STATISTICAL METHODS FOR THE STANDARDIZATION OF DIAGNOSTIC ASSAYS

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von

Andrea Ulrike Geistanger

Dortmund, Juli 2007

Gutachter:

Prof. Dr. Claus Weihs
Prof. Dr. Wolfgang Urfer

Tag der mündlichen Prüfung:

02. Juli 2007

Abstract

Diagnostic assays are measurement systems, measuring the concentration of analytes in human body liquids. To ensure the stability of the measured values over time, each diagnostic assays should be standardized against a so-called master sample. This is a sample with known concentration, which is measured by a very specific and precise measurement method. From this master copies are made subsequently, such that at the end of the chain a patient sample is measured on the standardized system.
A main problem for standardization systems of diagnostic assays is the definition of a master, which is stable, as analyte may be lost over time. Manufacturers of diagnostics assays as well as international organizations, especially the IFCC* have recognized the need for standardization systems of diagnostic assays that ensure stability.
Networks of laboratories are formed, which measure master samples with a reference measurement method and the averaged value of these measurements becomes the value of the master, the so-called assigned value. This value assignment is repeated after a certain time span for the next master sample, such that if the network is stable the continuity of master samples will be guaranteed.

In the context of such laboratory networks several statistical questions arise, which are discussed and answered throughout this thesis.

First it must be clear how the assigned value of the respective master and the uncertainty associated with this value is derived. The first part of the thesis examines a routine process of standardization within a laboratory network. The main sources of uncertainty within this process are revealed and how these sources have to be combined to obtain the uncertainty of the assigned value is shown. Especially the question how the uncertainty of the master is transferred to the uncertainty of the copies is discussed. A Bayesian model is presented which enables the inclusion of the uncertainty of the master within the calibration process. Based on a simulation study it is shown that this

*International Federation of Clinical Chemistry

model leads to much better results for the estimation of a measured value as well as its uncertainty, than the conventional calibration models. Further it is shown how so derived measured values should be combined to obtain the assigned value and its uncertainty.

The second part is dedicated to the identification of outliers in data of standardization networks. This is important for two reasons: one reason is that new laboratories may want to join a standardization network. As the network should ensure the stability of the master, the new laboratory must fit to the network. The other reason is that failures of measurements of the members of the network need to be detected before the assigned value is calculated. For both questions rules have to be defined which are valid for multiple value assignments.
The outlier identification for both tasks is based on robust estimation methods for linear mixed models. First it is shown how outlier identification rules for general linear mixed models can be defined. Afterwards two special cases, the one-way random effects model and the random coefficients model are regarded. Both are useful for the analysis of data from laboratory networks, the first one if only one sample within the network is regarded, the second if multiple samples are of interest. Finally an interpretation of the impact of these rules for allowable measurement deviations within a laboratory network is given.

The third part of the thesis deals with repeated method comparison studies which are necessary, if a standardization system is replaced by a new one. This might happen if global standardization system replace existing national systems, or if a more specific measurement method is established. In these cases, assigned values might change and recalculation formulas are needed for the transformation of new values into old ones and vice versa. Concepts are presented for the comparison of repeated method comparison studies.
Further the combination of these studies via hierarchical Bayesian models, to obtain a recalculation formula, is presented. The focus lies especially on the impact of the prior distributions on the results and the definition of appropriate prior distributions based on posterior predictive checks.

The presented statistical methods are applied to data of the IFCC network for standardization of HbA1c.

Acknowledgement

I would like to thank all people supporting me during the last three years for accomplishing this work.

First of all many thanks to my colleagues from the biostatistics department of Roche Diagnostics GmbH, Penzberg, especially to Christoph Berding, who gave me the possibility for doing this work and for many fruitful discussions in the context of statistics, standardization and uncertainty. To Sabine Arends and Stefan Schubert for discussions and cooperation in the field of standardization and uncertainty. To Wilhelm Kleider and Michael Pfeffer for software support with SAS. To Ursula Garczarek and Mareike Kohlmann for proof reading and comments to the text.

Further I want to thank Claus Weihs for various discussions on statistics and its interest in this new field of application.

Thanks to Cas Weykamp and Carla Siebelder from the IFCC network for standardization of HbA1c, for introducing the statistical questions of the network to me, for providing me their data and for their questions and comments to the proposed solutions.

Thanks to the colleagues of the R&D department of Roche Diagnostics GmbH, Fridl Lang and Eduard Vorberg for discussion on standardization of diagnostic test, test principles and clinical chemistry.

For the accommodation and hospitality during my stays at Dortmund I would like to thank family Hufnagel and Porfetye. Last but not least thanks to Monika and Volkmar Konnert for their love and to Martin Geistanger for his wonderful support, his love and patience.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The standardization of the measurement of length goes back to ancient Egypt, where about 3000 BC the first unit of length was defined. The "Royal Egyptian Cubit" was defined as the length of the forearm from the elbow to the tip of the Pharaoh ruling at that time plus the width of his palm. The master cubit was carved out of a block of granite to endure for all times. The workers at the building sites were supplied with cubits made of wood or granite and it was the responsibility of the architects to maintain them. All workers had to bring back their cubit sticks at each full moon to compare them to the master. The death penalty faced those who forgot this duty. [How03]
The standardization system of ancient Egypt already included the definition of the measurand, a master unit, and copies of this master unit, which were distributed to workers. Based on this standardization system the Egyptians were able to build their vast pyramids with high accuracy. These properties are the main parts of modern standardization systems, too.

Diagnostic assays are also measurement systems, measuring the concentration of analytes in human body liquids. However, in contrast to the measurement of length, standardization systems for many analytes are still not in place. For important analytes such as HbA1c ([JKB+02]), Cholesterol ([MKW+00]) or Total Thyroxine ([TUM+05]) working groups have been established to provide worldwide accepted masters for them. Their main idea is to form networks of laboratories, such that master samples are measured with a specific measurement method in these laboratories to obtain an assigned value of the masters. For the analysis of this data different statistical questions arise, which are discussed throughout the thesis.

In this chapter we explain basic elements of standardization systems for diagnostic

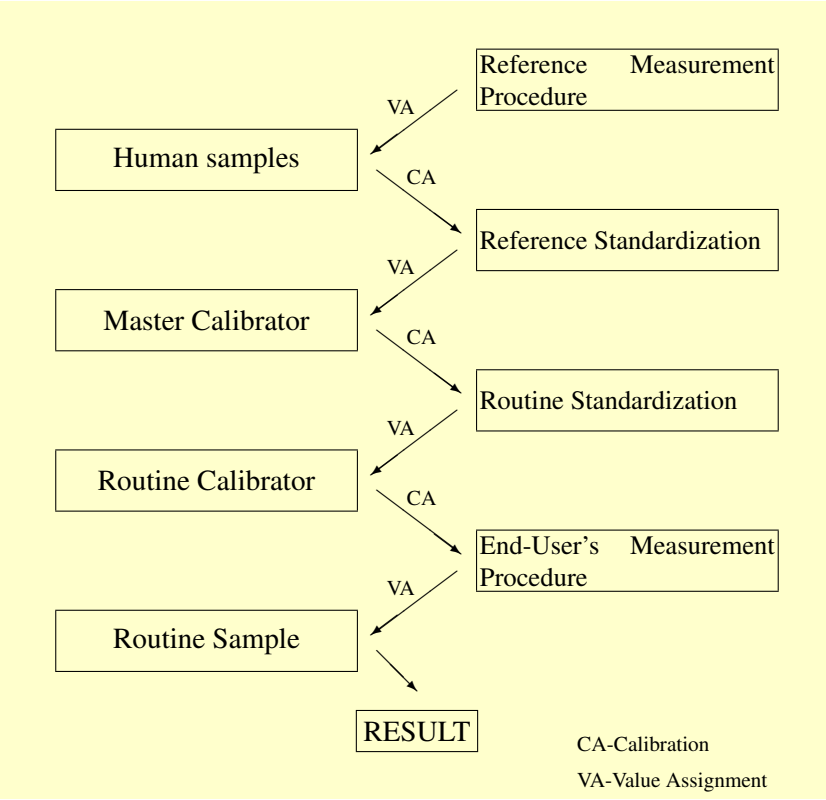
assays and introduce the three parts of the thesis - assigned value derivation, outlier identification and method comparison studies. Further on we present the IFCC* network for standardization of HbA1c†, as we apply the statistical methods to data of this network.

1.1 Standardization of diagnostic assays

An important element of today's medical diagnosis is the diagnostic assay that determines the concentration of a biologically meaningful analyte within a patient sample. It supports physicians with diagnostic information on the functional status of tissues or organs, as well as on infections and other diseases.
To achieve comparability of the results of the diagnostic assay over space and time, the results must be traceable to the highest possible reference system. This means that the result of a patient sample measured within a routine system should be comparable to the result of that sample, measured within the reference system. To achieve this traceability, copies of the reference system are derived through a so-called standardization cascade. In Figure 1.1, a theoretical standardization cascade for diagnostic assays is shown. The standardization cascade starts with a reference measurement procedure, which is able to measure the well defined analyte very precisely and specific. By means of this reference measurement procedure, values are assigned to human samples. In the second step, based on these human samples, copies, the so-called master calibrators, are derived. This cascade goes down, until the routine laboratory measures patient samples, to assign values to these samples. It is clear that going down the cascade the uncertainties of the assigned values increases. The calculation of the uncertainties at each level of the cascade is one of the statistical questions, which will be considered in this thesis. To be able to calculate the uncertainty of the values, one needs to know how these values are transferred from the higher metrological‡ level to the next lower level. There are several possible approaches, in this thesis we regard only the so-called sample reading approach.
Diagnostic assays do not directly measure the concentration of an analyte in human blood, but some indirect signal, e.g. a photospectrometrical signal. The relationship between this signal and the concentration of the analyte is described by the calibration function. Prior to the measurement, this calibration function needs to be set up based

*International Federation of Clinical Chemistry
†beta-N-terminal glycated hemoglobin A
‡Metrology is the science of measurement, not to be confounded with meteorology, the science of the weather.

Figure 1.1: **Scheme of the standardization cascade.**



on calibrators. These are samples with already known concentration. This concentration value is called assigned value of the calibrator. Hence, the measurement procedure can be divided into two stages: the calibration stage in which the calibration function is established and the sample reading stage, in which the unknown concentration of a sample is determined. This procedure is not only used for the measurement of patient samples, but also in the sample reading standardization approach: Samples with assigned values from the metrological higher level are used as calibrators of the next measurement method, and the concentration values of the samples of the next lower level are read from this calibration function.

To minimize systematic deviations of these readings, the readings take place in different laboratories and multiple measurements per laboratory. At the end of the day, these multiple measurements need to be combined to the assigned value of the respective sample.

The different laboratories of one level form a network, they can be viewed as the ref-

erence measuring system of a particular metrological level. For example, the IFCC network for standardization of HbA1c consists of up to 15 laboratories. Each laboratory measures the same set of human samples with the specific reference measurement method [JKB+02].

1.2 Assigned value derivation

In the first part of the thesis the calculation of the posterior distribution of the assigned value of a calibrator, derived via the sample reading approach, is discussed. In each standardization step not only a point estimate of the assigned value has to be determined, but also the uncertainty of this value. The "International Vocabulary of Metrology" ([VIM93]), published by the ISO, defines uncertainty as *"parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand."*
The "Guide to the Expression of Uncertainty in Measurement" ([GUM93]) defines the standard deviation of the assigned value as its uncertainty. [GUM93] proposes to use an error-propagation formula, based on a Taylor expansion of degree 1, to approximate the standard deviation of the assigned value (see e.g. [Die91] for its derivation). However, this approximation is doubtful in complex nonlinear situations.
Therefore, instead of working with the error-propagation formula we use a Bayesian approach for uncertainty calculation. In this way the posterior distribution of the assigned value is obtained. In our eyes this is the most natural way for uncertainty calculation, as we obtain automatically posterior distributions of the parameter of interest, instead of approximate confidence intervals. Moreover practitioners mostly interpret confidence intervals in a Bayesian way, instead of the frequentist approach to them ([Rub84]).
For the calculation of the posterior distribution of the assigned value, two questions must be answered. The first one is the question how the uncertainty of the calibrators used for sample reading can be included in the objective posterior distribution. This is important, as going down the standardization cascade, the uncertainty from the higher levels must be transferred to the lower ones. The second question is how to combine multiple measurements from different laboratories to obtain the assigned value.
In Chapter 3 we focus on the reading of a single measurement from the calibration function and define models of the calibration and sample reading step, where the uncertainty of the calibrators is incorporated. With this approach we will show, how the uncertainty of the assigned values of the samples of the higher metrological level is transmitted to

the posterior distribution of single measurements of samples of the next lower level.
In Chapter 4 we discuss how multiple measurements of the same sample from different laboratories can be combined to obtain the posterior distribution of the assigned value. This type of data is often modelled as a heteroscedastic one-way random effects model. We will compare this approach which our approach, where we incorporate the whole posterior distributions of the measurements, given by the sample reading process discussed in Chapter 3.

1.3 Outlier identification

The second part of the thesis is dedicated to the identification of outliers within data of laboratory networks. Within this data different types of outliers can occur. For instance, a whole laboratory can be regarded as an outlier, if its results over all samples are extreme in comparison to the measurements of the other laboratories. A single measurement within a laboratory can be regarded as an outlier, if it is extremely different to the other measurements within the particular laboratory.
The identification of outliers can serve different purposes: based on the identification of laboratories as outliers, rules can be defined for laboratories which want to join a network. For laboratories, being already members of the network, quality control rules can be defined.
The outlier identification is based on linear mixed models, where the laboratory effects are modelled as random effects. We will define rules for the identification of extreme laboratories based on the random effects, and rules for outliers within laboratories based on the residuals of the linear mixed models.
We regard two models for data from laboratory networks: the one-way random effects model and the random coefficients model. The first one is used if data from only one particular sample is regarded and one wants to detect extreme laboratories and extreme measurements within laboratories for this particular sample.
The second one is appropriate, if multiple samples are taken into account simultaneously. In this case the measurement behavior of the laboratories over multiple samples is analyzed.
[WG03] defined already outlier identification rules for the one-way random effects model, applied to data from laboratory networks. They developed a robust estimation method for this particular model. We generalize their ideas, such that they can be applied to more complex linear mixed models.
We generalize the identification rules, to adopt them to multi-dimensional random ef-

fects and we use a robust estimation method, based on t-linear mixed models, which can be applied to more general linear mixed models.
Our concepts are visualized by several examples, mostly from data of the IFCC network for standardization of HbA1c.

1.4 Method comparison studies

From time to time standardization systems are replaced. For instance a national standardization system is replaced by an international one, or a new more specific measurement method is introduced. In these cases the measured values of patient samples might change. To check to which extend this change occurs so-called method comparison studies are performed.
In method comparison studies a set of human samples is measured within two measurement methods, for instance in a measurement method based on the national standardization and a measurement method based on the international standardization. Hence, for each sample one obtains a pair of measured values, such that a regression line between both methods can be derived.
On one hand side this experiment should reveal, if the two methods are exchangeable, which is the case if the intercept of the regression line is near zero and the slope close to one. On the other side it provides a transformation rule for values of one method to the other, if the two methods are not exchangeable.
As method comparison studies are a well-known tool for diagnostic assays there is a lot of literature for the analysis of one particular study (see e.g. [RRR01], [MRR02], [PB83]). These articles discuss especially the best regression method to use. In most cases both methods are subject to error, hence errors-in-variable models are appropriate for the derivation of the regression line (see e.g. [CVN99], [Ful87] for an introduction to errors-in-variables models).
We work on two different issues for these experiments, especially when they are repeated after a certain time period or in different laboratories, such that we obtain repetitions of these studies.
The first issue concerns the equality between multiple regression lines. Even if the regression lines indicate that the methods are not exchangeable, it is of interest to know if this relationship is stable over time, or in different laboratories. We present a test developed by [LJZ04] to test whether two regression lines are equal.
Afterwards we extend this test to identify whether a new regression line is equal to a reference regression line. This might be another approach to show the exchangeability

of two methods over a specified concentration range.
The second issue discusses, how these multiple regression lines should be compared, to obtain an average regression line. To answer this we use a Bayesian approach for hierarchical linear models. We extend the usual Bayesian hierarchical linear model approach to incorporate the errors in both methods. Further we present a method to check the adequacy of different prior settings and provide a measure for the derivation of the most adequate prior.

1.5 The IFCC network for standardization of HbA1c

Many of the examples presented throughout the thesis base on data of the HbA1c standardization network, hence, we shortly present this network.
The measurement of HbA1c in percent of total hemoglobin (HbA1c and HbA0[§]) in human blood is the most important biomedical marker for long-term assessment of the glycemic status in patients with diabetes mellitus. Goals for therapy are set at specific HbA1c target values [DCC93]. The International Federation of Clinical Chemistry (IFCC) recognized the need for a reliable anchor of this major biomedical analyte and installed the IFCC Working Group on HbA1c standardization [HM96]. This group succeeded to develop a reference system of highest metrological order which has been approved by all member national societies of the IFCC [JKB$^+$02].
The components of the HbA1c reference system cover the upper part of the standardization cascade: HbA1c is defined on basis of it's molecular structure. Based on artificial HbA1c and HbA0 standards, primary calibrators for the reference measurement procedure are obtained. These primary calibrators are mixtures of the artificial standards, for a detailed explanation of their production process see [KAS$^+$06]. The reference measurement procedure is an approved reference method (enzymatic digestion followed by HPLC [JKB$^+$02]) and the secondary calibrators are whole blood panels to which values have been assigned with the reference method.
Each year a set of primary calibrators is produced. The set of primary calibrators is used to calibrate the reference method in the year after production. The reference measurement method, in which values are assigned to the secondary calibrators is not operated only in one laboratory but in different laboratories all over the world, forming the IFCC network for standardization of HbA1c.
The value assignment of the secondary calibrators takes place in so-called studies, which are performed twice a year. Within a study whole blood samples are shipped

[§]non-glycated hemoglobin A

to the member laboratories of the network and are measured by the reference method in each laboratory. Dependent on the study the network consists of $9 - 15$ member laboratories and up to 3 candidate laboratories. Candidate laboratories must prove their ability of performing the reference method and each member laboratory is controlled, too.
Different types of samples are measured in each study:

(i) Control samples with already known concentration of HbA1c, they serve as intra-laboratory control samples.

(ii) Primary calibrators with known percentage of HbA1c in total hemoglobin, derived from the production process of the calibrators. By measuring these samples within the network, assigned values from production are checked and stability problems are addressed.

(iii) Secondary calibrators, also called intercomparison samples, for the determination of their percentage of HbA1c.

Within each laboratory the measuring design is the following: The measuring method is an enzymatic digest, all samples are split into two digests, afterwards every sample is measured in two repetitions per digest.
The key task of the IFCC network for standardization of HbA1c is the assignment of HbA1c values to unknown samples with high accuracy and reliability. These values shall be used for the worldwide standardization of HbA1c.
However, nowadays there are national standardization networks for HbA1c in Europe, the Unites States and Japan. Hence, the introduction of the worldwide IFCC standardization scheme will cause a shift in reported HbA1c values, which are based on the national standardization networks.
Therefore, twice a year method comparison studies are launched between the IFCC and the other networks. A set of samples is measured based on the IFCC standardization and on the national standardizations. Afterwards a regression line between these values is fitted. Now, the obtained regression lines need to be compared, to be sure that the relationship between the networks has not changed. Further an average regression line has to be derived, such that IFCC values can be transformed in values of the national networks and vice versa.

Part I

Assigned value derivation

The goal of a standardization system is the determination of the concentration of an analyte in particular samples. As these samples will serve as calibrators for other samples, more effort has to be put in this value assignment, compared to the routine reading of patient samples. The so obtained concentration value of a calibrator is called assigned value. This value assignment can be done in different forms, we focus on the sample reading process.
The sample reading process is divided into two stages: Within the calibration stage a calibrator is used to establish the calibration function. In the reading stage the concentration of a new sample can be determined, by transferring, via the calibration function, the signal of the sample into a concentration value. If the new sample should serve as calibrator of the next standardization level, it is measured in different laboratories and repetitions within the laboratories, to account for laboratory specific effects.
In this part of the thesis we deal with the question how to derive the assigned value of a calibrator and its uncertainty within the sample reading process. We will express the uncertainty of the assigned value in terms of its posterior distribution. One part of its uncertainty is made up of the variation seen in the data due to the multiple measurements. But also variation sources due to the previous standardization steps have to be considered. Therefore we divide this task into two stages:
In the first stage we regard a single measurement of a sample and show how the posterior distribution of the measured value can be derived. We model explicitly the errors in the assigned values of the used calibrators, such that the uncertainty sources of the previous standardization steps are already incorporated in the posterior distribution. This is explained in detail in Chapter 3.
This sample reading problem goes back to [Eis39], who considered the problem of prediction from the inverse of a linear calibration function. [Kru67] derived an estimator, known as inverse estimator, as he considered the regression of the concentrations to the signals, but [Ber69] showed that this estimator is not consistent. However, all of these authors considered the assigned values of the calibrators to be error-free.
Milestones for the Bayesian approach to the sample reading problem are [Hoa70] and [HL81], who derived two different models for this problem. Both models do not take into account the errors of the assigned values. [RPWS91], [DS95], [GCS04] considered these errors, however they restricted the analysis to the estimation of the parameters of the calibration curve. We will expand the ideas of [DS95] and [Hoa70] to derive the uncertainty of the inverse predicted value by taking into account the errors of the assigned values of the calibrators.
In the second stage, we discuss how multiple measurements can be combined to obtain the assigned value of a calibrator. It is quite forward to model such a situation

as a one-way random effects model. In the case that the posterior distributions of the measurements are known we have a one-way random effects model with known error variances. In Chapter 4 we compare several approaches for the derivation of the assigned value of a calibrator based on different one-way random effects models.
In Chapter 2 we give a short introduction to Bayesian inference and simulation algorithms, which will be used throughout the thesis for the derivation of posterior distributions.

Chapter 2

Revision of Bayesian inference

In this chapter we review the main concepts and tools for bayesian inference, which will be used throughout the thesis. Besides a short introduction to Bayesian analysis we present simulation algorithms for the derivation of posterior distributions. Further we discuss model checking techniques, to reveal the adequacy of the model and the prior settings compared to the observed data.
Bayesian analysis combines the information on a parameter $\Theta \in \mathbb{R}^d$, contained in the observed data $\mathbf{y}$, with information that is available about Θ before the data is observed. This prior information is summarized in a prior distribution on Θ, denoted as $p(\Theta)$. The analysis results in the derivation of the posterior distribution of Θ, $p(\Theta|\mathbf{y})$, which is determined via Bayes's Theorem:

$$p(\Theta|\mathbf{y}) = \frac{p(\Theta) \cdot p(\mathbf{y}|\Theta)}{p(\mathbf{y})}, \tag{2.0.1}$$

where $p(\mathbf{y}|\Theta)$ denotes the probability density function of the observed data conditional on Θ. This function can also be viewed as a function of Θ and is referred to as likelihood function.
$p(\mathbf{y})$ denotes the marginal distribution of $\mathbf{y}$, given by

$$p(\mathbf{y}) = \int_{\Theta} p(\Theta)p(\mathbf{y}|\Theta)d\Theta.$$

The derivation of the marginal distribution is often cumbersome, as it may require multidimensional integration. Hence the calculation of the posterior distribution of the parameter via Bayes's Theorem is impractical for elaborated statistical models. In these cases, samples from the posterior distribution of the parameters can be obtained by simulation algorithms. They will be explained in detail in the next section.

2.1 Simulation algorithms

In this section we shortly introduce the acceptance-rejection sampling algorithm as well as the Metropolis-Hastings algorithm and its special case the Gibbs sampler. In order to keep this section as short as possible we will pass on convergence proofs and Markov Chain Monte Carlo theory. The interested reader is referred to [Tie94], [CG95], [CG92] and the book of [GRS96], which is full of applications.
The following notations will be used: Let π be the absolute continuous density function from which we want to sample, with $\pi(\Theta) = f(\Theta)/K$, where f is the unnormalized density and K the unknown normalizing constant.

2.1.1 Acceptance - rejection sampling

Acceptance-rejection sampling requires an envelope function $c \cdot h(\Theta)$, with h, a density from which we can simulate values, and with a constant c, such that

$$f(\Theta) \leq c \cdot h(\Theta), \ \forall \Theta \in \mathbb{R}^d.$$

The idea of the acceptance-rejection algorithm is to draw samples from h and each sampled point Θ is subject to an accept/reject test, i.e. each sampled point Θ is accepted with probability $f(\Theta)/(c \cdot h(\Theta))$. If the point is not accepted, it is discarded and sampling restarts, until one point is accepted. Hence, the algorithm is given by

Step 1 Sample Θ from h.

Step 2 Sample U from the uniform distribution on [0,1].

Step 3 If $U \leq \frac{f(\Theta)}{c \cdot h(\Theta)}$ accept Θ, else go to Step 1.

The crucial part for this method to be efficient is the selection of the constant c. Setting $c = \sup_\Theta \frac{f(\Theta)}{h(\Theta)}$ provides the most efficient acceptance rate, however the evaluation of this constant might be time-consuming, too.
In cases where the objective density function is proportional to a likelihood function times a prior, as given for the posterior distribution according to Bayes's Theorem, i.e.

$$\pi(\Theta) = p(\Theta|\mathbf{y}) \propto p(\mathbf{y}|\Theta) \cdot p(\Theta),$$

[SG92] propose to set the constant $c = p(\mathbf{y}|\hat{\Theta})$, where $\hat{\Theta}$ maximizes the likelihood function. Samples are drawn from the prior distribution of Θ. The likelihood function acts as a resampling probability, as the acceptance probability now becomes

$$\alpha = \frac{p(\Theta) \cdot p(\mathbf{y}|\Theta)}{p(\Theta) \cdot p(\mathbf{y}|\hat{\Theta})} = \frac{p(\mathbf{y}|\Theta)}{p(\mathbf{y}|\hat{\Theta})}.$$

Hence, those Θ of the prior with a high-likelihood are more likely to be retained in the posterior.
However, the sharper the likelihood is in contrast to the prior, the less efficient and slower becomes the algorithm.

2.1.2 Markov Chain Monte Carlo algorithms

Markov Chain Monte Carlo (MCMC) algorithms may be more efficient than acceptance-rejection sampling, but they do not produce independent samples but dependent samples, as the sampling distributions of a particular step depends on the results of the previous step. In this section, we introduce the Metropolis-Hastings algorithm and the Gibbs sampler, being a special case of the former. Further we discuss how hybrid algorithms can be constructed out of these.

Regarding the Metropolis-Hastings algorithm, we denote with $q(\Theta_n, \Theta_{n+1})$ the candidate generating density, that is, if the sampling process is at the point Θ_n, the density generates a point Θ_{n+1} from $q(\Theta_n, \cdot)$. In order that the generated chain converges to the searched target density $\pi(\Theta)$, the sampled point must be subject to an accept/reject test similar to the acceptance-rejection sampling. The acceptance probability is given by (see [CG95] for its derivation)

$$\alpha(\Theta_n, \Theta_{n+1}) = \begin{cases} \min\left\{\frac{\pi(\Theta_{n+1})q(\Theta_{n+1},\Theta_n)}{\pi(\Theta_n)q(\Theta_n,\Theta_{n+1})}, 1\right\} & \text{if } \pi(\Theta_n)q(\Theta_n, \Theta_{n+1}) > 0 \\ 1 & \text{otherwise.} \end{cases}$$

The other difference to the sampling-resampling algorithm is, that if the new sampled point is not accepted, the previous point is taken as sampled point in the sequence. In short, the algorithm may be written as:
For $n = 1, .., N$

Step 1 Sample $\tilde{\Theta}$ from $q(\Theta_n, \cdot)$.

Step 2 Sample U from the uniform distribution on [0,1].

Step 3 If $U \leq \alpha(\Theta_n, \tilde{\Theta})$ set $\Theta_{n+1} = \tilde{\Theta}$, else set $\Theta_{n+1} = \Theta_n$.

