad \left.+\,\boldsymbol{x}_{it}^{\text{e}\,T}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{x}_{it}^{\text{e}}-2\boldsymbol{x}_{it}^{\text{mis}T}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{x}_{it}^{\text{e}}+\boldsymbol{x}_{it}^{\text{mis}T}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{x}_{it}^{\text{mis}}\right])\\ &= \exp(-\frac{1}{2}\left[\boldsymbol{x}_{it}^{\text{mis}T}\underbrace{(\boldsymbol{\Sigma}_{e|o}^{-1}+\boldsymbol{\Sigma}_u^{-1})}_{\boldsymbol{\Sigma}^{*-1}}\boldsymbol{x}_{it}^{\text{mis}}-2\boldsymbol{x}_{it}^{\text{mis}T}\underbrace{(\boldsymbol{\Sigma}_{e|o}^{-1}\boldsymbol{\mu}_{e|o}+\boldsymbol{\Sigma}_u^{-1}\boldsymbol{x}_{it}^{\text{e}})}_{\boldsymbol{\mu}^*\boldsymbol{\Sigma}^{*-1}}+\boldsymbol{\mu}^{*T}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}^*\right]\\ &\propto \mathcal{N}(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*).\end{aligned}$$

The mean and covariance of this multivariate normal distribution are

$$\boldsymbol{\Sigma}^* = (\boldsymbol{\Sigma}_u^{-1}+\boldsymbol{\Sigma}_{e|o}^{-1})^{-1}, \boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*(\boldsymbol{\Sigma}_u^{-1}\boldsymbol{x}_{it}^{e}+\boldsymbol{\Sigma}_{e|o}^{-1}\boldsymbol{\mu}_{e|o}).$$

With the help of this full conditional, the E-step can be calculated. In the M- step the complete data parameters $\boldsymbol{\theta}^{(h+1)} = (\boldsymbol{\mu}^{(h+1)}, \Sigma^{(h+1)})$ are estimated by maximizing the complete data likelihood resulting from the E-step. So the crucial difference between MI and MO is the calculation of the E-step but the rest of the EMB algorithm is essentially the same as described in the MI section.

This was one example for a MO model with normal data but Blackwell et al. (2015b) allow all data generating processes which fulfil the property of statistical duality. These are for example the Poisson, Gamma and Inverse-Gamma distribution. The reason for overimputing multiply is that this way the uncertainty in the location of the unobserved latent values $\mathbf{x}_{it}$ is incorporated. When using MO 10 to 25 imputations are sufficient as long as the fraction of missing values is not too large or the degree of measurement error too high. Each of the resulting datasets contains different values for the missing and mismeasured values which vary more if the predictive ability of the model is small. The overimputed datasets can than be analysed and the results can be combined by the same methods as described for MI (see 3.1).

The MO approach was also added to the Amelia package which assumes that the data are multivariate normal. Amelia can be used for (over-)imputing datasets with measurement error and missing data. In the function `amelia()` we have to specify a prior for MO where one column describes the means of the mismeasured variables which are the values of the mismeasured variable itself. This represents the information from the measurement error model. The information from the observed variables is also used in MI so there is no special incorporation of it in MO. Moreover we have to make an assumption concerning the measurement error variance or the amount of the variance of the mismeasured variable which is described by the measurement error.

Blackwell et al. (2015a,b) described this method for normal (not longitudinal) data. But in reality there are also longitudinal studies with measurement error and missing data. We want to examine how multiple overimputation performs for this kind of data (longitudinal data with measurement error and missings) using a simulation study. In the next section we describe the model used in the simulation study.

4 The Model

In this section we want to describe the model that we use in the simulation study. The model is very simple because to our knowledge it was not tested before if MO works for longitudinal data, too.

We consider a longitudinal dataset with five baseline covariates and one response. The response is measured at several times per subject. We assume that a subject specific random intercept makes sense. That could mean for example that the propositi have different starting values for the blood pressure - some have a very high blood pressure at the beginning of the study, others have a low blood pressure and others are in between. This is a common scenario in clinical studies. For more information about random intercepts, random slopes and more general mixed models we refer to Verbeke and Molenberghs (2000).

The model can be written as follows:

$$y_{it} = \boldsymbol{\beta} \cdot \boldsymbol{x}_{it} + \beta_t t + b_i + \varepsilon_{it} \text{ or} \tag{21}$$

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \beta_4 \boldsymbol{x}_4 + \beta_5 \boldsymbol{x}_5 + \beta_t t + \boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{22}$$

where i is the index for the subjects, the indices t is the time index, the variable t is a value in $1, 2, ..., T$, ε is the noise and b_i describes the subject-specific random intercept. The random intercept shall be normally distributed with mean zero and variance σ_b. Moreover we assume that the distribution of $\boldsymbol{X}$ is multivariate normal with mean zero and variance-covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

On the diagonal, $\boldsymbol{\Sigma}$ contains the variances of the covariates p, $var(\boldsymbol{x}_p)$ and next to the diagonal the covariances of the covariates p, q with $p \neq q$, $cov(\boldsymbol{x}_p, \boldsymbol{x}_q)$. So we assume that the covariates are correlated. The response $\boldsymbol{y}$ is obtained by a linear combination of the covariates (see 22) Therefore the response is normally distributed, also.

If we would like to estimate this normal random intercept model we could simply use the Linear Mixed Model (LMM) approach. But we assume that the data are subject to missingness and measurement error. More specifically we assume that some individuals drop out of the study so that the response contains missing values. The first measurement of the response is always observed but if a later value of the response is missing, all following values for the corresponding individual are missing, too. Moreover we assume that missing values are contained in $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$. Additionally the first covariate $\boldsymbol{x}_1$ is subject to an normal classical additive measurement error. and the observed mismeasured variable is denoted by:

$$\boldsymbol{x}_1^e = \boldsymbol{x}_1 + \boldsymbol{u}, \tag{23}$$

where $\boldsymbol{u}$ is the measurement error (see 1) normally distributed with mean zero and variance σ_u^2. We assume that the measurement error variance is the same for all observations and that the IMMA assumption described in the MO section is fulfilled. So the distribution of the mismeasurement indicator does not vary if the missings or true unobserved values of the mismeasured data vary. For the missings it means that MAR has to be fulfilled. So the missingness probability just depends on the covariates which are fully observed ($\boldsymbol{x}_4$, $\boldsymbol{x}_5$) and on the response at time one since the response is always observed at time one.
Our observed model is:

$$\boldsymbol{y}^{\text{obs}} = \beta_0 + \beta_1 \boldsymbol{x}_1^{e,\text{obs}} + \beta_2 \boldsymbol{x}_2^{\text{obs}} + \beta_3 \boldsymbol{x}_3^{\text{obs}} + \beta_4 \boldsymbol{x}_4 + \beta_5 \boldsymbol{x}_5 + \beta_t t + \boldsymbol{b\varepsilon}, \tag{24}$$

If we would use a normal LMM (for example the R function `lmer` from the R-package lme4), the estimators would be biased because the rows that contains missings would be deleted (complete case analysis) and because of the measurement error in the first covariate (see 2.5.1).
To our knowledge this combination of problems (longitudinal data + missing covariates + drop-out in response + mismeasured covariate) was not studied before. We want to use the Multiple Overimputation approach for missing data and measurement error implemented in the package Amelia (Blackwell et al. (2011), Blackwell et al. (2015)). To check if this approach also works for longitudinal data with measurement error and missings in covariates and drop-outs in the response, we conduct a simulation study which we will describe in the next section.

5 Simulation

In this section we want to describe our simulation study, divided into three parts corresponding to the first three parts of this section. In the first part we explain how the true and observed dataset is generated. In the second part we describe the (over-) imputation of the mismeasured and missing values and in the third part we describe the analysis and the results. In the last part of this section we want to present a small sensitivity analysis where we vary the amount of measurement error, the time points and the number of missings and drop-outs. Moreover we examine how the results change if we do not specify the share of measurement error variance correctly. All simulations consist of 1000 runs.

5.1 Data Generation

To build the dataset which we described in the model section we generate the covariate data first. As described in the previous section, we assume that the covariate-vector $\boldsymbol{X}$ contains only baseline covariates and is multivariate normally distributed:

$$\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is a covariance-matrix with $diag(\boldsymbol{\Sigma}) = 1 = Var(\boldsymbol{x}_p), p \in \{1, 2, 3, 4, 5\}$ and the entries which are not on the diagonal are the Covariances $Cov(\boldsymbol{x}_p, \boldsymbol{x}_q) = 0.5$ with $p \neq q$ and $q \in \{1, 2, 3, 4, 5\}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}.$$

We draw $I = 1000$ samples from the above distribution since we want to have 1000 different subjects in our dataset and we need one value for each of the subjects. We assume that all covariates are baseline covariates. Therefore we have to duplicate the sampled values until we obtain $T = 5$ observations per subject. So all five values of a covariate for one subject are the same. The result of this first step is a dataset with 5 variables (covariates) and for each variable we do have 5000 observations. Then we incorporate a time variable with values in $\{1, 2, 3, 4, 5\}$ and an ID indicator with values between 1 and 1000 so we know which observations belongs to which subject.

In the next step we have to generate the response. This works straight forward since we just have to insert all values in model 22 plus an error term ε. We already know the covariate values but we still need the coefficients, the random intercept and the error term to calculate the response.

We define the coefficient vector $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\beta} = \Big(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_t\Big)^T = \Big(1, 10, 5, -8, 3, 1, 1\Big)^T \tag{25}$$

The random intercept is drawn from the normal distribution, one for each subject:

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \tag{26}$$

where we set the variance of the random intercept to $\sigma_b^2 = 3$ and $i \in 1, 2, ..., 1000$ are the subjects. The result is a random intercept vector b with dimension 5000×1 since we have to duplicate the values for each subject until we have 5 same random incept values per subject. The error term ε is drawn from a normal distribution, too, but there is one for each subject at each time point:

$$\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon). \tag{27}$$

We set the error term variance to $\sigma_\varepsilon = 2$. Now we can draw $T \cdot I = 5000$ error term values from this normal distribution.
The the response is calculated as follows:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \beta_4 \boldsymbol{x}_4 + \beta_5 \boldsymbol{x}_5 + \boldsymbol{b} + \boldsymbol{\varepsilon}. \tag{28}$$

When we combine the response with the dataset containing the covariates, the id-variable and the time-variable, we get a complete dataset. This complete dataset contains the true values without missings and measurement error. We can use the normal LMM to estimate the coefficients of this model. We do this to compare the coefficients from the true LMM with the coefficients we get using MO for the observed dataset. The generation of the observed data will be described now.
The next step is to generate the mismearued variable x_1^e as described in equation 23. Therefore we draw one measurement error for each subject from the normal distribution

$$u_i \sim \mathcal{N}(0, \sigma_u^2), \tag{29}$$

where $\sigma_u^2 = 0.5$ is the variance of the measurement error. With $\boldsymbol{u} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_{1000})$ we can calculate the observed mismeasured variable $\boldsymbol{x}_1^e = \boldsymbol{x}_1 + \boldsymbol{u}$.
Since we assume that some covariates are subject to missingness and the response is subject to drop out, we need to generate the missings and dop-outs for which we assume that they are MAR. Therefore we calculate missingness probabilities for $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{x}_3$ (the covariates which contain missing values) as follows

$$\pi_{ij} = \frac{exp(\alpha_{y,j} \cdot y_{i1} + \alpha_{4,j} \cdot x_{4,i1} + \alpha_{5,j} \cdot x_{5,i1} + \alpha_{0,j})}{(1 + exp(\alpha_{y,j} \cdot y_{i1} + \alpha_{4,j} \cdot x_{4,i1} + \alpha_{5,j} \cdot x_{5,i1} + \alpha_{0,j}))}, \tag{30}$$

where $j \in \{1, 2, 3\}$ is the index for the covariates subject to missingness. Since we want data which are MAR, π is determined by a constant $\alpha_{0,j}$ which can take different values for the different missing variables j and by the variables which are fully observed multiplied with a constant α_j which can vary for the different observed variables. So we can control the values of π and the influence of the observed variables on the missingness probabilities $\boldsymbol{\pi}_{ij}$. The fully observed variables are $\boldsymbol{x}_4, \boldsymbol{x}_5$ and y_{i1} since we assume that the first measurement of the response is never missing. We choose $\alpha(\cdot, j)$ so that we get the missing probabilities that we want, here around $\sum_{i=1}^{I} \pi_i = 0.1$. The values are

$$\alpha_{0,1} = -2.5, \alpha_{4,1} = 0.8, \alpha_{5,1} = 0.8, \alpha_{y,1} = -0.12, \text{ (for the missingness probabilty of } \boldsymbol{x}_1)$$
$$\alpha_{0,2} = -2, \alpha_{4,2} = 0.4, \alpha_{5,2} = -0.25, \alpha_{y,2} = -0.05, \text{ (for the missingness probabilty of } \boldsymbol{x}_2)$$
$$\alpha_{0,3} = -2.3, \alpha_{4,3} = -0.7, \alpha_{5,3} = 0.4, \alpha_{y,3} = 0 \text{ (for the missingness probabilty of } \boldsymbol{x}_3).$$

The calculation of the missingness probability π_{ity} for the response works similar but we need different values for different time points since the values of y are time-dependent:

$$\pi_{ity} = \frac{exp(\alpha_{y,y} \cdot y_{i,t-1} + \alpha_{4,y} \cdot x_{4,i1} + \alpha_{5,y} \cdot x_{5,i1} + \alpha_{0,y})}{(1 + exp(\alpha_{y,y} \cdot y_{i,t-1} + \alpha_{4,y} \cdot x_{4,i1} + \alpha_{5,y} \cdot x_{5,i1} + \alpha_{0,y}))}, \tag{31}$$

where the missingness probability in t depends on the value of y in $t-1$ and by assumption $\pi_{i1y} = 0 \; \forall \; i$.

Here we set the values of $\alpha_{\cdot,y}$ as follows

$$\alpha_{0,y} = -2.5, \alpha_{4,y} = 0.8, \alpha_{5,y} = 0.8, \alpha_{y,y} = -0.05.$$

With the resulting missingness probabilities we can generate a dummy variable $r_{x_1^e}$ by drawing from the binomial distribution $B(0, \pi_{ij})$ which takes the following values:

$$r_{i,x_1^e} = \begin{cases} 1, & \text{if } \boldsymbol{x}_1^e \text{ is missing for subject } i \\ 0, & \text{if } \boldsymbol{x}_1^e \text{ is observed for subject } i. \end{cases} \tag{32}$$

r_{i,x_2} and r_{i,x_3} are defined analogously. The dummy variables for the covariates only need to be dependent on the subject i and not on the time because the covariates are baseline covariates and have the same value over time regarding a subject.

Since we assume that $\boldsymbol{y}$ is subject to drop-out and always observed for $t = 1$, the values of $r_{it,y}$ have to be different for the different time values. So we need an indicator for missingness for each subject and each time point with $r_{i1,y} = 0 \; \forall \; i$ and $r_{it,y}$ for $t > 1$ is defined by:

$$r_{it,y} = \begin{cases} 1, & \text{if } \boldsymbol{x}_1^e \text{ is missing for subject } i \text{ and } t > 1 \\ 0, & \text{if } \boldsymbol{x}_1^e \text{ is observed for subject } i \text{ and } t > 1. \end{cases} \tag{33}$$

To generate $r_{it,y}$ we draw from the binomial distribution $B(0, \pi_{ity})$. But since we assume that $\boldsymbol{y}$ is subject to drop-out, we manipulate the vector $\boldsymbol{r}_{i,y}$ in a way that $r_{it,y} = 1$ if $r_{i(t-1),y} = 1$. With the above mentioned values of $\alpha_{\cdot,y}$ the drop-out probability is around 30%.

The next step is to replace all values of $\boldsymbol{x}_1^e, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{y}$ with missing values if the corresponding dummy variable $\boldsymbol{r}_{x_1^e}, \boldsymbol{r}_{x_2}, \boldsymbol{r}_{x_3}, \boldsymbol{r}_y$ takes the value one. We obtain an observed dataset which contains one covariate with measurement error and missings, two only with missings, two covariates which are observed correctly, one response with drop-out, an id-variable and a time-variable.

A naive method to analyse such a dataset is to do a complete case analysis where all rows with missings are dropped and the measurement error is ignored. Another naive method would be to correct for missings but not for measurement error by using multiple imputation. We will compare the results of these methods to the multiple overimputation approach in section 5.3. In the following part we will describe the data imputation using the MO approach.

5.2 Imputation

Since we have a dataset with missings and measurement error and we want to obtain good estimates of the coefficients, we need a method that can deal with both problems. To our knowledge there is no literature concerning the underlying combination of problems - longitudinal data with missing covariates, response drop-out and a mismeasured covariate. But in practical applications this is a common phenomenon. So we want to examine if MO works better than a complete case analysis and MI when we face this problems. The theory of the MO approach is described in section 3.2. Here the focus is on the practical imputation using MO with the statistical software R.

All covariates in our model are baseline covariates and hence take the same value for every time-point dependent on the subject. So first we (over-)impute just the covariates for the first time-point since otherwise the imputed values for each time point differ due to the use of the difference response values for (over-)imputation.

Since the dataset contains a mismeasured variable which we want to overimpute, we need to make some assumptions concerning the variance of the measurement error. In this simulation we know the variance of the measurement error but in reality it is mostly unknown. The R-package Amelia (Honaker et al. (2011)) which we use for (over-)imputation contains a function `moPrep()` to generate priors for MO of the mismeasured values. In `moPrep()` we can specify the measurement error directly or we can specify the error proportion $\frac{\sigma_u^2}{\sigma^2+\sigma_u^2}$ where $\sigma^2 = var(\boldsymbol{x}_1)$. We chose to set the error proportion because we think the knowledge of the error proportion is a more realistic assumption. Since we assume that the measurement error variance is $\sigma_u^2 = 0.5$ the error proportion is given by

$$\frac{\sigma_u^2}{\sigma^2 + \sigma_u^2} = \frac{0.5}{1 + 0.5} = \frac{1}{3}.$$

So the proportion of the measurement error variance on the total variance of $\boldsymbol{x}_1^e$ is around 33.3%. In reality we often do not know the error proportion exactly. Therefore we do a sensitivity analysis in section 5.4 where we consider the case that the error proportion is not specified correctly. The error proportion describes the share of variance of $\boldsymbol{x}_1^e$ due to measurement error and therefore we can also write it as $1 - \frac{\sigma^2}{\sigma^2 + \sigma e^2}$ where the second term represents the reliability ratio. Hence the assumption on the error proportion can be interpreted as assumption on the reliability ratio.

Since we only want to impute the covariates for $t = 1$ first, we create a subset of the observed data which only contains the rows for the first time point. Then we generate the priors for the overimputation with the function `moPrep` in the package Amelia. This yields a prior matrix with 4 columns and one row for each subject for whom $\boldsymbol{x}_1^e$ is not missing. The first column represents the rownumber of the data for which the covariate $\boldsymbol{x}_1^e$ is not missing, the second column contains the observed covariate $\boldsymbol{x}_1^e$ from the dataset, the third contains the prior mean and the fourth the prior variance. Here the prior mean are the corresponding values of x_1^e and if we specify the error proportion, the prior variance is calculated as follows [14]:

$$\text{prior variance } = \sqrt{var(x_1^{e,obs}) \cdot \frac{\sigma_e^2}{\sigma^2 + \sigma_e^2}},$$

where $var(x_1^{e,obs})$ is the variance of the mismeasured covariate without the missings. Additional to the priors matrix `moPrep` yields an overimputation vector. This describes which variables and which corresponding rows should be overimputed. It is just the same as the first two columns of the priors vector but we have to set it separately in the `amelia()` function which is used for imputation. Using the priors and the overimputation vector we impute the missing values and overimputes the mismeasured values for $t = 1$ using `amelia()`. For imputation we use the covariates $\boldsymbol{x}_1^{e,obs}, \boldsymbol{x}_2^{obs}, \boldsymbol{x}_3^{obs}, \boldsymbol{x}_4, \boldsymbol{x}_5$ and the observed response. `amelia()` uses the EMB algorithm described in sections 3.1 and 3.2. So it draws a bootstrap sample from the observed dataset for $t = 1$. Then it uses the EM algorithm to estimate the complete data parameters $\boldsymbol{\theta}$ and then these parameters are used to take draws from the predictive posterior distribution of the unobserved data to obtain the overimputations. The imputations of the missing values are draws from the distribution of the unobserved data conditional on the observed data and the parameters

[14] This formula is part of the `moPrep.default()` function in the R-package Amelia II.

$\boldsymbol{\theta}$.

The result of this (over-)imputation is an Amelia object. It contains k imputed datasets for the first time point without missings and corrected for measurement error.

Since the covariates are baseline covariates we can replace the values for the following time points by the values of the first. So we combine each imputed dataset in the Amelia object for the first time point with the observed dataset beginning from the second time point yielding k datasets without missings and measurement error for $t = 1$ but with missings and measurement error for $t > 1$. Then we can replace the covariate values beginning for $t > 1$ by the imputed values for $t = 1$. We obtain a manipulated Amelia object without missing covariates and corrected for measurement error. But the response is still subject to drop-out.

To impute the missing values of the response in each of the k datasets, we use `amelia()` again. But now we use it to impute the missing respons values in the manipulated Amelia object containing the k datatsets without measurement error and with imputed covariates. We do not have to specify a priors matrix or an overimputation vector because we assume that the response is measured correctly. For this, the regular MI algorithm for missing data is used. This yields again an Amelia object with k imputed dataset but now completely without missings. These imputed datasets can be analysed using all normal statistical methods which can be used for datasets without missings and measurement error. We will describe the analysis and results in the next part.

5.3 Analysis and Results

Since normal statistical methods can be used to analyse the (over-)imputed dataset, we can run a LMM for each overimputed dataset and combine the results by the function given in 2. Or we can use the `zelig()` function to directly combine the results. We described this function in the last part of section 3.1. We decided to use `zelig()` which has an option for LMM (`ls.mixed`) in the package ZeligMultilevel. The normal `summary` function does not work for this object class but with some modification it does [15]. We call this modified function `summary.MI`. In each of the 1000 simulation runs we estimate the coefficients for the overimputation model (MO) using `zelig()`. We want to compare MO with a complete case analysis (CC) (ignores missings and measurement error) and with MI (ignores measurement error). Therefore we also have to use CC and MI in each of the simulation runs. Moreover we estimate the coefficients of the true data (TD) which would be unobserved in reality. For the CC and TD analysis we use the R-function `lmer`

[15] The function can be retrieved on the following homepage: http://stackoverflow.com/questions/16571580/multi-level-regression-model-on-multiply-imputed-data-set-in-r-amelia-zelig-l . The website was called on 04.07.2015.

with a random intercept to estimate the coefficients of the LMMs and for MI we also use `zelig` with model `ls.mixed`.
Since we run the simulation $S = 1000$ times, we obtain 1000 coefficient vectors for each of the four approaches (TD, CC, MI, MO). To compare these approaches we use the MSE which we also split up in its components - Bias and Variance. The MSE over all simulations is given by

$$MSE(\boldsymbol{\beta}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^2, \tag{34}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector for one analysis method and $\boldsymbol{\beta}$ is the true coefficient vector (see 25). The MSE can also be written as a function of the bias and the variance of the coefficients:

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}) &= \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\beta}}^{(s)} - \boldsymbol{\beta}\right)^2 \\ &= \frac{1}{S}\sum_{s=1}^{S}\left[\left(\hat{\boldsymbol{\beta}}^{(s)} - \bar{\hat{\boldsymbol{\beta}}}\right) + \left(\bar{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}\right)\right]^2 \\ &= \underbrace{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\beta}}^{(s)} - \bar{\hat{\boldsymbol{\beta}}}\right)^2}_{Var} + \underbrace{\frac{1}{S}\sum_{s=1}^{S}\left(\bar{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}\right)^2}_{Bias^2}, \end{aligned} \tag{35}$$

where $\bar{\hat{\boldsymbol{\beta}}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\boldsymbol{\beta}}^{(s)}$ is the mean of the estimated coefficient vector over all simulations. We also compare the variance and the bias of the coefficient vectors resulting from the four analysis methods to examine whether the difference in MSE is induced by a difference in the variance or the bias. We calculate the Bias and Variance as follows:

$$Var(\hat{\boldsymbol{\beta}}) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{\beta}}^{(s)} - \bar{\hat{\boldsymbol{\beta}}}\right)^2 \tag{36}$$

$$Bias(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\bar{\hat{\boldsymbol{\beta}}} - \boldsymbol{\beta}\right)^2}. \tag{37}$$

Before comparing the MSE, Bias and Variance we want to take a look at the resulting coefficients itself. These are given in Table 1. We can see that especially the estimated coefficients for β_1 using the CC analysis and MI differs extremely from the true data estimate. Whereas the estimate with MO is really close to the estimate with the true data. For a better illustration we plotted the densities of the estimators. These are shown in figures 2 and 3.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_t$
true data	1.0013	10.0017	4.9987	-7.9994	3.0023	1.0028	1.0003
CC	1.5623	5.2540	5.6976	-6.8395	4.0235	1.9995	0.9932
MI	1.2815	5.4526	5.9023	-7.0787	3.9026	1.9017	0.9181
MO	1.3310	9.6462	5.0919	-7.9281	2.9911	0.9834	0.8430

Table 1: Means of the Coefficients over all Simulations.

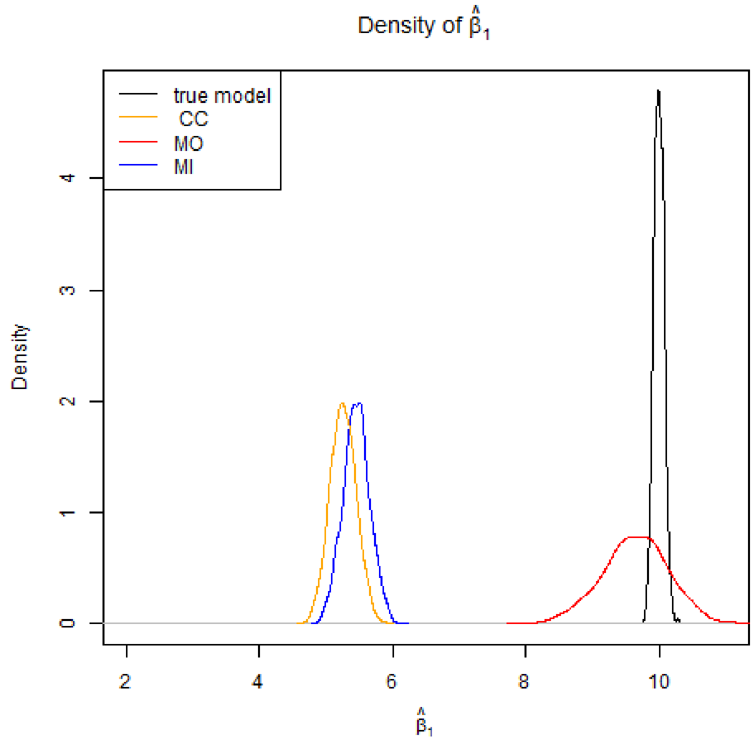


Figure 2: Density of β_1 for the true data, CC, MI and MO

Here we can see that on average the estimated β_1 using CC and MI is much lower than the other two. So the coefficient for the mismeasured covariate is underestimated when using the CC analysis. MO seems to perform much better although the variance is higher. For the other estimated coefficients (see figure 3) the CC analysis and MI overestimate the coefficient. The MO method leads again to higher variances but on average it estimates the coefficient better. To support this interpretation we calculated the MSE, the variance and the bias of the coefficients for all four analysis. The results are given in table A.1 in the appendix. We want to take a look at a plot of the MSEs since we think it is more intuitive than the values themselves (see figure 4). In the left part of figure 4 the MSE is plotted for the four different methods. It is remarkable that the MSE for $\hat{\beta}_1$ is extremely high if we use CC or MI. Since the other bars are comparatively small we cut the left graphic at MSE $= 2$. The result is shown in the right part of the same figure. If we

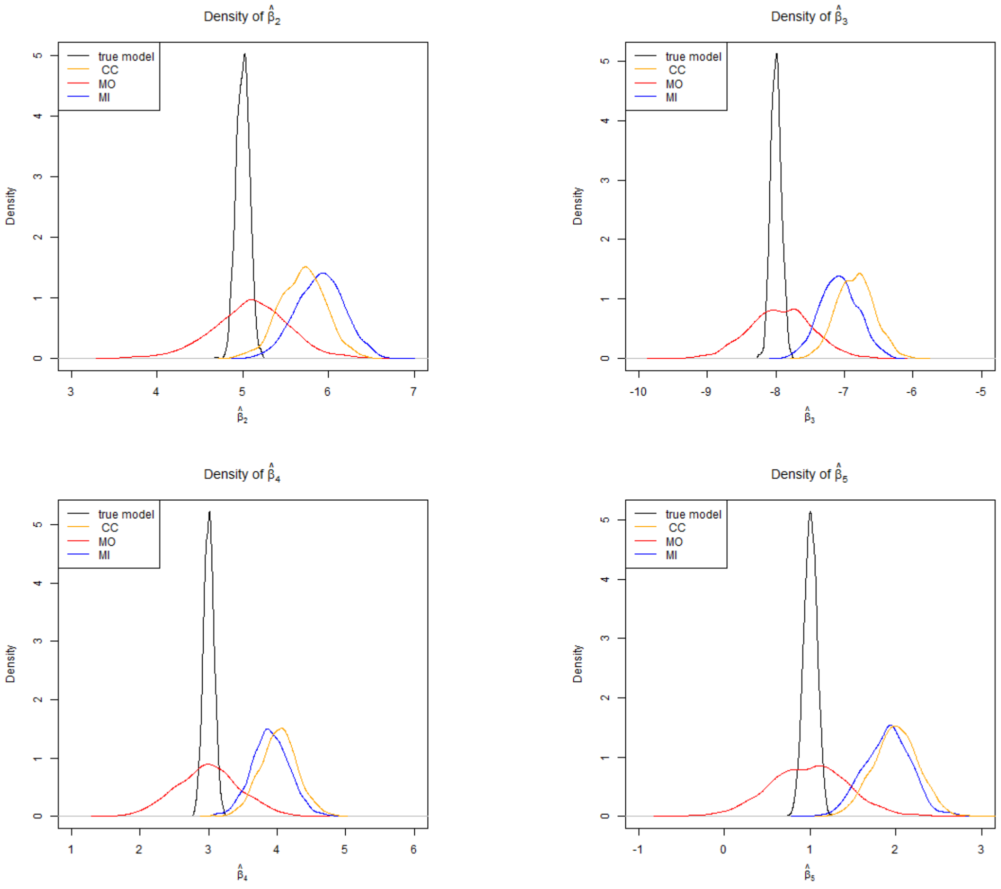


Figure 3: Density of β_2, β_3, β_4, β_5 for the true data, CC, MI and MO

use CC and MI, the MSE is always larger than with MO (except for β_t) and it takes the lowest values using the true data. The MSE with the true data is always around zero. That is how it should be and therefore it is a good indicator to show that we generated the data correctly. Using CC the measurement error and the missings are ignored and only the observed values and the mismeasured covariate are used in the analysis. This yields a strongly biased estimator. The same holds for MI with the difference that it corrects for missings but neither for measurement error and so the estimates are biased, too. To examine if the bias or the variance are responsible for the MSEs values we also calculated these two quantities (see figures 5 and 6).

The bias plot looks very similar to the MSE-plot for the CC and MI analysis - the bias for the five estimated covariate coefficients is always higher using MI and CC than using MO and lowest for the true data. The variance is very low using CC and MI. So we can conclude that the MSE for the CC and MI analysis mostly consists of the bias. The variance for the estimated coefficients using MO is nearly the same for the five baseline covariates and higher than for the other two methods but it is still small. For MO the

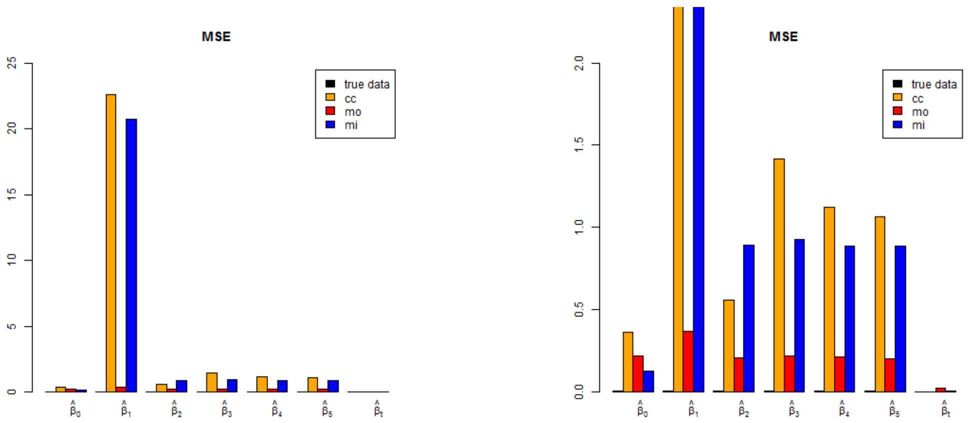


Figure 4: MSE of the estimated coefficients using the true data, CC, MI and MO

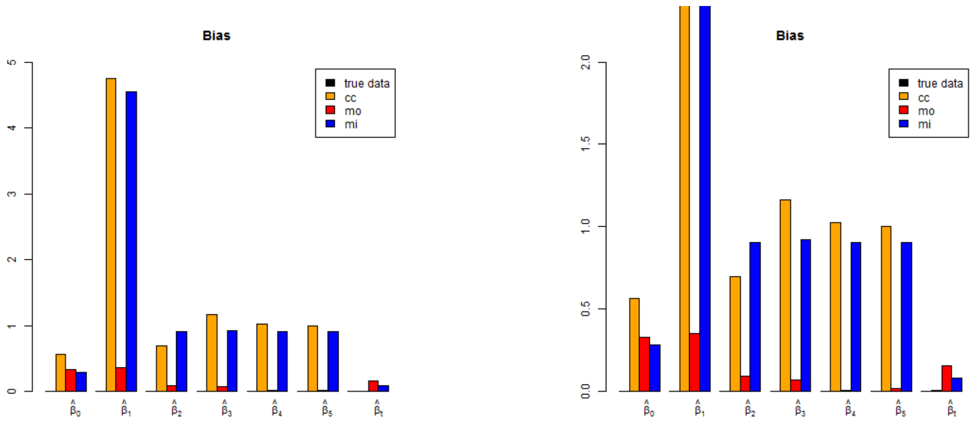


Figure 5: Bias of the estimated coefficients using the true data, CC, MI and MO

main part of the MSE of the baseline covariate coefficients is not the bias but the variance since the bias is smaller than 0.5 and so the squared bias which is contained in the MSE is even smaller. In this analysis MO performs much better than MI and CC in terms of bias and MSE. So we suppose if the assumptions we made here are fulfilled, for example that the data are multivariate normal, the measurement error proportion is known and it is not too high - here 33.3% - MO should be used instead of CC and even instead of MI. The improvement of the estimates is worth the additional computationally cost for using MO. Mostly the quantity of interest are the fixed effects even if the underlying model is a mixed model and so there are also random effects. Due to this interest we also concentrated on the fixed effects coefficients but we also want to take a look at the random effects. Our model only contains a random intercept for which we set the variance to 3. The estimated random intercept variances for the three analysis methods and the true data are given in figure 7. Using the CC analysis or MI the variance of the random intercept is

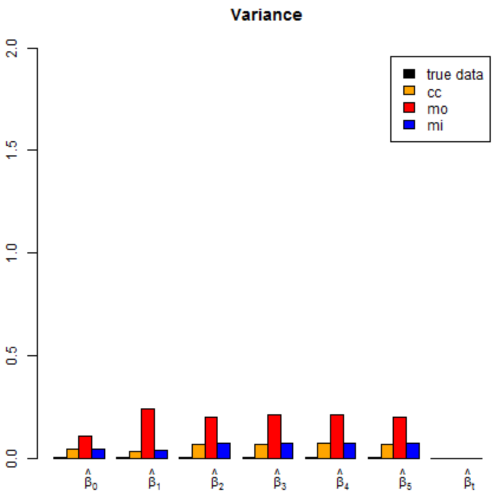


Figure 6: Variance of the estimated coefficients using the true data, CC, MI and MO

highly overestimated. Expectedly the random intercept variance is nearly three if the true data are analysed. Using MO the variance is also overestimated but not as much as with the two naive analysis methods. We conjecture that maybe a part of the measurement error variance is included in the variance of the random intercept since if we correct for measurement error (MO) the variance of the random intercept is much smaller as if we do not correct (CC, MI). Probably using MO the main part of the measurement error variance is removed due to the overimputation of the mismeasured values and only a rest is included in the random intercept variance. But if we do not correct for measurement error the whole measurement error variance is still present in the data and the naive methods maybe include this in the variance of the random intercept since this is the only way to deal with it. To support this thesis we compare random intercept variances for different amounts of measurement error in the sensitivity analysis.

In this section we made assumptions on the measurement error which are very strong since in reality we often do not know the exact amount of measurement error. In the following sensitivity analysis we want to examine if MO should still be preferred if we assume a measurement error which is higher or lower than the true one. As part of the sensitivity analysis we want to examine also why the variance is higher using MO. We suppose that the reason could be the additional variability due to the overimputation of the mismeasured values. This variability arises because we impute the missing values several times. If the additional variability is the reason for the higher variance, the variance should increase with the measurement error proportion. Moreover we want to vary the measurement error variance/ measurement error proportion itself, the missingness/drop-out probabilities and the number of measurements per subject.

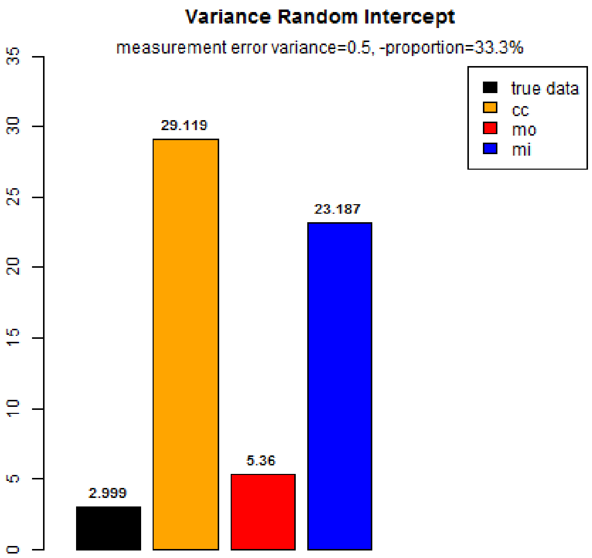


Figure 7: Variance of the random intercepts using the true data, CC, MI and MO

5.4 Sensitivity Analysis

We conduct a sensitivity analysis to examine if MO perfoms better than MI and CC for other scenarios, too. At first we want to vary the measurement error variance/the measurement error proportion. Additionally to the measurement error proportion of 33.3% (measurement error variance 0.5) we considered above, we now use the values pictured in table 2. For these four values we run the same simulation as described in the first two parts of this section. The only differences are the measurement error used to generate the mismeasured variable and the error proportion value inserted in the `moPrep()` function. The corresponding plots are given in A.2 in the appendix. With the smallest measurement error proportion (4.7%) the variance of the estimated baseline coefficients is much smaller than in the previous chapter where the measurement error proportion was higher. Moreover the difference between the naive approaches (CC, MI) which ignore the mea-

σ_u^2 (measurement error variance)	$\frac{\sigma_u^2}{\sigma_x^2+\sigma_u^2}$ (measurement error proportion)
0.05	4,7$\%$
0.1	10$\%$
1	50$\%$
2	66,7$\%$

Table 2: Different values for measurement error variance and measurement error proportion used in the sensitivity analysis.

surement error (and the missings) and the MO approach is much smaller, too. But for the mismeasured covariate the densities of the estimates with CC and MI still not overlay the density of the estimate with the true data. For the other covariates the difference between the densities for the naive and MO approaches is not clear anymore. The MSE, bias and variance plots clarify this result. The MSE and bias are much lower for the naive approaches and for MO, too. The variance is very small for all approaches. So if the measurement error proportion is very small it seems that MO would still yield better results for the mismeasured variable and that the higher variance with MO in general may occur because the mismeasured variable is less informative if the measurement error proportion increases. To support this conjecture we want to describe the results with the other mentioned amounts of measurement error. The variance of the estimated coefficient increases with increasing measurement error proportion and the distances between the densities of the estimated coefficients for the naive methods and MO get higher, too. Even with a measurement error proportion of 66.7% the MSE and bias for the mismeasured variable using MO is still much lower than with CC or MI. For the other baseline covariates the MSE and bias are lower, too if MO is used instead of MI or CC. The densities for the mismeasured variable using MO covers the true value of β_1 in all cases. Using CC or MI the true value is not covered. For the other baseline covariates and low measurement error proportion all three methods (CC, MI, MO) cover the true value but MO is always better since the peak of the corresponding density is very near to the true value. The other densities are shifted to the right. The shift is getting larger with increasing amount of measurement error so that with a measurement error proportion of 33.3% (see5.3) or larger the densities using MI and CC does not cover the true value. The MSE and bias also support the conjecture that MO should be preferred if a covariate is mismeasured since bias and MSE are always lower or the same using MO as if we were using CC or MI. So if we have longitudinal data with missings, drop-out and a mismeasured covariate whose measurement error proportion is known and not larger than $2/3$ we would recommend to use MO. For a higher measurement error proportion we cannot make a statement since we did not tested it.

A question that may arise is what happens if we do not know the exact measurement error proportion and assume a value which is too high or too low. To examine this problem we run our simulation again but we change the measurement error proportion which is inserted in the `moPrep()` function to generate the priors for MO. The true measurement error proportion is always 33.3% which corresponds to a measurement error of 0.5 since the variance of the true $\boldsymbol{x}_1$ does not vary. Therefore this measurement error of 0.5 is used to generate the mismeasured variable in all six cases shown in table 3. So we run six simulations with different assumed measurement error proportions used for MO. Using

true ME proportion	assumed ME proportion	under-/overestimation
33.3%	4.7%	- 28.6 %-points or $0.14 \cdot 33.3$
33.3%	10%	- 23%-points or $0.3 \cdot 33.3$
33.3%	20%	- 13%-points or $0.6 \cdot 33.3$
33.3%	40%	+ 17 %-points or $1.2 \cdot 33.3$
33.3%	50%	+ 27 %-points or $1.5 \cdot 33.3$
33.3%	66.7%	+ 33%-points or $2 \cdot 33.3$

Table 3: Different values for the measurement error (ME) proportion used in the sensitivity analysis.

the measurement error proportions given in the first three rows of table 3, a lower measurement error proportion is assumed than the true value. In the other three scenarios the measurement error proportion is overestimated. The plots for this simulations are shown in appendix A.3. The under-/overestimation of the measurement error proportion leads to a shift of the MO density. If the measurement error is underestimated, the density for $\hat{\beta}_1$ is shifted to the left and the shifts gets larger the higher the underestimation of the measurement error proportion. The result is that the density of the coefficient of the mismeasured variable using MO is very close to the densities using MI and CC if the assumed measurement error proportion is nearly 29%-points lower than the true proportion which is one seventh of it. This is intuitive since MI is the extreme case of assuming a measurement error proportion of zero. The variance of the estimates remains the same but the MSE and bias increase for MO. For MI and CC the MSE and bias are the same as with known measurement error since these approaches ignore the measurement error. For the other baseline covariates the densities of the estimated coefficients shift to the right to compensate the shift to the left of the density of $\hat{\beta}_1$. Using MO the bias and MSE increase also if the difference between the assumed (underestimated) and true measurement error proportion gets higher. So if the measurement error proportion is underestimated, this may have a huge influence on the estimates and the corresponding bias/mse. Even if the assumed measurement error proportion is greater than the half of the true proportion the density for β_1 using MO does not cover the true value but is shifted far to the left.