The question arises how the candidate generating density should be chosen.
One common choice is the random walk, where $q(\Theta_n, \Theta_{n+1}) = \tilde{q}(\Theta_{n+1} - \Theta)$, with $\tilde{q}$ being usually a multivariate normal distribution centered at Θ_n and appropriate scale parameter. For example, the implementation of the random walk in the R function `MCMCmetrop1R` in the package **MCMCpack** [MQ06] uses the approximate Hessian matrix for the determination of the scale parameter.
Another option, suggested by [CG94], is useful for target densities, which can be written as

$$\pi(\Theta) \propto p(\mathbf{y}|\Theta) \cdot p(\Theta).$$

Here $p(\mathbf{y}|\Theta)$ is uniformly bounded and $p(\Theta)$ is a density from which we can sample. By setting $q(\Theta_n, \Theta_{n+1}) = p(\Theta_{n+1})$, the acceptance probability becomes

$$\alpha(\Theta_n, \Theta_{n+1}) = \min\left\{\frac{p(\mathbf{y}|\Theta_{n+1})}{p(\mathbf{y}|\Theta_n)}, 1\right\}.$$

Other candidate-generating densities e.g. the independence chain or pseudo-dominating densities are introduced in [Tie94].

A special case of the Metropolis-Hastings algorithm is the Gibbs sampler (see [CG95] for a proof of this relationship). The d-dimensional parameter vector $\Theta \in \mathbb{R}^d$ is divided into q subvectors. Denote with Θ^q the qth - subvector of Θ and with Θ^{-q} the parameter vector, where the qth subvector is discarded. Further define the full conditional distribution of Θ^q as $p(\Theta^q|\Theta^{-q}, \mathbf{y})$. If we can sample from the defined full conditional distributions, we have a very efficient simulation algorithm, the so called Gibbs sampler. The Gibbs sampler cycles through the full conditional distributions and draws the values conditional on the others. [CG92] proof that the so generated sequences converge to the posterior distribution of Θ.
The high efficiency is given, as every draw is accepted. However, if a lot of full conditional distributions are used, convergence can be slow.

For models, where not all full conditional distributions have a closed form, so called hybrid algorithms can be used: For full conditionals with known closed form these distributions are used, for the remaining a Metropolis-Hastings algorithm is taken.

2.2 How many iterations?

When starting a simulation algorithm, we need to specify the number of iterations of the algorithm, which we denote with n.
For MCMC algorithms the draws are not independent, thus it is useful to save only every kth value to reduce autocorrelation between the samples. This technique is known as thinning of the chains, but to apply thinning, the value k must be determined.
Further on the algorithms need starting values, so the first m draws might be affected from them. Therefore it is necessary not to take these values into account.
[RL96] developed methods for the determination of the numbers n, k, m for a single long chain based on a pilot sample. The determination is guided by the idea that the precision for the value of interest, derived from the posterior distribution, should be specified. For example, if the real interest lies in the median of the distribution it is clear that less samples are necessary, than for the derivation of the 0.975 quantile, given the same precision. Precision is defined as follows: The estimator of the posterior probability $P(U \leq u|\mathbf{y})$ should be included within the interval $\pm r$ with probability s, where U is a function of Θ and u is the quantile of interest.
The number of thinning iterations is determined by regarding the sequence

$$Z_t = \begin{cases} 1 & \text{if } U_t < u \\ 0 & \text{otherwise,} \end{cases}$$

which is a binary 0-1 process, but no Markov process, due to higher dependence in Z_t. Regarding however the sequences Z_t^k for $k = 1, 2, ..,$ consisting of every kth iteration of the original process, will result in a Markov process choosing k reasonably large. Thus, based on model fitting criteria, the appropriate thinning factor can be found.
For the determination of the number of burn-in iterations to be discarded, [RL96] regard the probability of

$$P(Z_m^k = i|Z_0^k = j),$$

which should be smaller than a predefined ε. Using standard results of Markov chain theory they derive

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

where α is the probability of changing from the first state to the second state and β is the probability of changing from the second state to the first state. Taking the number of thinning iterations into account we have $m = m^* \cdot k$.
The estimate of $P(U \leq u|\mathbf{y})$ is given by $\bar{Z}_n^k = 1/n \sum_t Z_t^k$, which is approximately nor-

mally distributed with mean q and variance

$$\frac{1}{n}\frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}.$$

Thus the requirement

$$P(q-r \leq \bar{Z}_n^k \leq q+r) = s$$

implies that

$$n^* = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3}\left\{\frac{\Phi^{-1}(1/2(s+1))}{r}\right\}^2,$$

where Φ is the standard normal cumulative distribution function. Thus we have $n = n^* \cdot k$ as number of total iterations.
These methods are implemented in the statistic software R2.3.1 [R D06] in the function `raftery.diag` of the **coda** package [PBCV06].

[GR92] pointed out that the lack of convergence can also be assessed from multiple independent sequences by calculating the potential scale reduction factor $\hat{R}$. Consider, we have $j \cdot n$ draws of a random variable from j independent chains and n repetitions per chain. From the j chains we can make j different inferences and compare them to the inference obtained from mixing the j chains together.
Suppose we are interested in a scalar summary ψ (e.g. mean or variance) from the target distribution. The potential scale reduction factor is defined as the ratio of "between interval length" and "mean within interval length" of the empirical confidence interval of ψ. This means that from each individual chain the length of the centered $(1-\alpha)\%$ interval is calculated as well as the mean length. On the other side the length of the centered $(1-\alpha)\%$ interval is calculated from the entire $j \cdot n$ simulated values:

$$\hat{R} = \frac{\text{length of total-sequence interval}}{\text{mean length of the within-sequence interval}}.$$

If $\hat{R}$ is large, this suggests that either "the between interval length" can be decreased by further simulations or that "the mean within interval length" will be increased, since the simulated sequences have not made a full tour of the target distribution. If $\hat{R}$ is close to 1, we can conclude that each chain of the n simulated observations is close to the target distribution.
To monitor convergence dependent on the run-length, [BG98] propose the following graphical approach: The j chains are divided into batches of length b. $\hat{R}(i)$ is calculated based on the second half of the observations of a sequence of length $2ib$, for $i = 1, .., j/b$.

Afterwards $\hat{R}$ as well as the nominator and denominator are plotted against $2ib$. [BG98] pointed out that besides the converge of $\hat{R}$ towards 1, it is important to check whether the two individual terms stabilize as functions of n.
OpenBUGS2.2.0 [STBL05] calculates for every parameter this convergence statistic in terms of the width of the central 80% interval. The statistics are calculated in bins of length 50 (see [STBL05] for details).

2.3 Model checking

The posterior distributions of a Bayesian analysis depend on the prior distributions and the likelihood function. A way to check if the inferences drawn from these posteriors are appropriate, is to combine the observed data with data predicted from these distributions. This enables us to compare different prior settings and/or likelihood functions. This method, called posterior predictive check, was proposed and applied by [Gut67], [Rub81] and [Rub84]. [GCSR04] present this method in detail and give some applications to examples.
Posterior predictive checks may be used for two purposes - to assess whether the assumed model is appropriate and to find the most adequate prior distributions for the analysis.
We introduce the following notations: Let $\mathbf{y}$ be the observed data and Θ the parameter vector of the model. Define by $\mathbf{y}^r$ the replicated data, i.e. the data that would be observed, if the experiment is carried out a second time. The posterior predictive distribution is then given by

$$p(\mathbf{y}^r|\mathbf{y}) = \int p(\mathbf{y}^r|\Theta) \cdot p(\Theta|\mathbf{y})\, d\Theta.$$

The discrepancy between the model and the data will be measured by test quantities $T(\mathbf{y}, \Theta)$, being summaries of parameters and data. By these, data and simulations from the posterior predictive distribution are compared.
[GCSR04] propose to compare the lack-of-fit by the p-value of the test quantity, defined by

$$Pr(T(\mathbf{y}^r, \Theta) \leq T(\mathbf{y}, \Theta)).$$

P-values close to 0.5 mean that the predicted data from posterior distribution lead to the same inferences as the observed data. P-values close to zero or 1 indicate that the derived posterior distributions do not fully explain the observed data.
However, p-values are influenced by the spread of the test statistic. Hence, in our eyes

this measure lacks an important detail: When comparing different prior distributions, we may obtain approximately the same p-values, all close to 0.5 for different prior definitions. But the spread of the distribution of the test statistic may be very different. In this case we would like to choose the prior setting, which provides the best compromise between still acceptable p-value and the spread of the distributions. Therefore we suggest another measure of discrepancy, namely a mean-square error of the test statistic, defined by

$$MSE_T = \int (T(\mathbf{y}^r, \Theta) - T(\mathbf{y}, \Theta))^2 \, d(\Theta, \mathbf{y}^r).$$

In practice we have n simulations from the posterior distribution of the parameter vector Θ. The ith simulated value of y^{ri} is obtained by drawing from the sampling distribution of $\mathbf{y}$, given the simulated value of the parameter vector Θ.
The p-value of interest is the proportion of the n simulations, for which

$$T(\mathbf{y}^{ri}, \Theta^i) \leq T(\mathbf{y}, \Theta^i), \ \forall i = 1, .., n.$$

MSE_T is obtained by

$$MSE_T = \frac{1}{n} \sum_{i=1}^{n} (T(\mathbf{y}^{ri}, \Theta^i) - T(\mathbf{y}, \Theta^i))^2.$$

Chapter 3

Sample reading

In this chapter we show, how the posterior distribution of a single sample, which is read of a calibration function, can be determined. Especially we regard the situation when the assigned values of the calibrators are subject to error. We introduce first the statistical model and estimation algorithms. At the end of the chapter we discuss a simulation study, to emphasize the need of the incorporation of the errors of the assigned values of the calibrators.
Sample reading can be formalized as a two stage process:

(i) In the calibration stage we observe assigned values x_i and signals y_i of the calibrators, which are estimates of the respective true values u_i and η_i. For these true values the signal-to-concentration relation, expressed as a calibration function, holds

$$\eta_i = f(u_i, \Theta).$$

The aim of this stage is to obtain an estimator for the parameter vector Θ of the calibration function.

(ii) In the second stage a sample is set on the measurement system, for which it is assumed that the same calibration function holds for its true values, with the goal of predicting the true concentration u_0, given an observed signal value y_0.

We present first the Bayesian approach to sample reading of [Hoa70] and propose an extension of the later, to include also the uncertainty of the assigned values of the calibrators in the model. As in this second case no closed forms of the posteriors can be derived, we present an MCMC algorithm for the simulation of the posteriors.

3.1 The Hoadley model

[Hoa70] was the first to present a Bayesian model for the sample reading problem of a linear calibration function. Denote with $\mathbf{Y}_1 = (y_1, .., y_n)'$ the obtained signals of the calibrators and with $\mathbf{Y}_2 = (y_{n+1}, .., y_{n+m})$ the obtained signals of the sample. The Hoadley model then reads

$$\begin{aligned} E(y_i) &= b_0 + b_1 \cdot x_i \\ y_i &\sim N(E(y_i), \sigma_y^2), \ \forall i = 1, .., n \\ y_i &\sim N(Y_0, \sigma_y^2), \ \forall i = n+1, .., n+m \\ Y_0 &= b_0 + b_1 \cdot x_0, \end{aligned} \tag{3.1.1}$$

which requires the definition of a prior distribution $p(\Theta = (b_0, b_1)', \sigma_y^2, x_0)$. Hoadley regards a general form of the prior distribution, i.e.

$$p(\Theta, \sigma_y^2, x_0) \propto p(\Theta, \sigma_y^2) \cdot p(x_0)$$

and uses the non-informative Jeffrey's prior for $p(\Theta, \sigma_y^2) \propto \sigma_y^{-2}$. For this case he showed that the posterior of x_0 is proportional to

$$p(x_0|\mathbf{Y}_1, \mathbf{Y}_2) \propto p(x_0) \cdot L(x_0),$$

with $L(x_0)$ being a kind of likelihood function. But as $L(x_0)$ is not integrable, it is necessary that $p(x_0)$ is a proper density function to obtain a sensible posterior distribution. For $m = 1$ he deduces that setting

$$p(x_0) \sim t_{n-3}(0, (n+1)/(n+3))$$

leads to

$$p(x_0|\mathbf{Y}_1, \mathbf{Y}_2) \sim t_{n-2}\left(\hat{x}_I, \left(\frac{n+1+\hat{x}_I^2/R}{F+n-2}\right)\right),$$

where F is the F-statistic used for testing that $b_1 = 0$. Further $R = F/(F+n-2)$ and $\hat{x}_I$ is the inverse estimator for x_0, which would be obtained by regression of X on Y.
[HL81] also presents a Bayesian model for the sample reading from a linear calibration function. The main differences between both models are that [Hoa70] demands a prior distribution for the unknown concentration of the sample x_0, whereas [HL81] demands a prior for the mean of the signals of the unknown sample, Y_0, as the second stage of their model is based on the inverse calibration function. However, in the discussion of

this paper several authors (see [Law81], [Hil81], [Lwi81], [Orb81]) point out, that it is more appropriate to assign a prior distribution to x_0 than to Y_0.
We agree with this view, as in most situations it will be possible to assign at least a uniform prior to the concentration of the sample between 0 and the maximal expected amount in human body liquids. But signals of the assays reaction will and can vary from laboratory to laboratory in unknown amounts. Hence, we proceed by expanding the Hoadley model to an arbitrary calibration function and by incorporating in the model the errors in the assigned values of the calibrators.

3.2 Incorporation of calibrator uncertainty

Many authors dealing with the sample reading problem exclude the uncertainty of the assigned values of the calibrators based on the argument, that this uncertainty is much smaller than the signal uncertainty (see [Eno99]). This might be true for routine assays, however for reference measurement methods this is not. Reference measurement methods have very precise and specific signals, however they mostly need calibrators, too. Their calibrators are artificially produced and their uncertainty is often at least as big as the uncertainty of the signals, (see e.g. [KAS+06] for the situation within the IFCC network for standardization of HbA1c). It is clear that in these cases the uncertainty of the assigned values of the calibrators is not longer to be omitted. In Section 3.3.1, we give an example of the impact of measurement error in the assigned values for the estimation of the parameters of the calibration function.
Expanding the Hoadley-Model to a general calibration function and modelling explicitly the measurement error of the assigned values leads to the following model:

$$\begin{aligned} \eta_i &= f(u_i, \Theta) \\ y_{ij}|u_i &\sim N(\eta_i, \sigma_y^2),\ \forall i = 1, .., n, j = 1, .., J_i \\ x_i &\sim N(u_i, \sigma_{x_i}^2),\ \forall i = 1, .., n \\ y_i &\sim N(\eta_0, \sigma_y^2),\ \forall i = n+1, .., n+m \\ \eta_0 &= f(u_0, \Theta). \end{aligned} \tag{3.2.1}$$

Now the signal of each calibrator is measured multiple times J_i, such that y_{ij} denotes the measured signal of calibrator i and repetition j. The observed assigned values are modelled as random variables, which vary around their true value u_i, with known variance $\sigma_{x_i}^2$. It is appropriate to assume the variance $\sigma_{x_i}^2$ known, as this value is derived in the previous standardization step of the calibrators.

In Model (3.2.1) a prior distribution

$$p(\Theta, \sigma_y^2, u_0, u_i, i = 1, .., n)$$

needs to be specified.
The measurement error model of the assigned value is sometimes assumed to be multivariate normal, to account for the correlation between these values (see e.g. [RPWS91], [KAS+06]).

3.2.1 MCMC algorithms

We will use Markov Chain Monte Carlo updating schemes to obtain samples of the posterior distribution of the parameters of Model (3.2.1).
According to Bayes Theorem (2.0.1) the full posterior distribution of Model (3.2.1) is proportional to

$$p(\Theta, \sigma_y^2, u_0, \mathbf{u}|\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{x}) \quad \propto \quad p(\mathbf{Y}_1|\mathbf{u}, \Theta, \sigma_y^2) \cdot p(\mathbf{x}|\mathbf{u}) \cdot p(\mathbf{Y}_2|\Theta, \sigma_y^2, u_0) \cdot p(\mathbf{u}, \Theta, \sigma_y^2, x_0).$$

Although the above full conditional posterior is correct from a statistical point of view, there is a problem from the practical point of view. The updating scheme for the parameters of the calibration function and the measurement error variance incorporates not only the calibration stage, but also the reading stage. It is obvious, that for each read sample, the knowledge about these parameters increases, which means, that for samples read at the end of the calibration period the posterior distribution would be sharper than for samples read at the beginning. However, in practice this is not done. Each reading step is only based on the information obtained in the initial calibration step.
To treat all samples equally within a calibration period, the above model can be divided in the two explicit stages, the calibration stage and reading stage. For each sample only the information on the calibration parameters and measurement error variance from the calibration stage is incorporated in the reading stage.
Hence, the full posterior of the calibration stage becomes

$$p(\Theta, \sigma_y^2, \mathbf{u}|\mathbf{Y}_1, \mathbf{x}) \quad \propto \quad p(\mathbf{Y}_1|\mathbf{u}, \Theta, \sigma_y^2) \cdot p(\mathbf{x}|\mathbf{u}) \cdot p(\mathbf{u}, \Theta, \sigma_y^2), \tag{3.2.2}$$

and for the reading stage

$$p(\Theta, \sigma_y^2, u_0|\mathbf{Y}_2) \quad \propto \quad p(\mathbf{Y}_2|u_0, \Theta, \sigma_y^2) \cdot p(u_0, \Theta, \sigma_y^2). \tag{3.2.3}$$

The priors of the reading stage $p(\sigma_y^2)$, $p(\Theta)$ are the posteriors of the calibration stage. Although for each read sample we obtain a posterior distribution for σ_y^2 and Θ, this information is not included in the reading of the next sample.

The function `MCMCmetrop1R` of the **MCMCpack** package [MQ06] in R2.3.1 [R D06] provides possibilities for the implementation of a random walk algorithm (see Section 2.1.2), where all parameters are updated in one step. In the calibration step, we obtain simulated values from the full conditional distribution (3.2.2) and in the reading step from (3.2.3).
However, closed forms of the prior distribution of the parameter in the reading step are needed for this approach. But as these are derived in the calibration step, only simulated values of them are available. One way to use this algorithm nevertheless is to approximate these distributions by closed forms distributions, and to estimate there parameters from the simulated values.
This algorithm is quite fast, so that we could perform a simulation study with different parameter settings for the linear calibration case, which is presented in detail in Section 3.3.2.

The other possibility is to divide the parameter vector of the calibration stage into three subvectors σ_y^2, Θ and $\mathbf{u}$ and to perform a hybrid Metropolis-Hastings algorithm based on the conditional posterior distributions

$$\begin{aligned} p(\sigma_y^2|\Theta, \mathbf{u}, \mathbf{Y}_1, \mathbf{x}) &\propto p(\mathbf{Y}_1|\mathbf{u}, \Theta, \sigma_y^2) \cdot p(\sigma_y^2) \\ p(\Theta|\sigma_y^2, \mathbf{u}, \mathbf{Y}_1, \mathbf{x}) &\propto p(\mathbf{Y}_1|\mathbf{u}, \Theta, \sigma_y^2) \cdot p(\Theta) \\ p(\mathbf{u}|\sigma_y^2, \Theta, \mathbf{Y}_1, \mathbf{x}) &\propto p(\mathbf{Y}_1|\mathbf{u}, \Theta, \sigma_y^2) \cdot p(\mathbf{x}|\mathbf{u}) \cdot p(\mathbf{u}). \end{aligned}$$

However, at least for the last conditional posterior no closed form exists, such that in this case a Metropolis-Hastings step is necessary.
For the reading step the parameter vector is divided into the subvectors σ_y^2, Θ, u_0, with conditional posterior distributions

$$\begin{aligned} p(\sigma_y^2|\Theta, u_0\mathbf{Y}_2) &\propto p(\mathbf{Y}_2|u_0, \Theta, \sigma_y^2) \cdot p(\sigma_y^2) \\ p(\Theta|\sigma_y^2, u_0, \mathbf{Y}_2) &\propto p(\mathbf{Y}_2|u_0, \Theta, \sigma_y^2) \cdot p(\Theta) \\ p(u_0|\sigma_y^2, \Theta, \mathbf{Y}_2) &\propto p(\mathbf{Y}_2|u_0, \Theta, \sigma_y^2) \cdot p(u_0). \end{aligned}$$

For the reading step, this approach has the advantage that there is no need to approximate the prior distributions of σ_y^2 and Θ, as they are given by sampled values from the

calibration step. So the updating is based on a Metropolis-Hastings step, as described in Section 2.1.2, where the candidate value is sampled from the prior distribution and the acceptance probability is based on the ratios of the likelihoods.

3.3 Linear calibration case

We will now examine in more detail the influence of measurement errors in the assigned values of the calibrators for the linear calibration case. First we give a small example and afterwards present a simulation study for different data situations.

3.3.1 Example

To clarify the need for modelling the measurement errors in the assigned values of the calibrators, we regard the following example:
Suppose the true concentration values of the calibrators are given by $\mathbf{u} = (3, 6, 9, 12, 15)$. The observed assigned values $\mathbf{X}$ have the variance-covariance matrix $\mathbf{\Sigma_X} = 1 \cdot \mathbf{I}_5$, such that we observe assigned values $\mathbf{x} = (5.54, 6.31, 7.76, 12.90, 15.73)$, instead of the true values $\mathbf{u}$.
The expected mean of the signals is given by $\eta = 0 + 1 \cdot \mathbf{u}$, the variance of the observed signals is $\sigma_y^2 = 0.1$. Note that the variance of the errors of the assigned values is 10 times higher than the variance of the errors of the signals. For each assigned value, we have four repeated measured signals. We are especially interested in the measured value of a sample with true concentration value $u_0 = 5$.
We analyze the data by explicitly modelling the measurement error of the assigned values according to the algorithm described in Section 3.2.1, as well as by ignoring this error according to the basic Hoadley model. We regard the posterior distributions of the parameters of the calibration function b_0, b_1, the variance of the signals σ_y^2 and the concentration of the read sample u_0. These posterior distributions are calculated with the `MCMCmetrop1R` function of package **MCMCpack** [MQ06] in R2.3.1 [R D06] for both models.
To determine the number of iterations of the algorithm, we use the method of [RL96], which is explained in Section 2.2. A burn-in of 1000 iterations, and 10.000 iterations taken thereafter are sufficient for the estimation of the 0.05 and 0.95 quantiles of the posterior distributions with a precision of ±0.025. In the calibration step, we assign an $InvGamma(0.001, 0.001)$ prior to σ_y^2 and a $N_2((0, 1)', 10^3 \cdot \mathbf{I}_2)$ prior to the coefficients of the calibration function.
In Figure 3.1 the signals are plotted against the true values $\mathbf{u}$ (black) and the observed

values **x** (red). Three regression lines are plotted, too. The dashed black line is the least-squares regression line of the observed signals vs. the true concentration values of the calibrators. This is the regression line one obtains with error-free calibrator values. The solid black line is the regression line estimated based on the observed signals and observed calibrator values, by taking the measurement error of the observed calibrator values into account. The solid red line is estimated based on these values too, however the measurement errors of the calibrator values are not considered. When modelling the measurement error explicitly, normal priors with mean given by the observed values **x** and variance 4 are assigned to each unknown u_i. The priors of the $u_i's$ are chosen such that the order of the calibrators is still in place. A uniform prior between $(1, 16)$ is assigned to the unknown u_0.
In Table 3.1, the median as well as the 0.05 and 0.95 quantiles of the MCMC samples of the reading step are given. Taking the measurement error in the assigned values not into account, leads especially to an overestimation of the variance of the signals, which in turn leads to much too wide confidence intervals for the read value of the sample.
To examine this situation in more detail for different situations of the measurement errors of the assigned values, we make a simulation study. It is explained in the next section.

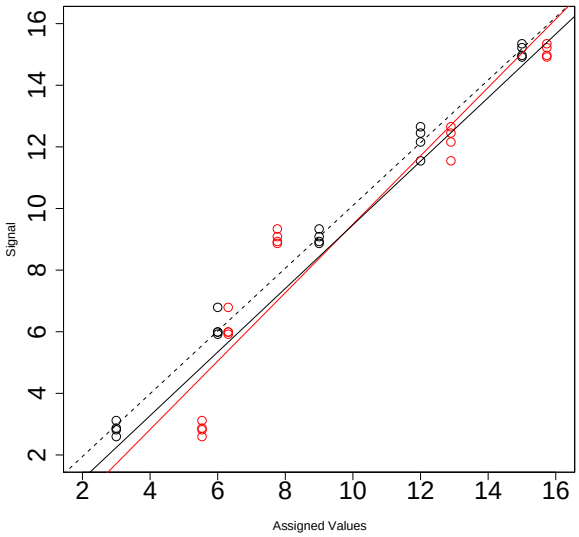
Table 3.1: **Results of the linear calibration example, without and with taking the measurement error of the assigned values into account.**

Parameter	True Value	Without measurement error			With measurement error		
		$q_{0.05}$	Median	$q_{0.95}$	$q_{0.05}$	Median	$q_{0.95}$
σ_y^2	0.1	1.19	1.93	3.70	0.06	0.11	0.23
b_0	0	-2.26	-0.84	0.49	-3.78	-1.58	0.53
b_1	1	0.90	1.03	1.16	0.89	1.11	1.34
u_0	5	3.22	5.66	7.94	4.72	5.95	6.95

3.3.2 Simulation Study

For the linear calibration case we examine different data situations, dependent on the slope of the linear calibration function, the relationship between the variance of the errors of the assigned values and the signals, and the correlation of the assigned values of the calibrators. We simulate 12 different data situations, listed in the first column of Table 3.2, and analyze them (i) according to the calibration model with measurement

Figure 3.1: **Plot of the signals versus true values of the calibrators (black) and signals versus observed values of the calibrators (red) with estimated linear calibration functions.**



error, with the algorithm given in Section 3.2.1 and (ii) according to the calibration model with not accounting for the error in the assigned values of the calibrators.
The goal of this simulation study is to compare both models for each parameter situation in terms of bias, coverage and length of the confidence intervals of the parameters of the calibration function and the read sample.
All data sets are simulated according to the following algorithm:

(i) Generation of assigned values of the calibrators, given the true values of the calibrators $u_i,\ i = 1, .., n$ and the variance-covariance matrix $\mathbf{\Sigma_x}$:

$$\mathbf{x} \sim N_n(\mathbf{u}, \mathbf{\Sigma_x}).$$

(ii) Generation of the signals of the calibrators, given the true values of the calibrators $u_i,\ i = 1, .., n$, the parameters of the calibration function b_0, b_1 and the variance of the signals σ_y^2 :

$$\begin{aligned} \eta_i &= b_0 + b_1 \cdot u_i \\ y_{ij} &\sim N(\eta_i, \sigma_y^2),\ \forall i = 1, .., n, j = 1, .., J. \end{aligned}$$

(iii) Generation of the signal of the read sample, given the true values of the sample u_0, the parameters of the calibration function b_0, b_1 and the variance of the signals σ_y^2 :

$$\begin{aligned} \eta_0 &= b_0 + b_1 \cdot u_0 \\ y_0 &\sim N(\eta_0, \sigma_y^2). \end{aligned}$$

For all data sets, the assigned values of the calibrators are set to $\mathbf{u} = (3, 6, 9, 12, 15)$, the variance of the signals $\sigma_y^2 = 0.01$ and the intercept of the calibration function $b_0 = 0$. Four independent signals are generated for each calibrator. The true value of the read sample is set to $u_0 = 5$. The other parameters of the simulation vary, according to

$$b_1 \in \{1, 5\},\ Cor(x_i, x_j) \in \{0, 0.8\}\ \forall i \neq j,\ \sigma_x^2 \in \{0.01, 0.1, 1\},$$

resulting in 12 different data situations. For each of these settings, 100 data sets are simulated and analyzed. The posterior distributions are calculated with the `MCMCmetrop1R` function of package **MCMCpack** [MQ06] in R2.3.1 [R D06] for both models.