To detect if maybe overestimation of the measurement error should is harmless than underestimation, we also examine the influence of overestimating the measurement error proportion. If the measurement error proportion is assumed to be higher than the true unknown value, the density of $\hat{\beta}_1$ using MO shifts to the right, in contrast to the case of underestimating the measurement error proportion. The greater the overestimatation the greater is the shift to the right. But even if the assumed measurement error proportion is twice as high as the true value, the density of $\hat{\beta}_1$ using MO covers the true value very good since the shift is only small. In comparison to the case where the true value is the assumed value the MSE and bias increase but the variance changes barely. If the true and assumed measurement error proportion is 66.7% the variance is much higher. (It seems that the assumed value does not have a great influence on the variance.) The density of the other baseline covariate coefficients using MO are shifted to the left but also just a little bit. Here the variance increases a bit more as for the $\hat{\beta}_1$ and the MSE and bias increase as well. If we compare the under- and overestimation it is remarkable that the overestimation seems to have a lower influence on the MSE and bias of the estimates. So we would recommend to rather overestimate than underestimate the measurement error proportion. With overestimation the results are still very good and seem to be reliable. This was perhaps the most interesting part of the sensitivity analysis since often the measurement error proportion is not known exactly but the results let us suggest that even if the measurement error proportion is not estimated or assumed correctly the coefficients may be reliable.

Besides varying the measurement error proportion and the assumption about it, we also changed the number of measurements per subject and hold the measurement error proportion constant at 33% to examine if this influences the results. We run the simulation for two additional numbers of measurement per subject: 2 and 10. The difference between the results is very small and we cannot identify any patterns.

The last part of this simulation study addresses the influence of the missing data/drop-out probability. In addition to the missingness/drop-out probability used in the study in main simulation (missingness probability around 10%, drop-out probability around 30%), we run simulations with the values shown in table 4. The corresponding plots of the results are given in appendix A.5. The differences between the results is very small if we vary the amount of drop-outs. But if we vary the missingness probability, changes are identifiable. With the higher missingness probability (30%), the density of $\hat{\beta}_1$ is shifted to the left using MO, using CC or MI only the variance increases a little bit. For the other coefficients the density shifts a little bit to the right using MO. The Bias and the variance for the baseline coefficients increases using MO and therefore the MSE, too. But no matter what missingness or drop-out probability is set, MO performs always better in terms of bias and MSE.

missingness probability (covariates)	drop-out probability (response)
around 10%	around 15%
around 30%	around 15%
around 30%	around 30%

Table 4: Different values for measurement error variance and measurement error proportion used in the sensitivity analysis.

In the previous parts of the sensitivity analysis, we only concentrated on the estimated baseline coefficients (fixed effects) and their bias, MSE and variance. Now we want to take a look at the variance of the random intercept for different measurement error variances /proportions. In section 5.3 we constructed the hypothesis that if we use CC, MI and MO, a part of the measurement error variance will be included in the random intercept variance. To support this thesis we also calculated and plotted the random intercept variance for the measurment error proportions used in the sensitivity analysis. The plots are given in appendix A.6. Please be careful since the y-axis have different length. With increasing measurement error variance the variance of the random intercept increases, too, using CC, MI and MO. Expectedly the variance of the random intercept using the true data is always around three. Using MO the variance increases only slightly, with the lowest value of measurement error (0.05) and a measurement error of 0.1 it is around four, if the measurement error is 1, the variance is 6.6 and with the highest measurement error it is 8.5. So the random intercept variance is 2.3 times higher with the largest measurement error variance than with the lowest. In comparison using the naive approaches the random intercept variance is 6.3 (MI) or 7.5 (CC) for the lowest measurement error variance. So even with a very small measurement error variance it is much higher than with MO. With the highest amount of measurement error the random intercept variance is around six times larger (37 with MI and 46 with CC) than with the lowest measurement error variance and much larger than with MO. This results support our thesis that a part of the measurement error variance will be included in the random intercept variance. So if data are subject to measurement error the random intercept variance is overestimated and this overestimation increases with increasing measurement error variance.

6 Discussion

The combined problem of measurement error, missing data often arises in empirical longitudinal studies but most analysts only correct for missing data since there was no implemented easy-to-use method. We conducted a simulation study to examine if the multiple overimputation approach proposed by Blackwell et al. (2015a,b) can be used to simultaneously correct for missing data and measurement error if the dataset is longitudinal. We generated multivariate normally distributed data with baseline covariates and a longitudinal response where some covariates are subject to missingness and one is subject to both, missingness and measurement error. The response is subject to dropout. We analysed this dataset using multiple overimputation combined with a linear mixed model with random intercept and compared the estimated baseline coefficients to those which we obtained using multiple imputation or a complete case analysis and to the analysis results for the linear mixed model with random intercept of the true complete dataset. We also conducted a sensitivity analysis where we examine the influence of different mea-

surement error proportions, different numbers of measurements per subject and different missingness and drop-out probabilities on the estimated baseline coefficients. Moreover we relaxed the assumption of a known measurement error proportion in the sensitivity analysis and assumed different values for the measurement error proportion.
For all chosen known measurement error proportions MO performs better than the naive approaches in terms of bias and MSE of the estimated baseline coefficients. With increasing measurement error proportion, bias and variance increase if MO is used and therefore the MSE increases, too. Using the naive approaches the MSE increases too but especially because of an increasing bias. The variance increases only slightly. Even if the measurement error proportion is above 50%, which was named as critical border by Blackwell et al. (2015a), the true values of the coefficients are met in some simulation runs and therefore the density of the estimated baseline coefficients covers still the true value if MO is used (although such a high measurement error porportion would be hard to justify in practice (Blackwell et al. (2015a))). If instead the naive approaches are used the true coefficient value of the mismeasured variable is not even covered for the smallest tested measurement error proportion. So MO should be preferred if the measurement error is known and the measurement error proportion seems to have an influence on the reliability of the estimates. The drop-out and missingness probabilities could also influence the results. If the missingness probability increases, the MSE increases, too. This increase of the MSE can be ascribe to the increase of the bias since the variance increase is very low. But the true value is always covered if MO is used. In comparison the drop-out probability and the number of measurements per subject seems to have slightly no influence on the estimated baseline coefficients. Since we estimated a LMM with random intercept we also obtain estimates for the random intercept variance. The random intercept variance increases with increasing measurement error proportion and therefore we conjecture that a part of the measurement error variance is included in the random intercept variance due to the analyses and so the random intercept variance is overestimated even if MO is used. But the overestimation is much lower with MO than with MI or CC. Several assumptions are necessary for these analyses. One is that the measurement error proportion is known or estimated exactly which is no realistic assumption. Therefore we also conducted a sensitivity analysis where we varied the assumed measurement error proportion for a fixed true measurement error proportion. The result is that MO is never worse than CC or MI but if the assumed measurement error variance is highly underestimated the results using MO, MI and CC does not differ very much but MO is still slightly better for the mismeasured variable. In comparison if the measurement error is highly overestimated, MO still performs very well since the true coefficient value is covered even for the mismeasured variable and therefore the MSE is less higher than with a highly underestimated

measurement error proportion. So we recommend to rather assume a measurement error proportion which is too high than to underestimate it.

Besides the measurement error and the missing data we created a perfect dataset where all variables are multivariate normally distributed which is a necessary assumption for using MO but according to Blackwell et all. (2015b) it is also robust to categorical variables measured with error. Moreover we created the missings drop-out so that the data are missing at random which is also necessary for using MO. In reality there are also often datasets which are not missing at random. The performance of MO facing not missing at random data was not tested yet but we assume that it will not work well since a specification of the missingness process is not possible. The measurement error probability is also not allowed to depend on the missing values if MO is used.

Since, to our knowledge, the performance of multiple overimputation was not tested before for longitudinal data, we generated a very simple multivariate normally distributed longitudinal dataset with five baseline covariates and one response. Only one covariate is subject to measurement error. Since now we know that MO performs well for this simple scenario, more complicated simulation studies could be conducted, for example with categorical variables or with more than one mismeasured variable, especially the response. Moreover we estimated a linear mixed model with random intercept, only. So it would be interesting to test the MO approach for a model with random slope, also. Other distributions for the response and therefore other models like generalized mixed models could also be part of further research.

The result of the conducted simulation study is that MO is a good method for dealing with measurement error and missing data in a longitudinal dataset at least if the necessary assumptions are fulfilled. In contrast to the coefficients the random intercept variance needs to be handled carefully since it increases with increasing measurement error variance. The coefficients are estimated very well with a moderate measurement error (a third of the whole variance of the variable). Even if the measurement error proportion is not known exactly, MO can perform very well regarding the coefficients, especially if the measurement error variance is assumed to be higher than the true one. With these findings, we recommend to rather overestimate the measurement error proportion since a high underestimation yields nearly the same results for the baseline coefficients as MI or CC. A big advantage of MO, besides its good performance, is that it is implemented in R and therefore easy-to-use. With this method, it is much easier than before to correct for both measurement error and missing data. Therefore we hardly recommend to correct for measurement error, too, because it improves the results a lot, especially if the measurement error is vaguely known.

Literature

Bityukov SI, Smirnova VV and Taperechkina VA (2008): Statistically dual distributions and estimation of the parameters

Blackwell M, Honaker J, King G (2015a): A Unified Approach to Measurement Error and Missing Data: Overview and Applications, *Sociological Methods & Research*, **1-39**

Blackwell M, Honaker J, King G (2015b): A Unified Approach to Measurement Error and Missing Data: Details and Extensions, *Sociological Methods & Research*, **1-28**

Buonaccorsi JP (2010): Measurement Error Models, Methods and Applications *Chapman & Hall/CRC*

Carroll RJ, Rupert D and Stefanski LA (1995): Measurement Error in Nonlinear Models, *Chapman & Hall/CRC*, 2nd Edition

COOK J, STEFANSKI LA (1995): A simulation extrapolation method for parametric measurement error models, *J. Am. Statist. Assoc.*, 89, 1314–28.

Dempster AP, Laird NM, Rubin DB (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38.

Fuller WA (1987): Measurement Error Models, *Joahn Wiley & Sons*

Hansen (1982): Large sample properties of generalized method of moments estimators, *Econometrica*, 50, **1029-1054**

Honaker J, King G (2010): What to do about Missing Values in Time-Series Cross-Section Data, *American Journal of political Sciences*, 54(2), **561-581**

Honaker J, King G, Blackwell M (2011): Amelia II: A Program for Missing Data, *Journal of Statistical Software*, 45(7), **1-47**, http://www.jstatsoft.org/v45/i07/.

Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005): Missing-data methods for generalized linear models: A comparative review, *Journal of the American Statistical Association*, 100, 332-346.

Ibrahim JG, Chu H, Chen MH (2012): Missing Data in Clinical Studies: Issues and Methods, *Jounal of Clinical Oncology*, 30(26), 3297-3303.

Ibrahim JG and Molenbergh G (2009): Missing data methods in longitudinal studies: a review, *Test (Madr.)*, 18(1), 1-43.

Imai K, King G, and Lau O (2007): Zelig: Everyone's Statistical Software, `http://GKing.harvard.edu/zelig`.

Imai K, King G, and Lau O(2008) Toward A Common Framework for Statistical Analysis and

Development.", *Journal of Computational and Graphical Statistics*, 17(4), 892-913

King G, Honaker J, Joseph A, Scheve K (2001): Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation, *American Political Science Review*, 95(1, March), 49–69.

Lipsitz SR, Ibrahim JG and Zha LP (1999): A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood, *Journal of the American Statistical Association*, 94, 1147-1160

Little R, Rubin D (2002): Statistical Analysis with Missing Data, *Wiley Series in Probability and Statistics.*

Liu W, Wu Lang (2007): Simultaneous Inference for Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors and Missing Responses, *Biometrics*, 63, 342–350.

Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G (2014): Handbook of Missing Data Methodology, *Chapman & Hall/CRC*

Rubin D (1987): Multiple Imputation for Nonresponse in Surveys, *New York, Wiley.*

Robins JM, Rotnitzky A, Zhao LP (1995): Analysis of Semiparametric Regression Models for Outcomes in the Presence of Missing Data, *Journal of American Statstical Association*, 90(429), 106-121.

Schafer JL (1997): Analysis of Incomplete Multivariate Data, *London: Chapman and Hall.*

Schafer JL, Olsen MK (1998): Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective, *Multivariate Behavioral Researchs*, 33(4), 545–571.

Verbeke G, Molenberghs G (2000): Linear Mixed Models for Longitudinal Data, *Springer Series in Statistics.*

Yi GY (2008): A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates, *Biostatistics*, 9(3), 501–512.

Yi GY, Liu W, Wu Lang (2011): Simultaneous Inference and Bias Analysis for Longitudinal Data with Covariate Measurement Error and Missing Responses, *Biometrics*, 67, 67–75.

Yi GY, Ma Y, Carrol RJ (2012): A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error, *Biometrika*, 99(1), 151–165.

A Tables and Figures

A.1 MSE, Bias and Variance for the estimated coefficients using the true data, CC and MO

	β_0	β_1	β_2	β_3	β_4	β_5	β_t
TD MSE	0.00504	0.00573	0.00563	0.00573	0.00555	0.00563	0.00018
TD Var	0.00504	0.00573	0.00562	0.00573	0.00554	0.00562	0.00018
TD Bias	0.00130	0.00173	0.00131	0.00064	0.00234	0.00276	0.00039
CC MSE	0.36474	22.5617	0.55594	1.41887	1.12022	1.06630	0.00038
CC Var	0.04859	0.03718	0.06929	0.07221	0.07272	0.06733	0.00033
CC Bias	0.56227	4.74600	0.69760	1.16045	1.02348	0.99948	0.00677
MI MSE	0.12648	20.7177	0.89034	0.92603	0.88836	0.88584	0.00753
MI Var	0.04725	0.03926	0.07613	0.07726	0.07359	0.07285	0.00082
MI Bias	0.28148	4.54735	0.90234	0.92129	0.90265	0.90166	0.08194
MO MSE	0.22154	0.36986	0.21007	0.21783	0.21201	0.20404	0.02511
MO Var	0.11201	0.24467	0.20162	0.21266	0.21193	0.20376	0.00048
MO Bias	0.33096	0.35383	0.09192	0.07191	0.00890	0.01664	0.15696

A.2 Different measurement error proportions

A.2.1 measurement error proportion: 4.7%

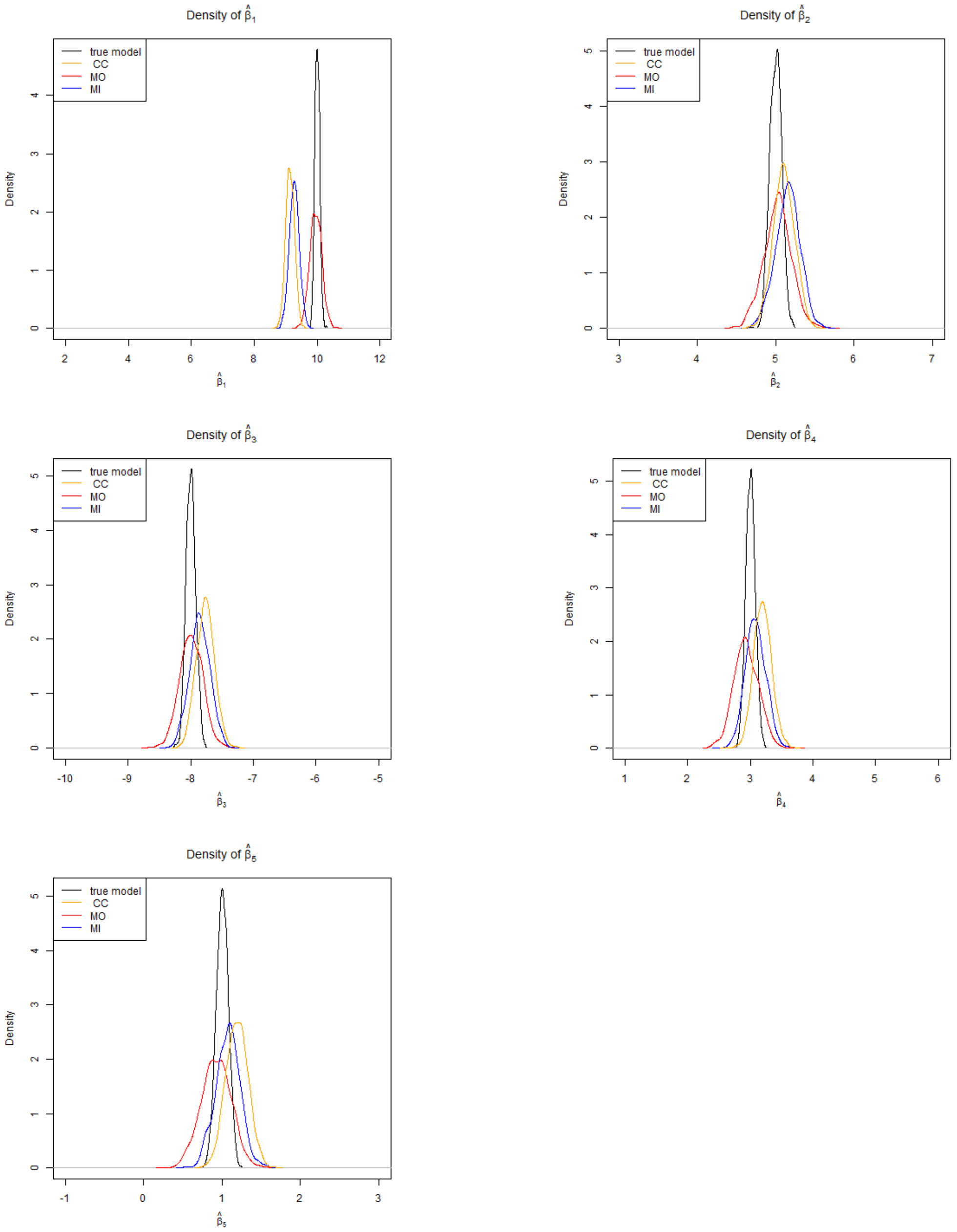


Figure 8: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is 4.7%

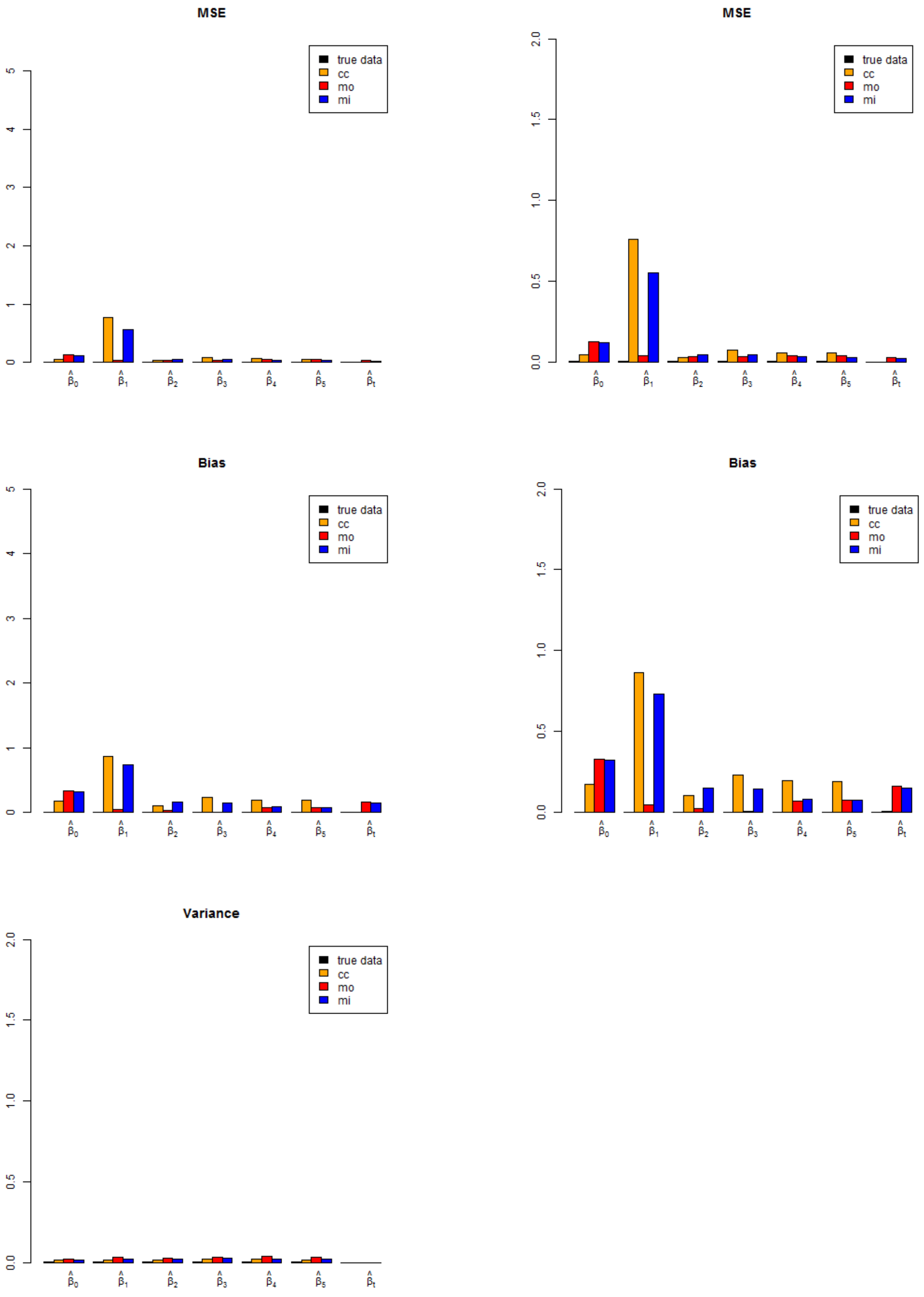


Figure 9: MSE. Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is 4.7%

A.2.2 measurement error proportion: 10%

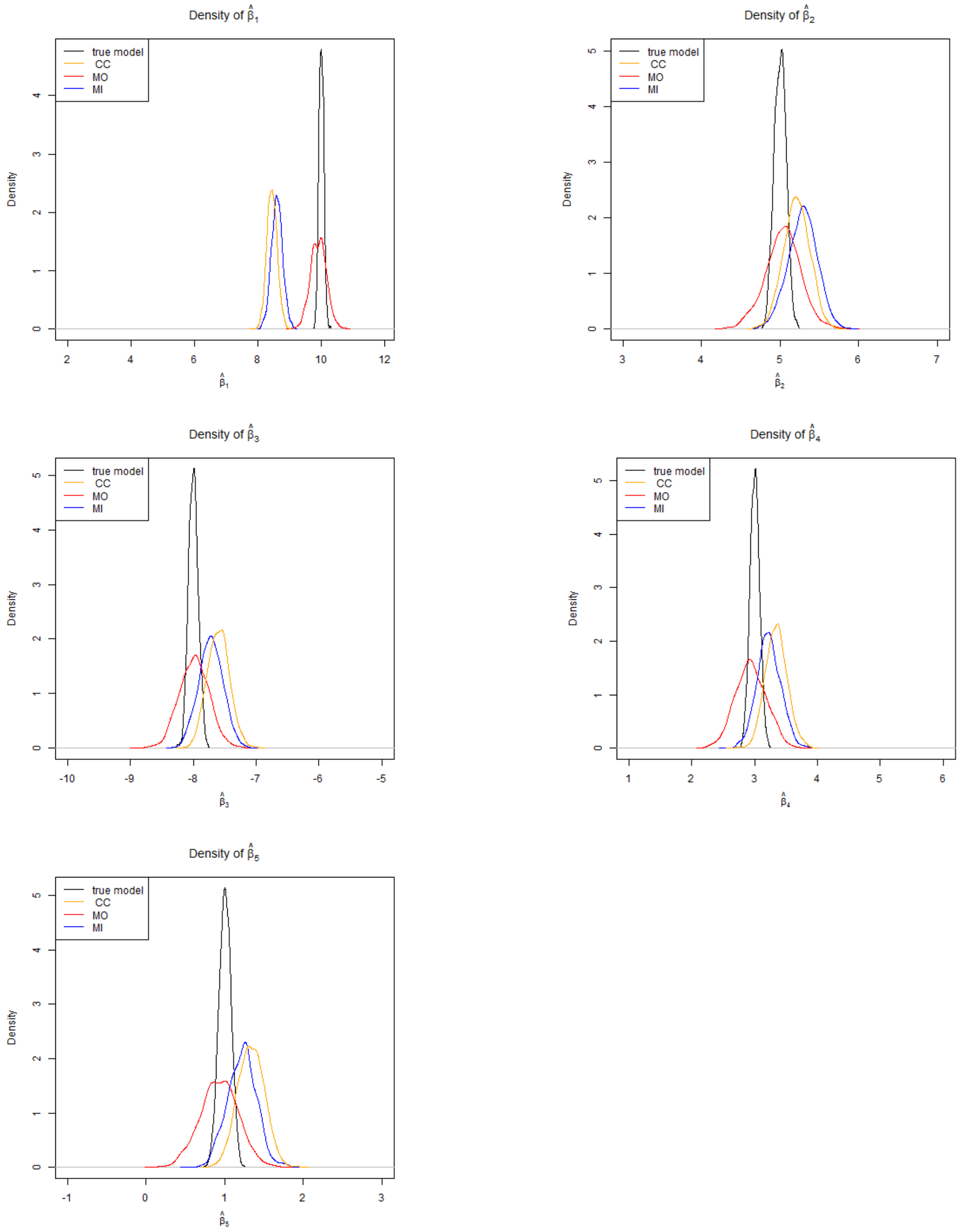


Figure 10: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is 10%

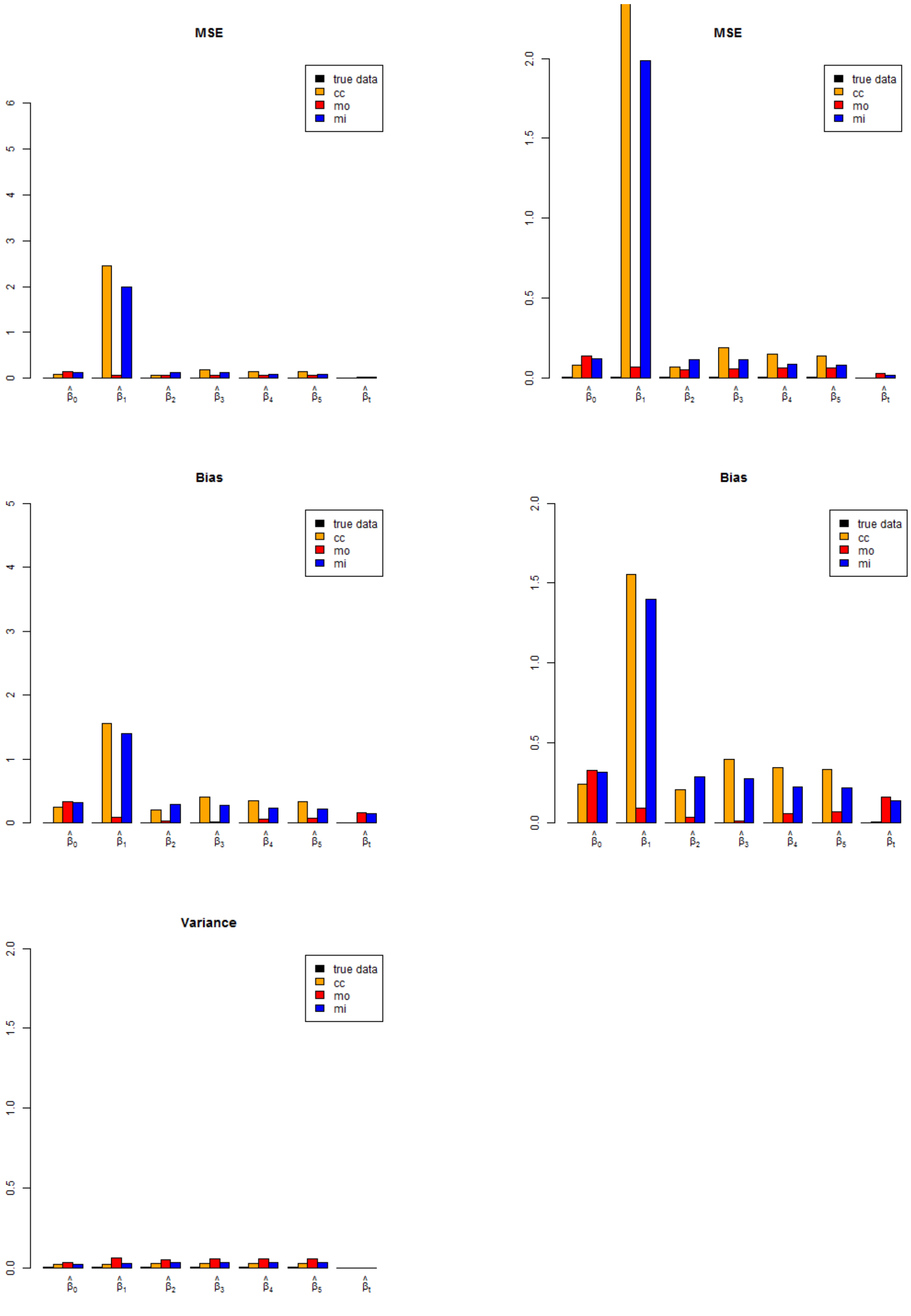


Figure 11: MSE, Bias, Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is 10%

A.2.3 measurement error proportion: 50%

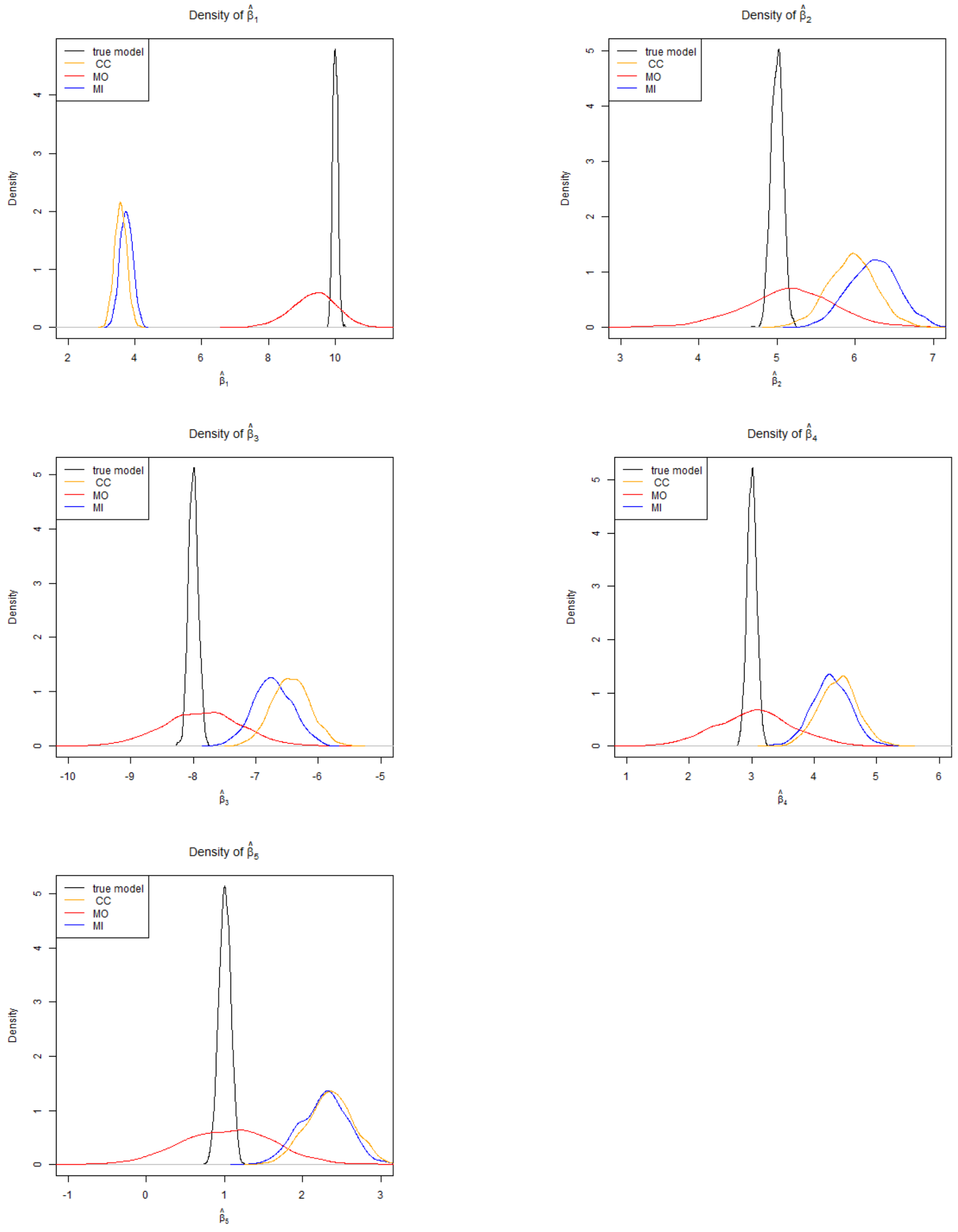


Figure 12: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is 50%

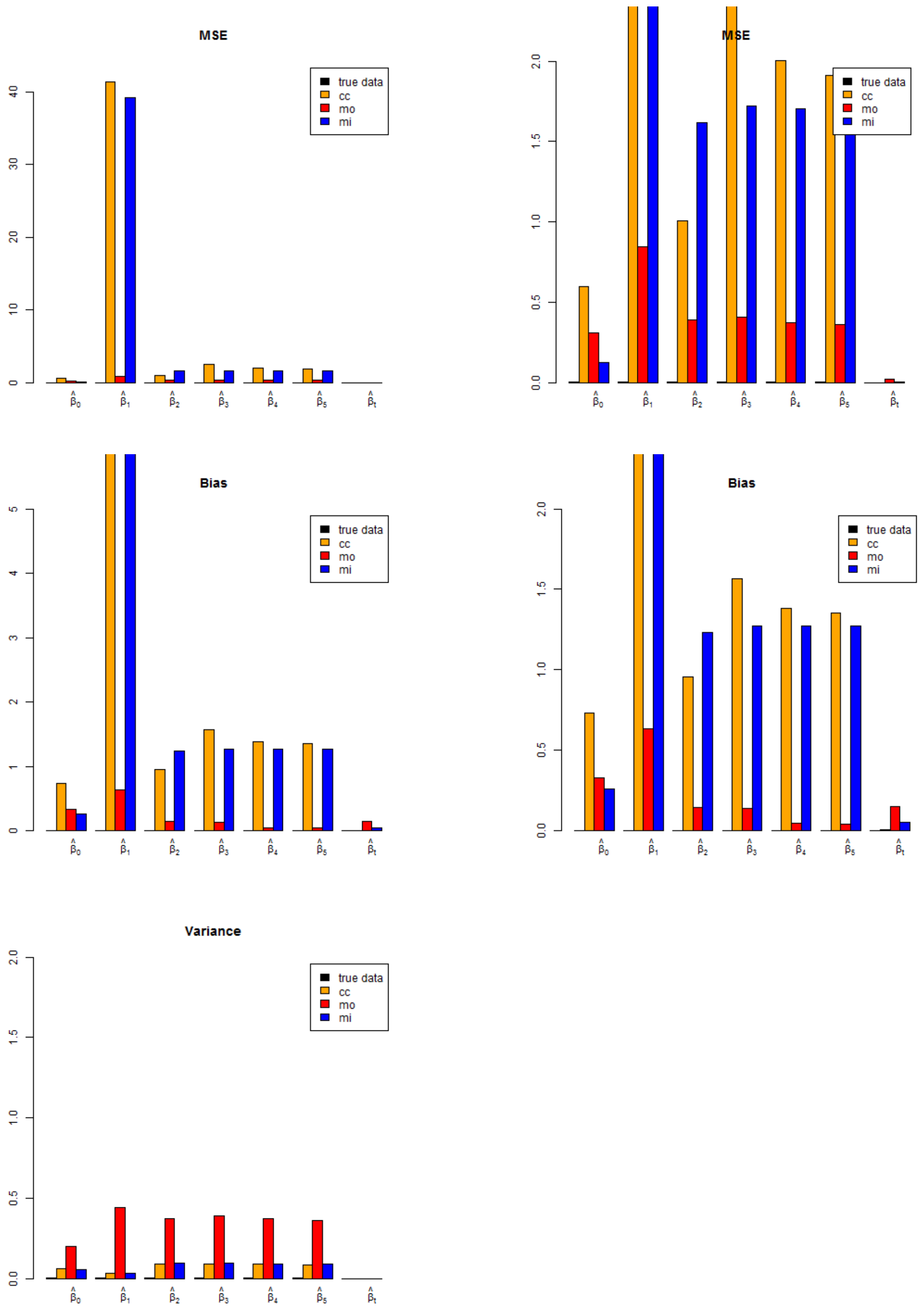


Figure 13: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is 50%

A.2.4 measurement error proportion: 66.7%

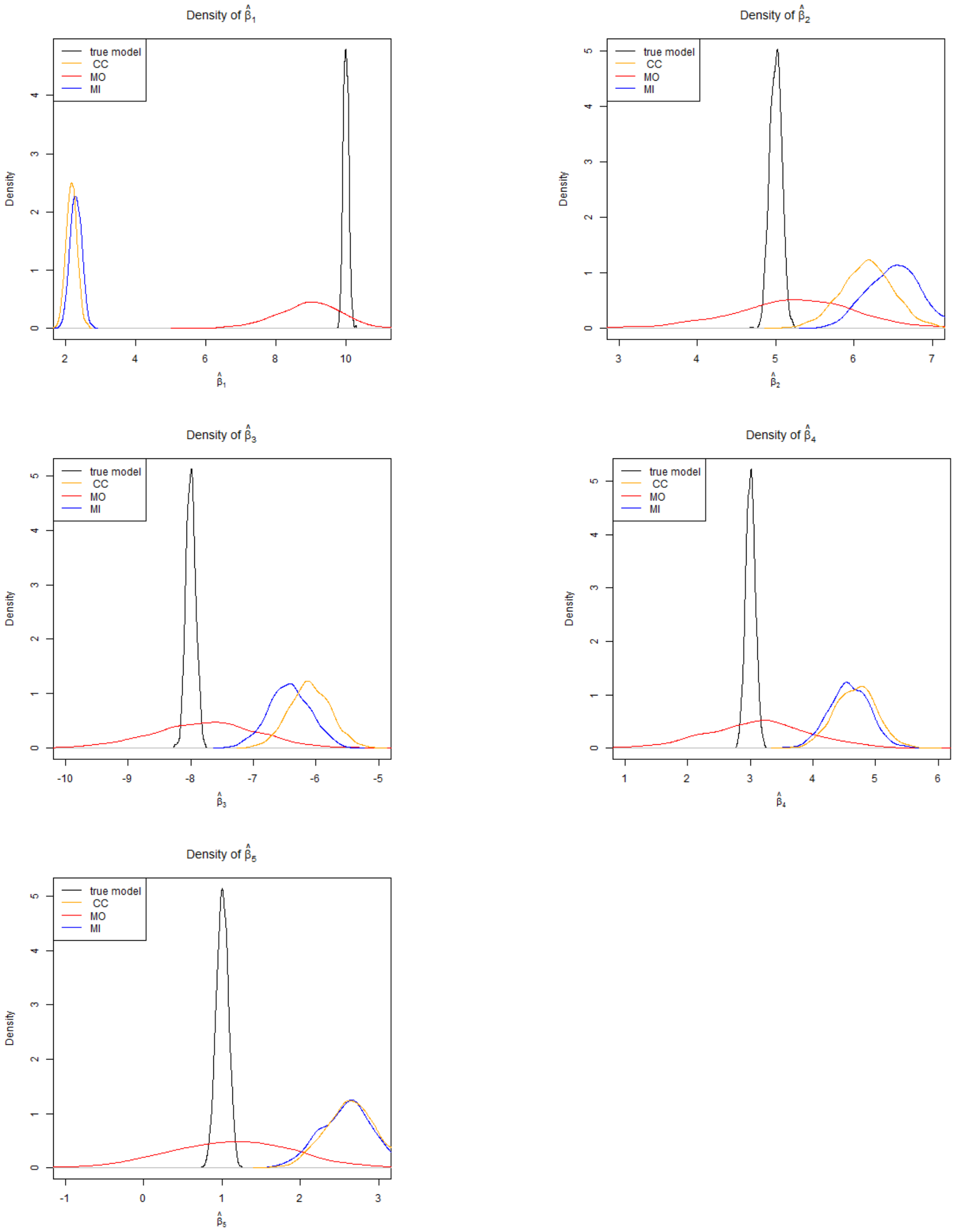


Figure 14: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is 66.7%

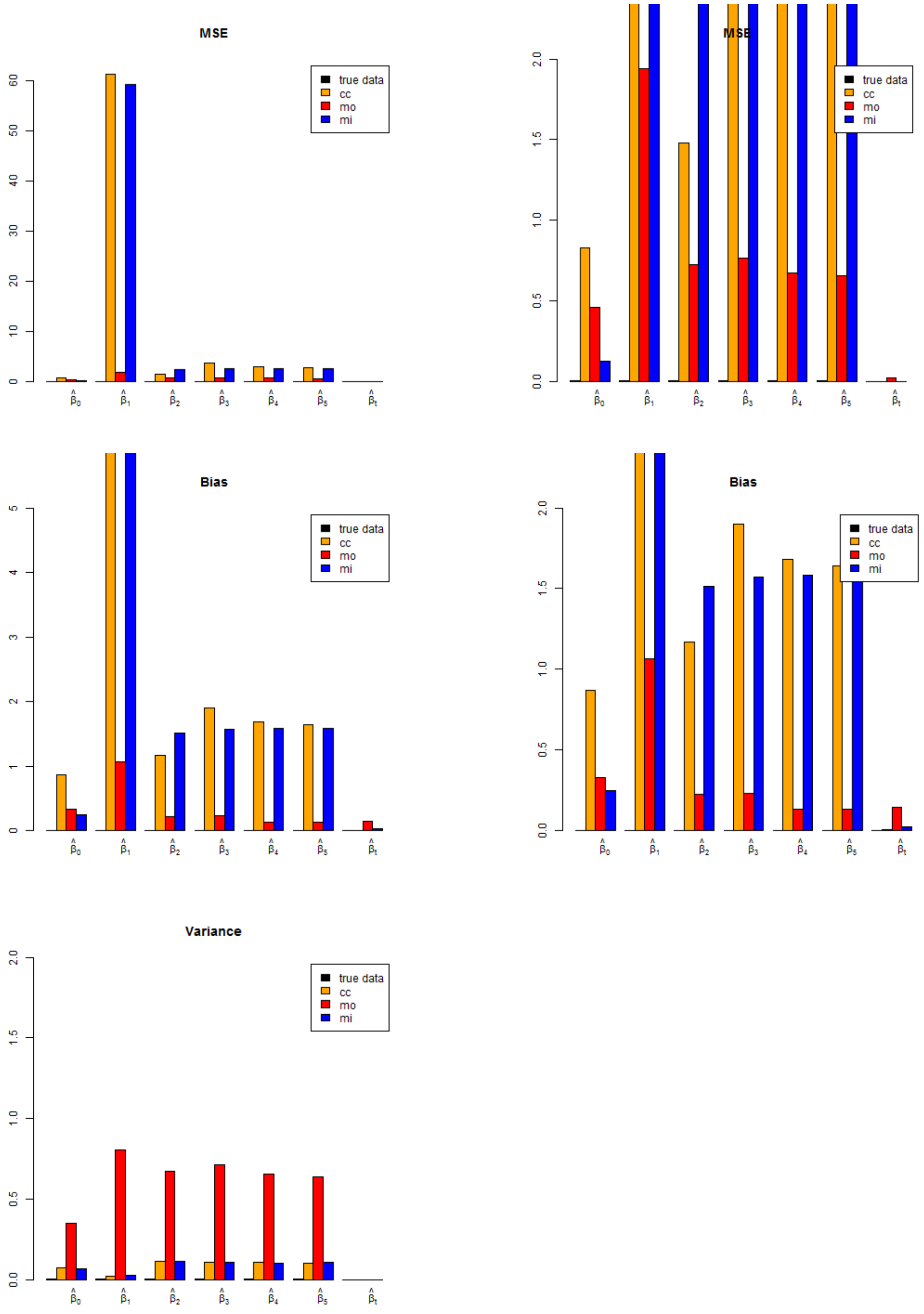


Figure 15: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is 66.7%

A.3 Unknown measurement error proportion (true measurement error proportion: 33.3%)