For each MCMC algorithm, the first 1000 samples are used as burn-in, afterwards 10.000 samples are saved. In Table 3.2, the bias, the coverage of the 90% empirical confidence interval and the length of this interval are given for the parameters $\sigma_y^2, b_0, b_1, u_0$, derived in the reading step. For each data set the median of the samples of the posterior distributions is taken as point estimator of the parameter. The mean of these estimates averaged over the 100 data sets minus the true value becomes the bias for each data situation. The 90% empirical confidence interval is defined as the interval between the 5% and 95% empirical quantiles of the samples. The coverage is calculated by counting how many times the 90% empirical confidence interval covers the respective true value. The length of the confidence interval is the averaged length over the 100 data sets.

Summarizing Table 3.2, we see that for all data situations taking the errors of the assigned values not into account, leads to an overestimation of σ_y^2. This becomes worse with increasing σ_x^2. The overestimation is also greater if there is no correlation between the assigned values.
The coverage of the confidence intervals for σ_y^2, b_0 and b_1 drops very fast with increasing σ_x^2 and is even less for a steeper slope. The coverage of the confidence interval of u_0 seems to be stable in cases of no correlation between the assigned values, however this is especially due to the high overestimation of σ_y^2. For correlated assigned values, the effect of overestimation of σ_y^2 is less, hence the coverage of the confidence intervals for u_0 drops with increasing σ_x^2, too.
Regarding the case of steeper calibration curve ($b_1 = 5$ in the simulation of the data sets) we note that these effects are amplified.

If the errors of the assigned values are taken into account we obtain nearly in all cases an unbiased estimation of σ_y^2.
The coverage of the confidence intervals for σ_y^2, b_0, b_1 and u_0 is stable in situations without correlation and drops slightly in the correlation case.
The confidence intervals of u_0 are sometimes even shorter than in the case of not accounting for the errors of the assigned values with approximately the same coverage. This is due to the large overestimation of σ_y^2, if the errors of the assigned values are not modelled explicitly.
In summary we can say that the coverage results are quite satisfactory for all data situations, if the errors of the assigned values are taken into account, except for the last data situation ($\sigma_x^2 = 1$, $Cor(x_i, x_j) = 0.8$, $b_1 = 5$), although even in this situation the results are better than for the simple algorithm. In this case u_0 is underestimated, because of the overestimation of b_0. This leads to the smaller coverage.

This simulation study shows the necessity of modelling the errors in the assigned values and to incorporate this knowledge in the calibration and sample reading stage. The need for this modelling even grows, the steeper the calibration function becomes. A focus in the improvement of diagnostic assays is nowadays to obtain a steeper calibration function, as in this case the errors of the signals are attenuated. We showed however, that the steeper the calibration function, the more important become the errors of the assigned values. Only if they are modelled in the right way, a steeper calibration function might be an improvement of the diagnostic assay.
The incorporation of the measurement errors of the assigned values in the calibration process can easily be done by MCMC simulation algorithms, to obtain the posterior distribution for a single measured value of a sample.

Table 3.2: **Bias, coverage and length of the 90% empirical confidence interval for the linear calibration, without and with taking the error of the assigned values into account.**

Simulation		Without measurement error				With measurement error			
		σ_y^2	b_0	b_1	u_0	σ_y^2	b_0	b_1	u_0
$\sigma_x^2 = 0.01$	Bias	-0.01	-0.02	0.000	-0.02	-0.01	-0.02	0.001	-0.02
$Cor(x_i, x_j) = 0$	Coverage	84%	83%	77%	90%	87%	87%	90%	92%
$b_1 = 1$	Length CI	0.14	0.60	0.060	1.19	0.14	0.69	0.086	1.21
$\sigma_x^2 = 0.1$	Bias	-0.08	-0.02	0.001	0.02	-0.01	-0.01	0.000	0.02
$Cor(x_i, x_j) = 0$	Coverage	55%	76%	70%	95%	88%	94%	95%	93%
$b_1 = 1$	Length CI	0.22	0.75	0.075	1.50	0.14	1.24	0.132	1.35
$\sigma_x^2 = 1$	Bias	-0.59	-0.06	0.012	0.00	-0.01	0.07	-0.003	-0.03
$Cor(x_i, x_j) = 0$	Coverage	5%	52%	46%	93%	90%	86%	84%	88%
$b_1 = 1$	Length CI	0.85	1.45	0.150	2.88	0.15	3.00	0.310	2.11
$\sigma_x^2 = 0.01$	Bias	-0.18	-0.05	0.005	0.00	-0.01	-0.06	0.010	0.00
$Cor(x_i, x_j) = 0$	Coverage	27%	61%	64%	97%	80%	94%	97%	97%
$b_1 = 5$	Length CI	0.35	0.93	0.093	0.37	0.15	1.83	0.336	0.43
$\sigma_x^2 = 0.1$	Bias	-1.94	0.02	-0.010	0.01	-0.00	0.08	-0.015	0.01
$Cor(x_i, x_j) = 0$	Coverage	0%	52%	52%	94%	83%	87%	92%	91%
$b_1 = 5$	Length CI	2.49	2.45	0.25	0.962	0.14	5.26	0.619	0.73
$\sigma_x^2 = 1$	Bias	-4.63	-0.87	0.114	0.07	-0.11	0.24	-0.010	-0.02
$Cor(x_i, x_j) = 0$	Coverage	0%	26%	23%	75%	63%	77%	79%	79%
$b_1 = 5$	Length CI	5.81	3.76	0.37	1.57	0.71	11.83	1.39	1.69
$\sigma_x^2 = 0.01$	Bias	-0.01	-0.01	0.000	-0.01	-0.01	-0.02	0.000	-0.01
$Cor(x_i, x_j) = 0.8$	Coverage	85%	83%	92%	89%	85%	89%	93%	92%
$b_1 = 1$	Length CI	0.13	0.59	0.059	1.17	0.14	0.69	0.084	1.26
$\sigma_x^2 = 0.1$	Bias	-0.01	-0.12	0.004	0.07	-0.00	-0.13	0.004	0.08
$Cor(x_i, x_j) = 0.8$	Coverage	88%	50%	90%	78%	91%	81%	98%	85%
$b_1 = 1$	Length CI	0.14	0.60	0.060	1.20	0.13	1.15	0.090	1.50
$\sigma_x^2 = 1$	Bias	-0.14	-0.11	0.012	-0.01	-0.01	-0.12	0.009	0.01
$Cor(x_i, x_j) = 0.8$	Coverage	28%	34%	59%	63%	84%	76%	90%	76%
$b_1 = 1$	Length CI	0.30	0.86	0.086	1.72	0.15	2.31	0.170	2.26
$\sigma_x^2 = 0.01$	Bias	-0.04	-0.04	0.004	0.00	-0.00	-0.07	0.009	0.00
$Cor(x_i, x_j) = 0.8$	Coverage	74%	45%	74%	78%	89%	89%	95%	94%
$b_1 = 5$	Length CI	0.17	0.65	0.070	0.26	0.14	1.77	0.290	0.51
$\sigma_x^2 = 0.1$	Bias	-0.37	0.20	-0.003	-0.04	-0.01	0.16	-0.002	-0.03
$Cor(x_i, x_j) = 0.8$	Coverage	14%	31%	45%	53%	87%	84%	93%	84%
$b_1 = 5$	Length CI	0.59	1.20	0.120	0.47	0.16	4.67	0.410	0.94
$\sigma_x^2 = 1$	Bias	-3.19	-1.24	0.024	0.23	-0.07	-1.29	-0.003	0.28
$Cor(x_i, x_j) = 0.8$	Coverage	1%	12%	43%	39%	74%	56%	84%	56%
$b_1 = 5$	Length CI	4.05	3.06	0.310	1.25	0.37	10.39	1.070	1.80

Chapter 4

Combining Multiple Measurements

In Chapter 3 we showed, how the posterior distribution of a single measurement of a sample is derived. This posterior distribution contains already the errors of the assigned values of the calibrators.
In standardization networks the derivation of the assigned value of a sample is based on multiple measurements of this sample, obtained in different laboratories and multiple measurements per laboratory. In this way it is accounted for laboratory specific effects which would otherwise introduce a bias in the estimation of the assigned value of a sample. In this chapter we discuss, how these multiple measurements, made in different laboratories should be combined to estimate the assigned value of the sample.
This problem was first considered by [Coc37] and revised several times in [YC54], [Coc54] and [RKC81]. They all modelled this data situation as a one-way random effects model and examined maximum-likelihood estimation techniques.
We will follow another way in this chapter, which is appropriate if not only the measurement results are known, but also the posterior distribution of these results.

4.1 One-way random effects models

Suppose we have $i = 1, .., I$ laboratories with $j = 1, .., J_i$ repetitions per laboratory. Denote with Y_{ij} the measurement of the concentration of a sample made in laboratory i and repetition j, further with $\bar{Y}_{i.} = 1/J_i \cdot \sum_j Y_{ij}$ the mean of the measurements of laboratory i and with $\bar{Y}_{..} = 1/I \cdot \sum_i \bar{Y}_{i.}$ the overall mean.
Each laboratory has an individual effect on the measurements, such that we model this

data as a one-way random effects model:

$$Y_{ij} = \mu + a_i + \varepsilon_{ij},\ i = 1, .., I,\ j = 1, .., J_i.$$

The parameter μ denotes the true concentration of the sample, a_i the effect of laboratory i and ε_{ij} the error term. The laboratory effects $a_i,\ i = 1, .., I$, have mean zero and variance σ_a^2 - the so-called between-laboratory variance. The errors terms have mean zero and variance $\sigma_{\varepsilon_i}^2$ - called the within-laboratory variances. In most applications laboratory effects and error terms are assumed normally distributed.
Dependent on the number of replicates within the laboratories and the within-laboratory variances different one-way random effects models can be formulated.
We will shortly present the two most common models and estimation techniques, afterwards we introduce our estimation approach. This new approach takes the whole posterior distribution of each single measurement into account.
The three approaches are compared based on a simulation study, which is explained in detail in Section 4.2.

If the number of replicates in the laboratories is equal and the within-laboratories variances are equal, too, i.e.

$$J_i = J,\ \sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2,\ \forall\ i = 1, .., I,$$

we speak about the balanced, homoscedastic one-way random effects model.
In this model all observations are treated equally for the estimation of the unknown parameters $(\mu, \sigma_a^2, \sigma_\varepsilon^2)$. [SCM92] give a broad overview of the different estimation techniques for these parameters. The most common approach is based on the ANOVA table of the model, which is given in Table 4.1, and the derivation of the expected mean squares.

Table 4.1: **ANOVA Table for the homoscedastic one-way random effects model.**

Sum of Squares	DF	Mean Squares	Expected Mean Squares
$SSA = J\sum_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$I - 1$	$MSA = \frac{SSA}{I-1}$	$J\sigma_a^2 + \sigma_\varepsilon^2$
$SSE = \sum_i\sum_j(Y_{ij} - \bar{Y}_{i.})^2$	$I(J-1)$	$MSE = \frac{SSE}{I(J-1)}$	σ_ε^2

The mean squares are unbiased estimators of the expected mean squares. Hence, setting the two expressions equal and solving the resulting equations for the σ's gives unbiased estimators for σ_a^2 and σ_ε^2. This leads to

$$\hat{\sigma}_a^2 = \frac{MSA - MSE}{J}, \qquad \hat{\sigma}_\varepsilon^2 = MSE. \tag{4.1.1}$$

An unbiased estimator for μ is given by $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ (see [SCM92]), with variance

$$\sigma^2(\hat{\mu}) = \frac{1}{I}\sigma_a^2 + \frac{1}{IJ}\sigma_\varepsilon^2.$$

Under normality assumptions a two-sided $1 - \alpha$ $(0 < \alpha < 1)$ coverage interval for μ is given by

$$\bar{Y}_{\cdot\cdot} \pm t_{I-1,1-\alpha/2} \cdot \hat{\sigma}(\hat{\mu}),$$

where $t_{I-1,1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the t-distribution with $I - 1$ degrees of freedom.

Coverage intervals for σ_ε^2 and σ_a^2 can also be derived. For σ_ε^2 an exact coverage interval is available, as $SSE/\sigma_\varepsilon^2 \sim \chi^2_{I(J-1)}$. Hence, a two-sided $1 - \alpha$ coverage interval for σ_ε^2 is

$$\left[\frac{SSE}{\chi^2_{I(J-1),1-\alpha/2}}, \frac{SSE}{\chi^2_{I(J-1),\alpha/2}}\right].$$

For σ_a^2 only approximate coverage intervals are available. [BG92] derive such intervals based on the Cornish-Fischer expansion [FC60]. The interested reader is referred to [BG92] for more details on this issue.

The unbalanced, heteroscedastic one-way random effects model is given, if the number of replicates within the laboratories as well as the within-laboratory variance are not equal in all laboratories. [RKC81] give an overview of estimators for $(\mu, \sigma_a^2, \sigma_{\varepsilon_i}^2)$ for this case.
[MP70] and [PM82] developed a simplified iterative estimation algorithm for the parameters of this model, called the Mandel-Paule algorithm. It is widely used to combine several sets of measurements from different laboratories. Therefore we will present this algorithm in short and compare it to our proposed algorithm.
The Mandel-Paule algorithm is based on the following ideas: In the heteroscedastic

one-way random effects model the variance of an individual measurement is given by

$$Var(Y_{ij}) = \sigma_a^2 + \sigma_{\varepsilon_i}^2,\ \forall i = 1,..,I,\ j = 1,..,J_i$$

and the correlation between two measurements of the same laboratory by

$$Cor(Y_{ij}, Y_{ik}) = \sigma_a^2,\ \forall i = 1,..,I,\ j,k = 1,..,J_i,\ j \neq k.$$

Therefore, the variances of the laboratory means calculate to

$$Var(\bar{Y}_{i.}) \quad = \quad \sigma_a^2 + \frac{1}{J_i}\sigma_{\varepsilon_i}^2,\ \forall i = 1,..,I.$$

Hence, the estimator of the parameter μ is a weighted means statistic

$$\tilde{\mu} = \frac{\sum_i w_i \cdot \bar{Y}_{i.}}{\sum_i w_i}, \tag{4.1.2}$$

with weights given by

$$w_i = \frac{1}{Var(\bar{Y}_{i.})} = \frac{1}{\sigma_a^2 + \frac{1}{J_i}\sigma_{\varepsilon_i}^2}.$$

The within-laboratory variances are estimated in advance from the repeated measurements in each laboratory by

$$\tilde{\sigma}_{\varepsilon_i}^2 = \frac{1}{J_i - 1} \cdot \sum_{j=1}^{J_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2.$$

As the estimator of μ requires a plug-in estimator of σ_a^2, too, an iterative estimation algorithm is set up.
Because

$$F(\sigma_a^2) = \sum_{i=1}^{I} \frac{\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}{\sigma_a^2 + \frac{1}{J_i}\sigma_{\varepsilon_i}^2} \sim \chi_{I-1}^2,$$

we have $E\left(F(\sigma_a^2)\right) = I - 1$, which motivates the estimating equation

$$F(\sigma_a^2) = I - 1 \quad \text{or} \quad G(\sigma_a^2) = F(\sigma_a^2) - (I - 1) = 0. \tag{4.1.3}$$

The searched σ_a^2 can be found by using a truncated Taylor expansion of G.
More precisely for a given value $\sigma_{a_t}^2$ we want to find an adjustment $d\sigma_a^2$, such that

$G(\sigma^2_{a_t} + d\sigma^2_a) = 0$. Using the truncated Taylor expansion of G around $\sigma^2_{a_t}$, we have

$$G(\sigma^2_{a_t} + d\sigma^2_a) \approx G(\sigma^2_{a_t}) + \left(\frac{\partial G}{\partial \sigma^2_a}\right)_{\sigma^2_{a_t}} d\sigma^2_a = 0.$$

Hence it follows that

$$d\sigma^2_a = -\frac{G(\sigma^2_{a_t})}{\left(\frac{\partial G}{\partial \sigma^2_a}\right)_{\sigma^2_{a_t}}}.$$

Having everything together we define the Mandel-Paule algorithm by:
Start with an initial value for σ^2_a, say $\sigma^2_{a_0}$.

Step 1 Compute weights

$$w_{i_t} = \sigma^2_{a_{t-1}} + \frac{1}{J_i}\sigma^2_{\varepsilon_i}$$

and

$$\mu_t = \frac{\sum_i w_{i_t} \cdot \bar{Y}_{i.}}{\sum_i w_{i_t}}.$$

Step 2 Compute

$$d\sigma^2_{a_t} = \frac{F(\sigma^2_{a_{t-1}}) - (I-1)}{\sum_i w^{-2}_{i_t} \cdot (\bar{Y}_{i.} - \mu_t)^2}.$$

Step 3 If $d\sigma^2_{a_t} > \varepsilon$ compute

$$\sigma^2_{a_t} = \sigma^2_{a_{t-1}} + d\sigma^2_{a_t},$$

set $t = t + 1$ and go to Step 1.
Else stop, set $\tilde{\mu} = \mu_t$ and $\tilde{\sigma}^2_a = \sigma^2_{a_{t-1}}$.

Mandel and Paule [PM82] approximate the variance of $\tilde{\mu}$ by

$$Var(\tilde{\mu}) \approx \frac{1}{\sum_{i=1}^{I} w_i},$$

by ignoring the variation within the weights. An $1 - \alpha$ coverage interval for μ can then be defined by

$$\tilde{\mu} \pm t_{I-1,1-\alpha/2} \cdot \sqrt{Var(\tilde{\mu})}$$

where $t_{I-1,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the t-distribution with $I-1$ degrees of freedom.
[RV98] and [RBV00] showed, that the Mandel-Paule algorithm can be seen as a special form of maximum-likelihood or restricted maximum-likelihood estimation.

In our eyes both models lack an important detail. Each single measurement is regarded as a single value, the repeated measurements within a laboratory give information on the error variance of the measurements within the laboratory.
However, if we use the Bayesian approach to sample reading developed in Chapter 3, we do not have only a point estimate of each single measurement, but a whole posterior distribution for it. From this distribution one can calculate the variance of the errors, such that we can speak in this case of an unbalanced one-way random effects model with known error variance.
We write this model in following way:

$$\begin{aligned} Y_{ij}|\mu, a_i &\sim p(Y_{ij}),\ \forall i = 1,..,I,\ j = 1,..,J_i, \\ E(Y_{ij}|\mu, a_i) &= \mu + a_i,\ \forall i = 1,..,I,\ j = 1,..,J_i, \\ a_i|\sigma_a^2 &\sim N(0, \sigma_a^2),\ \forall i = 1,..,I. \end{aligned} \tag{4.1.4}$$

The likelihood function of the model is given by the product of the posterior distributions of the measurements, i.e.

$$p(\mathbf{Y}|\mu, \mathbf{a}) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} p(Y_{ij}).$$

The full posterior distribution of this model is proportional to

$$p(\mu, \sigma_a^2|\mathbf{Y}) \propto p(\mathbf{Y}|\mu, \mathbf{a}) \cdot p(\mathbf{a}|\sigma_a^2) \cdot p(\mu, \sigma_a^2).$$

If the likelihood functions $p(Y_{ij})$ are approximated by normal distributions and the priors for μ and σ_a^2 are defined as full conditional priors, Gibbs sampling can be used to obtain samples of the posterior. Otherwise Metropolis-Hastings algorithms may be adequate.
In the following section we compare the three estimation approaches based on a small simulation study.

4.2 Comparison of the models

In this section we compare the outcomes of a balanced one-way classification model with (i) homoscedastic within-laboratory variance, (ii) heteroscedastic within-laboratory variance and (iii) known variance of each measurement, for the estimation of μ and σ_a^2. We regard the 12 different data situations already introduced in Section 3.3.2, for the errors of the assigned values of the calibrators. For each data situation 8 laboratories are simulated, with four repeated measurements per laboratory. 100 data sets are regarded for each data situation.

In each laboratory the same set of calibrators is used. Each calibrator has an assigned value, that carries already an uncertainty. Similar to the simulations in Chapter 3 there are 5 calibrators with true values $\mathbf{u} = (3, 6, 9, 12, 15)$. Dependent on the data situation the variance of the assigned values is $\sigma_x^2 \in \{0.01, 0.1, 1\}$ and their correlation $Cor(x_i, x_j) \in \{0, 0.8\}$. In each laboratory a linear calibration function ($b_0 = 0, b_1 \in \{1, 5\}$) is derived based on the signals and assigned values of the calibrators. The variance of the signals is set to 0.01. From each calibration function a sample with true concentration value of 5 is read.

As we apply for the calibration and the reading the Bayesian model established in Chapter 3, we do not only obtain a point estimate of the concentration of the sample, but the whole posterior distribution. The posterior distribution of the measurements is derived based on the MCMC algorithm of Section 3.2.1, where the errors of the assigned values of the calibrators are taken into account.

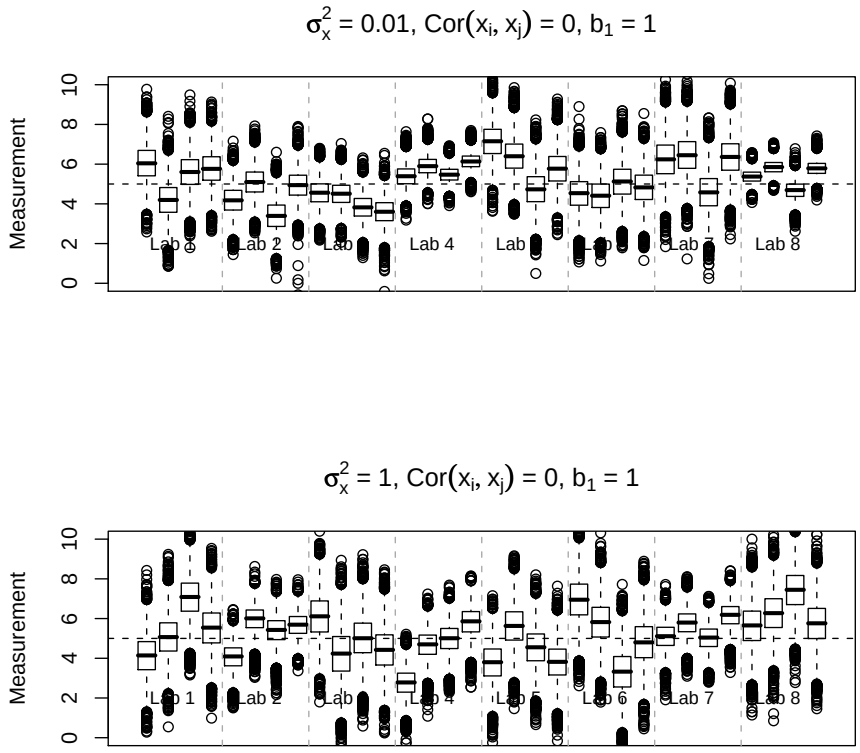
To include a laboratory effect in this data, we draw for each laboratory a laboratory effect a_i, from a normal distribution with mean 0 and variance 0.5. Afterwards, the derived posterior distributions are shifted by this specific laboratory effect.

Finally, heteroscedasticity is included in the data by assigning to each laboratory a further error variance between 0.1 and 1, such that to each sample of the posterior distributions a normally distributed error term with mean zero and the respective variance is added.

In Figure 4.1 a randomly selected data set for the first and third data situation is shown. For the estimation of the parameters μ and σ_a^2 by the homoscedastic or heteroscedastic one-way random effects models only the point estimator (given by the median of the posterior distribution) of the measured values is used, whereas in the third model more information of the posterior distributions is incorporated.

For this simulation study we approximate the posteriors by a normal distribution, with variance estimated from the simulated samples and mean given by the median of them. All three models are fit with the algorithms described in Section 4.1.

Figure 4.1: **Boxplots of the posterior distributions of a read sample in 8 laboratories and 4 measurements within each laboratory. Two data situations, dependent on the variance structure of the assigned values of the calibrators is shown**



To fit the third model MCMC algorithms based on OpenBUGS2.2.0 [STBL05] and **BRugs** [Lig06] are used. The prior of μ is uniform between 0 and 16, whereas the priors of the unknown between-laboratory variance is chosen as $InvGamma(0.001, 0.001)$. A burn-in of 1000 simulations is used, afterwards every 50th sample is saved, until 10.000 values are obtained. Three different chains are run, such that the convergence of the chains could be assessed by the potential scale reduction factor (see Section 2.2).
In Table 4.2 the bias, the coverage of the 95% confidence intervals and the length of the confidence intervals are given. All three models give unbiased estimates for μ. Regarding the estimates for the between-laboratory variance we note that the Bayesian model underestimates this variance. Neither do the derived confidence intervals cover the true value with enough confidence. Regarding the confidence intervals for μ we note that the Bayesian model has the lowest coverage, but which is still acceptable, whereas the other two models have sometimes even a higher coverage than 95%. On the other side the confidence intervals of the third model are much smaller than the one of the other models. Note especially that in these intervals the uncertainty of the calibrators of the reading step is already included.
Regarding the different data situations we see that the behavior concerning the length of the confidence intervals of μ is the same as already seen in Chapter 3. If the uncertainty of the assigned values of the calibrators, or their correlation increases, they become wider. A steeper slope results in smaller confidence intervals.
In summary we see that the uncertainty information of the measurement process can easily be incorporated into the combination of multiple repeated measurements, to obtain reasonable estimates of the assigned value of a calibrator and its uncertainty.