A.3.1 Assumed measurement error proportion: 66.7%

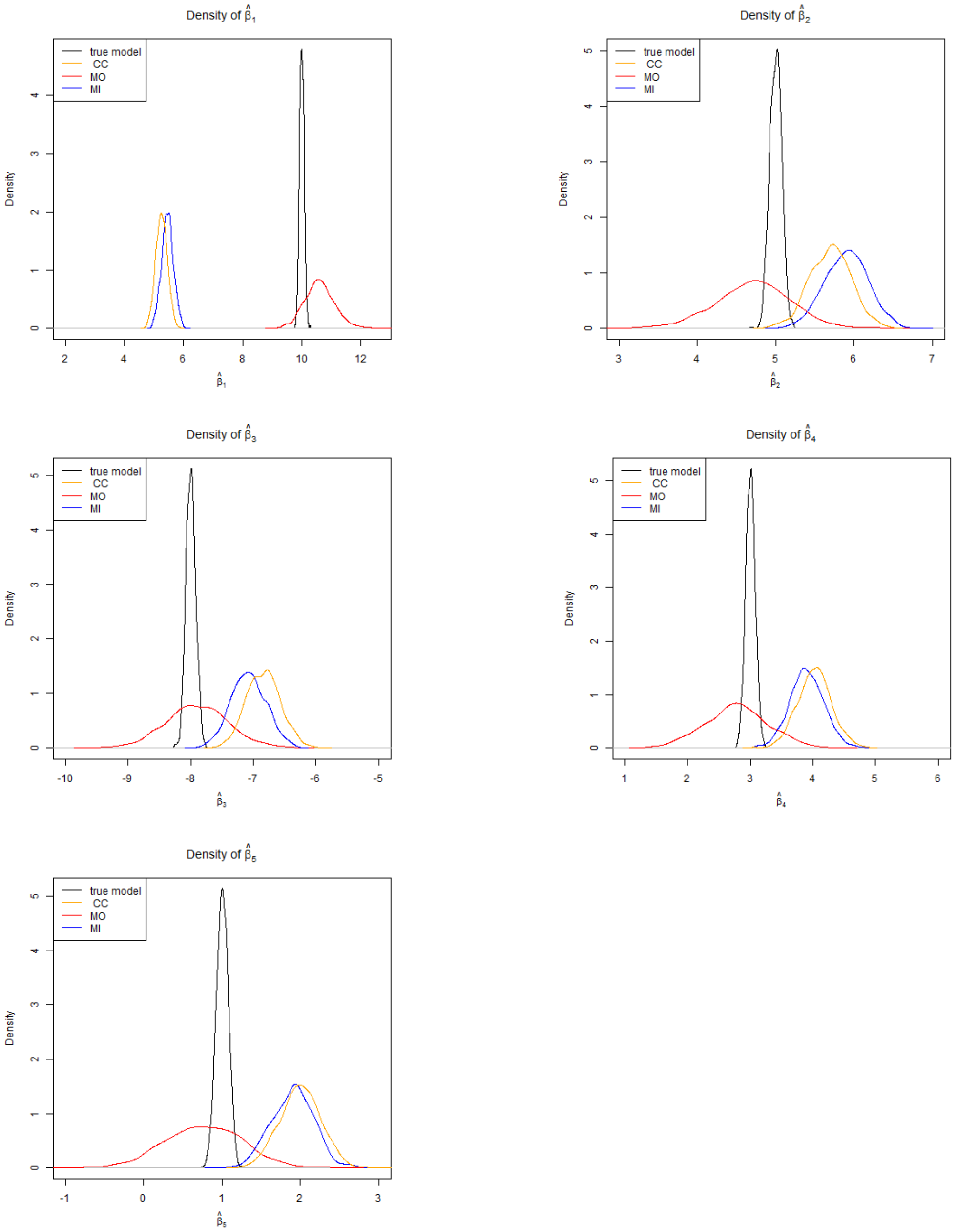


Figure 16: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 66.7% instead of 33.3%

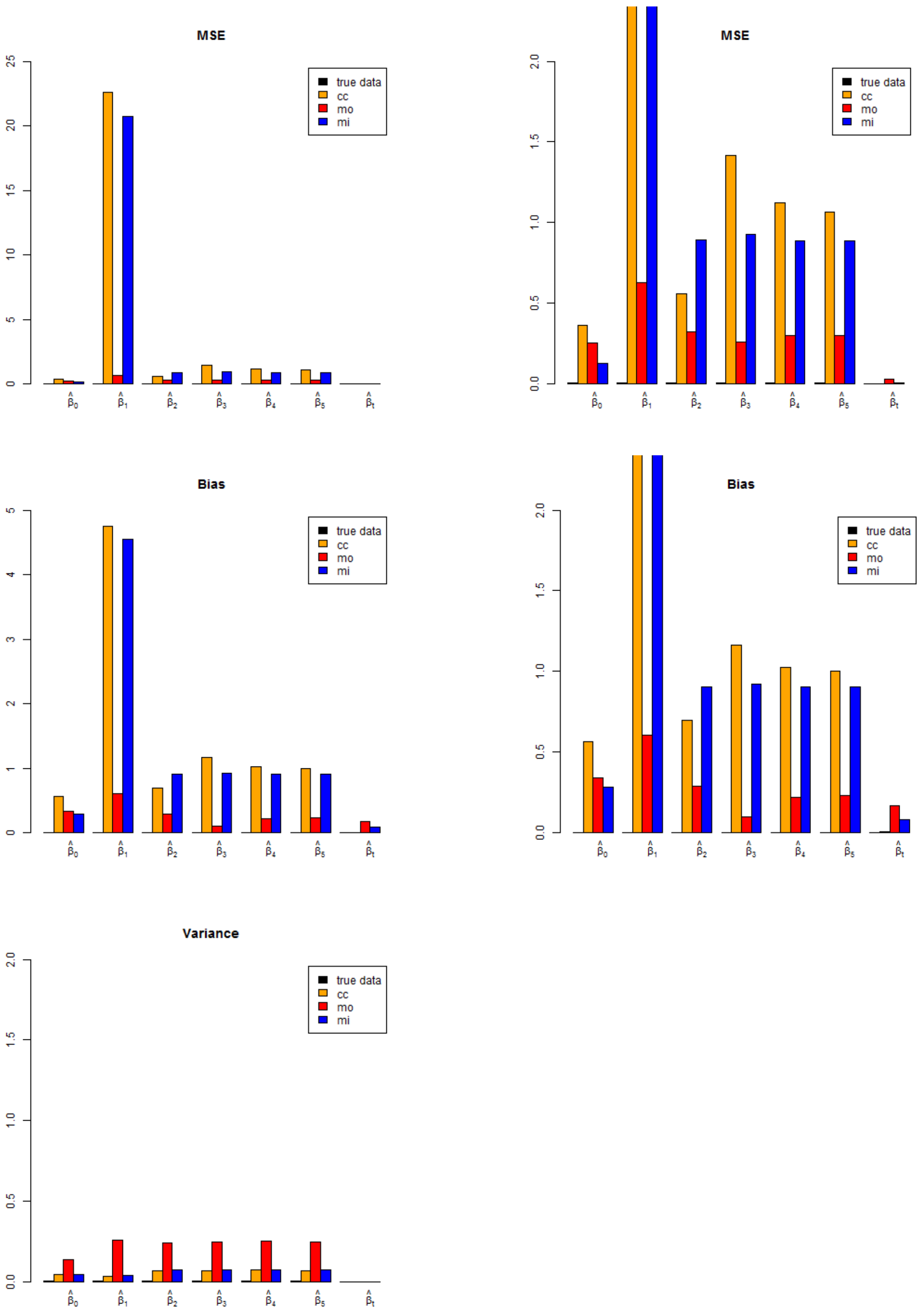


Figure 17: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 66.7% instead of 33.3%

A.3.2 Assumed measurement error proportion: 50%

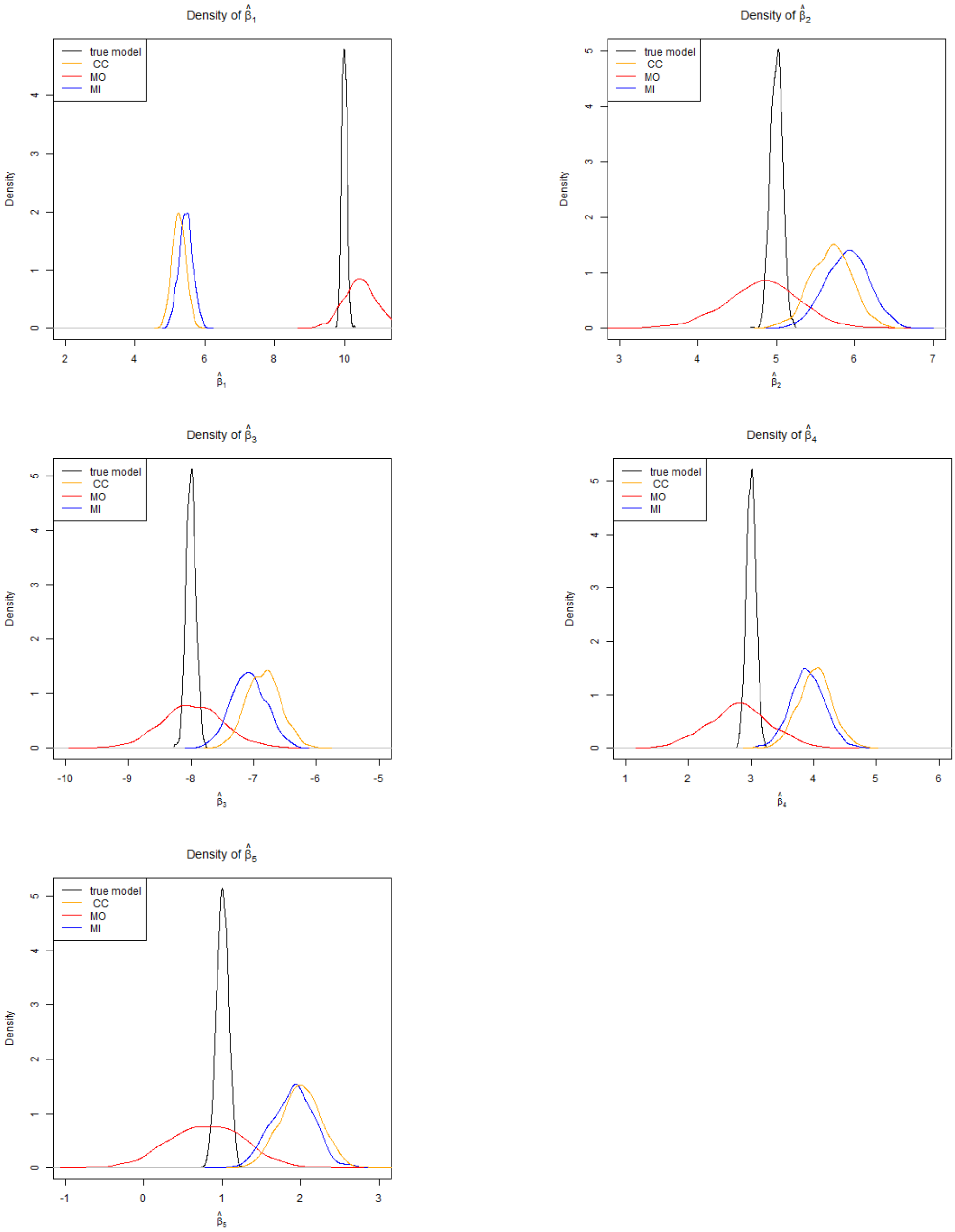


Figure 18: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%

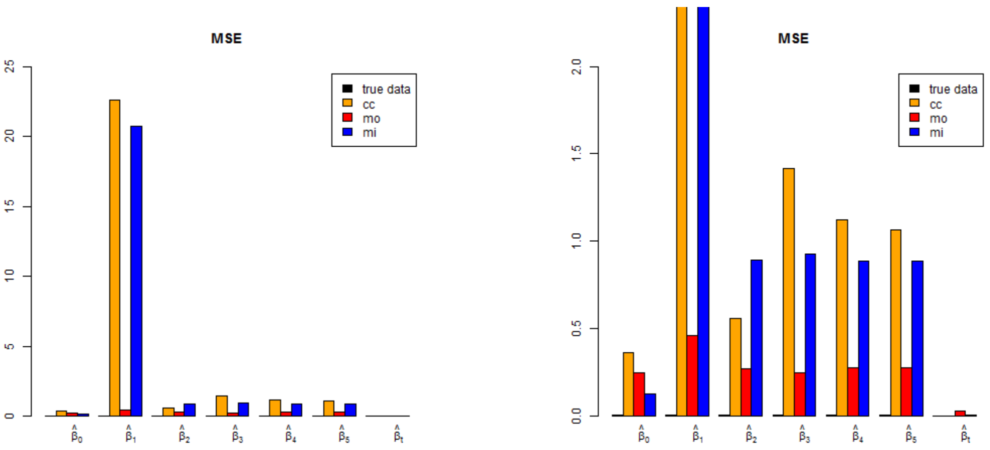
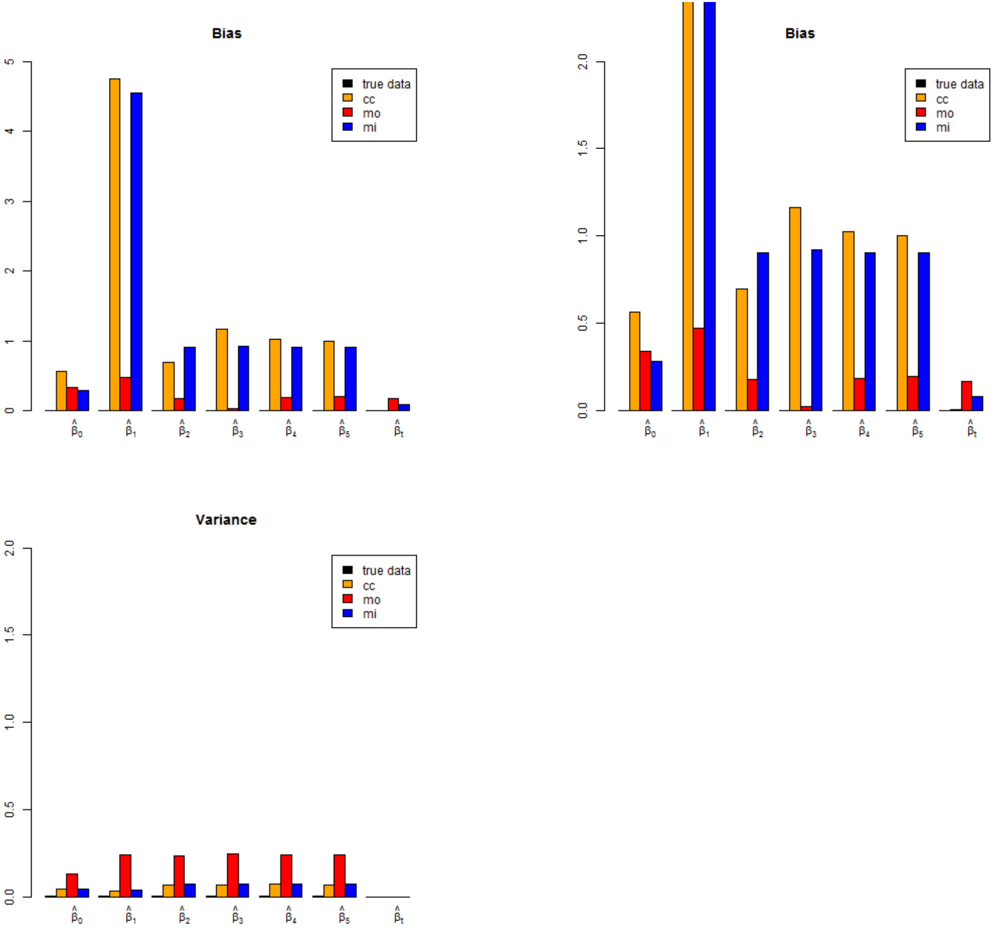


Figure 19: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%



A.3.3 Assumed measurement error proportion: 40%

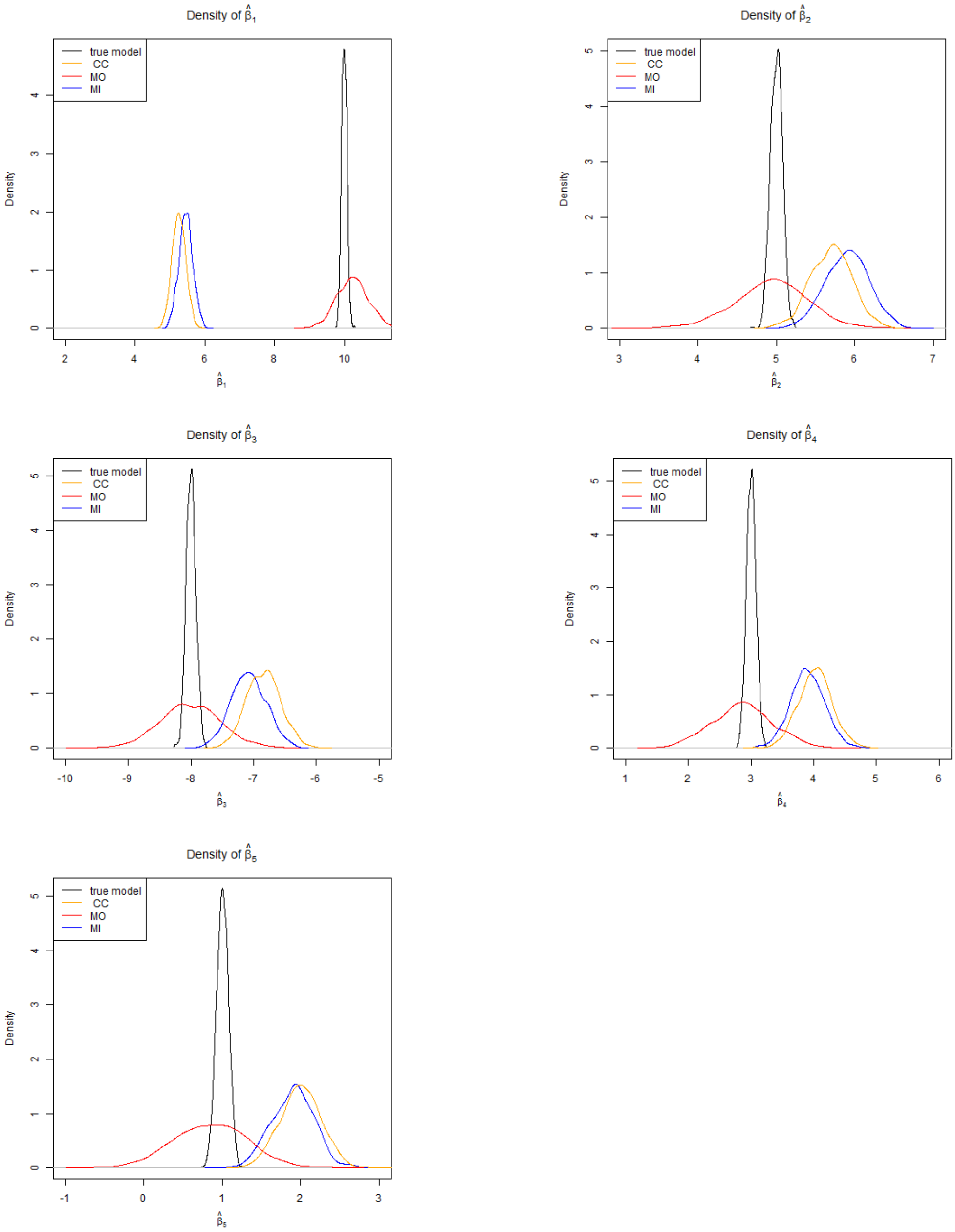


Figure 20: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%

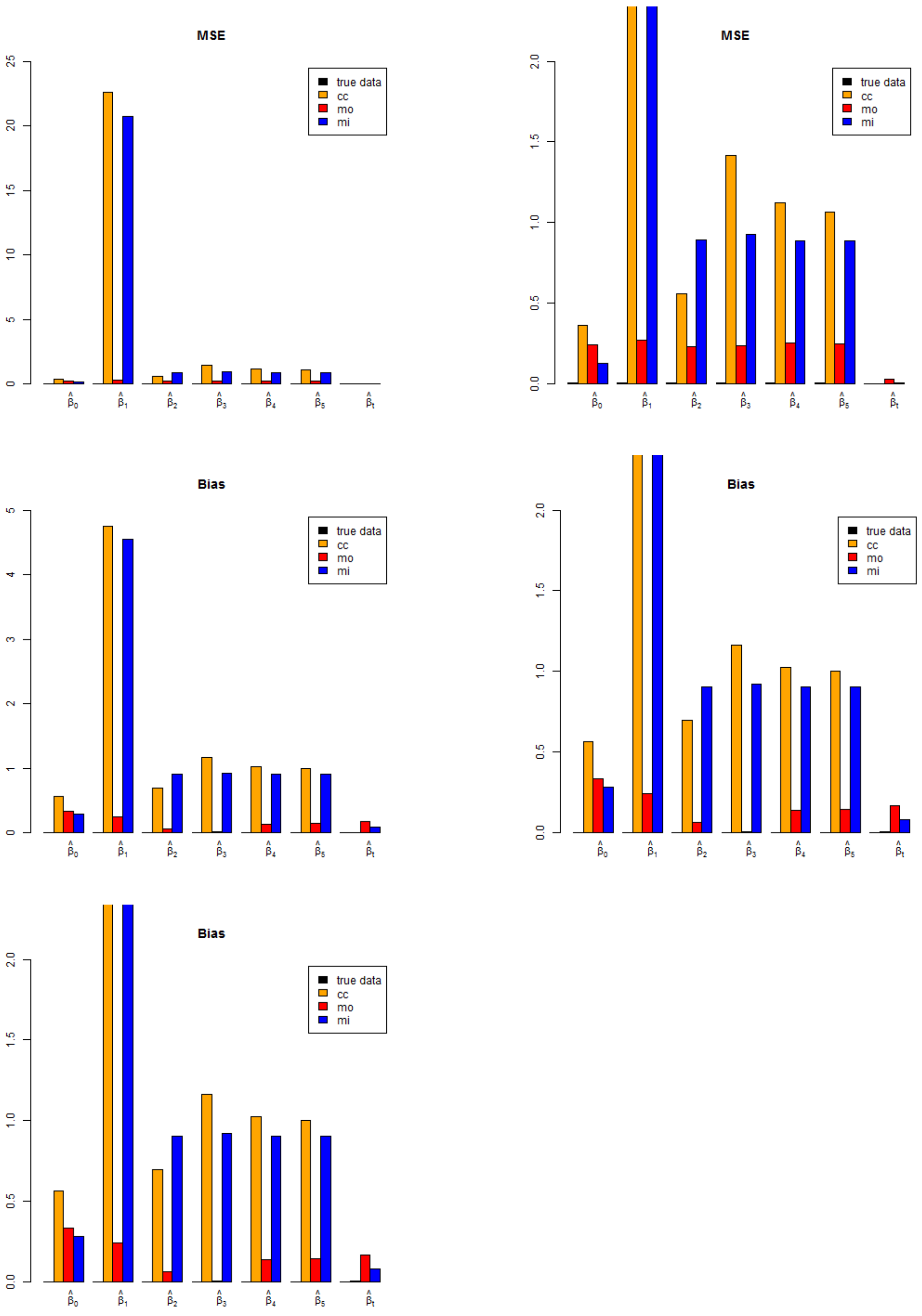


Figure 21: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 40% instead of 33.3%