Table 4.2: **Bias, coverage and length of the 95% empirical confidence intervals for the parameters μ and σ_a^2 from the different one-way random effects models.**

Simulation		Homoscedastic model		Heteroscedastic model		Model with known $\sigma^2_{\varepsilon_{ij}}$	
		μ	σ_a^2	μ	σ_a^2	μ	σ_a^2
$\sigma_x^2 = 0.01$	Bias	-0.04	0.01	-0.04	-0.01	-0.05	-0.12
$Cor(x_i, x_j) = 0$	Coverage	97%	95%	96%	96%	96%	89%
$b_1 = 1$	Length CI	1.30	1.74	1.30	1.74	1.23	1.17
$\sigma_x^2 = 0.1$	Bias	0.03	-0.07	0.03	-0.09	0.03	-0.15
$Cor(x_i, x_j) = 0$	Coverage	95%	84%	95%	82%	92%	87%
$b_1 = 1$	Length CI	1.37	1.98	1.37	1.98	1.29	1.22
$\sigma_x^2 = 1$	Bias	-0.05	-0.03	-0.06	-0.04	-0.06	-0.15
$Cor(x_i, x_j) = 0$	Coverage	97%	90%	98%	90%	97%	88%
$b_1 = 1$	Length CI	1.41	2.07	1.43	2.07	1.32	1.27
$\sigma_x^2 = 0.01$	Bias	-0.01	-0.09	-0.01	-0.09	-0.01	-0.19
$Cor(x_i, x_j) = 0$	Coverage	94%	92%	93%	90%	93%	78%
$b_1 = 5$	Length CI	1.37	1.96	1.37	1.96	1.30	1.21
$\sigma_x^2 = 0.1$	Bias	-0.04	0.07	-0.03	0.06	-0.04	-0.06
$Cor(x_i, x_j) = 0$	Coverage	97%	87%	95%	88%	95%	92%
$b_1 = 5$	Length CI	1.21	1.54	1.21	1.54	1.13	1.08
$\sigma_x^2 = 1$	Bias	-0.01	0.029	-0.01	0.01	0.00	-0.13
$Cor(x_i, x_j) = 0$	Coverage	95%	91%	97%	90%	95%	87%
$b_1 = 5$	Length CI	1.35	1.90	1.38	1.90	1.26	1.20
$\sigma_x^2 = 0.01$	Bias	-0.03	-0.02	-0.03	-0.03	-0.03	-0.13
$Cor(x_i, x_j) = 0.8$	Coverage	96%	93%	96%	91%	96%	88%
$b_1 = 1$	Length CI	1.33	1.83	1.32	1.83	1.25	1.19
$\sigma_x^2 = 0.1$	Bias	0.06	0.001	0.07	0.01	0.05	-0.11
$Cor(x_i, x_j) = 0.8$	Coverage	94%	87%	95%	89%	92%	90%
$b_1 = 1$	Length CI	1.33	1.82	1.35	1.82	1.24	1.19
$\sigma_x^2 = 1$	Bias	0.03	0.02	0.03	-0.02	0.03	-0.22
$Cor(x_i, x_j) = 0.8$	Coverage	100%	95%	100%	94%	98%	79%
$b_1 = 1$	Length CI	1.51	2.30	1.59	2.30	1.41	1.35
$\sigma_x^2 = 0.01$	Bias	-0.04	0.01	-0.04	-0.03	-0.03	-0.11
$Cor(x_i, x_j) = 0.8$	Coverage	96%	86%	94%	89%	93%	89%
$b_1 = 5$	Length CI	1.27	1.69	1.28	1.69	1.20	1.13
$\sigma_x^2 = 0.1$	Bias	0.01	-0.01	0.01	-0.02	0.01	-0.13
$Cor(x_i, x_j) = 0.8$	Coverage	95%	90%	97%	90%	92%	87%
$b_1 = 5$	Length CI	1.31	1.80	1.32	1.80	1.23	1.16
$\sigma_x^2 = 1$	Bias	0.20	0.09	0.22	0.07	0.20	-0.18
$Cor(x_i, x_j) = 0.8$	Coverage	95%	97%	95%	95%	92%	85%
$b_1 = 5$	Length CI	1.48	2.11	1.58	2.11	1.33	1.26

Part II

Outlier identification

The second part of the thesis deals with the identification of outliers within data from laboratory networks. Let us remind the structure of data from laboratory networks: Samples of different concentration are sent to each laboratory where multiple readings per sample are made. Hence, we have a hierarchy in the data: the lowest level consists of several observations within a laboratory, from one or more samples. The individual laboratories form the next higher level. Hence, different types of outliers within the data can be defined.
A laboratory is an outlier within the network if its measurements of all samples are different, compared to the measurements of the other laboratories. A single measurement of a sample within a laboratory is an outlier, if it is extreme compared to the other measurements of this specific sample in the respective laboratory.
[Man95] addressed already this issue, having the idea to model data from multiple samples, measured within a laboratory network as a linear model for each laboratory. However, he takes not into account the inherent variability between laboratories. Linear mixed models, especially the random coefficients model, close this gap: The coefficients of each laboratory are seen as random draws from a multivariate normal distribution, which marks the variability between laboratories. In consequence, laboratories with extreme coefficients can be regarded as outliers within the network.
The linear mixed model approach can also be used to identify extreme single measurements of individual samples, in cases that single samples are regarded. [WG03] propose a robust one-way random effects model for this setting. But, the robust estimation methods used therein are especially applicable to this model.
In Chapter 5, maximum-likelihood estimation of linear mixed models is introduced and outlier identification rules for this model class are developed. But, working with normal-linear mixed models, may lead to masking or swamping effects for such identifications, as maximum-likelihood estimation is highly influenced by extreme observations. Masking occurs if outliers are not identified due to the incorrect estimation of parameters. Swamping means that data points, which are no outliers, are falsely identified (see [GD93]). In Chapter 6, we present a robust estimation approach based on linear mixed models with t-distributed random effects. This robust estimation approach can be used for the random coefficients model and for the one-way random effects model, such that both identification tasks can be performed based on this estimation method.
Finally in Chapter 7, we present how outlier identification rules for laboratory networks can be derived that hold for different studies. If these rules should be applied to different studies over the course of time, the dispersion parameters of the linear mixed models must be set in advance, such that the same limits hold. We present how these parameters can be derived from historical data and give an interpretation of them.

Chapter 5

Linear mixed models

Linear mixed models are a powerful tool, if data from different subjects is gathered and one wants to model subject-specific effects as well as population effects. In case of data from laboratory networks these subjects are the individual laboratories. Therefore we will refer from now on to laboratory effects. [Man95] has been the first to model data from multiple samples, measured within a laboratory network, as a linear model. However he does not introduce a variation structure between laboratories.
In this chapter we present some general tools and techniques for dealing with linear mixed models. For more detailed discussions on this large issue we refer to [VM00] or [PB00].
Furthermore we discuss two particular linear mixed models and their application to laboratory network data.

5.1 General model

Suppose I laboratories contribute to the data. For laboratory i, the general form of the linear mixed model is given by (see [LW82])

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\beta_i + \varepsilon_i \\ \beta_i &\sim N_k(0, \mathbf{D}) \\ \varepsilon_i &\sim N_{n_i}(0, \mathbf{R}_i), \end{aligned} \tag{5.1.1}$$

where $\mathbf{Y}_i \in \mathbb{R}^{n_i}$ denotes the data vector from laboratory i, $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ is the design matrix of the fixed effects $\mathbf{b} \in \mathbb{R}^p$, $\mathbf{Z}_i \in \mathbb{R}^{n_i \times k}$ the design matrix of the random effects $\beta_i \in \mathbb{R}^k$. The random effects are assumed to be normally distributed with mean 0 and

variance-covariance matrix $\mathbf{D}$. Finally ε_i is the residual vector of laboratory i, normally distributed with mean 0 and variance-covariance matrix $\mathbf{R}_i$, independent of β_i. Furthermore it is assumed that the random variables of different laboratories are independent from each other.
The variance-covariance matrix $\mathbf{R}_i$ of the errors allows the modelling of correlation structures among the errors. To avoid the estimation of many parameters, it is often assumed that this matrix is known up to a laboratory-specific constant. For our applications we will set $\mathbf{R}_i = \sigma_\varepsilon^2 \cdot \mathbf{I}_{n_i}$.
We introduce the linear mixed model with random effects and errors being multivariate normally distributed, which is the most common approach. The normal distribution allows for an easy derivation of maximum-likelihood estimators of the parameters. However, as this estimation is sensitive to the presence of outliers, we will introduce a robust estimation procedure in Chapter 6.
It follows from (5.1.1) that conditional on the random effects β_i, $\mathbf{Y}_i$ is normally distributed with mean $\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\beta_i$ and variance-covariance matrix $\mathbf{R}_i$. The marginal density function of $\mathbf{Y}_i$ is then

$$\mathbf{Y}_i \quad \sim \quad N_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{V}_i\right), \tag{5.1.2}$$

where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$. Note that the marginal model follows from the general linear mixed model (5.1.1) but this is not true the other way round. Model (5.1.1) is much more restrictive, as all variance-covariance matrices $\mathbf{D}$, $\mathbf{R}_i$, need to be positive definite. In the marginal model this needs only to be true for the matrix $\mathbf{V}_i$.
Denote with $\Theta = (\mathbf{b}, \gamma)$ the vector of the unknown parameters, where γ contains all dispersion parameters. Let $\Omega = \Omega_\mathbf{b} \times \Omega_\gamma$ be the parameter space of the fixed effects and dispersion parameters. Regarding model (5.1.1), Ω_γ is the set of values, such that $\mathbf{D}$ and $\mathbf{R}_i$ are positive definite. For the marginal model, Ω_γ comprises all values for γ, such that the $\mathbf{V}_i$'s are positive definite.

For the analysis of data from laboratory networks two particular models are of interest.
The one-way random effects model is appropriate, if data from one particular sample are regarded. We introduced this model already in Chapter 4, where we focus on the estimation of the overall sample mean and its uncertainty. Now we are interested in the estimation of the variance parameters and the prediction of the laboratory effects.

The one-way random effects model is written as

$$Y_{ij} = \mu + a_i + \varepsilon_{ij},\ \forall\ i = 1, .., I,\ \ j = 1, .., n_i$$

or in matrix notation $\forall\ i = 1, .., I$

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{1}_{n_i}\mu + \mathbf{1}_{n_i}a_i + \varepsilon_i \\ a_i &\sim N(0, \sigma_a^2) \\ \varepsilon_i &\sim N_{n_i}(0, \sigma_{\varepsilon_i}^2 \cdot \mathbf{I}_{n_i}), \end{aligned} \tag{5.1.3}$$

where $\mathbf{Y_i}$ is the vector of measurements of laboratory i. The parameter μ denotes the overall sample mean, the a_i's the laboratory effects, and ε_i the normally distributed error vector of laboratory i. The variance of the laboratory effects is referred to as between-laboratory variance, whereas $\sigma_{\varepsilon_i}^2$ as within-laboratory variance.
The one-way random effects model is well studied (see e.g. [SCM92]) and closed forms of the estimators of the model parameters are available. However, it can also be viewed as a special linear mixed model, which is treated like a general linear mixed model.
[WG03] developed an outlier-identification technique for this particular model, where extreme laboratories as well as extreme observations within a laboratory, can be detected. They use some robust estimation procedure, derived from the closed forms of the estimators of the parameters and defined inlier- as well as outlier-regions for the estimated laboratory effects and errors.
We will extend their ideas to other linear mixed models, and we present a robust estimation method which can be applied to general linear mixed models, too.

The second model is the random coefficients model. It is useful if data from multiple samples that are distributed over the whole concentration range are considered. This approach can be used for the approval of laboratories to participate within a network. It can be seen as a further extension of the ideas of [Man95].
The main idea is, that on one hand side to each sample k an overall sample concentration C_k can be assigned. On the other side we have measurements M_{ikj} of this sample in laboratory i. We regard the differences $M_{ikj} - C_k,\ \forall\ i, j, k$ and model them as a linear mixed model dependent on the C_k's. [Man95] shows that this is equivalent to model a sample-specific and laboratory-specific effect for the differences as well as an interaction effect between sample and laboratory.

The linear mixed model can be written in matrix notation as

$$\begin{aligned}
\mathbf{Y}_i = \begin{pmatrix} M_{i11} - C_1 \\ M_{i12} - C_1 \\ \vdots \\ M_{iK1} - C_K \\ \vdots \\ M_{iKn_i} - C_K \end{pmatrix} &= \begin{pmatrix} 1 & C_1 \\ 1 & C_1 \\ \vdots & \vdots \\ 1 & C_K \\ \vdots & \vdots \\ 1 & C_K \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} 1 & C_1 \\ 1 & C_1 \\ \vdots & \vdots \\ 1 & C_K \\ \vdots & \vdots \\ 1 & C_K \end{pmatrix} \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} + \varepsilon_i, \ \forall\, i = 1, .., I \\
\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} \sigma^2_{\beta_0} & \rho\sigma_{\beta_0}\sigma_{\beta_1} \\ \rho\sigma_{\beta_0}\sigma_{\beta_1} & \sigma^2_{\beta_1} \end{pmatrix} \right) \\
\varepsilon_i &\sim N_{n_i}\left(0, \sigma^2_\varepsilon \cdot \mathbf{I}_{n_i}\right),
\end{aligned} \tag{5.1.4}$$

where K denotes the number of measured samples, L the number of repeated measurements per sample, $(b_0, b_1)'$ the overall deviation from the overall sample location, which is expected to be close to zero.
We are interested in the prediction of the laboratory specific coefficients $(\beta_{0i}, \beta_{1i})'$, which can be seen as a laboratory specific systematic effect β_{0i} and proportional effect β_{1i}. They are the measure, the approval of laboratories within a network is based on. Further the estimation of the variance-covariance matrix $\mathbf{D}$ and of the variance of the errors σ^2_ε is of interest.

5.2 Maximum-likelihood estimation

The basic approach for the estimation of the parameters of a linear mixed model is maximum-likelihood estimation, based on the maximization of the likelihood of the marginal model with respect to Θ. The log-likelihood function of model (5.1.2) is given by

$$\begin{aligned}
L_{ML} &= \ln\left(\prod_{i=1}^{I} \left\{ (2\pi)^{-n_i/2} |\mathbf{V}_i|^{-1/2} \cdot \exp\{ -\frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \right\} \right) = \\
&= -\frac{1}{2} \sum_{i=1}^{I} \left(n_i \ln(2\pi) + \ln|\mathbf{V}_i| + (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b}) \right)
\end{aligned} \tag{5.2.1}$$

[LW82] showed that the MLE of $\mathbf{b}$ conditional on γ is given by

$$\hat{\mathbf{b}}(\gamma) = \left(\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i. \tag{5.2.2}$$

Hence, after replacing $\mathbf{b}$ by (5.2.2) the MLE of γ is obtained by maximizing (5.2.1). This approach is quite forward, when the estimation of $\mathbf{b}$ and γ is considered simultaneously by maximizing the joint likelihood. The maximization of (5.2.1) is usually done by Newton-Raphson based procedures, as e.g. in SAS PROC MIXED or in R with the `lme` function of the **nlme** package. For a detailed discussion of SAS PROC MIXED see [MSLW96], the reference for the **nlme** package is [PB00].
The maximum-likelihood estimation of the variance components does not take into account the loss of degrees of freedom, due to the estimation of the fixed effects. Accounting for this loss [Har74] introduced the restricted maximum-likelihood function (REML):

$$L_{REML}(\theta) = \left|\sum_{i=1}^{I} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right|^{-1/2} + L_{ML}(\theta). \tag{5.2.3}$$

In the examples that follow we estimate the parameters by REML estimation.

5.3 Best linear unbiased prediction

To detect extreme laboratories in the data of laboratory networks the random effects of both models must be estimated. They are estimated by best linear unbiased predictors (BLUP). According to [Rob91] this means:
"*BLUP estimates of the realized values of the random variables [..] are linear in the sense that they are linear functions of the data [..]; unbiased in the sense that the average value of the estimate is equal to the average value of the quantity being estimated; best in the sense that they have minimum mean squared error within the class of linear unbiased estimators; and predictors to distinguish them from estimators of fixed effects*".
It can be shown (e.g. [SCM92]) that the best linear unbiased predictor (BLUP) of the random effect β_i is given by

$$\hat{\beta}_i = BLUP(\beta_i) = \mathbf{D}\mathbf{Z}_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}). \tag{5.3.1}$$

Note that the BLUP's are functions of the dispersion parameters of the linear mixed model, they cannot be estimated independent of them. As the best linear unbiased predictor depends on the unknown parameters of the model, estimates of these parameters are plugged in the above equation for predicting the random effects.
The BLUP of the random effects is a shrinked estimator of the laboratory effects towards the overall mean. This means that its components are less spread then the least-square estimator, which will be obtained, if the components of β are regarded as fixed effects.
Maximum-likelihood estimation and best linear unbiased prediction, based on the normal distribution is sensitive to the presence of outliers within the data, see e.g. [PLW01]. As the important aim in the analysis of laboratory network data is the detection of deviating laboratories, we have to investigate more into this issue. In Chapter 6, we present an algorithm for the robust estimation of the parameters in the linear mixed model and compare both fitting strategies, by applying them to different data from laboratory networks.

5.4 Outliers in linear mixed models

The identification of extreme data is a very important point in the analysis of laboratory network data. It leads on one hand side to the formulation of rules for the approval of laboratories as members of the network. On the other side it points towards data which should not enter the calculation of the assigned value of a sample.
In linear models, residuals are often regarded to detect extreme observations. This concept is extended to the random effects of linear mixed models.
According to the formulation of the linear mixed model, certain types of outliers can be defined.

(i) Laboratory i is defined as location outlier, if its specific random effect β_i deviates from the majority of the random effects β_j, $j = 1, ..i-1, i+1, .., I$.

(ii) If the variance-covariance matrix of the errors is modelled as $\mathbf{R}_i = \sigma_\varepsilon^2 \cdot \mathbf{R}$, $\forall\ i = 1, .., I$ laboratory i is defined as scale outlier, if the laboratory-specific variation is extremely different from the pooled laboratory variation.

(iii) An observation within a laboratory is defined as location-outlier, if it is extreme in comparison to the other observations within this laboratory.

The given outlier definitions can be translated to the two linear mixed models in the following way: For the one-way random effects model we have

(i) Laboratory i as location-outlier, if all measurements of the specific sample are higher or lower than the measurements in the other laboratories.

(ii) Laboratory i as scale-outlier, if the variation of its measurements of the specific sample is higher than the variation of the measurements in the other laboratories.

(iii) A location-outlier within a laboratory, as an extreme single measurement within this laboratory.

For the random coefficients model we can translate these definitions to:

(i) Laboratory i is a location-outlier, if its systematic and/or proportional effect are quite different to these effects of the other laboratories.

(ii) Laboratory i is a scale-outlier, if the variation of the measurements around the laboratory-specific regression line is higher than for the other laboratories.

(iii) A location-outlier within a laboratory, is a single measurement, which deviates extremely from the laboratory-specific regression line.

[WG03] show how these types of outliers can be detected within a one-way random effects model. The basic ideas are to estimate the parameters of the model as well as the predictors of the random effects in a robust way. Afterwards inlier-regions are defined, in which the respective effects are supposed to lie with a given confidence. Effects outside these regions are considered as outliers.
We extend the concept of outlier/inlier regions to the class of linear mixed models, defined in (5.1.1). The definition of the inlier-regions is based on the decomposition of the residual sum of squares. One can show that the residual sum of squares

$$\delta_{t_i}^2 = (\mathbf{Y}_i - \mathbf{X}_i\mathbf{b})'(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}),$$

can be split up in the sum of squares of the random effects $\delta_{\beta_i}^2$ and the sum of squares of the conditional residuals $\delta_{\varepsilon_i}^2$. More precisely we have

$$\begin{aligned}\delta_{t_i}^2 &= (\mathbf{Y}_i - \mathbf{X}_i\mathbf{b})'(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}) = \\ &= \beta_i'\mathbf{D}^{-1}\beta_i + (\mathbf{Y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\beta_i)'\mathbf{R}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\beta_i) \\ &= \delta_{\beta_i}^2 + \delta_{\varepsilon_i}^2. \end{aligned} \tag{5.4.1}$$

In linear mixed models with normally distributed random effects and residuals, the distributions of these quadratic forms are known ([Sea71]):

$$\delta_{\beta_i}^2 \sim \chi_k^2 \qquad \delta_{\varepsilon_i}^2 \sim \chi_{n_i}^2.$$

As $E(\delta^2_{\beta_i}) = k$ it follows that $E(\delta^2_{\beta_i}/k) = E(\delta^2_{\varepsilon_i}/n_i) = 1$.
Outlier-regions for laboratories as location-outliers, as well as laboratories as scale-outliers can be defined based on the statistics $\delta^2_{\beta_i}/k$, $\delta^2_{\varepsilon_i}/n_i$.
According to the ideas of [GD93], the specified confidence level needs to be adjusted to the number of laboratories under consideration. For a given pre-specified level α and I laboratories under consideration, this can be done by setting $\alpha_I = 1 - (1-\alpha)^{(1/I)}$, in analogy to multiple testing strategies.

Hence, we define the following outlier identification rules for $0 \leq \alpha \leq 1$:

(i) Laboratory i is an $\alpha-$location-outlier, if

$$\frac{\delta^2_{\beta_i}}{k} > \frac{q^k_{1-\alpha_I}}{k},$$

where $q^k_{1-\alpha_I}$ denotes the $1-\alpha_I$ quantile of the χ^2_k distribution.

(ii) Laboratory i is an $\alpha-$scale-outlier, if

$$\frac{\delta^2_{\varepsilon_i}}{n_i} > \frac{q^{n_i}_{1-\alpha_I}}{n_i},$$

where $q^{n_i}_{1-\alpha_I}$ denotes the $1-\alpha_I$ quantile of the $\chi^2_{n_i}$ distribution.

Under the assumption that the residuals have the same variance and are uncorrelated, i.e. $\mathbf{R}_i = \sigma^2_\varepsilon \cdot \mathbf{I}_{n_i}$, $\delta^2_{\varepsilon_i}$ can be split up into the individual squares of the residuals. Then we can examine, if the scale differences are caused by a single extreme measurement, or due to a higher variation of all measurements within the respective laboratory.
Define ε_{ij} the jth component of the conditional residual vector $(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\beta_i)$. Then we have

$$\delta^2_{\varepsilon_i} = \frac{1}{\sigma^2_\varepsilon}\sum_{j=1}^{n_i}\varepsilon^2_{ij},$$

with

$$\frac{1}{\sigma^2_\varepsilon}\varepsilon^2_{ij} \sim \chi^2_1.$$

Hence, we define the measurement y_{ij} within laboratory i as $\alpha-$location-outlier, if

$$\frac{1}{\sigma^2_\varepsilon}\varepsilon^2_{ij} > q^1_{1-\alpha_n},$$

where $q^1_{1-\alpha_n}$ denotes the $1-\alpha_n$ quantile of the χ^2_1 distribution and $n = \sum n_i$.
The above outliers rules are defined on the following considerations: For a prespecified $\alpha \in (0,1)$

$$\begin{aligned} P(\exists\, i : \delta^2_{\beta_i} > l_\beta) &= \alpha \\ P(\exists\, i : \delta^2_{\varepsilon_i} > s_\varepsilon) &= \alpha \\ P(\exists\, i,\ j : \frac{1}{\sigma^2_\varepsilon}\varepsilon^2_{ij} > l_\varepsilon) &= \alpha, \end{aligned}$$

under model (5.1.1).
Under the assumption that the random effects and residuals are normally distributed the critical values $l_\beta, s_\varepsilon, l_\varepsilon$ become the respective Chi-square quantiles. It should be clear that when applying these rules to estimates of the dispersion parameters and predictors of the random effects this might no longer be true. Nevertheless we use these quantiles as a first approximation to the right critical values. Further research is needed to determine more appropriate values, for example through simulation studies.

Clearly in the presence of outliers, a robust estimation procedure for linear mixed models would be more appropriate. The robust estimation procedure for the one-way random effects model of [WG03] cannot be extended to more general models. [PLW01] developed an EM-algorithm for the robust fitting of linear mixed models, based on the t-distribution. The random effects and residuals are no longer assumed normally distributed, but t-distributed with appropriate degrees of freedom. Hence, the parameter space is extended to the degrees of freedom of the multivariate t-distribution. These can either be estimated from the data, or set in advance, corresponding to a tuning parameter for the robustness of the algorithm. The algorithm and its application to the models, which appear in the context of laboratory networks, is presented in Chapter 6.
In the last section of this chapter we present the application of the different outlier rules to the one-way random effects model and to the random coefficients model, based on the standard non-robust maximum-likelihood method.

5.5 Outlier identification by normal-linear mixed models

In this section we present the application of the outlier identification rules to the one-way random effects model as well as to the random coefficients model for data from

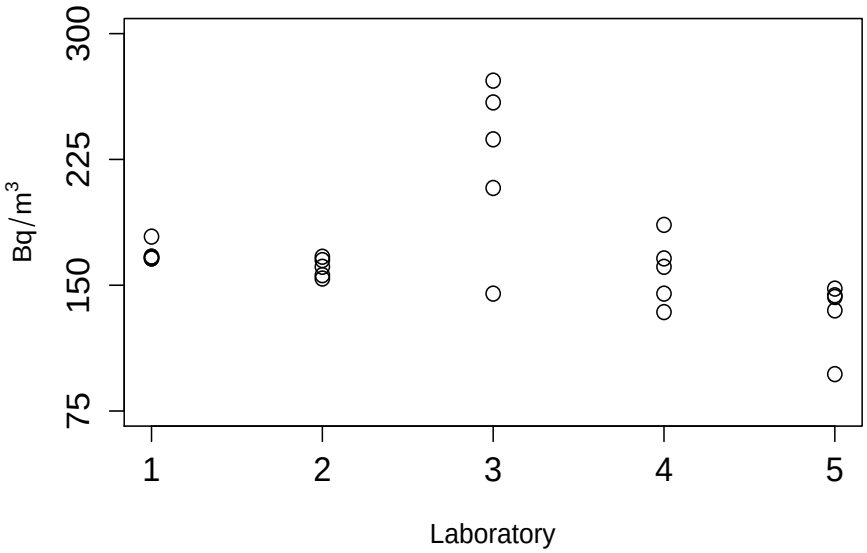
laboratory networks. Parameters of the models are estimated by REML estimation based on the linear mixed models with normally distributed random effects and errors.

5.5.1 One-way random effects model

For the one-way random effects model, we regard two examples, the first one is the data from an intercomparison trial of radon measurements, given in [WG03]. [WG03] demonstrated on this example the concepts of their outlier identification procedures, based on robust estimators of the parameters and outlier regions for (i) laboratories as location-outliers, (ii) laboratories as scale-outliers and (iii) measurements within laboratories as location-outliers.
They define outlier regions for the random effects and residuals, as well as for the individual within-laboratory variance. The limits of the outlier regions are found by simulation and they are defined such that the pre-specified confidence levels over all laboratories, or measurements are expected to hold. However, this procedure is specially designed for the one-way random effects model and can not be extended to general linear mixed models.
In a first step we want to apply our definitions of outliers, as given in Section 5.4, to the radon data, to see if we receive comparable results. In Figure 5.1, the radon measurements, obtained in 5 laboratories and 5 measurements within each laboratory, are shown. [WG03] identified with a confidence level of $\alpha = 0.1$ (i) laboratory 3 as a location-outlier within the laboratories, (ii) the lowest and highest observations of laboratory 3 and the lowest observation of laboratory 5 as location-outliers within these laboratories. No laboratories are identified as scale-outliers.
Modelling the data as a one-way random effects model with normally distributed random effects and errors and estimating the between-laboratories variance and within-laboratories variance by REML leads to $\hat{\sigma}_a^2 = 986.4$, and $\hat{\sigma}_\varepsilon^2 = 689.6$. These estimates are much higher than the variances, estimated in a robust way by [WG03], being $\tilde{\sigma}_a^2 = 87.723$, and $\tilde{\sigma}_\varepsilon^2 = 83.456$. This is mostly due to the very high variation of laboratory 3, with empirical variance of 2548.
Setting $\alpha = 0.1$, the limits for the outlier statistics defined in Section 5.4 are

$$\begin{aligned} q^K_{1-\alpha_I}/k &= q^1_{1-\alpha_5}/1 = 5.33, \\ q^{n_i}_{1-\alpha_I}/n_i &= q^5_{1-\alpha_5}/5 = 2.66, \\ q^1_{1-\alpha_n}/1 &= q^1_{1-\alpha_{25}}/1 = 8.19. \end{aligned}$$

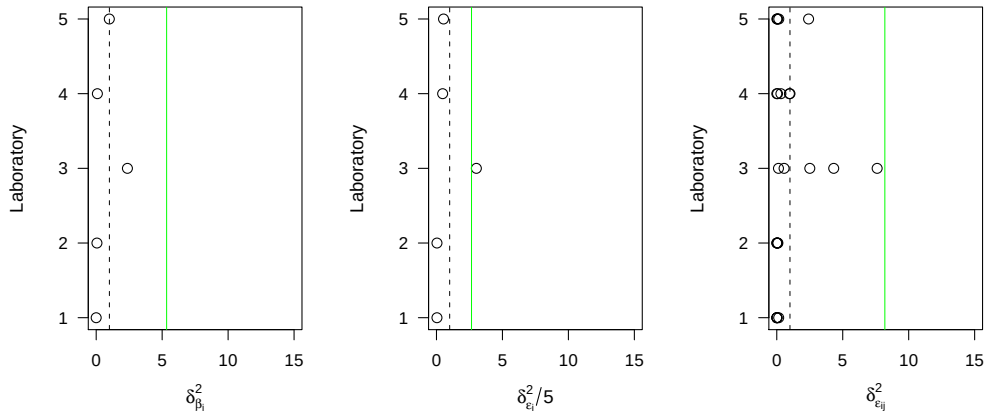
Figure 5.1: **Radon measurements from [WG03], obtained in 5 laboratories and 5 measurements per laboratory.**



Applying the outlier identification rules from Section 5.4, we do not identify laboratory 3 as location-outlier, neither the two lowest measurements of laboratory 3 and 5 as location-outliers within these laboratories, due to the much higher variance estimates. See Figure 5.2 for the calculated outlier statistics $\delta^2_{\beta_i}$, $\delta^2_{\varepsilon_i}/5$, $\delta^2_{\varepsilon_{ij}}$ and the respective limits indicated as green line. The outlier statistic of the residual vector of laboratory 3, $\delta^2_{\varepsilon_3}$, is outside the respective limit. Regarding the individual residuals of laboratory 3, it seems that this is mostly due to the higher variation, not due to an extreme single measurement within this laboratory.

The second example for the application of the one-way random effects model to data from an individual sample measured in different laboratories, comes from the IFCC network for standardization of HbA1c. As already mentioned in Section 1.5, in each standardization study samples are distributed among different laboratories. These laboratories measure each sample in two so-called digests (sample preparation steps) and make two repetitions per digest. However, in our analysis the digest effect is not taken into account, as the number of factor-specific observations would become too low. Further, various analyses ([KBA+06], [KAS+06]) have also shown that the digest effect is mostly negligible, compared to the variation between and within the laboratories. Therefore, we pool the data from each laboratory and treat them as four repeated measurements. Within the graphics the two different digests are indicated by different colors.
As an example we regard a calibrator sample (CAL) as well as an intercomparison sample (ICS), both measured within the Orlando 2 study. In Figure 5.3, the scatterplots of both samples are given. For the CAL sample, laboratory 16 measures substantially higher than the other laboratories, and the highest observation of laboratory 16 is quite different from the three other measurements within this laboratory.
For the ICS sample, laboratory 10 seems to measure lower than the other laboratories. Regarding the variation of the measurements within the different laboratories, we observe a quite higher one in laboratory 13, as well as in laboratory 16.
We fit a one-way random effects model to the data from these two samples, to inspect whether these exploratory observations can be confirmed.
For the CAL sample, the estimated between-laboratory variance is $\hat{\sigma}^2_a = 0.0086$ and the within-laboratory variance $\hat{\sigma}^2_\varepsilon = 0.0022$, whereas for the ICS sample, the estimated variances are $\hat{\sigma}^2_a = 0.0024$, and $\hat{\sigma}^2_\varepsilon = 0.0064$. Note that for the ICS sample the within-laboratories variance is much higher than for the CAL sample, although both samples are from the same concentration range. This is an indication that this estimator is influenced by the much higher variation in laboratory 13 and 16.
Regarding the outlier identification statistics (with $\alpha = 0.05$), given in Figure 5.4, with

Figure 5.2: **Outlier identification statistics for radon measurements from [WG03] based on the normal-linear mixed model. The outlier limits ($\alpha = 0.1$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



limits printed in green, we note that for the CAL sample laboratory 16 has the highest distance of the laboratory effect, however it is still within the limits. The variation of laboratory 16 is also not significantly higher than for the other laboratories.

For the ICS sample, laboratory 16 is identified as a scale-outlier due to its highest measurement, which is also extreme compared to the other measurements of laboratory 16. Laboratory 13 is not conspicuous in regard to its variation, neither laboratory 10 for its location.

5.5.2 Random coefficients model

The random coefficients model is a useful statistical method to analyze the measurement behavior of laboratories within a network. For the IFCC network for standardization of HbA1c this is an important question in each study. Candidate laboratories need to be approved and laboratories being already member need to be re-approved. We choose data from two studies as examples, the Kyoto 1 study and the Orlando 2 study.
In the Kyoto 1 study 13 laboratories participated. Each laboratory measured 5 intercomparison samples in two digests and two repetitions per digest. Similar to the one-way random effects model, we pool data from a single laboratory and sample, as the digest effect is negligible. In Figure 5.5, the plot of the differences between the individual

Figure 5.3: **Measurements of [%] HbA1c from two samples measured within the IFCC network for standardization of HbA1c. The different colors indicate the two different digests in each laboratory.**

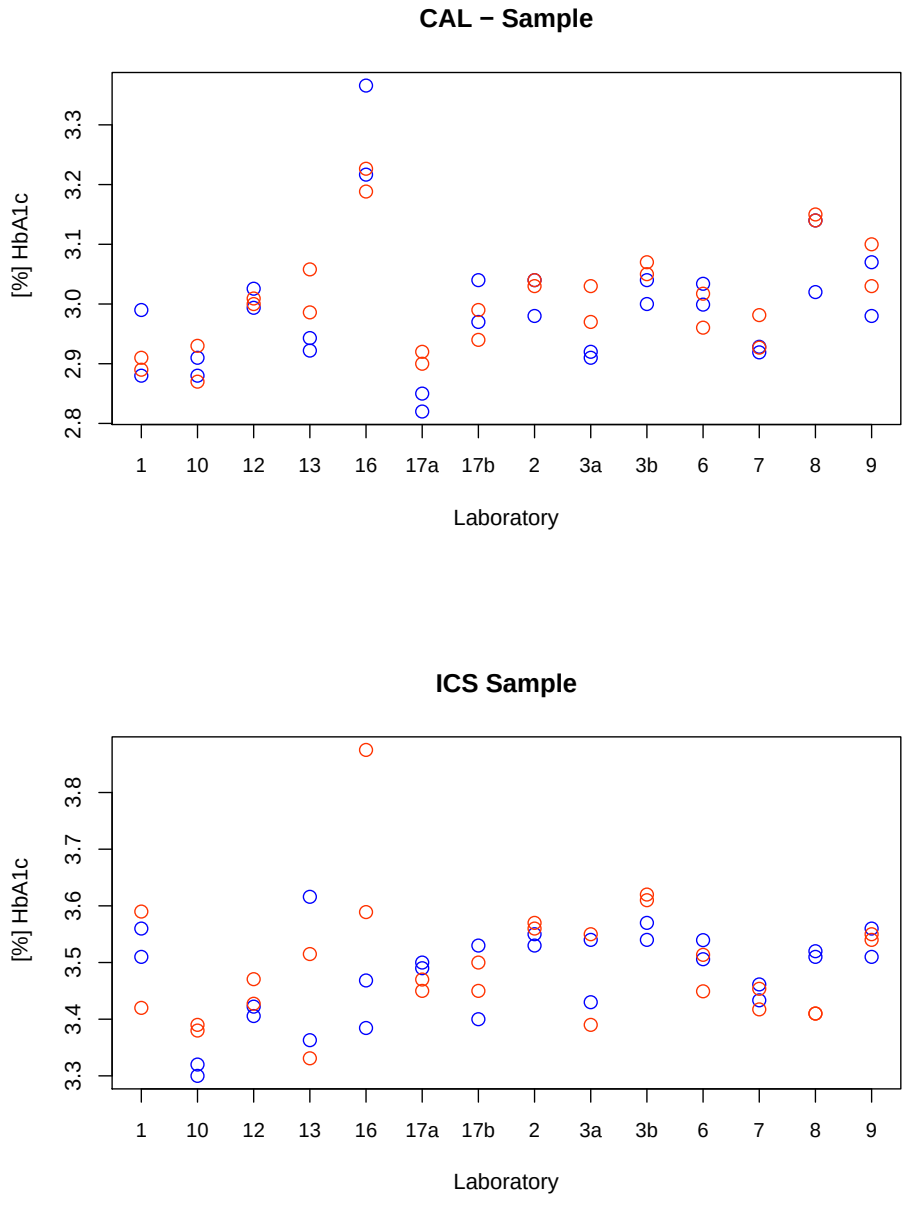
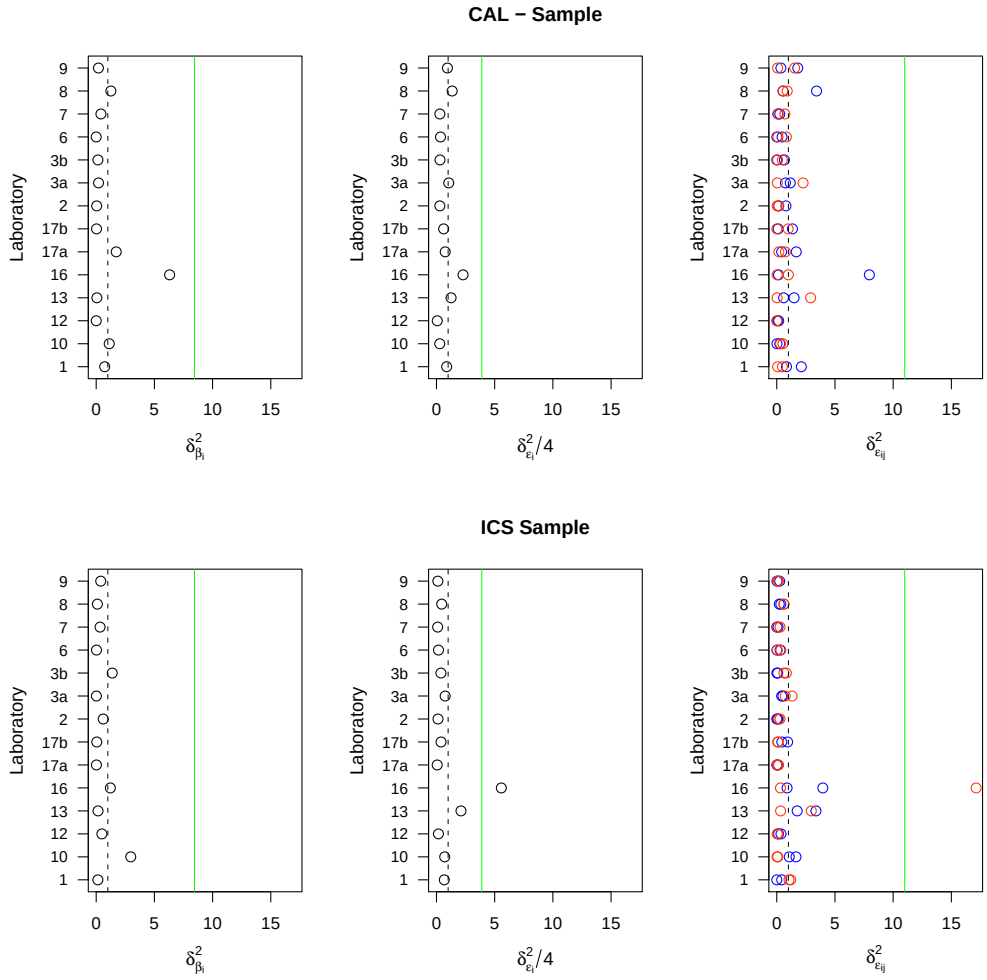


Figure 5.4: **Outlier identification statistics for the CAL and ICS sample of the IFCC network for standardization of HbA1c, based on the normal linear mixed model. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



measurements and the overall median against the overall median are given for each laboratory. Besides the data points, the least-square regression line per laboratory is displayed.

Examining the scatterplots in Figure 5.5, there is some variation between the relationship of the differences and the overall medians among the laboratories. Especially laboratory 15_c, has a very high slope and measures quite differently from the other laboratories within the study. The assumed linear relationship fits the data in each laboratory quite well, except for laboratory 12. Here is the variation of the residuals quite higher than in the other laboratories.

In Figure 5.6, the estimated coefficients and residuals are plotted for each laboratory. As already noted in Figure 5.5, the coefficients of laboratory 15_c are far away from the point (0,0), but also the coefficients of laboratory 10. The residuals of laboratory 12 show a much higher variation than for the other laboratories. But are these observations also significant in terms of the defined outlier identification rules?

In Figure 5.7, the calculated outlier identification statistics are given, together with the limits based on the adjusted 0.95 quantiles of the respective Chi-Square distributions. Examining these, neither laboratory 15_c nor laboratory 10 are identified as location-outliers within the laboratories. Laboratory 12 is identified as scale-outlier, as the variation of the residuals is higher than in the other laboratories. The estimates of the variance-covariance matrix of the coefficients, as well as the estimates of the variance of the residuals are highly influenced by the data of laboratories 10, 12 and 15_c. Hence, the resulting estimated values are too high and lead to masking effects. We will see in Section 6.4 that applying a robust estimation procedure to this data leads to more reliable results.

In the Orlando 2 study, 14 laboratories participated, each laboratory measured 5 intercomparison samples in two digests and two repetitions per digest. In Figure 5.8, the plots of the differences between individual measurements and the overall median against the overall median for each laboratory are given. Differences in the relationship between the laboratories are observable, although there seems to be not such a highly deviating laboratory as laboratory 15_c in the Kyoto 1 study. The variation of the residuals seems to be highest in laboratory 3a.

Figure 5.9 shows the estimated coefficients and residuals for each laboratory and Figure 5.10 the outlier statistics, based on the estimators of the normal linear mixed model. The variation of the residuals is higher in laboratory 3a than in the other laboratories, and laboratory 16 shows an extreme residual. Based on the outlier identification statistics, no laboratory is detected as location-outlier. Laboratory 3a is a scale-outlier due to the higher variation of its residuals.

Figure 5.5: **Plot of the differences between individual measurements and overall median against the overall median for each laboratory from the Kyoto 1 study. The solid lines are linear least-square fits, the dashed line indicates the zero line.**

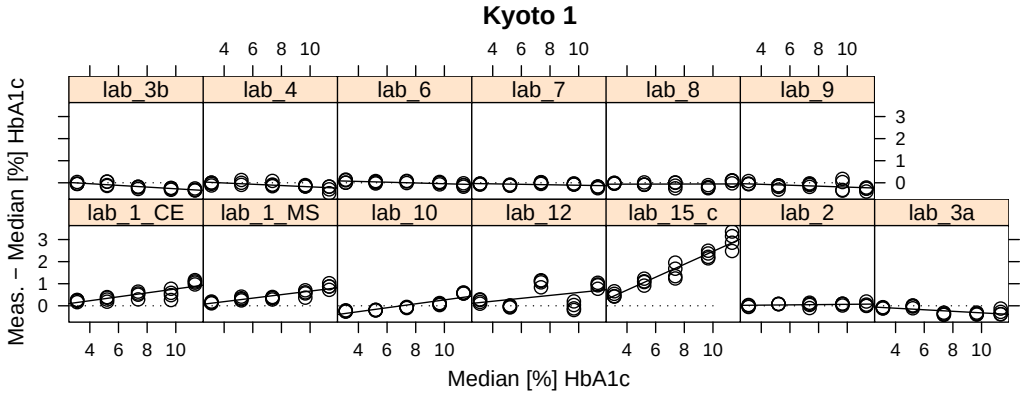


Figure 5.6: **Plot of the systematic and proportional effect of each laboratory for the Kyoto 1 study estimated by the normal random coefficients model. The plot on the right hand side shows the residuals for each laboratory; different colors indicate the different samples.**

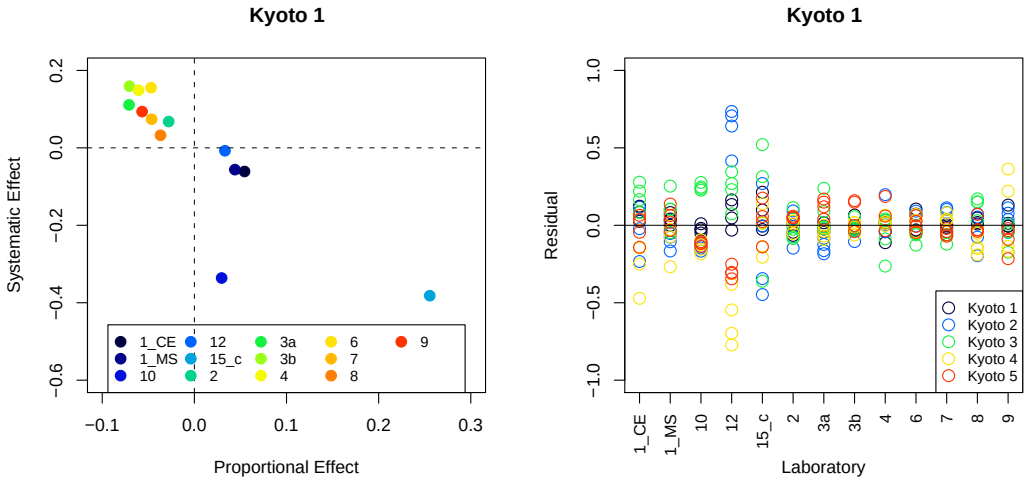


Figure 5.7: **Outlier identification statistics for the Kyoto 1 study based on the normal-linear mixed model. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**

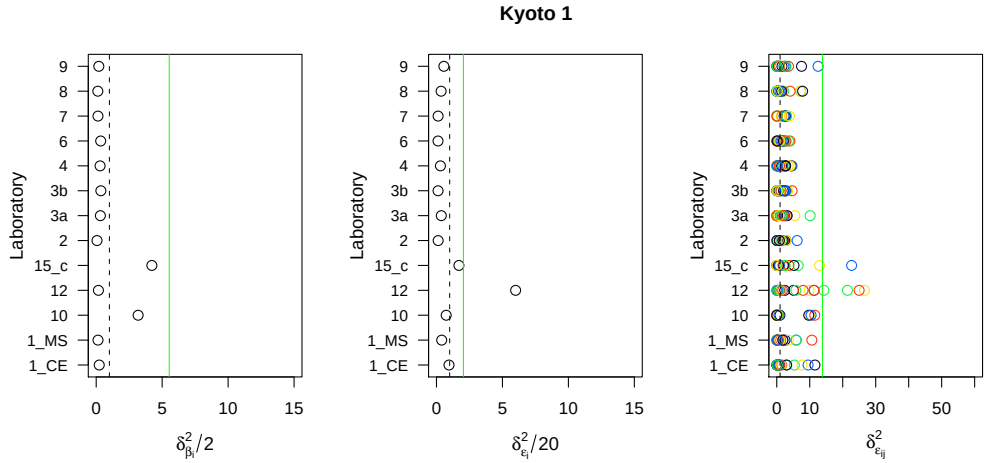


Figure 5.8: **Plot of the differences between individual measurements and overall median against the overall median for each laboratory from the Orlando 2 study. The solid lines are linear least-square fits, the dashed line indicates the zero line.**

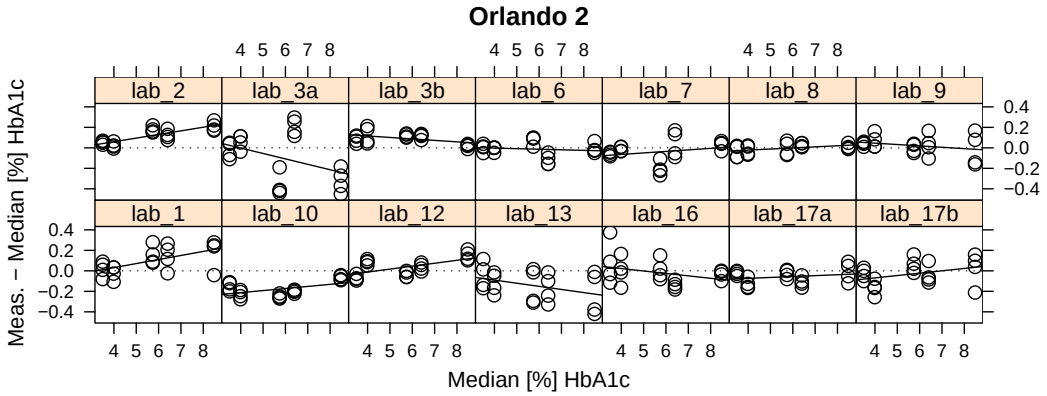


Figure 5.9: **Plot of the systematic and proportional effect of each laboratory for the Orlando 2 study estimated by the normal random coefficients model. The plot on the right hand side shows the residuals for each laboratory; different colors indicate the different samples.**

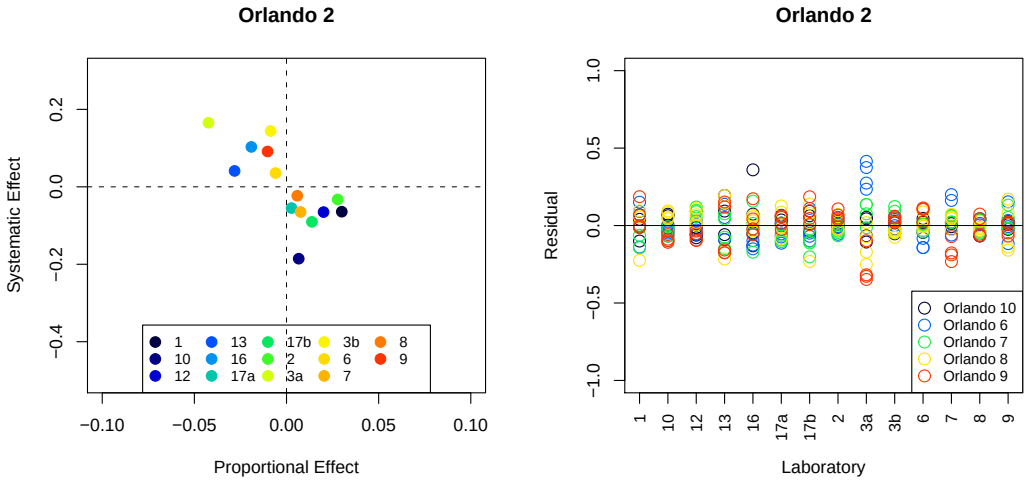
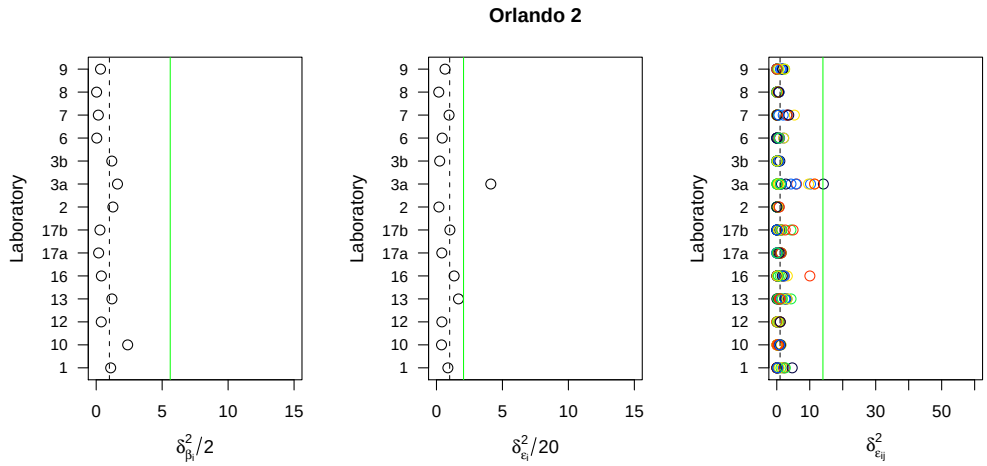


Figure 5.10: **Outlier identification statistics for the Orlando 2 study based on the normal linear mixed model. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



Chapter 6

Robust estimation in linear mixed models

The use of the multivariate normal distribution in the formulation of the linear mixed model, leads straightforward to maximum-likelihood estimation of the parameters (see Section 5.2). However, estimation procedures based on the normal distribution will be inefficient if outlier are present in the data (see for example [PLW01]). Outlier identification procedures based on these estimates are sensible to masking or swamping effects, as shown in Section 5.5. Robust estimation procedures avoid these problems as outlying data are downweighted during the estimation. Hence, estimates of the parameters of the distributions are not influenced in a disproportional way.
[PLW01] present an algorithm for the robust estimation of linear mixed models based on the multivariate t-distribution. That is, the multivariate normal distribution is replaced by the multivariate t-distribution, either with known or unknown degrees of freedom. This replacement results in downweighting extreme data, dependent on the degrees of freedom.

6.1 The t-linear mixed model

The normal-linear mixed model (5.1.1), may also be written in the following form:

$$\begin{pmatrix} \mathbf{Y}_i \\ \beta_i \end{pmatrix} \sim N_{n_i+k}\left(\begin{pmatrix} \mathbf{X}_i\mathbf{b} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{Z}_i\mathbf{D} & \mathbf{D} \end{pmatrix}\right), \; i = 1, .., I. \tag{6.1.1}$$

To introduce robustness into the estimation procedure, the multivariate normal distribution is replaced by the multivariate t-distribution with ν_i degrees of freedom, resulting

in

$$\begin{pmatrix} \mathbf{Y}_i \\ \beta_i \end{pmatrix} \sim t_{n_i+k}\left(\begin{pmatrix} \mathbf{X}_i\mathbf{b} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{Z}_i\mathbf{D} & \mathbf{D} \end{pmatrix}, \nu_i\right),\ i = 1, .., I. \tag{6.1.2}$$

The marginal distribution of the observations is then given by

$$\mathbf{Y}_i \sim t_{n_i}(\mathbf{X}_i\mathbf{b}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i, \nu_i),\ i = 1, .., I. \tag{6.1.3}$$

According to the definition of the multivariate t-distribution, Model 6.1.2 (see e.g. [GCSR04]) may be written in the form of a hierarchical model

$$\begin{aligned} \begin{pmatrix} \mathbf{Y}_i \\ \beta_i \end{pmatrix} | \tau_i &\sim N_{n_i+k}\left(\begin{pmatrix} \mathbf{X}_i\mathbf{b} \\ 0 \end{pmatrix}, \frac{1}{\tau_i}\begin{pmatrix} \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{Z}_i\mathbf{D} & \mathbf{D} \end{pmatrix}\right) \\ \tau_i &\sim Gamma\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right),\ i = 1, .., I, \end{aligned}$$

or by taking into account the hierarchical structure of the random effects

$$\begin{aligned} \mathbf{Y}_i | \beta_i, \tau_i &\sim N_{n_i}\left(\mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\beta_i, \frac{1}{\tau_i}\mathbf{R}_i\right) \\ \beta_i | \tau_i &\sim N\left(0, \frac{1}{\tau_i}\mathbf{D}\right) \\ \tau_i &\sim Gamma\left(\frac{\nu_i}{2}, \frac{\nu_i}{2}\right),\ i = 1, .., I. \end{aligned} \tag{6.1.4}$$

Model (6.1.4) provides the basis for the natural implementation of an expectation-maximization (EM) algorithm for the maximum-likelihood estimation of the parameters.
Another form of the t-linear mixed model is written as

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\beta_i + \varepsilon_i \\ \beta_i &\sim t_k(0, \mathbf{D}, \nu_i) \\ \varepsilon_i &\sim t_{n_i}(0, \mathbf{R}_i, \nu_i),\ i = 1, .., I. \end{aligned} \tag{6.1.5}$$

This form allows the identification of laboratories as location-outliers, as well as laboratories as scale-outliers based on the outlier statistics defined in Section 5.4.
The mean of the observations is given by $E(\mathbf{Y}_i) = \mathbf{X}_i\mathbf{b}$ and for $\nu_i > 2$

$$Var(\beta_i) = \frac{\nu_i}{\nu_i - 2}\mathbf{D},\ Var(\varepsilon_i) = \frac{\nu_i}{\nu_i - 2}\mathbf{R}_i.$$

This results in two differences between the normal-linear mixed model and the t-linear mixed model:
First, depending on the different degrees of freedom for the laboratories, the random effects are allowed to have different variations. However, to avoid the estimation of a large number of parameters in what follows we set $\nu_i = \nu \; \forall i$.
[PLW01] restrict the different degrees of freedom to groups of subjects which are known in advance. For our application we can not define groups of laboratories in advance, therefore we take the same degrees of freedom for all laboratories. Additionally we are restricted to a small number of laboratories, which makes the estimation of the degrees of freedom even more difficult.
The matrices $\mathbf{D}$ and $\mathbf{R}_i$ have different meanings in both models. In the normal-linear mixed model they are the variance-covariance matrices of the random effects and residuals, respectively. In the t-linear mixed model they must be multiplied by the factor $\frac{\nu}{\nu-2}$ for obtaining the respective variance-covariance matrices.
To apply the EM algorithm for the estimation of the model parameters the conditional distributions of $\beta_i|\mathbf{Y}_i$, as well as $\tau_i|\mathbf{Y}_i$ must be known. Some calculus leads to

$$\begin{aligned} \beta_i|\mathbf{Y}_i \quad \sim \quad & t_k \Big(\mathbf{D}\mathbf{Z}_i'(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)^{-1}(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}), \\ & \mathbf{D} - \mathbf{D}\mathbf{Z}_i'(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)^{-1}\mathbf{Z}_i\mathbf{D}, \nu\Big) \end{aligned} \tag{6.1.6}$$

and

$$\tau_i|\mathbf{Y}_i \sim Gamma\left(\frac{\nu + n_i}{2}, \frac{\nu + \delta_i^2(\mathbf{b}, \mathbf{D}, \mathbf{R}_i)}{2}\right), \tag{6.1.7}$$

where $\delta_i^2(\mathbf{b}, \mathbf{D}, \mathbf{R}_i)$ refers to the residual sum of squares, given by

$$\delta_i^2(\mathbf{b}, \mathbf{D}, \mathbf{R}_i) = (\mathbf{Y}_i - \mathbf{X}_i\mathbf{b})'(\mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i' + \mathbf{R}_i)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}).$$

The split of the residual sum of squares into a part considering only the random effects and another considering the residuals, as shown in (5.4.1), holds, too. Therefore we can easily calculate the outlier identification statistics defined in Section 5.4, based on the estimators from the t-linear mixed model algorithm.