A.3.4 Assumed measurement error proportion: 20%

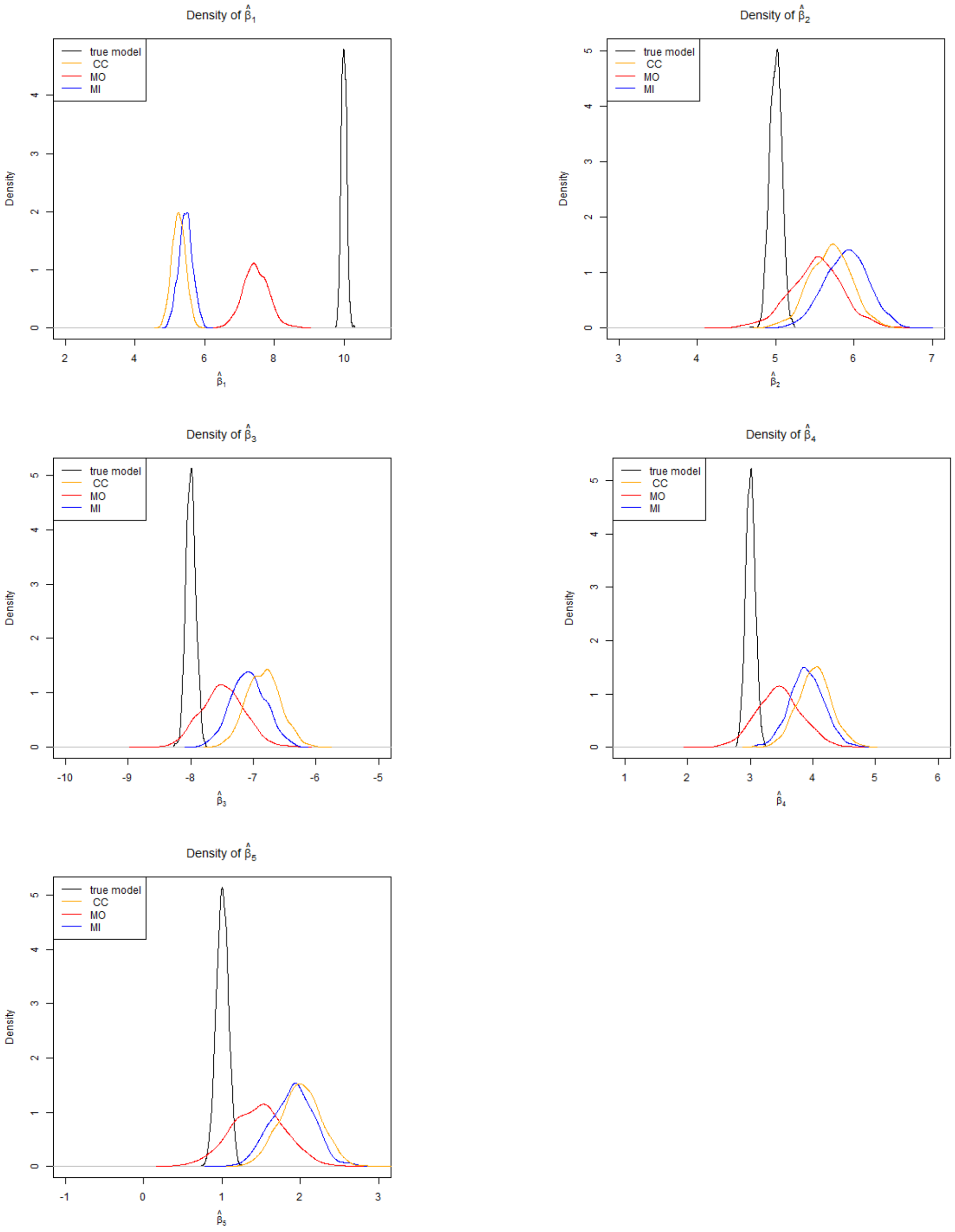


Figure 22: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%

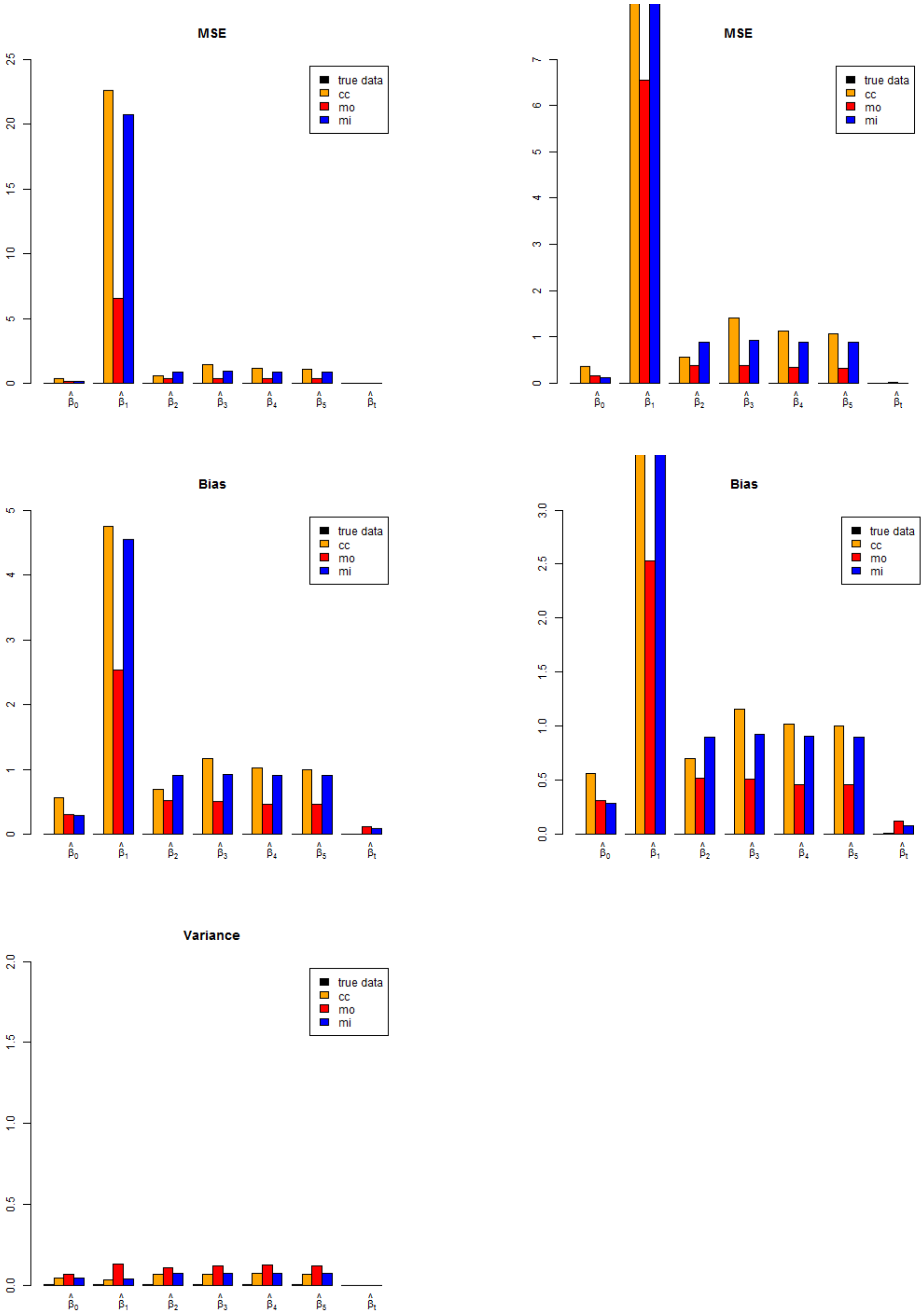


Figure 23: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 20% instead of 33.3%

A.3.5 Assumed measurement error proportion: 10%

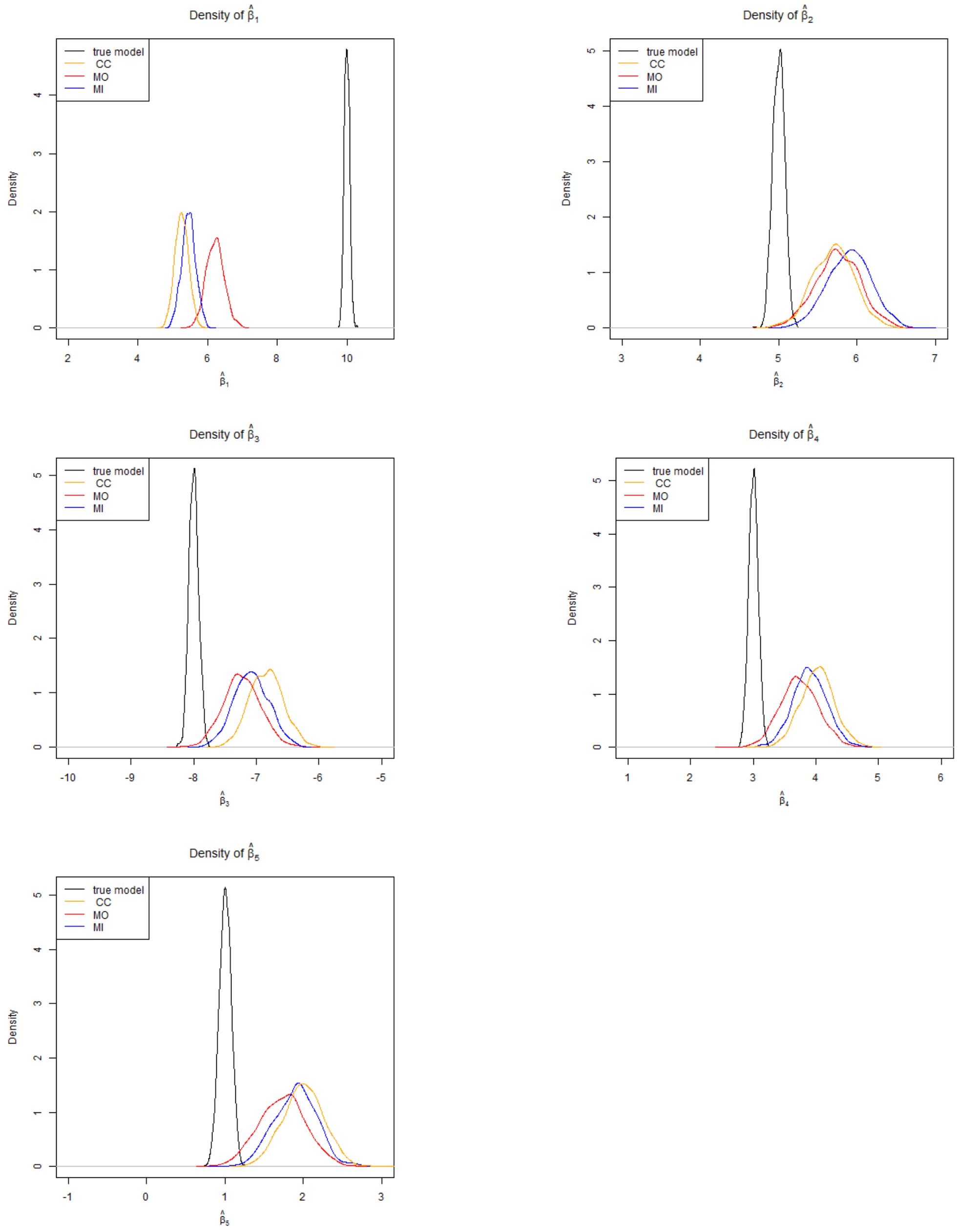


Figure 24: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%

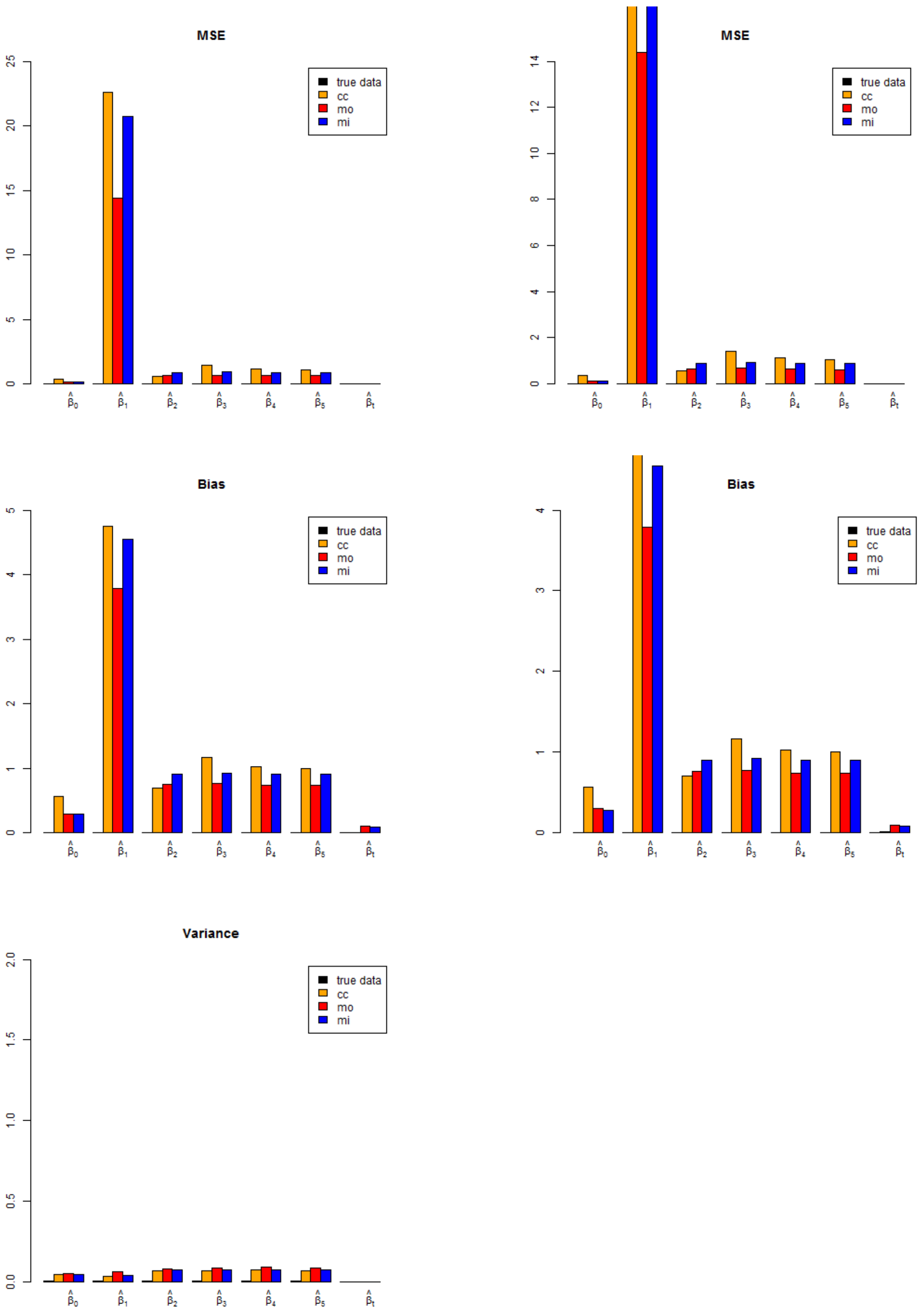


Figure 25: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 10% instead of 33.3%

A.3.6 Assumed measurement error proportion: 4.7%

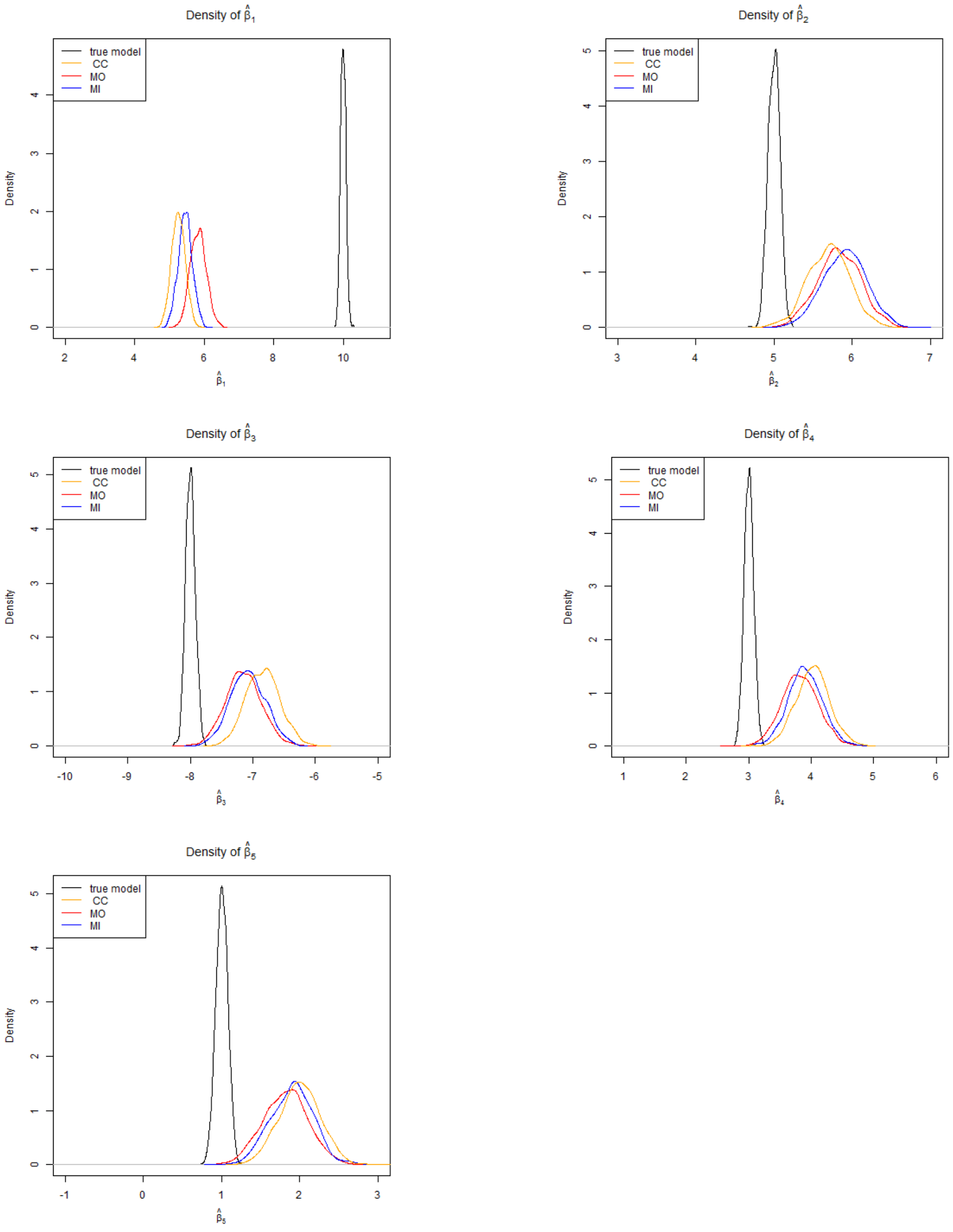


Figure 26: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO if the measurement error proportion is assumed to be 50% instead of 33.3%

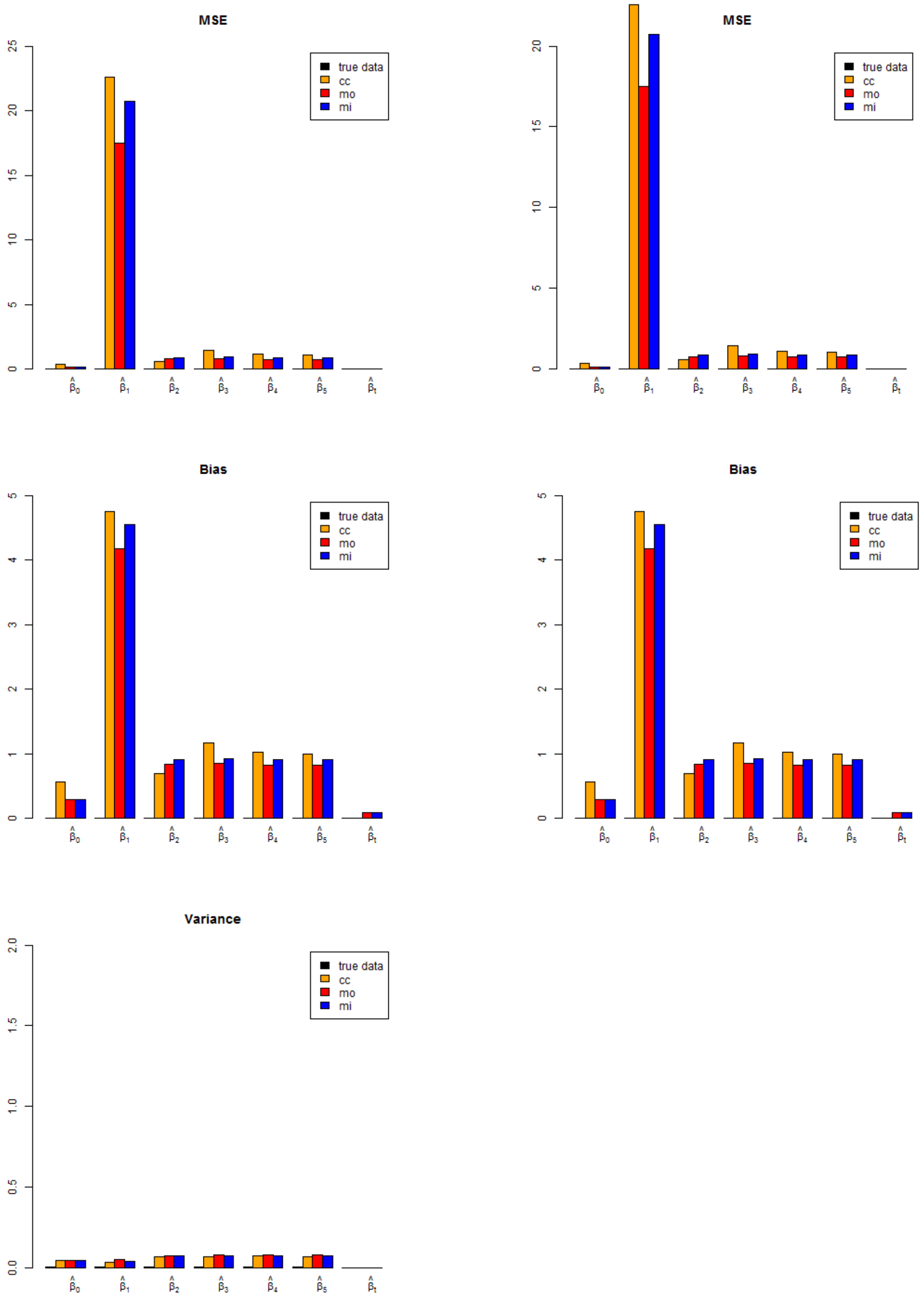


Figure 27: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO if the measurement error proportion is assumed to be 4.7% instead of 33.3%