Note that the conditional mean of $\tau_i|\mathbf{Y}_i$ is given by

$$E(\tau_i|\mathbf{Y}_i) = \frac{\nu + n_i}{\nu + \delta_i^2(\mathbf{b}, \mathbf{D}, \mathbf{R}_i)},$$

i.e. it decreases with δ_i^2. In comparison to the normal-linear mixed model, the t-linear mixed model allows each laboratory to have its own scale τ_i, which is unobservable and needs to be imputed from the data. The different individual scales result in different weights for the estimation of the model parameters. As $E(\tau_i|\mathbf{Y_i})$ decreases with δ_i^2, laboratories with larger residual sum of squares will have less weight in the determination of the parameter estimates. On the other side, the influence of the residual sum of squares on the scales τ_i is controlled by the degrees of freedom ν; the smaller ν, the larger the influence of δ_i^2 on τ_i. Hence, setting the degrees of freedom in advance is equal to the definition of a robustness parameter.

6.2 The expectation-maximization algorithm

The expectation-maximization (EM) algorithm ([DLR77]) is an iterative algorithm for models with incomplete data. The observed data vector $\mathbf{y}$ is viewed as being incomplete and is regarded as an observable function of the so-called complete data.
In that sense in the t-linear mixed model (6.1.4), both β_i and τ_i are treated as missing, though of course they are never observable in a data sense.
Denote with $\mathbf{y}_c = (\mathbf{y}', \mathbf{y}_m')'$ the complete data vector, with $g_c(\mathbf{y}_c, \Theta)$ the probability density function of the random vector $\mathbf{Y}_c$, corresponding to the complete-data vector $\mathbf{y}_c$, and with Θ the parameter vector defined on the parameter space Ω.
The complete-data log-likelihood function that could be formed if $\mathbf{y}_c$ were fully observable, is given by

$$\ln L_c(\Theta) = \ln g_c(\mathbf{y}_c, \Theta).$$

But as this log-likelihood is not observable, it is replaced by its conditional expectation given $\mathbf{y}$, using the current fit of Θ.
Let $\Theta^{(0)}$ be some initial value of Θ. In the first iteration, the expectation step (E-step) requires the calculation of

$$Q(\Theta; \Theta^{(0)}) = E_{\Theta^{(0)}} (\ln L_c(\Theta)|\mathbf{y}) .$$

In the maximization step (M-step) $Q(\Theta; \Theta^{(0)})$ is maximized with respect to Θ over the parameter space Ω. In the (k+1)th iteration, the E-step and M-step are defined as:

E-step: Calculate $Q(\Theta; \Theta^{(k)}) = E_{\Theta^{(k)}}(\ln L_c(\Theta))$

M-step: Choose $\Theta^{(k+1)} \in \Omega$ such that

$$Q(\Theta^{(k+1)}; \Theta^{(k)}) \geq Q(\Theta; \Theta^{(k)}) \; \forall \Theta \in \Omega.$$

[DLR77] showed, that after every EM iteration, the log-likelihood of the incomplete data increases, i.e.

$$L(\Theta^{(k+1)}) \geq L(\Theta^{(k)}).$$

Thus, convergence is obtained for a sequence of likelihood values that are bounded above.
The M-step in the EM algorithm is difficult to implement, it will often be useful to replace it with a sequence of constraint maximization steps (CM-steps). In each CM-step $Q(\Theta; \Theta^{(k)})$ is maximized over Θ, in order to update some of the elements of Θ while the other elements of Θ are fixed. This is known as the ECM algorithm ([MR93]). In [Bil98] a good introduction is given to the application of the EM algorithm to mixed models.

6.3 ECM-algorithms for t-linear mixed model

We describe two algorithms for the fitting of the t-linear mixed model, both presented in [PLW01]. The first one regards the β_i as well as τ_i as missing values and results in closed form estimators. It is presented to clarify the scheme of the ECM-algorithm. The second one integrates the β_i out, resulting in a computationally more intensive CM-step. However, this step can easily be solved with standard statistic software, hence we decided to use this algorithm to fit the t-linear mixed models to our data.

6.3.1 Algorithm with β_i and τ_i missing

Formulation (6.1.4) of the t-linear mixed model, leads to some straightforward application of the ECM algorithm. The coefficients β_i and scales τ_i are viewed as missing data, although they are never observable in a data sense.
Let $\mathbf{y} = (\mathbf{y}_1', .., \mathbf{y}_I')'$, $\beta = (\beta_1', .., \beta_I')'$, and $\tau = (\tau_1, .., \tau_I)'$. Furthermore we assume that the variance structure of the errors is known up to a constant and equal over all laboratories, i.e. $Var(\varepsilon) = \sigma^2 \mathbf{R}_i$, where $\mathbf{R}_i$ are known matrices. In our applications they will be the identity matrix. All parameters of the model are collected within the parameter vector $\Theta = (\mathbf{b}, \mathbf{D}, \sigma^2, \nu)$.

The log-likelihood for the complete data in the t-linear mixed model with unknown degrees of freedom is

$$\begin{aligned} L(\mathbf{b}, \mathbf{D}, \sigma^2, \nu|\mathbf{y}, \beta, \tau) &= \ln\left(\prod_{i=1}^{I} P_\Theta(\mathbf{Y}_i|\beta_i, \tau_i) \cdot P_\Theta(\beta_i|\tau_i) \cdot P_\Theta(\tau_i)\right) = \\ &= L_1(\mathbf{b}, \sigma^2|\mathbf{y}, \beta, \tau) + L_2(\mathbf{D}|\beta, \tau) + L_3(\nu|\tau) + const., \end{aligned}$$

with

$$\begin{aligned} L_1(\mathbf{b}, \sigma^2|\mathbf{y}, \beta, \tau) &= -\sum_{i=1}^{I} \frac{n_i}{2} \ln(\sigma^2) - \sum_{i=1}^{I} \frac{\tau_i}{2\sigma^2} trace(\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\beta_i)(\mathbf{y}_i - \mathbf{Z}_i\beta_i)') \\ &+ \sum_{i=1}^{I} \frac{\tau_i}{\sigma^2} \mathbf{b}'\mathbf{X}_i'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\beta_i) - \sum_{i=1}^{I} \frac{\tau_i}{2\sigma^2} \mathbf{b}'\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\mathbf{b} \\ L_2(\mathbf{D}|\beta, \tau) &= -\frac{I}{2} \ln(|\mathbf{D}|) - \frac{1}{2} trace\left(\mathbf{D}^{-1} \sum_{i=1}^{I} \tau_i \beta_i \beta_i'\right) \\ L_3(\nu|\tau) &= -\sum_{i=1}^{I} \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \ln(\tau_i) + \frac{\nu}{2}\left(\ln\left(\frac{\nu}{2}\right) + \ln(\tau_i) - \tau_i\right). \end{aligned}$$

The E-step requires the calculation of the conditional mean of the log-likelihood function, using the current fit of Θ denoted as $\hat{\Theta}$. Based on the three terms of the log-likelihood function and the conditional distributions of $\tau_i|\mathbf{y}_i$ and $\beta_i|\mathbf{y_i}$, we have

$$\begin{aligned} E(L_1(\mathbf{b}, \sigma^2|\mathbf{y}, \beta, \tau)|\mathbf{y}, \Theta = \hat{\Theta}) &= -\sum_{i=1}^{I} \frac{n_i}{2} \ln(\sigma^2) - \sum_{i=1}^{I} \frac{\hat{\tau}_i}{2\sigma^2} \mathbf{b}'\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\mathbf{b} \\ &+ \sum_{i=1}^{I} \frac{\hat{\tau}_i}{\sigma^2} \mathbf{b}'\mathbf{X}_i'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i) \\ &- \sum_{i=1}^{I} \frac{1}{2\sigma^2} trace\left(\mathbf{R}_i^{-1}\left(\hat{\tau}_i(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i)(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i)' + \mathbf{Z}_i\hat{\Omega}_i\mathbf{Z}_i'\right)\right) \\ E(L_2(\mathbf{D}|\beta, \tau)|\mathbf{y}, \Theta = \hat{\Theta}) &= -\frac{I}{2} \ln(|\mathbf{D}|) - \frac{1}{2} trace\left(\mathbf{D}^{-1} \sum_{i=1}^{I} \left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right)\right) \\ E(L_3(\nu|\tau)|\mathbf{y}, \Theta = \hat{\Theta}) &= \sum_{i=1}^{I} \left(\frac{\nu}{2}\left(\ln\left(\frac{\nu}{2}\right) + E(\ln\tau_i|\mathbf{y}, \hat{\Theta}) - \hat{\tau}_i\right)\right. \\ &\left. -E(\ln(\tau_i)|\mathbf{y}, \hat{\Theta}) - \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right), \end{aligned}$$

where $\hat{\tau}_i = E(\tau_i|\mathbf{y}, \Theta = \hat{\Theta})$, $\hat{\beta}_i = E(\beta_i|\mathbf{y}, \Theta = \hat{\Theta})$ and $\hat{\Omega}_i = \hat{\tau}_i \cdot Cov(\beta_i|\mathbf{y}, \Theta = \hat{\Theta})$.

Note that for the derivation of the conditional mean the following property of the mean is used: For two random variables A, B we have $E(A \cdot B) = E(A) \cdot E(B) + Cov(A, B)$ from which it follows that $E(AB^2) = E(A) \cdot E(B)^2 + E(A) \cdot Var(B) + Cov(A, B^2)$.

The conditional expectations and variance are derived from the conditional distributions of $\tau_i|\mathbf{y}$ and $\beta_i|\mathbf{y}$ (see (6.1.7) and (6.1.6)), resulting in

$$\hat{\tau}_i = E(\tau_i|\mathbf{y}, \Theta = \hat{\Theta}) = \frac{\hat{\nu} + n_i}{\hat{\nu} + \delta_i^2(\hat{\mathbf{b}}, \hat{\mathbf{D}}, \hat{\sigma}^2)} \tag{6.3.1}$$

$$\hat{\beta}_i = E(\beta_i|\mathbf{y}, \Theta = \hat{\Theta}) = \hat{\mathbf{D}}\mathbf{Z}_i'(\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i' + \hat{\sigma}^2\mathbf{R}_i)^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}) \tag{6.3.2}$$

$$\hat{\Omega}_i = \hat{\tau}_i \cdot Cov(\beta_i|\mathbf{y}, \Theta = \hat{\Theta}) = \hat{\mathbf{D}} - \hat{\mathbf{D}}\mathbf{Z}_i'(\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i' + \hat{\sigma}^2\mathbf{R}_i)^{-1}\mathbf{Z}_i\hat{\mathbf{D}}. \tag{6.3.3}$$

Hence the steps of the EM-algorithm for fitting the t-linear mixed model are the following:

E-step: Given $\Theta = \hat{\Theta}$, compute $\hat{\tau}_i$, $\hat{\beta}_i$ and $\hat{\Omega}_i$, for $i = 1, .., I$ using (6.3.1), (6.3.2), (6.3.3).

M-step: This step is divided in 4 constrained maximization steps.

CM-step 1: For updating $\hat{\mathbf{b}}$, we fix $\sigma^2 = \hat{\sigma}^2$ and maximize $E(L_1(\mathbf{b}, \hat{\sigma}^2|\mathbf{y}, \beta, \tau)|\mathbf{y}, \Theta = \hat{\Theta})$ over $\mathbf{b}$, leading to

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{I} \frac{\hat{\tau}_i}{\hat{\sigma}^2}\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\right)^{-1} \sum_{i=1}^{I} \frac{\hat{\tau}_i}{\hat{\sigma}^2}\mathbf{X}_i'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i),$$

as

$$\frac{\partial E(L_1)}{\partial \mathbf{b}} = \sum_{i=1}^{I} \frac{\hat{\tau}_i}{\hat{\sigma}^2}\mathbf{X}_i'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i) - \sum_{i=1}^{I} \frac{\hat{\tau}_i}{\hat{\sigma}^2}\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\mathbf{b}.$$

CM-step 2: Fix $\mathbf{b} = \hat{\mathbf{b}}$ and update σ^2, by maximizing $E(L_1(\hat{\mathbf{b}}, \sigma^2|\mathbf{y}, \beta, \tau)|\mathbf{y}, \Theta = \hat{\Theta})$

over σ^2, which results in

$$\begin{aligned}\hat{\sigma}^2 &= \left(\sum_{i=1}^{I} trace\left(\mathbf{R}_i^{-1}(\hat{\tau}_i(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i)(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i)' + \mathbf{Z}_i\hat{\Omega}_i\mathbf{Z}_i')\right)\right) / \sum_{i=1}^{I} n_i \\ &- \left(\sum_{i=1}^{I} 2\hat{\tau}_i\hat{\mathbf{b}}'\mathbf{X}_i'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{Z}_i\hat{\beta}_i) - \sum_{i=1}^{I}\hat{\tau}_i\hat{\mathbf{b}}'\mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\hat{\mathbf{b}}\right) / \sum_{i=1}^{I} n_i \\ &= \sum_{i=1}^{I}\left(\hat{\tau}_i(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}} - \mathbf{Z}_i\hat{\beta}_i)'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}} - \mathbf{Z}_i\hat{\beta}_i) + trace(\hat{\Omega}_i\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i)\right) / \sum_{i=1}^{I} n_i.\end{aligned}$$

The last equation holds as $trace(\mathbf{Abb}') = \mathbf{b}'\mathbf{Ab}$ and by combining the terms to a quadratic form.

CM-step 3: The closed form of the updated $\hat{\mathbf{D}}$, maximizing $E(L_2(\mathbf{D}|\beta,\tau)|\mathbf{y},\Theta = \hat{\Theta})$ over $\mathbf{D}$, is obtained by taking the partial derivative of $E(L_2(\mathbf{D}|\beta,\tau)|\mathbf{y},\Theta = \hat{\Theta})$ with respect to $\mathbf{D}^{-1}$. Based on the identities (A.0.6) and (A.0.7) given in Appendix A, we obtain

$$\begin{aligned}\frac{\partial E(L_2(\mathbf{D}|\beta,\tau)|\mathbf{y},\Theta\hat{\Theta})}{\partial\mathbf{D}^{-1}} &= \frac{I}{2}(2\mathbf{D} - diag(\mathbf{D})) \\ &- \frac{1}{2}\left(2\sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right) - diag\left(\sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right)\right)\right) \\ &= \left(I\mathbf{D} - \sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right)\right) \\ &- \frac{1}{2}diag\left(I\mathbf{D} - \sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right)\right).\end{aligned}$$

Setting this partial derivative to zero, implies that

$$I\mathbf{D} - \sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right) = 0,$$

from which it follows that

$$\hat{\mathbf{D}} = \frac{1}{I}\sum_{i=1}^{I}\left(\hat{\tau}_i\hat{\beta}_i\hat{\beta}_i' + \hat{\Omega}_i\right).$$

CM-step 4: Updating $\hat{\nu}$ by maximizing $E(L_3(\nu|\tau)|\mathbf{y},\Theta = \hat{\Theta})$ over ν would result in a

one-dimensional search of

$$\hat{\nu} = arg\max_{\nu} \sum_{i=1}^{I} \left(\frac{\nu}{2} \left(\ln\left(\frac{\nu}{2}\right) + E\left(\ln(\hat{\tau}_i)|\mathbf{y}, \Theta = \hat{\Theta}\right) - \hat{\tau}_i \right) - \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right).$$

However, $E\left(\ln(\hat{\tau}_i)|\mathbf{y}, \Theta = \hat{\Theta}\right)$ does not have a closed form and therefore convergence might be very slow. To circumvent this problem, this step is transformed in a constraint maximum-likelihood step, i.e. ν is found as the value that maximizes the constrained likelihood over the degrees of freedom, with $\mathbf{b}, \mathbf{D}, \sigma^2$ fixed at their current estimates. The constrained likelihood is computed using the marginal model of the observations, that is

$$\mathbf{Y}_i \sim t_{n_i}\left(\mathbf{X}_i\mathbf{b}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{R}_i, \nu\right).$$

This results in the log-likelihood function

$$L(\nu; \mathbf{y}) = \sum_{i=1}^{I} \left(\ln\left(\Gamma\left(\frac{\nu + n_i}{2}\right)\right) - \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \frac{\nu}{2}\ln(\nu) - \frac{\nu + n_i}{2}\ln\left(\nu + \delta_i^2(\hat{\mathbf{b}}, \hat{\mathbf{D}}, \hat{\sigma}^2)\right)\right).$$

The updated $\hat{\nu}$ is the value, which maximizes this log-likelihood function over the parameter space of ν. This requires only a one-dimensional search, too, but all terms are written in closed form.

When the degrees of freedom are known, or set in advance, CM-step 4 of the EM-algorithm is omitted and the known ν is used in place of $\hat{\nu}$ in the remaining steps.

6.3.2 Algorithm with τ_i missing

In this algorithm the β_i are integrated out of the complete data likelihood, hence only the τ_i are treated as missing data. This results in a more complex CM-step, however this can be solved using standard statistical software, which facilitates its implementation.
We present the derivation of the algorithm for the case that the degrees of freedom of the t-distribution are set in advance. If they are also estimated from the data, CM-step 4 of the above algorithm is simply added.
The log-likelihood function of the complete data $(\mathbf{y}, \tau)'$ is based on the marginal distri-

bution of the observations, given in (6.1.3):

$$\begin{aligned} L(\mathbf{b}, \mathbf{D}, \sigma^2|\mathbf{y}, \tau) &= L_1(\mathbf{b}, \mathbf{D}, \sigma^2|\mathbf{y}, \tau) + const, \\ L_1(\mathbf{b}, \mathbf{D}, \sigma^2|\mathbf{y}, \tau) &= -\frac{1}{2}\sum_{i=1}^{I} \ln|\mathbf{V}_i| + \tau_i(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}), \end{aligned}$$

where $\mathbf{V_i} = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \sigma^2\mathbf{R}_i$.
It follows that

$$E(L_1(\mathbf{b}, \mathbf{D}, \sigma^2|\mathbf{y}, \tau)|\mathbf{y}, \Theta = \hat{\Theta}) = -\frac{1}{2}\sum_{i=1}^{I} \ln|\mathbf{V}_i| + \hat{\tau}_i(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}),$$

where $\hat{\tau}_i = E(\tau_i|\mathbf{y}, \Theta = \hat{\Theta})$ is defined in (6.3.1).
Hence, the algorithm is given by the following two steps:

E-step: Given $\Theta = \hat{\Theta}$, compute $\hat{\tau}_i$, defined in (6.3.1).

C-step: For fixed $\hat{\tau}$, update $E(L_1(\mathbf{b}, \mathbf{D}, \sigma^2|\mathbf{y}, \tau)|\mathbf{y}, \Theta = \hat{\Theta})$, which is equivalent to maximum-likelihood estimation in the normal-linear mixed model, $\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_i\mathbf{b} + \mathbf{Z}_i\beta_i + \varepsilon_i,\ i = 1, .., I$, where $\tilde{\mathbf{y}}_i = \sqrt{\tau_i}\mathbf{y}_i$ and $\tilde{\mathbf{X}}_i = \sqrt{\tau_i}\mathbf{X}_i$ (see (5.2.1) for the log-likelihood function of the normal linear mixed model). This can easily be done e.g. in the SAS software [SAS06] with PROC MIXED or in the R software [R D06] with the `lme` function of the **nlme** package [PBD+06].

Estimators of the random effects are obtained by BLUP according to Formula (5.3.1)

$$\hat{\beta}_i = \hat{\mathbf{D}}\mathbf{Z}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{b}}).$$

6.3.3 Derivation of the starting values

As the EM algorithm is an iterative algorithm, we need to compute starting values for the parameters. By fitting for each laboratory separate regression models and using method of moments estimators, we derive the initial values of the fixed effects and of the dispersion parameters. More concretely, let $\check{\beta}_i$ be the estimates of the coefficients and $\check{\sigma}_i^2$ of the residual variance, both derived by least-square regression for each laboratory individually. Initial values for the fixed effects and dispersion parameters are obtained

as

$$\begin{aligned}
\mathbf{b}^{(0)} &= \frac{1}{I}\sum_{i=1}^{I}\check{\beta}_i, \\
\mathbf{D}^{(0)} &= \frac{1}{I-1}\sum_{i=1}^{I}(\mathbf{b}^{(0)}-\check{\beta}_i)(\mathbf{b}^{(0)}-\check{\beta}_i)', \\
\sigma^{2(0)} &= \frac{1}{I}\sum_{i=1}^{I}\check{\sigma}_i^2.
\end{aligned}$$

In the case that the degrees of freedom also need to be estimated, we set $\nu^{(0)} = 20$, according to the recommendations of [PB00].

In the next section we apply the ECM algorithm of Section 6.3.2 to the data and models, which were already presented in Section 5.5.

6.4 Outlier identification by t-linear mixed models

In Section 5.5, we presented examples involving data from standardization networks, which were modelled as normal-linear mixed models. Especially we discussed the one-way random effects and random coefficients model. We inspected that the parameter estimators were very sensitive to the presence of outliers, such that the outlier identification statistics of Section 5.4 were subject to masking effects. In this section we regard the same data, but this time we model them by t-linear mixed models to achieve a more robust estimation of the parameters.

6.4.1 One-way random effects model

Formulating the one-way random effects model (5.1.3) in terms of a t-linear mixed model according to (6.1.5), we have

$$\begin{aligned}
\mathbf{Y}_i &= \mu\cdot\mathbf{1}_{n_i} + a_i\cdot\mathbf{1}_{n_i} + \varepsilon_i,\ i = 1,..,I, \\
a_i &\sim t_1(0,\sigma_a^2,\nu) \\
\varepsilon_i &\sim t_{n_i}(0,\sigma_\varepsilon^2\cdot\mathbf{I}_{n_i},\nu).
\end{aligned}$$

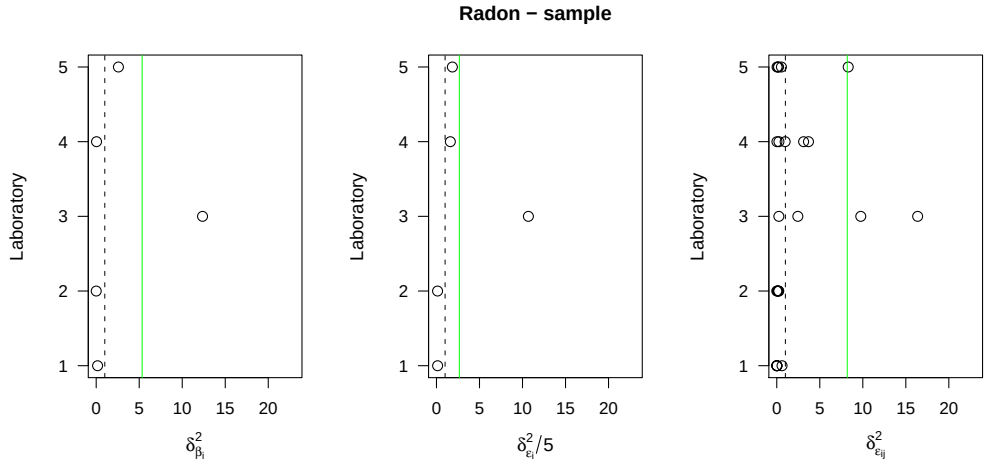
The parameters of this model are the fixed effect μ, the between-laboratories variance σ_a^2, the within-laboratories variance σ_ε^2 and the degrees of freedom ν. The degrees of freedom can either be estimated from the data or set in advance. In the second case they are robustness tuning parameters; the smaller ν, less influence is given to extreme values (see Section 6.1).
For the example of radon measurements from [WG03], we obtain the following estimators of the parameters, in the case that the degrees of freedom are estimated from the data, too: $\hat{\mu} = 163$, $\hat{\sigma}_a^2 = 115.6$, $\hat{\sigma}_\varepsilon^2 = 116.4$ and $\hat{\nu} = 1.26$.
The estimator of the fixed effect has the same interpretation as in the normal-linear mixed model, so that these estimators are directly comparable. Based on the normal-linear mixed model the estimator is 169, whereas the median-based estimator of [WG03] results in 161. Hence, the estimator of the t-linear mixed model is less influenced by the high results of laboratory 3 than the estimator of the normal-linear mixed model.
As already mentioned in Section 6.1, the estimators $\hat{\sigma}_a^2, \hat{\sigma}_\varepsilon^2$ are not directly comparable, as in the t-linear mixed model the variance of the random effects and the residuals are

$$Var(a_i) = \nu/(\nu-2) \cdot \sigma_a^2, Var(\varepsilon_{ij}) = \nu/(\nu-2) \cdot \sigma_\varepsilon^2$$

and exist only for $\nu > 2$. In our case, the estimate of ν is less than 2, because there are only 5 laboratories with quite different behaviors.
We refitted the data, this time setting $\nu = 4$, to avoid the problem of $\nu < 2$. Now the estimate $\hat{\mu} = 162$, which is very close to the median-based estimate. The variances are estimated as $\hat{\sigma}_a^2 = 227.6$ and $\hat{\sigma}_\varepsilon^2 = 198.9$. Comparing these with the estimates of the normal-linear mixed model, we have $\hat{Var}(a_i) = 2 \cdot \hat{\sigma}_a^2 = 455.2$, compared to 986.4 and $\hat{Var}(\varepsilon_{ij}) = 2 \cdot \hat{\sigma}_\varepsilon^2 = 387.8$, compared to 689.6 from the normal-linear mixed model. As expected, the estimation via the t-linear mixed model results in lower variances of the random effects and residuals.
Regarding the outlier identification statistics with $\alpha = 0.1$ (see Figure 6.1), laboratory 3 is now clearly identified as location- and scale-outlier, as well as the lowest and highest measurement of laboratory 3. The lowest measurement of laboratory 5 is also identified as location-outlier within this laboratory. Hence, based on these rules we found the same outlier pattern within the radon data as [WG03].

For the CAL and ICS sample of the IFCC network for standardization of HbA1c the estimates of the parameters of the t-linear mixed model are given in Table 6.1, fitted with unknown degrees of freedom as well as by setting $\nu = 4$. For the CAL sample the estimated degrees of freedom are $\hat{\nu} = 9.24$, therefore we observe some differences between the two estimation approaches. Fixing $\nu = 4$ leads to smaller estimators of σ_a^2

Figure 6.1: **Outlier identification statistics for radon measurements from [WG03] based on the t-linear mixed model with $\nu = 4$. The outlier limits ($\alpha = 0.1$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



and σ^2_ε, but to higher variances of the random effects and residuals. For the ICS sample the estimated degrees of freedom are $\nu = 3.03$, hence the results of both algorithms do not differ substantially.

For the CAL sample the estimate of the between-laboratory variance is much smaller than derived via the normal-linear mixed model approach. This shows that the influence of laboratory 16 on this estimate is smaller. The within-laboratory variances of the ICS and CAL sample are now comparable.

Regarding the outlier identification statistics, plotted in Figure 6.2 (based on the estimates of the model with fixed degrees of freedom $\nu = 4$) for the CAL sample, laboratory 16 is identified as location-outlier and the highest observation of laboratory 16 as location-outlier within this lab. Regarding the ICS sample, laboratory 16 and laboratory 13 are identified as scale-outliers. Laboratory 16 is a scale-outlier due to its highest measurement, which is also extreme in comparison to the other measurements of laboratory 16. Laboratory 13 shows a high variation among its measurements, but none of them is extreme compared to the others. Laboratory 10 has the largest distance of the random coefficients, but it is still within the allowable range.

Figure 6.2: **Outlier identification statistics for the CAL and ICS sample of the IFCC network for standardization of HbA1c, based on the t-linear mixed model with $\nu = 4$. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**

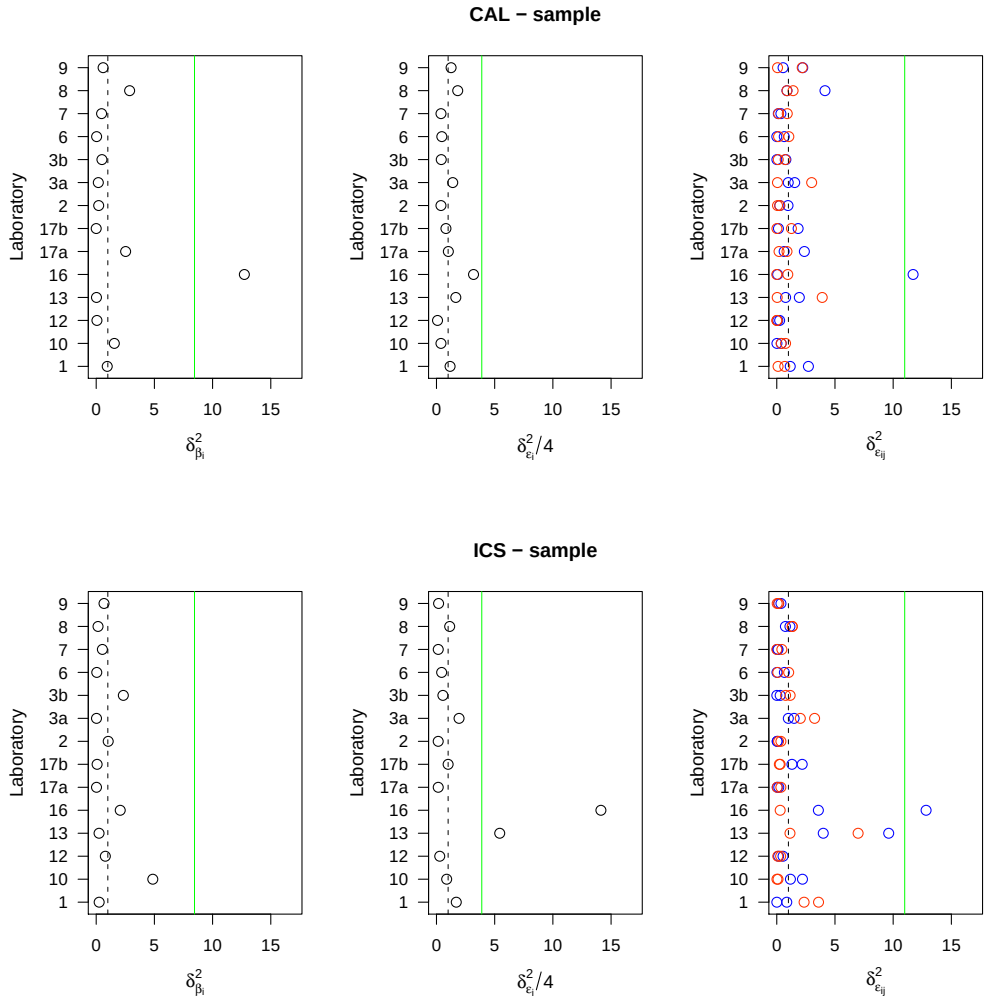


Table 6.1: **Estimated parameters of the t-linear mixed model for the CAL and ICS sample, with (i) unknown degress of freedom and (ii) setting the degrees of freedom to 4.**

Sample	t-model	$\hat{\mu}$	$\hat{\sigma}_a^2$	$\hat{\sigma}_\varepsilon^2$	$\hat{\nu}$	$\widehat{Var}(a_i)$	$\widehat{Var}(\varepsilon_{ij})$
CAL	ν unknown	2.99	0.0056	0.0018	9.24	0.0071	0.0023
	$\nu = 4$	2.99	0.0045	0.0016	4	0.0089	0.0032
ICS	ν unknown	3.49	0.0027	0.0022	3.03	0.0079	0.0066
	$\nu = 4$	3.49	0.0027	0.0025	4	0.0053	0.0049

6.4.2 Random coefficients model

In this section we apply the t-linear mixed model to the data from the Kyoto 1 and Orlando 2 study. For the Kyoto 1 study there is at least one laboratory, which shows high deviation from the other laboratories, as seen in Figure 5.5. However, based on the variance estimators from the normal-linear mixed model, this deviation is still acceptable.
In Figure 6.3, the estimated coefficients as well as the residuals, based on the t-linear mixed model, are plotted for each laboratory. The coefficients of laboratory 15_c are far away from the point (0,0), but also the coefficients of laboratory 10. The residuals of laboratory 12 show a much higher variation than the residuals of the other laboratories. The estimated coefficients from the t-linear mixed model are further shrinked towards the mean compared to those from the normal-linear mixed model.
The calculated outlier statistics are given in Figure 6.4, together with the limits based on the adjusted 0.95 quantiles of the respective Chi-Square distributions. Examining those of the Kyoto 1 study, laboratory 15_c and laboratory 10 are clearly identified as location-outliers within the laboratories. Laboratory 12 and laboratory 15_c are identified as scale-outlier, as the variation of their residuals is higher than in the other laboratories. One measurement of laboratory 1_CE is also extreme in comparison to the other measurements within this laboratory.
The results of the Orlando 2 study are quite similar to the Kyoto 1 study. Therefore we show only the plot of the outlier identification statistics. This time laboratory 3a is clearly a scale-outlier and a measurement of laboratory 16 is extreme compared to the other measurements within this laboratory.
We examined, that the robust estimation of the parameters of a linear mixed model leads to more reliable estimators of the involved variances, and therefore to better outlier identification based on the statistics defined in Section 5.4.

As threshold for the outlier limits, we choose the adjusted quantiles of the respective Chi-Square distributions. Under the assumption of the normal-linear mixed model, the outlier statistics follow this distribution. However, as we calculate the empirical sum of squares based on estimates of the random effects and residuals, this assumption need no longer to hold. It rests for further research to define more appropriate thresholds for these statistics for example based on simulation studies.

Table 6.2: **Estimates of the parameters of the t-linear mixed model for the the Kyoto 1 and Orlando 2 study, with (i) unknown degrees of freedom and (ii) setting the degrees of freedom to 4.**

Study	t-model	$\hat{\mathbf{b}}$	$\hat{\mathbf{D}}$	$\hat{\sigma}^2_\varepsilon$	$\hat{\nu}$
Kyoto 1	ν unknown	$\begin{pmatrix} 0.014 \\ -0.003 \end{pmatrix}$	$\begin{pmatrix} 0.0193 & -0.0054 \\ -0.0054 & 0.0023 \end{pmatrix}$	0.0100	2.38
	$\nu = 4$	$\begin{pmatrix} 0.010 \\ -0.001 \end{pmatrix}$	$\begin{pmatrix} 0.0203 & -0.0058 \\ -0.0058 & 0.0025 \end{pmatrix}$	0.0103	4
Orlando 2	ν unknown	$\begin{pmatrix} -0.048 \\ -0.010 \end{pmatrix}$	$\begin{pmatrix} 0.0142 & -0.0016 \\ -0.0016 & 0.0004 \end{pmatrix}$	0.0065	4.31
	$\nu = 4$	$\begin{pmatrix} -0.048 \\ 0.010 \end{pmatrix}$	$\begin{pmatrix} 0.0142 & -0.0016 \\ -0.0016 & 0.0004 \end{pmatrix}$	0.0064	4

Figure 6.3: **Plot of the systematic and proportional bias of each laboratory for the Kyoto 1 study estimated by the t-linear mixed model with $\nu = 4$. The plot on the right hand side shows the residuals for each laboratory; different colors indicate the different samples.**

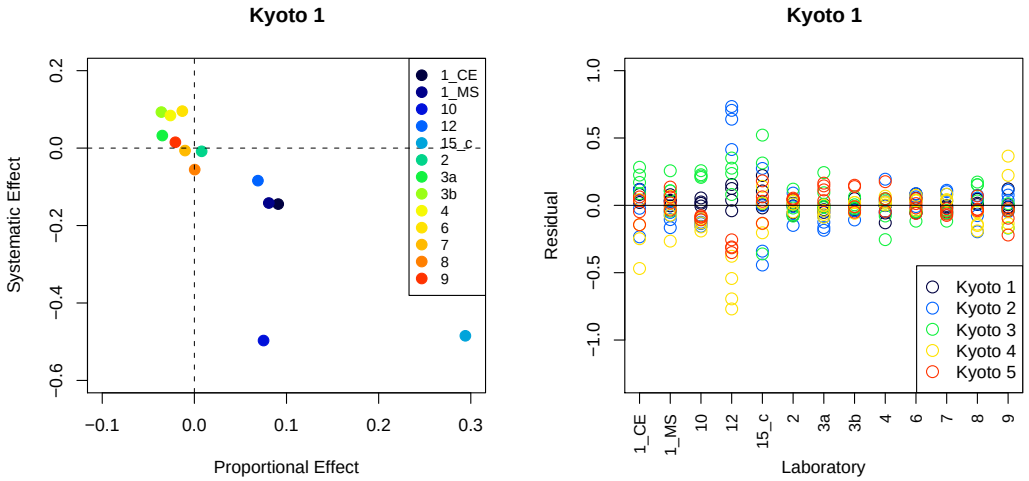


Figure 6.4: **Outlier identification statistics for the Kyoto 1 study based on the t-linear mixed model with $\nu = 4$. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**

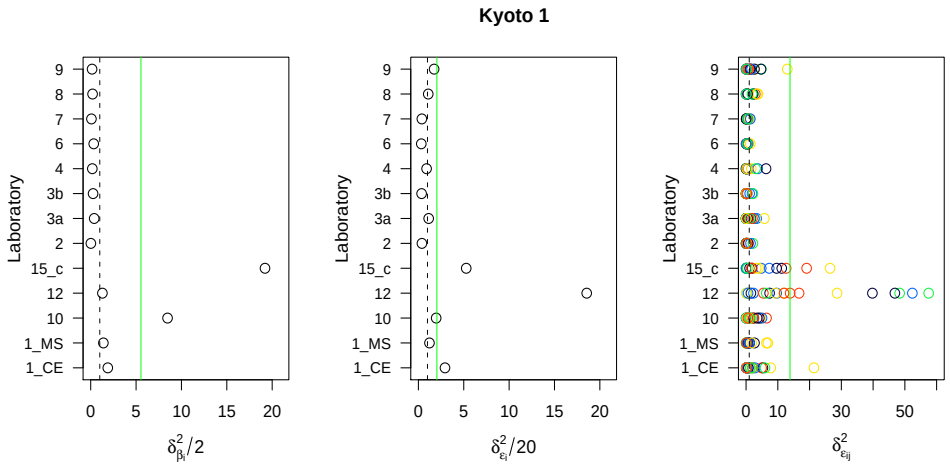
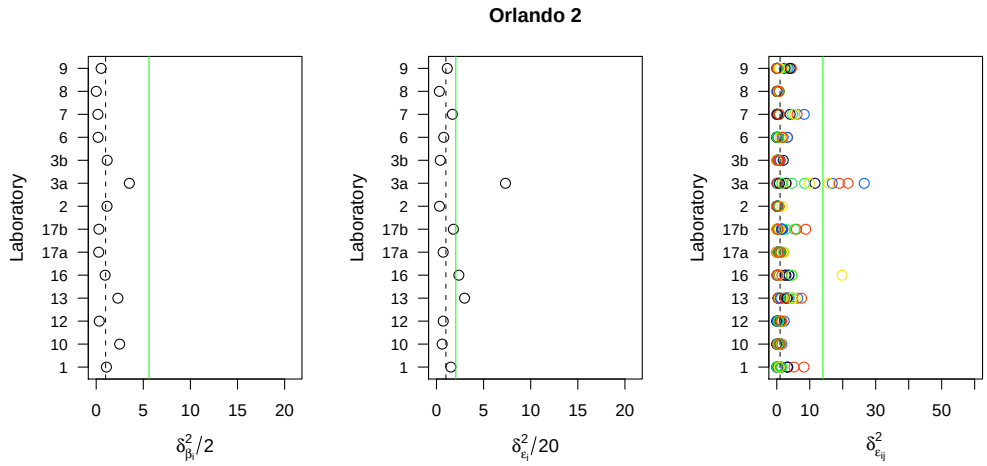


Figure 6.5: **Outlier identification statistics for the Orlando 2 study based on the t-linear mixed model with $\nu = 4$. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



Chapter 7

Quality control with linear mixed models

In Chapter 6, we demonstrated how outliers in linear mixed models can be detected, when the dispersion parameters of a linear mixed model are estimated based on the data at hand. However, this is not appropriate for quality control procedures, as these procedures are applied to a multitude of data-sets, which should all be treated equally. For quality control the acceptable quality needs to be defined in advance, for example based on preliminary experiments. Afterwards measurements are judged based on this quality.
We define quality in terms of acceptable variation, such that the outlier identification rules for linear mixed models can be applied. This means that the dispersion parameters for the definition of the outlier regions are set in advance.
In this chapter we show how these quality parameters can be derived and how they can be interpreted. We restrict ourself to data from the IFCC network for standardization of HbA1c and the one-way random effects and random coefficients model.

7.1 Quality control with one-way random effects models

Within a study of the IFCC network for standardization of HbA1c multiple samples are measured, each within each member laboratory. The measured results of a sample need to be combined to obtain the assigned values of each sample. Data from each sample needs to be scanned for extreme laboratories or extreme measurements within a laboratory.

In Section 6.4 we showed how this can be done by modelling this sort of data by a one-way random effects model. The model parameters are estimated in a robust way and the outlier identification statistics of Section 5.4 can then be applied. The dispersion parameters of the one-way random effects model are σ_a^2 - the between-laboratory variance, which describes the variation between the location of the network laboratories, and σ_ε^2 - the within-laboratory variance, describing the variation of the repeated measurements within the laboratories.
For the definition of a quality control rule, the two variances must be set in advance and the outlier identification statistics are based upon these. The definition of these parameters is either fully guided by external demands, e.g. by medical decision rules, or they are derived from preliminary experiments.
For the IFCC network for standardization of HbA1c data from 6 studies were already available, so that we decided to use these data to derive suitable variances.

7.1.1 Derivation of the dispersion parameters

The quality control rules for single samples are applied to different samples, with different percentages of HbA1c, ranging from 3% to 15%. So the first question to answer is if the between-laboratories variance and within-laboratories variance depend on the percentage of HbA1c within a particular sample.
In the 6 studies under consideration, 38 samples with an artificial matrix* and 63 whole blood samples were measured. For each sample we estimate the parameters σ_a^2 and σ_ε^2, based on the t-linear mixed model algorithm with 4 degrees of freedom, as this algorithm will also be used for quality control.
Afterwards we plot the estimated between- and within-laboratory standard deviations against the estimated concentration of the samples. We fit a linear regression line with intercept and without intercept to this data. Thus, we analyze how the standard deviations depend on the percentages of HbA1c within a sample.
Two questions are to be answered:

(i) Are there differences between these regression lines for samples with different matrices?

(ii) Is either the intercept or slope significant?

If the intercept is not significant for this variance function, we will be in the situation of a constant coefficient of variation over the measurement range, a well known behav-

*In clinical chemistry the matrix of a sample are all substances of the sample, except the analyte under consideration.

ior in clinical chemistry. If the slope is not significant, the standard deviation will stay constant over the measurement range.
In Figures 7.1 and 7.2, the estimated between-laboratories standard deviations and within-laboratories standard deviations are plotted versus the percentage of HbA1c in each sample. The standard deviations grow with growing amount of HbA1c and the linear modelling of the relationship is appropriate.

Regarding the estimates given in Table 7.1 in detail, we observe that both intercept and slope are significant in all cases except one. Only for the between-laboratory standard deviation for whole blood samples the intercept is not significant with a p-value of 0.258. For the between-laboratory standard deviation the functions differ for the two types of samples, whereas for the within-laboratory standard deviation they are equal. The linear fit is better for the within-laboratory standard deviation. This is an expected effect, as it comprises only the variation within the laboratories, whereas the between-laboratory variance is a mixture of several variation sources.

Based on the derived variance functions we can define quality control rules for samples

Table 7.1: **Estimates and fit statistics for the functions of the between-laboratory standard deviation and within-laboratory standard deviation dependent of the percentage of HbA1c.**

	Sample	Model	Coefficient	Estimate	p-value
σ_a	Whole Blood	With. Int.	Intercept	0.020	0.258
			Slope	0.016	$7 \cdot 10^{-8}$
		No Int.	Slope	0.018	$< 2 \cdot 10^{-16}$
	Artificial	With. Int.	Intercept	0.071	$1 \cdot 10^{-3}$
			Slope	0.009	$6 \cdot 10^{-4}$
		No Int.	Slope	0.017	$1 \cdot 10^{-4}$
σ_ε	Whole Blood	With. Int.	Intercept	0.025	$3 \cdot 10^{-6}$
			Slope	0.006	$2 \cdot 10^{-12}$
		No Int.	Slope	0.010	$< 2 \cdot 10^{-16}$
	Artificial	With. Int.	Intercept	0.025	$6 \cdot 10^{-3}$
			Slope	0.006	$3 \cdot 10^{-7}$
		No Int.	Slope	0.009	$< 2 \cdot 10^{-16}$

measured within the IFCC network for standardization of HbA1c:

(i) Fit an one-way random effects model with the t-linear mixed model algorithm with

Figure 7.1: **Estimated between-laboratories standard deviations versus the percentage of HbA1c for whole blood and artificial samples of the IFCC network for standardization of HbA1c. The red line is the fitted least-square regression with intercept, the green line without intercept.**

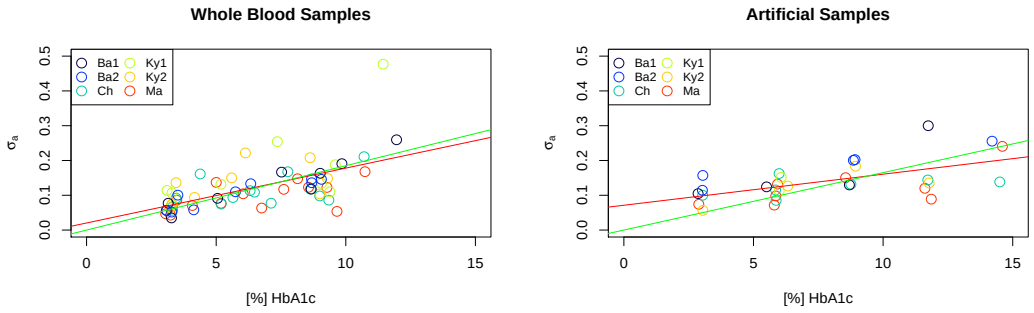
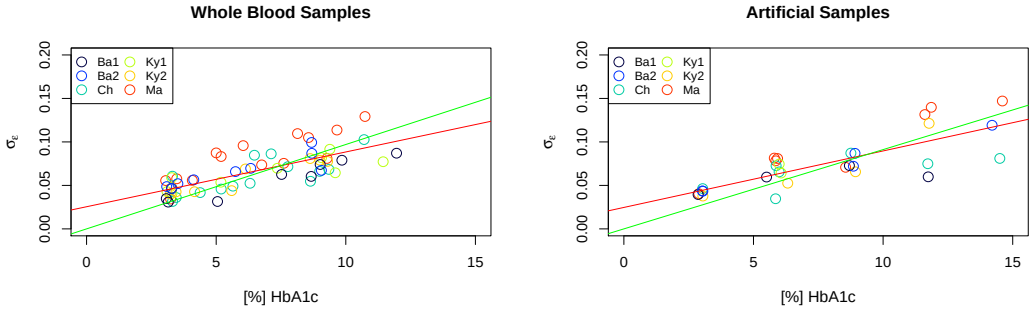


Figure 7.2: **Estimated within-laboratories standard deviations versus the percentage of HbA1c for whole blood and artificial samples of the IFCC network for standardization of HbA1c. The red line is the fitted least-square regression with intercept, the green line without intercept.**



4 degrees of freedom for data from each sample.

(ii) Based on the type of sample (artificial or whole blood sample) calculate the sample specific σ_a^2 and σ_ε^2, dependent on the estimated fixed effect $\hat{\mu}$:

$$\begin{aligned} \sigma_a &= \begin{cases} 0.018 \cdot \hat{\mu}, & \text{for whole blood sample} \\ 0.071 + 0.009 \cdot \hat{\mu}, & \text{for artificial sample.} \end{cases} \\ \sigma_\varepsilon &= 0.025 + 0.006 \cdot \hat{\mu}. \end{aligned} \tag{7.1.1}$$

(iii) Calculate the outlier statistics for the random effects and residuals, given the calculated σ_a^2 and σ_ε^2.

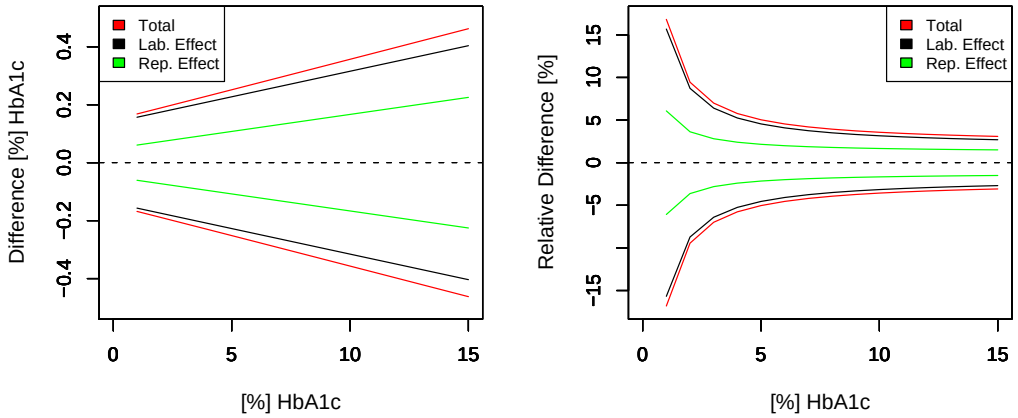
(iv) If the calculated outlier statistics exceed the respective Chi-square quantiles, identify the laboratory or observation as not conforming to the defined quality.

To provide an easy interpretation of this QC-rule, we compute the parameters σ_a^2, σ_ε^2 for given percentages of HbA1c ranging from 1% HbA1c to 15% HbA1c for artificial samples. Based on these, the 0.025 and 0.975 quantiles of the respective zero-mean normal distribution are calculated. These quantiles represent extreme values of the laboratory effects and measurements within laboratories, which are still accepted by the QC-rule. We calculate also the quantiles of a normal distribution with mean 0 and variance given by the sum of both variances, representing the allowable total variation of a single measurement. Finally, these quantiles are set in relation to the amount of HbA1c to express the effects in percentages of this amount. The plot of the quantiles is given in Figure 7.3.
The allowable variation increases with increasing percentage of HbA1c in the samples, from 0.2% HbA1c in the lower measurement range, to 0.4% HbA1c in the higher range. In terms of relative differences it becomes narrower for samples in the higher measurement range: for samples with more than 5% HbA1c, the allowable variation for the relative differences is less than 5%.

7.1.2 Application of the QA-rules to the CAL and ICS sample

In this section we will show the application of the quality control rules to the CAL and ICS sample, which were already introduced in Section 6.4. The CAL sample is an artificial sample, being a mixture of pure HbA1c and HbA0. (See [KAS+06] for a detailed explanation of the production process of calibrator samples.) The ICS sample is a whole blood sample.

Figure 7.3: **Absolute and relative deviation of the laboratory effects, repetition effects and total variation allowed by the quality control rule for artificial samples.**

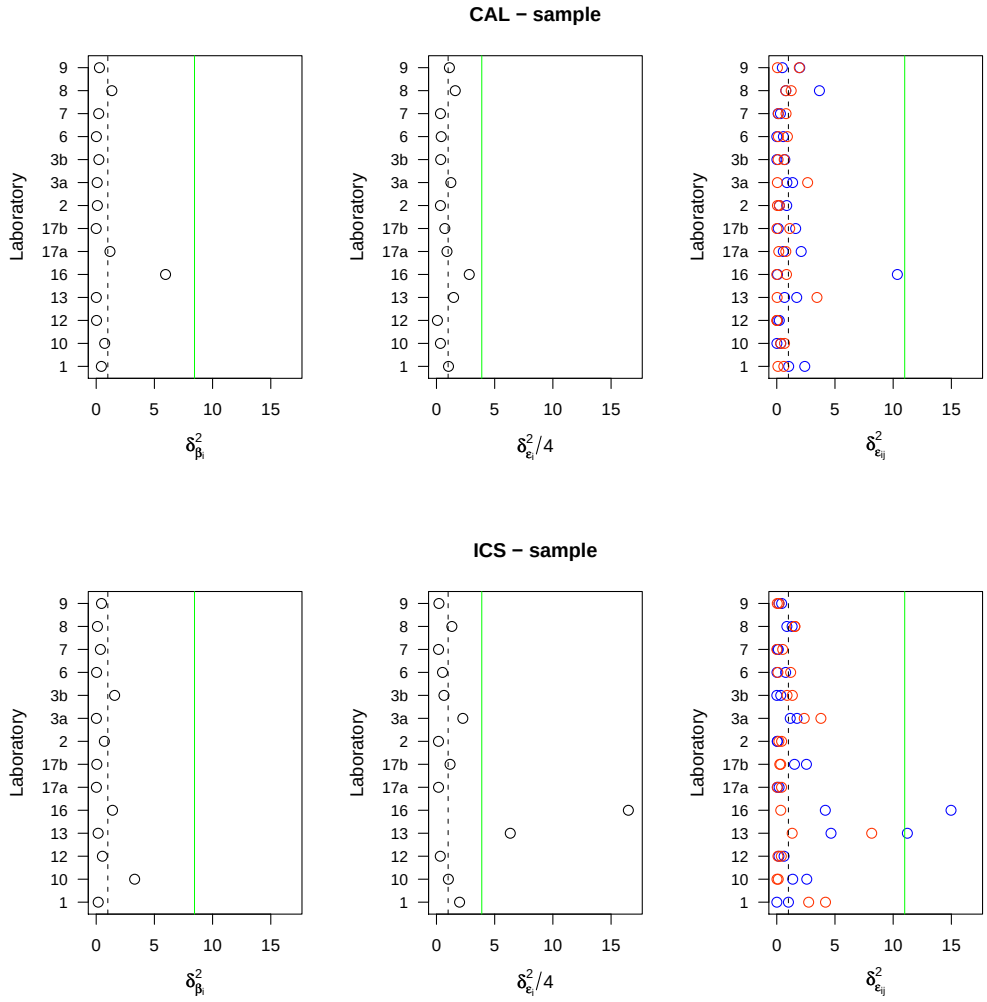


Fitting the one-way random effects model based on the t-linear mixed model algorithm with $\nu = 4$ to the data of both samples, we obtain $\hat{\mu}_{CAL} = 2.99$ and $\hat{\mu}_{ICS} = 3.49$. Based on these estimates we calculate σ_a^2 and σ_ε^2, according to the variance functions derived in Section 7.1.1, leading to

$$\tilde{\sigma}_a^2(CAL) = 0.0097, \quad \tilde{\sigma}_\varepsilon^2(CAL) = 0.0018,$$
$$\tilde{\sigma}_a^2(ICS) = 0.0039, \quad \tilde{\sigma}_\varepsilon^2(ICS) = 0.0021.$$

Comparing these with the estimates of the variance parameters based on the t-linear mixed model with $\nu = 4$, given in Table 6.1, we note that they are a little bit higher. This leads to smaller outlier identification statistics, i.e. it might be that laboratories or measurements which were identified as outliers based on the t-linear mixed model, will still agree with the quality defined by the quality control rules. Figure 7.4 shows these outlier identification statistics for both samples. For the CAL sample, neither a laboratory is detected as location-outlier nor as scale-outlier, further no extreme measurements within a laboratory are found. For the ICS sample the same extreme laboratories and measurements are identified as in Section 6.4.