A.4 Different number of measurements per subject

A.4.1 Two measurements per subject

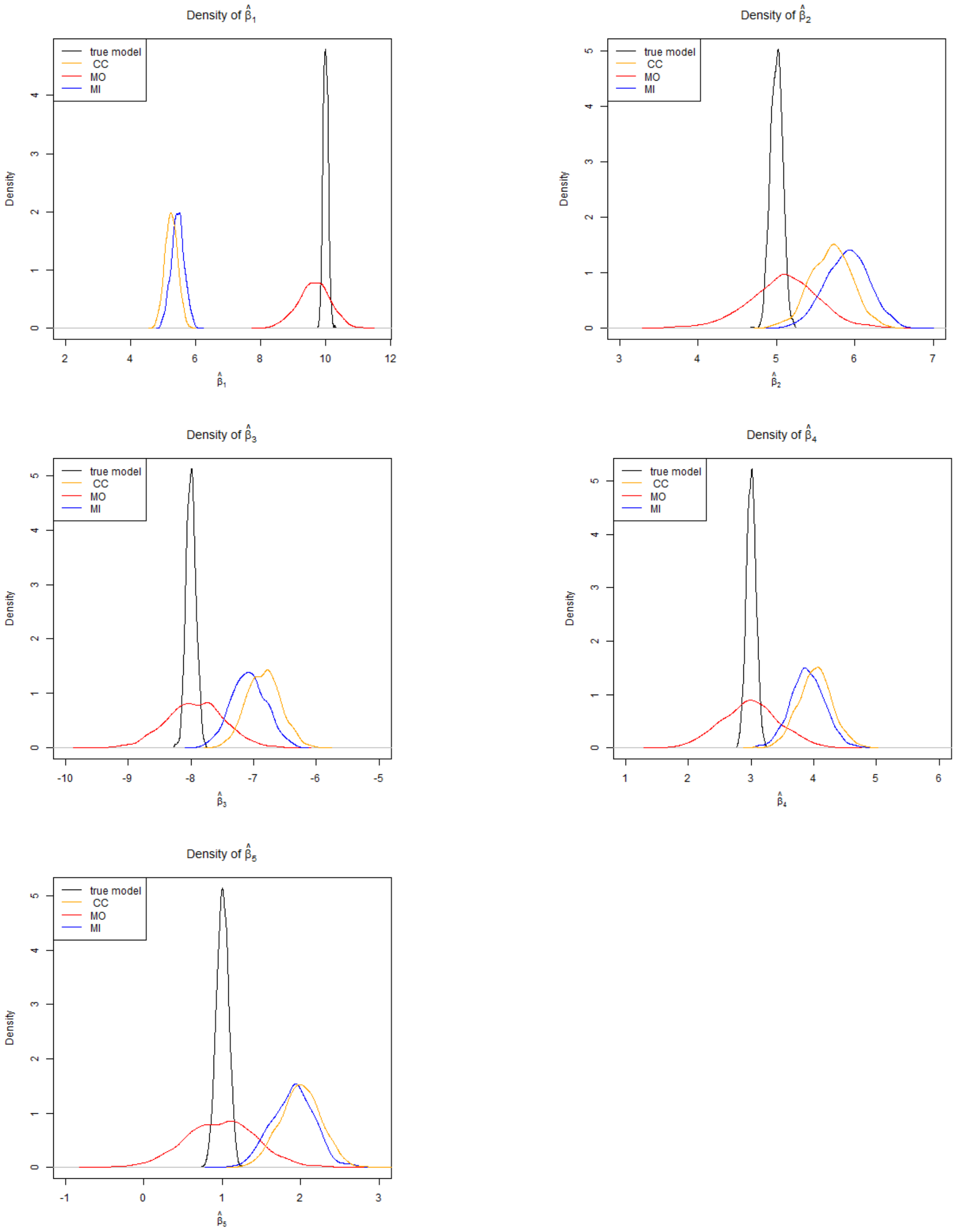


Figure 28: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO with two measurements per subject

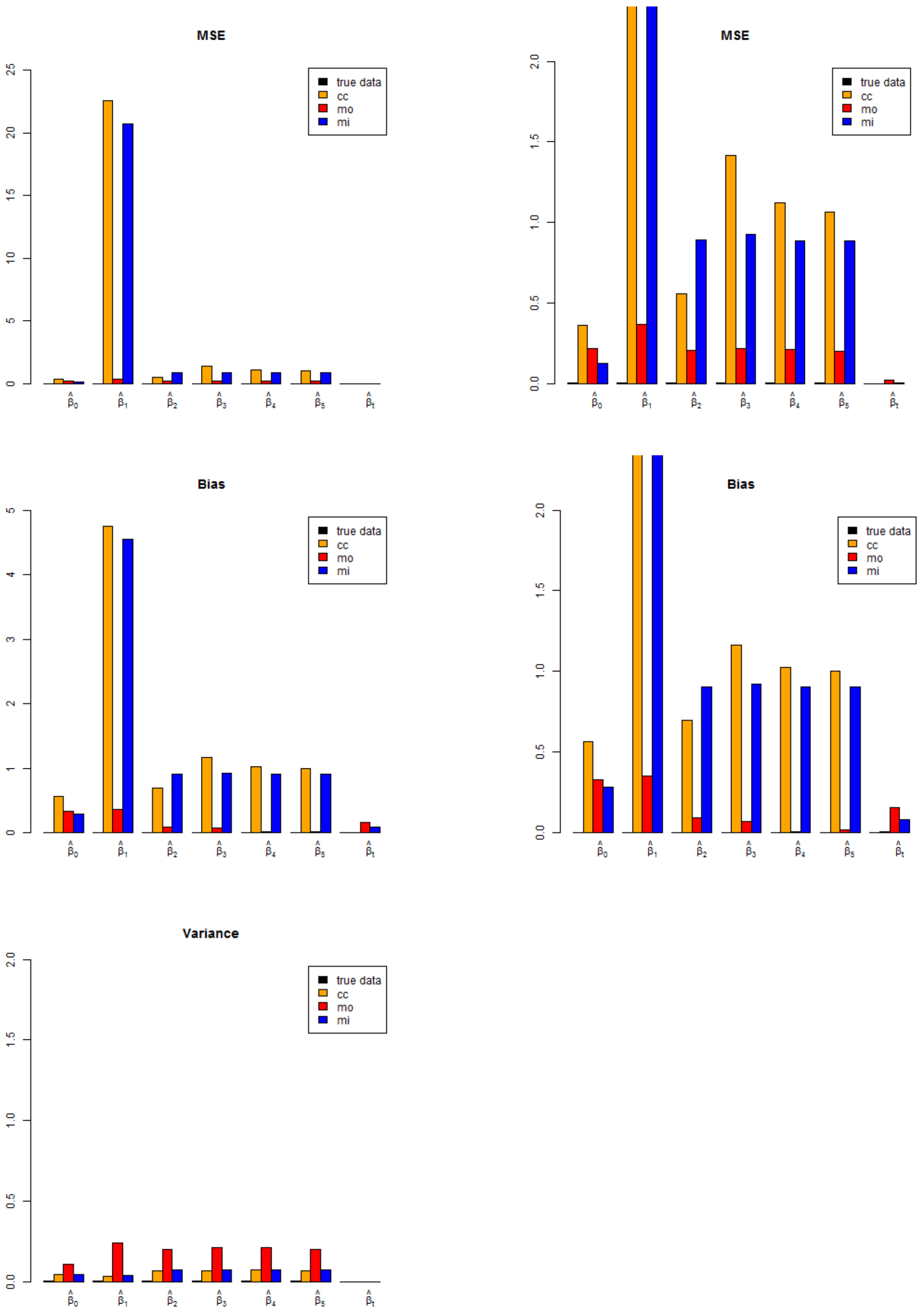


Figure 29: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO with two measurements per subject

A.4.2 Ten measurements per subject

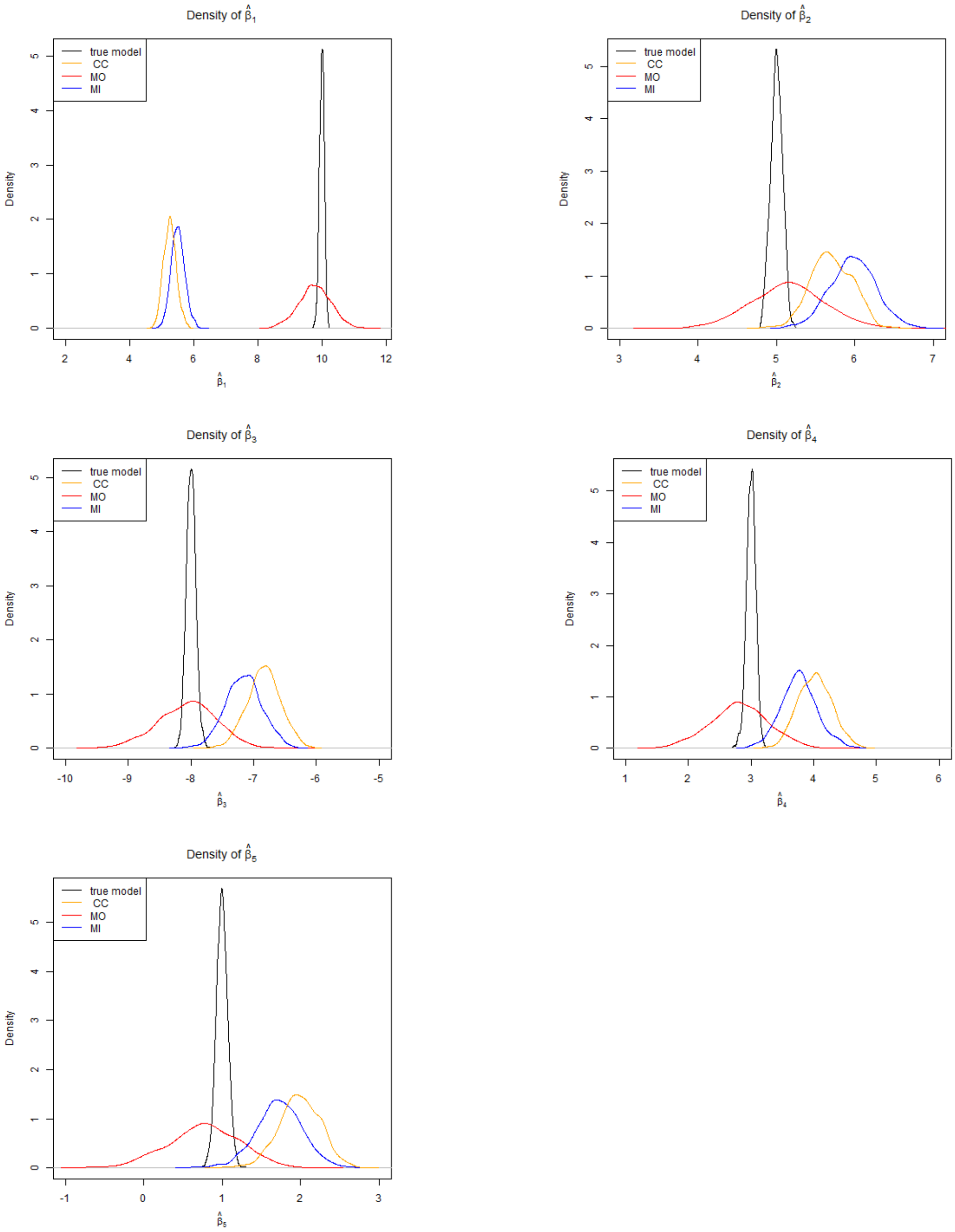


Figure 30: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO with ten measurements per subject

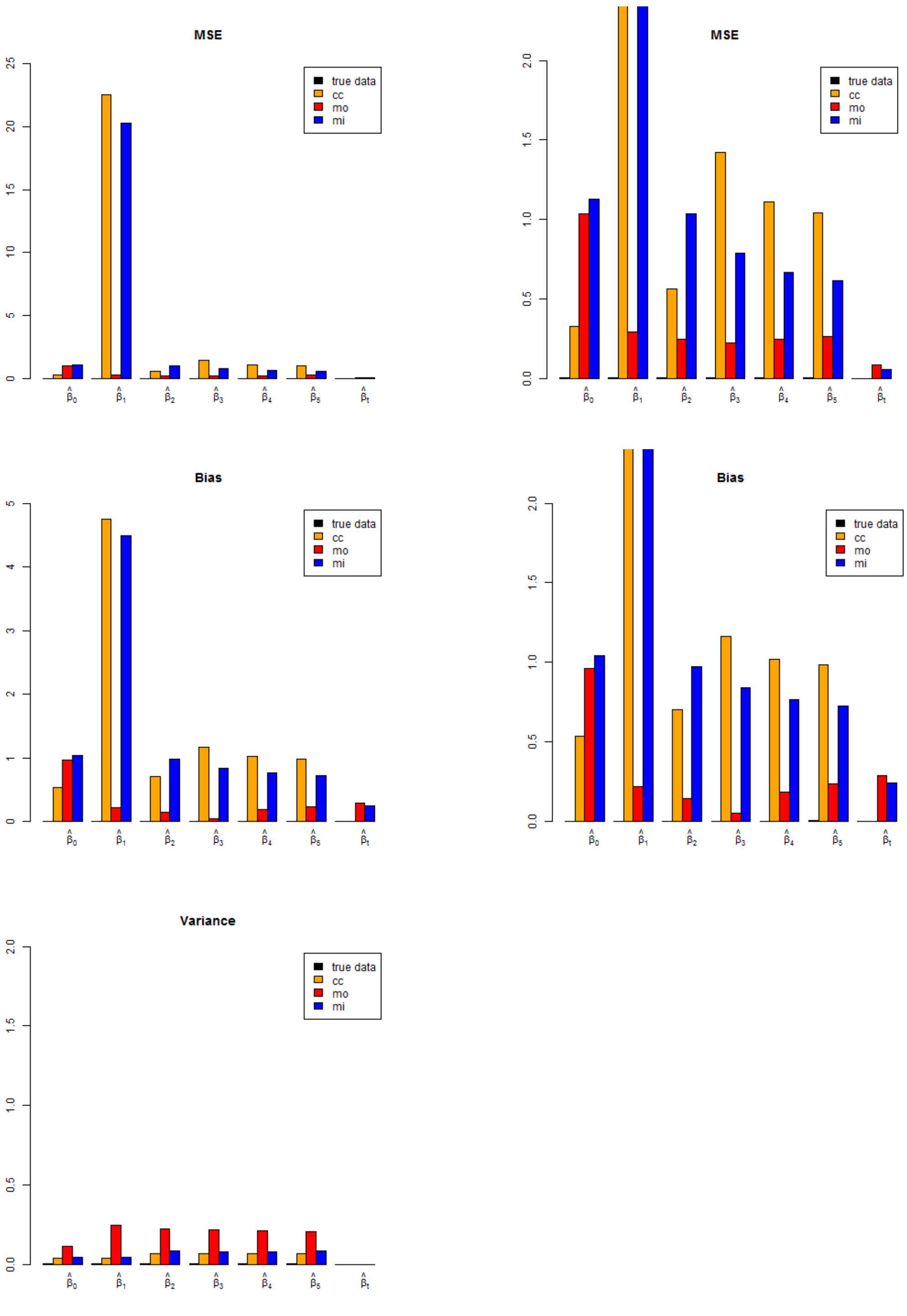


Figure 31: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO with ten measurements per subject

A.5 Different missingness and dropout probabilities

A.5.1 missingness probability around 10%, drop-out probability around 15%

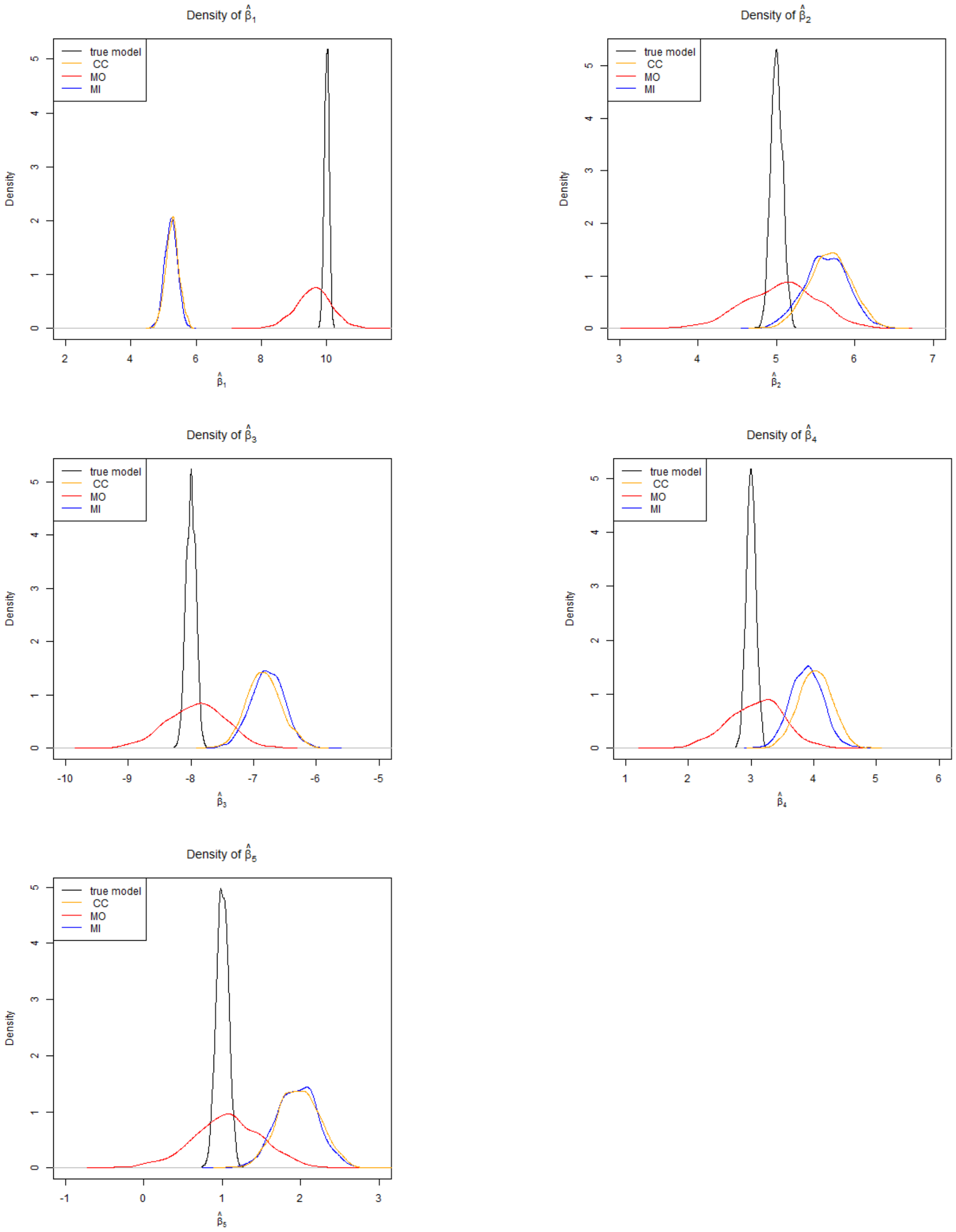


Figure 32: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO with missingness probability around 10% and drop-out probability around 15%

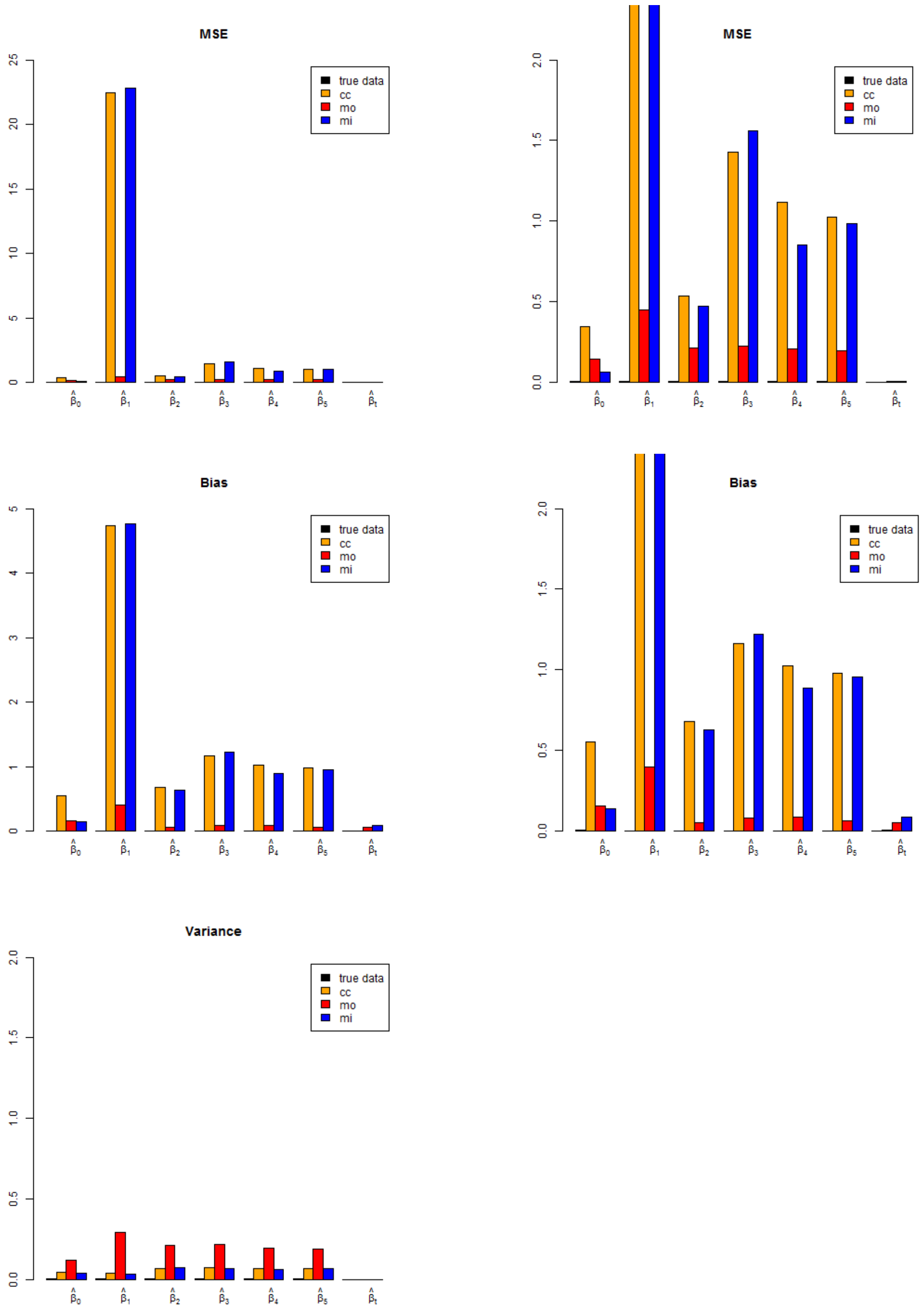


Figure 33: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO with missingness probability around 10% and drop-out probability around 15%

A.5.2 missingness probability around 30%, drop-out probability around 15%

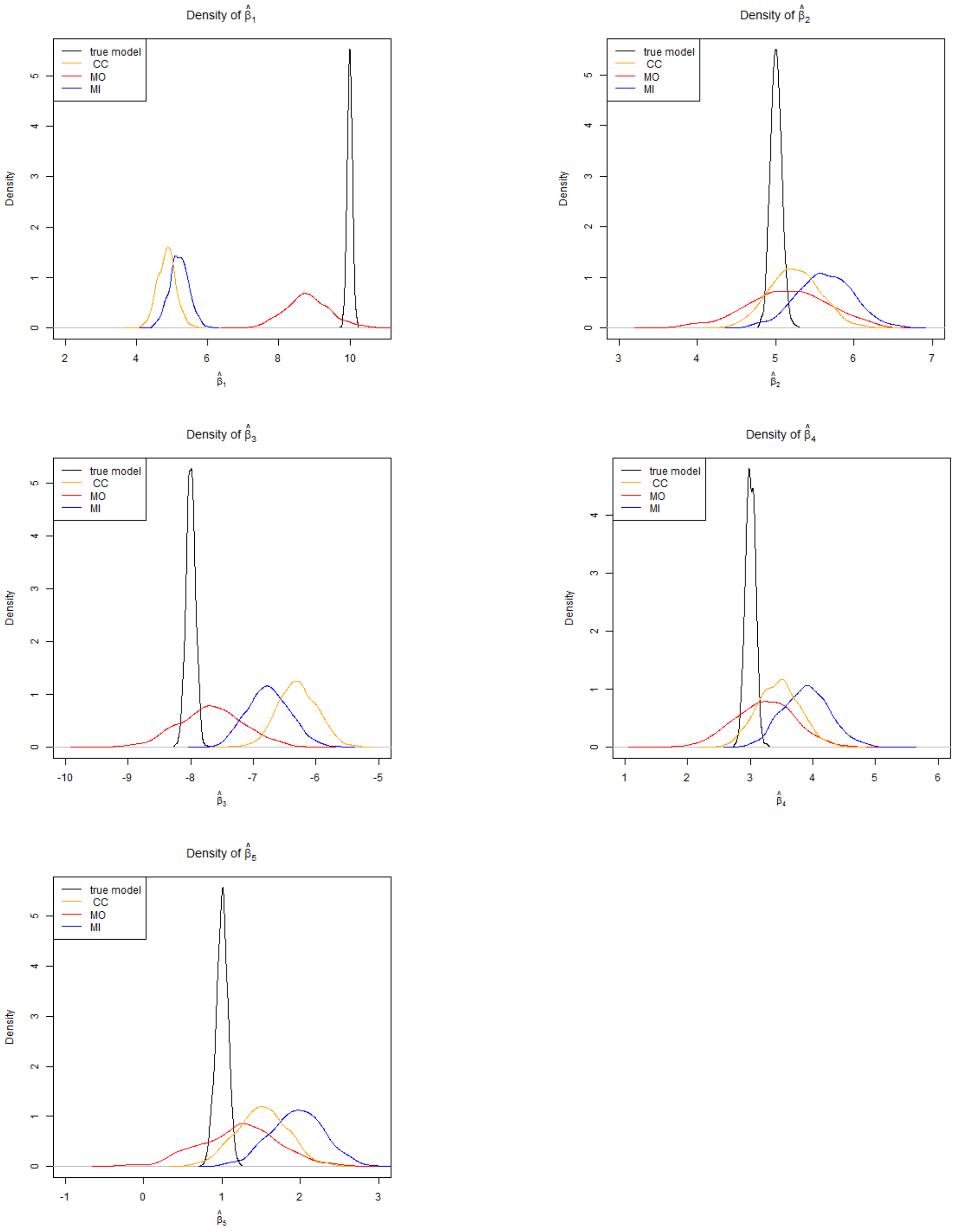


Figure 34: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO with missingness probability around 30% and drop-out probability around 15%

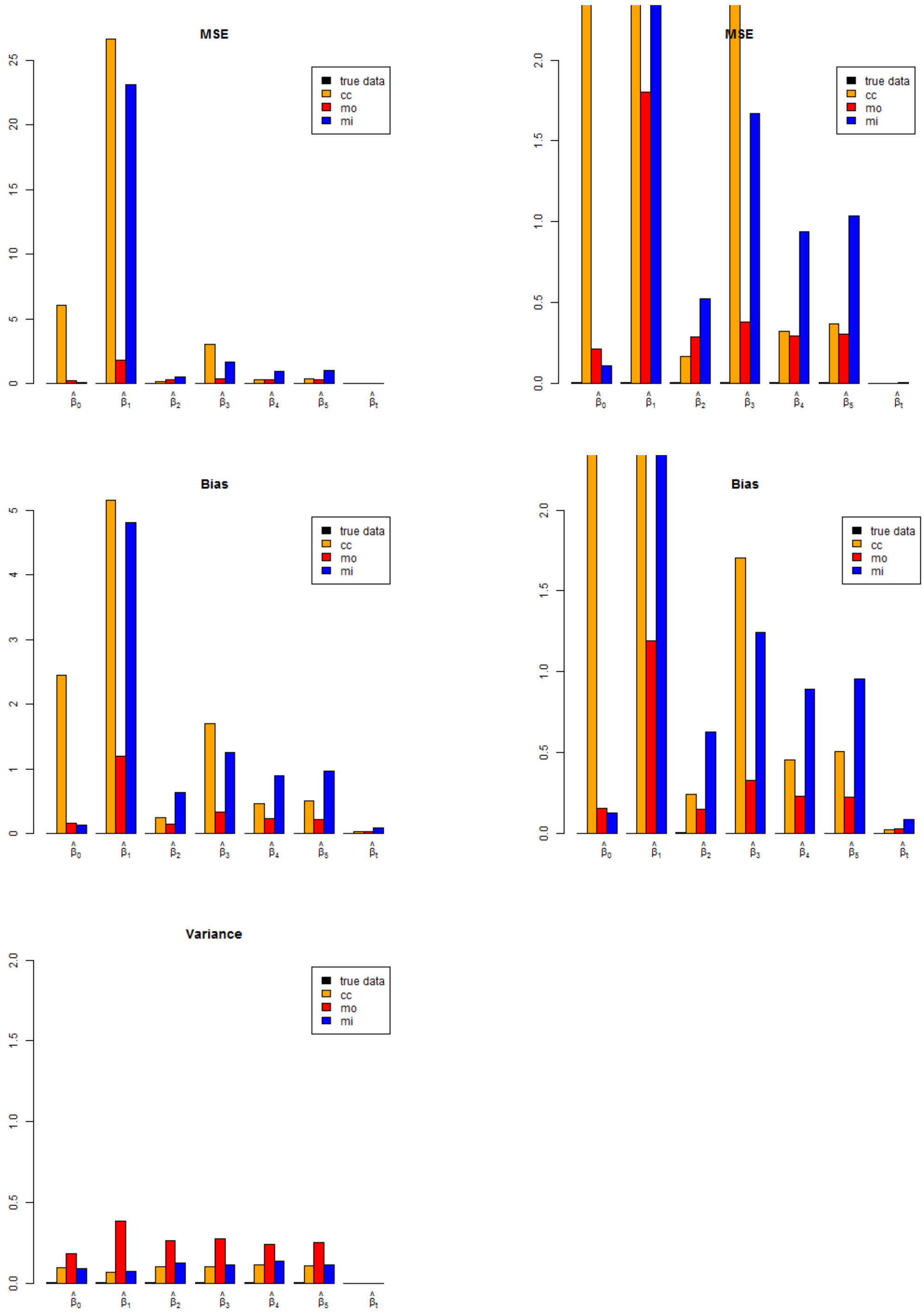


Figure 35: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO with missingness probability around 30% and drop-out probability around 15%

A.5.3 missingness probability around 30%, drop-out probability around 30%

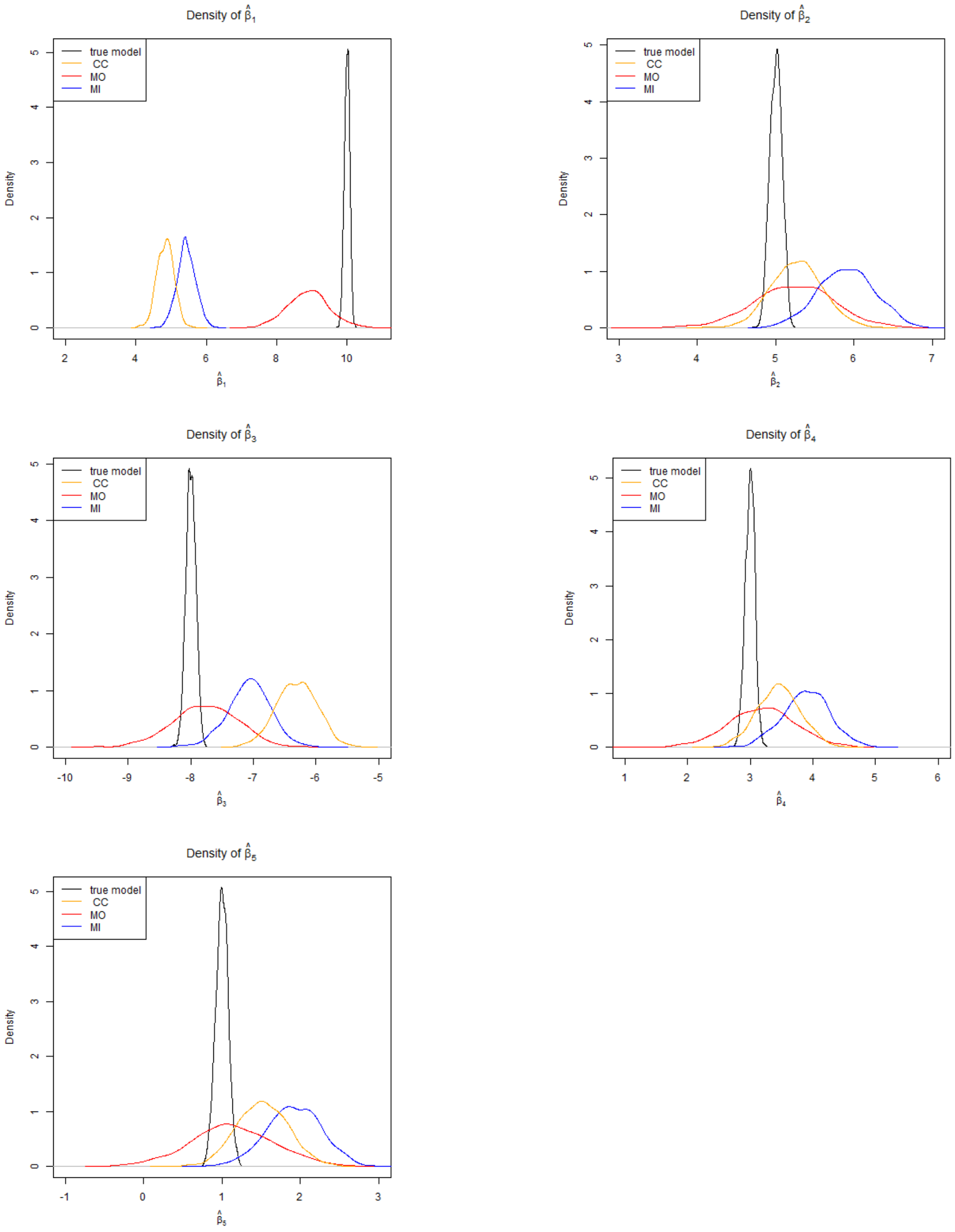


Figure 36: Density of β_1, β_2, β_3, β_4, β_5 for the true data, CC, MI and MO with missingness probability around 30% and drop-out probability around 30%

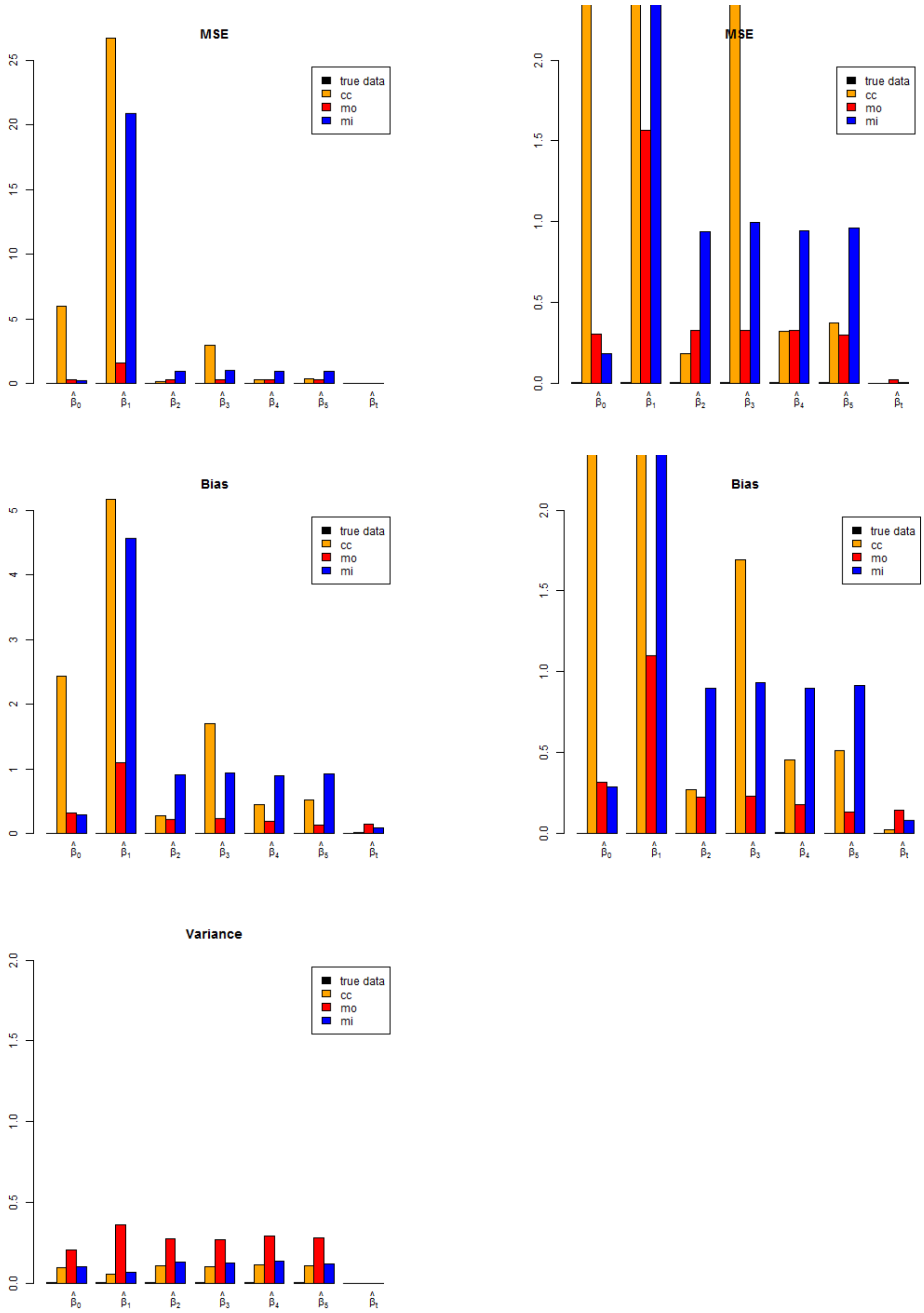


Figure 37: MSE, Bias and Variance of the estimated coefficients using the true data, CC, MI and MO with missingness probability around 30% and drop-out probability around 30%

A.6 Variances of the Random Intercept for different measurement error variances/proportions

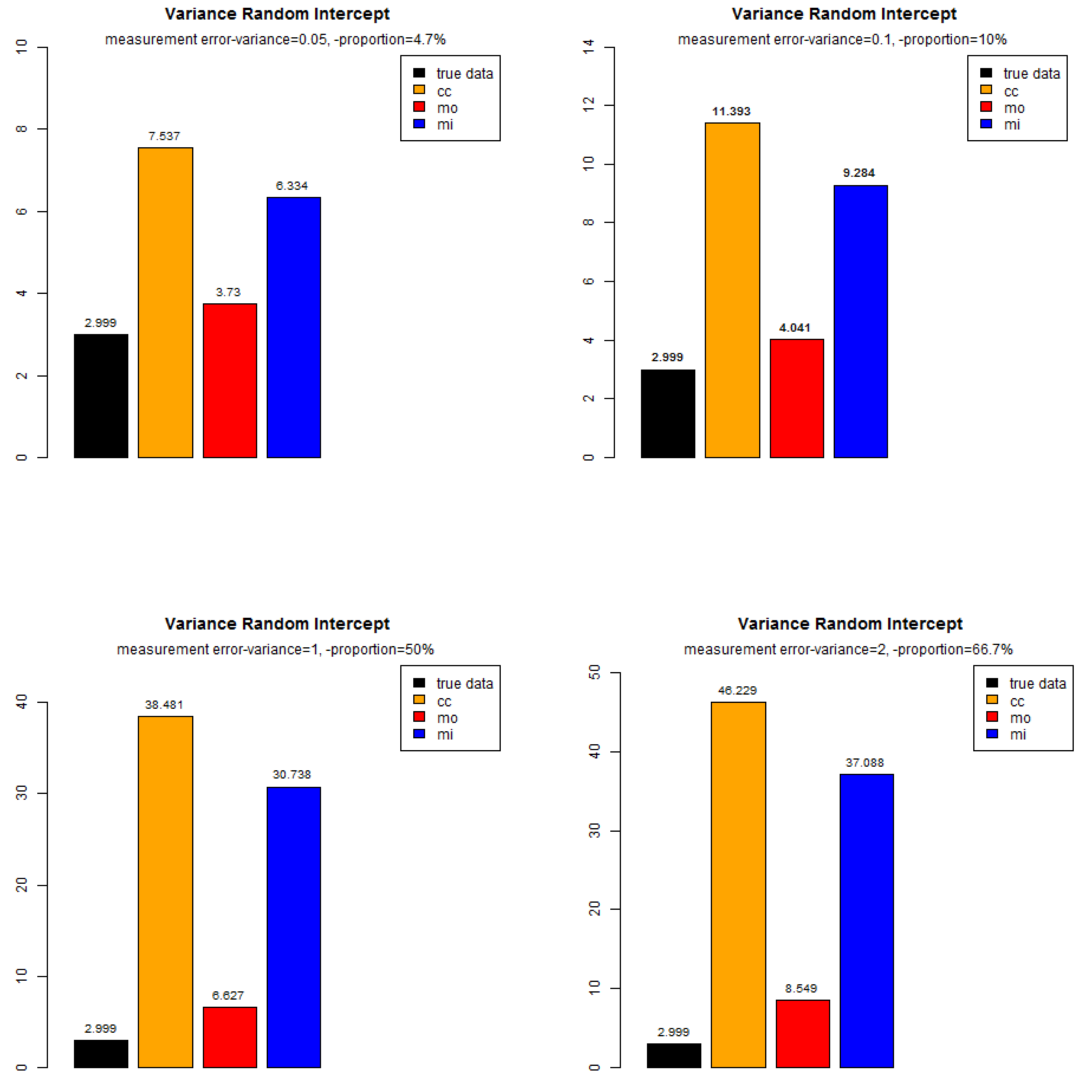


Figure 38: Variance of the random intercept using the true data, CC, MI and MO for the 4 different amounts of measurement error vrainace/proportion used in the sensitivity analysis

Statement against Plagiarism

I declare that I have written this thesis completely by myself and did not use neither other sources nor ressources except the listed ones. Moreover, I have not handed in this thesis elsewhere and also have not published it yet.

Lisa Baganz