Figure 7.4: **Outlier identification statistics for the CAL and ICS sample of the IFCC network for standardization of HbA1c, based on the quality control rules. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



7.2 Quality control with the random coefficients model

After the data from a study of the IFCC network for standardization of HbA1c is collected, one has to decide which laboratories and candidate laboratory are approved as members of the network. The random coefficients model is appropriate for the judgement of the measurement behavior of the laboratories over the whole measurement range, as shown in Section 6.4. The definition of quality control rules for laboratory approval, which hold for multiple studies, requires to set the dispersion parameters of the random coefficients model in advance.

7.2.1 Derivation of the dispersion parameters

The dispersion parameters of the random coefficients model are the matrix $\mathbf{D}$ and σ_ε^2 - the variance of the residuals. They can be defined by external requirements, but this starts to become difficult, for the definition of $\mathbf{D}$. In Section 7.2.2, we will give some interpretation of this matrix and show its impact on the variation of the coefficients of the random coefficients model.
Another possibility to define this matrix is to derive the dispersion parameters from older studies, similar to what we did in Section 7.1.1.
We regard the 6 studies, which were already used for the derivation of the variance function for the quality control rules for single samples. For each study the t-linear mixed model with $\nu = 4$ is fitted to obtain estimates of $\mathbf{D}$ and σ_ε^2. The estimate $\hat{\mathbf{D}}$ is a 2×2 matrix. The diagonal elements correspond to the variances of the β_i and the off-diagonal element is the covariance of these coefficients, in the normal-linear mixed model. Based on these, we calculate the correlation between the coefficients. The estimates for each study are listed in Table 7.2. To derive reference parameters for the quality control rules, we take the medians of the variances and of the correlations over all studies under consideration. The reference covariance is then calculated from the combination of the reference variances and correlation. Based on these 6 studies, the dispersion parameters $\tilde{\mathbf{D}}$, and $\tilde{\sigma}_\varepsilon^2$ used for the quality control rules are set to

$$\tilde{\mathbf{D}} = \begin{pmatrix} 0.0157 & -0.0023 \\ -0.0023 & 0.0006 \end{pmatrix}, \quad \tilde{\sigma}_\varepsilon^2 = 0.0087.$$

We define the quality control rules for the approval of laboratories from the IFCC network for standardization of HbA1c based on these dispersion parameters:

(i) Fit a random coefficients model by the t-linear mixed model algorithm with $\nu = 4$ for the data from the ICS samples of one study.

Table 7.2: **Estimates of the dispersion parameters of the random coefficients model calculated by the t-linear mixed model algorithm for 6 studies of the IFCC network for standardization of HbA1c.**

Study	$\hat{\mathbf{D}}[1,1]$	$\hat{\mathbf{D}}[2,2]$	ρ	$\hat{\sigma}^2_\varepsilon$
Ma	0.0031	0.0002	-0.6332	0.0122
Ch	0.0219	0.0004	-0.9147	0.0094
Ky1	0.0203	0.0025	-0.8010	0.0103
Ky2	0.0044	0.0010	-0.1323	0.0080
Ba1	0.0247	0.0007	-0.7717	0.0075
Ba2	0.0111	0.0004	-0.8804	0.0067
Median	0.0157	0.0006	-0.7863	0.0087

(ii) Calculate the outlier identification statistics for the laboratory effects and residuals with $\tilde{\mathbf{D}}$, and $\tilde{\sigma}^2_\varepsilon$.

(iii) If the calculated outlier statistics exceed the respective Chi-square quantiles, identify the laboratory or measurement as not conforming to the defined quality.

In the next section, we will give some interpretation for these quality control rules, to clarify the impact of the matrix $\tilde{\mathbf{D}}$ for the defined quality.

7.2.2 Interpretation of the variance-covariance matrix

The identification of laboratories as location-outliers in studies of the IFCC network for standardization of HbA1c depends strongly on the specified variance-covariance matrix $\tilde{\mathbf{D}}$. However, it is difficult to interpret the impact of different definitions of this matrix based on the elements of this matrix, especially as the correlation between both coefficients is important. To visualize the impact of the specified variance-covariance matrix, we regard the following steps for the matrices estimated in the 6 studies, as well as for the derived variance-covariance matrix.
The set

$$\{\beta \in \mathbb{R}^2 \mid \beta_i' \mathbf{D}^{-1} \beta_i < c\}$$

describes an elliptic region in the two-dimensional plane centered at $(0,0)'$. Laboratories, with estimated coefficients outside this region are considered as location-outliers according to the defined quality control rule. We approximate the border of this plane by 200 datapoints and transform each point into a regression line. These lines vary around

the zero line. In Figure 7.5, the elliptic region and the transformed regression lines are given for the quality control rule. Within the measurement range of the HbA1c[%] assay, i.e. $0\% - 15\%$, we calculate the maximal deviation from zero for the regression lines as well as the maximal relative deviation on a grid of 0.5%. With these calculations, we visualize the maximal relative deviation profile over the whole measurement range. Based on this profile the appropriate variance-covariance matrix could be chosen, too.
In Figure 7.6 the maximal deviation and relative deviation profiles are shown for the estimated variance-covariance matrices from the 6 studies, as well for $\tilde{\mathbf{D}}$. The deviation profiles based on $\tilde{\mathbf{D}}$ are displayed as red lines, named "QC" within the plot. Examining the absolute deviation plot, there is a minimum in the absolute deviations for each study, ranging from 0% HbA1c to 5% HbA1c. This is due to the negative correlation between the intercept and the slope. Regarding the absolute deviations based on the reference variance-covariance matrix $\tilde{\mathbf{D}}$, the maximal allowable absolute deviation ranges from 1 % HbA1c for samples with an amount of HbA1c around 5%, to 3% HbA1c in the upper measuring range.
The plot of the maximal relative differences shows that relative deviations of 6% for samples with an amount of 5% HbA1c to 10% HbA1c are still allowable. For samples with a higher amount the relative deviations are a little bit higher. For samples with percentage of HbA1c higher than 3%, relative deviations stay below 10%. For the studies of the IFCC network for standardization of HbA1c there was an older rule for the approval of laboratories as members of the network. It stated that laboratories were not approved, if more than three samples within one study deviated with their laboratory-mean more than 6% from the overall mean. This caused problems, as the number of samples changed from study to study and also, because samples were treated equally over the whole measurement range. In Figure 7.3 we see that for samples in the lower measuring range the relative deviation can exceed 6%. The laboratory approval rule based on the random coefficients model takes automatically this into account.

7.2.3 Application of the quality control rules

In this section we apply the quality control rule defined in Section 7.2.1, to the data of the Kyoto 1 and Orlando 2 study.
The aim is to identify laboratories as location-outliers based on the laboratory effects, as well as laboratories as scale-outliers according to the variation of the laboratory-specific residuals. The laboratory effects are derived from the t-linear mixed model with $\nu = 4$. The outlier identification statistics are calculated with the reference matrix $\tilde{\mathbf{D}}$ and $\tilde{\sigma}^2_\varepsilon$,

Figure 7.5: **Two-dimensional elliptic region and transformed regression lines for the quality control rule based on $\tilde{D}$.**

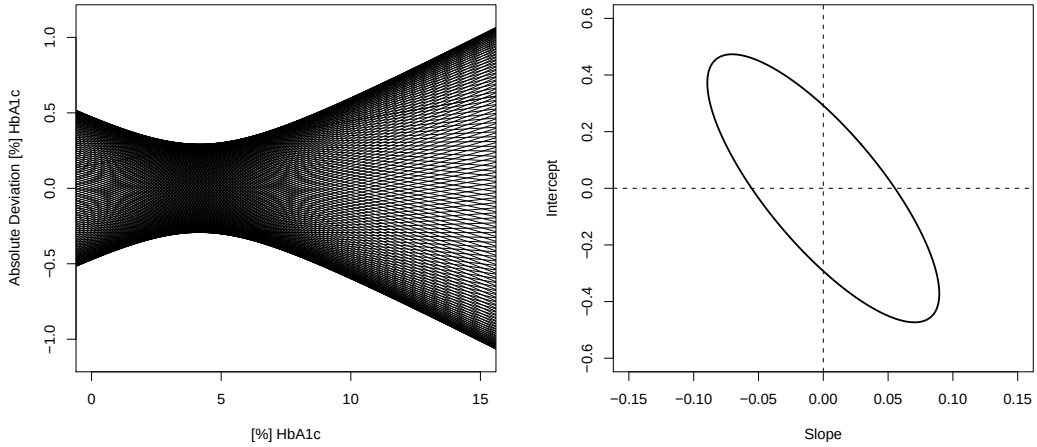
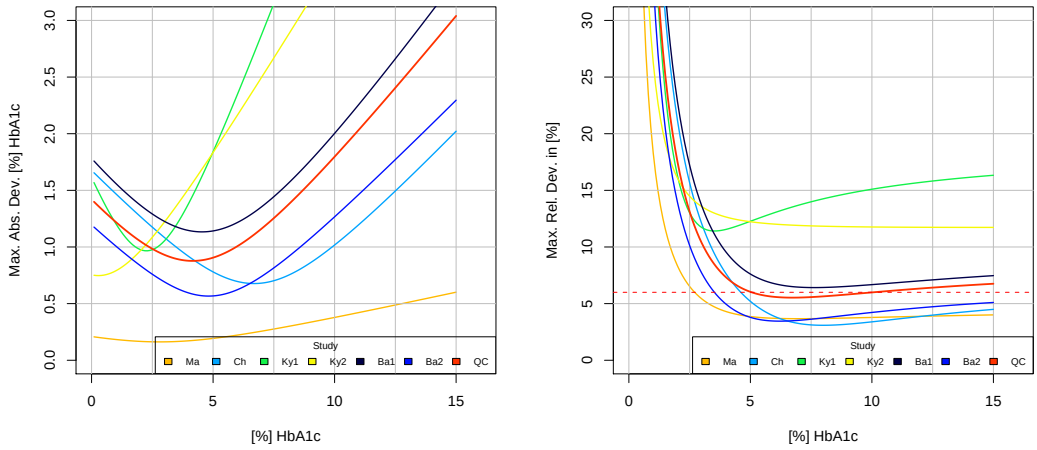
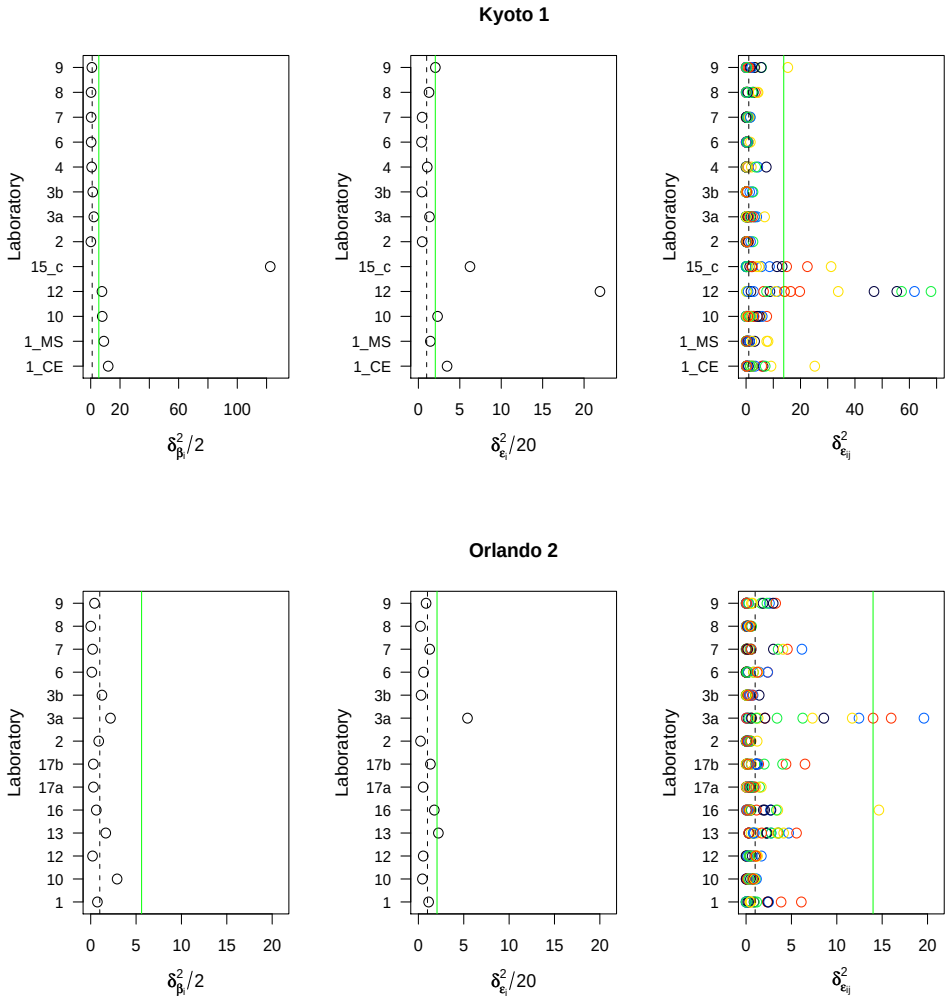


Figure 7.6: **Absolute and relative deviation profiles for the quality control rule for the identification of laboratories as location-outliers based on the random coefficients model.**



derived in Section 7.2.1.
Comparing the diagonal elements of the reference matrix $\tilde{\mathbf{D}}$, with those estimated from the data of the respective study, we note that they lie between the values estimated for the Kyoto 1 study and those estimated for the Orlando 2 study. The same holds for the variance of the residuals. In Figure 7.7 the outlier statistics are shown for both studies. Within the Kyoto 1 study, 5 laboratories are not approved, compared to two laboratories if the outlier statistics are calculated based on the estimated dispersion parameters from the Kyoto 1 study. This dispersion is the widest compared with all other studies. Perhaps one could also conclude that the whole study failed, as 5 out of 13 laboratories could not achieve the requested quality. Laboratory 15_c and laboratory 12 are also identified as scale-outliers, both due to a higher residual variation.
For the Orlando 2 study, we obtain a similar picture as in Section 6.4. No laboratory is a location-outlier, laboratory 3a is a scale-outlier due to a higher residual variance, laboratories 13 and 16 are borderline.

Figure 7.7: **Outlier identification statistics for the Kyoto 1 study based on the quality control rule. The outlier limits ($\alpha = 0.05$) are indicated as green line, the expected value of the outlier statistics as dashed line.**



Part III

Method comparison studies

Method comparison studies are an often used method to analyze the equivalence between two measurement methods, measuring devices or standardization systems.
To compare for instance two measurement methods a set of human samples is measured within both methods. Hence, for each sample one obtains a pair of measured values, such that a regression line between both methods can be derived, in the case that the linearity assumption holds.
On one hand side this experiment can reveal, if the two methods are exchangeable. If the intercept of the regression line is near zero and the slope close to one, one can argue that both methods are equivalent, or that they can be used exchangeable. If the two methods are based on different standardization systems, for example a national and an international one, the exchangeability is often not given. However, based on the though derived regression line a transformation rule from one method to the other can be defined. By means of this rule patient values of one method can be recalculated in values of the other method and vice versa.
As method comparison studies are a well-known tool for diagnostic assays there is a lot of literature for the analysis of one particular experiment (see e.g. [RRR01], [MRR02], [PB83]). These articles discuss especially the best regression method to use. In most cases both methods are subject to error, hence errors-in-variable models are appropriate for the derivation of the regression line (for more details see e.g. [CVN99], [Ful87]).
In cases, where one is interested in a transformation rule from one method to the other, it occurs that this experiment is repeated after a certain time and one would like to examine, if this transformation rule has changed. This is equal to the comparison of two or more regression lines over a certain interval. In Chapter 8 we will derive a test statistic for this comparison. The test statistic is based on a proposed test of [LJZ04] for the least-square regression case. We extend this test approach to be able to compare a new regression line with a reference regression line. This might be a new approach to show the exchangeability of two methods.
In Chapter 9 we present how multiple regression lines can be combined, to obtain an average regression line. We will use Bayesian hierarchical linear models, which are extended to incorporate the errors in both methods. Bayesian approaches require the specification of prior distributions, which might be difficult to assess. Hence, it is important to check the sensitivity of the results for different prior distributions as well as to determine the most appropriate prior distribution. We present a method to check the adequacy of different prior settings and provide a measure for the derivation of the most adequate prior. Based on data of the IFCC network for standardization of HbA1c we show the influence of the prior distributions and the model assumptions on the posterior distribution of the parameters of the averaged regression line.

Chapter 8

Comparison of regression lines

If method comparison experiments are repeated after a certain time, it is sometimes necessary to compare these sequentially obtained regression lines. We will base the comparison of regression lines on the construction of simultaneous confidence bounds for the differences of the predicted values over a given concentration range. [Spu99] developed exact confidence bounds for a contrast of regression lines under the assumption of equal design matrices. However, this assumption is not fulfilled for method comparison studies, as different samples are used in each study. Therefore we base our work on [LJZ04], where a test statistic with simulated probability density function is proposed. With this approach multiple regression lines can be compared even with unequal design matrices.
In a first step, we present the approach for the comparison of multiple regression lines. Afterwards we extend this setting to compare a new regression line with a reference regression line. At the end of this chapter we provide two examples to show the utility of both approaches in the context of standardization of diagnostic assays.

8.1 Multiple regression lines

First we consider the comparison of multiple regression lines, developed by [LJZ04]. Suppose we have k linear regression models, given by

$$\mathbf{Y}_i = \mathbf{X}_i\mathbf{b}_i + \varepsilon_i,\ i = 1, .., k,$$

where $\mathbf{Y}_i = (Y_{i1}...Y_{in_i})'$ is the vector of the response variable of regression model i, $\mathbf{b}_i \in \mathbb{R}^p$ the coefficient vector of model i and ε_i the vector of the residuals, with all residuals being independent normally distributed with mean zero and variance σ^2(independent of

the model). Finally, let $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ be the design matrix of model i with full rank. Hence, $\mathbf{X}_i'\mathbf{X}_i$ is nonsingular and the least-squares estimator of $\mathbf{b}_i$ is given by $\hat{\mathbf{b}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_i$. As the residuals are normally distributed, the estimator $\hat{\mathbf{b}}_i$ is also normally distributed with mean $\mathbf{b}_i$ and variance $Var(\hat{\mathbf{b}}_i) = \sigma^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}$.
The variance of the residuals is estimated by the pooled mean squares error over all models, i.e.

$$\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i - 2}SSE_i = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i - 2}(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_i)'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}_i).$$

The comparison of the regression lines is based on the set of simultaneous confidence bands for

$$\mathbf{x}'\mathbf{b}_i - \mathbf{x}'\mathbf{b}_j, \ \forall\ (i,j) \in \Lambda, \ \forall\ \mathbf{x} \in \mathbb{S},$$

where Λ is an index set that determines the comparison of interest. If all pairwise comparisons are of interest then $\Lambda = \{(i,j) \mid 1 \leq i \neq j \leq k\}$. If only the comparison of one particular regression line $i^\star$ with the others is of interest than $\Lambda = \{(i,j) \mid i = i^\star, 1 \leq j \leq k, j \neq i^\star\}$.
The set $\mathbb{S}$ denotes the p-dimensional region, on which the comparison of the regression lines is of interest:

$$\mathbb{S} = [c_{min}(1), c_{max}(1)] \times [c_{min}(2), c_{max}(2)] \times ... \times [c_{min}(p), c_{max}(p)].$$

If an intercept is included in the model $[c_{min}(1), c_{max}(1)] = 1$.
The variance of $\mathbf{x}'\hat{\mathbf{b}}_i - \mathbf{x}'\hat{\mathbf{b}}_j$ is given by

$$Var(\mathbf{x}'\hat{\mathbf{b}}_i - \mathbf{x}'\hat{\mathbf{b}}_j) = \sigma^2\mathbf{x}'\left((\mathbf{X}_i'\mathbf{X}_i)^{-1} + (\mathbf{X}_j'\mathbf{X}_j)^{-1}\right)\mathbf{x} = \sigma^2\mathbf{x}'\triangle_{ij}\mathbf{x}.$$

Therefore, the following set of simultaneous confidence intervals is constructed:

$$\mathbf{x}'\mathbf{b}_i - \mathbf{x}'\mathbf{b}_j \ \in \ \mathbf{x}'\hat{\mathbf{b}}_i - \mathbf{x}'\hat{\mathbf{b}}_j \pm c \cdot \hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}}, \ \forall\ (i,j) \in \Lambda, \ \forall\ \mathbf{x} \in \mathbb{S}.$$

The critical value c must be chosen such that the confidence level of this set of simultaneous confidence bands is $1 - \alpha$, $(0 < \alpha < 1)$ i.e. $P(T < c) = 1 - \alpha$, where

$$T = \sup_{(i,j)\in\Lambda}\sup_{\mathbf{x}\in\mathbb{S}}\frac{\mid \mathbf{x}'\left((\hat{\mathbf{b}}_i - \mathbf{b}_i) - (\hat{\mathbf{b}}_j - \mathbf{b}_j)\right) \mid}{\hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}}}.$$

Since the critical value c will be determined by simulating the distribution of T, another representation of T will be adopted from which a simulation algorithm can be developed easily.
The matrix $\triangle_{ij}$ is symmetric and positive-definite, hence there exists the Cholesky decomposition of the matrix $\triangle_{ij}$, i.e. there exist a nonsingular matrix $\mathbf{P}_{ij} \in \mathbb{R}^{p\times p}$ such that

$$\triangle_{ij} = \mathbf{P}'_{ij}\mathbf{P}_{ij},\ \forall\ (i,j) \in \Lambda.$$

Let $\mathbf{Z}_i \in \mathbb{R}^{p\times 1},\ i = 1,..,k$, be independent normally distributed random vectors, independent of $\hat{\sigma}^2$, with distribution $\mathbf{Z}_i \sim N(0,(\mathbf{X}'_i\mathbf{X}_i)^{-1})$. Define

$$\mathbf{Z}_{ij} = (\mathbf{P}'_{ij})^{-1}(\mathbf{Z}_i - \mathbf{Z}_j),\ \forall\ (i,j) \in \Lambda. \tag{8.1.1}$$

Based on these definitions the distribution of T is the same as the distribution of

$$\begin{aligned} & \sup_{(i,j)\in\Lambda}\ \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid \mathbf{x}'(\mathbf{Z}_i - \mathbf{Z}_j)\mid}{(\hat{\sigma}/\sigma)\sqrt{\mathbf{x}'\mathbf{P}'_{ij}\mathbf{P}_{ij}\mathbf{x}}} \\ = & \sup_{(i,j)\in\Lambda}\ \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid (\mathbf{P}_{ij}\mathbf{x})'\mathbf{Z}_{ij}\mid}{(\hat{\sigma}/\sigma)\sqrt{(\mathbf{P}_{ij}\mathbf{x})'\mathbf{P}_{ij}\mathbf{x}}} \\ = & \sup_{(i,j)\in\Lambda} Q_{ij}\frac{\parallel \mathbf{Z}_{ij}\parallel}{\hat{\sigma}/\sigma}, \end{aligned} \tag{8.1.2}$$

where

$$Q_{ij} = \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid (\mathbf{P}_{ij}\mathbf{x})'\mathbf{Z}_{ij}\mid}{\parallel \mathbf{P}_{ij}\mathbf{x}\parallel \cdot \parallel \mathbf{Z}_{ij}\parallel}.$$

For $p \geq 3$, [LJZ04] propose a gradient projection algorithm to directly calculate

$$W_{ij} = \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid \mathbf{x}'(\mathbf{Z}_i - \mathbf{Z}_j)\mid}{(\hat{\sigma}/\sigma)\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}}},\ \forall (i,j) \in \Lambda,$$

by solving the optimization problem:
Maximize $\forall (i,j)$

$$\begin{aligned} & f(\mathbf{x}) = \frac{\left(\mathbf{x}'(\mathbf{Z}_i - \mathbf{Z}_j)\right)^2}{(\hat{\sigma}/\sigma)^2\mathbf{x}'\triangle_{ij}\mathbf{x}} \\ \text{subject to} \quad & -x_i \leq -c_{min}(i),\ i = 1,..,p \\ & x_i \leq c_{max}(i),\ i = 1,..,p. \end{aligned}$$

The interested reader is referred to [LJZ04] for details about this algorithm.
However, if the regression model has only two coefficients, say an intercept and a slope, Q_{ij} can be calculated based on geometrical reformulations.
Denote in this case the column vectors of $\mathbf{P}_{ij}$ by $(\mathbf{p}_{ij}^1, \mathbf{p}_{ij}^2)$ and define the set

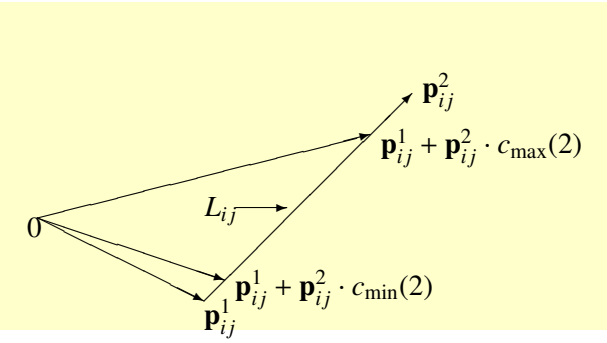
$$L_{ij} = \left\{ \mathbf{p}_{ij}^1 + x_2 \cdot \mathbf{p}_{ij}^2 : x_2 \in [c_{min}(2), c_{max}(2)] \right\}.$$

Then Q_{ij} can be written as

$$\begin{aligned} Q_{ij} &= \sup_{\mathbf{w} \in L_{ij}} \frac{|\mathbf{w} \cdot \mathbf{Z}_{ij}|}{||\mathbf{w}|| \cdot ||\mathbf{Z}_{ij}||} \\ &= \sup_{\mathbf{w} \in L_{ij}} \cos(\delta(\mathbf{w}, \mathbf{Z}_{ij})), \end{aligned}$$

where $\delta(\mathbf{w}, \mathbf{Z}_{ij})$ denotes the angle between $\mathbf{w}$ and $\mathbf{Z}_{ij}$. Therefore, the calculation of Q_{ij} simplifies to the determination of the smallest angle between either $\mathbf{w}$ and $\mathbf{Z}_{ij}$ or $\mathbf{w}$ and $-\mathbf{Z}_{ij}$. This restricted search in the two-dimensional space is even easier. By a simple

Figure 8.1: **The set** L_{ij}.



geometrical argument (see also Figure 8.1), it is evident that if $\mathbf{Z}_{ij}$ or $-\mathbf{Z}_{ij}$ is in the cone spanned by $\mathbf{p}_{ij}^1 + \mathbf{p}_{ij}^2 \cdot c_{\min}(2)$ and $\mathbf{p}_{\mathbf{ij}}^{\mathbf{1}} + \mathbf{p}_{\mathbf{ij}}^{\mathbf{2}} \cdot c_{\max}(2)$ then $Q_{ij} = 1$, otherwise

$$Q_{ij} = \max \left\{ \frac{| \, (\mathbf{p}_{ij}^1 + \mathbf{p}_{ij}^2 \cdot c_{\min})' \mathbf{Z}_{ij} \, |}{\| \, (\mathbf{p}_{ij}^1 + \mathbf{p}_{ij}^2 \cdot c_{\min}) \, \| \cdot \| \, \mathbf{Z}_{ij} \, \|}, \frac{| \, (\mathbf{p}_{ij}^1 + \mathbf{p}_{ij}^2 \cdot c_{\max})' \mathbf{Z}_{ij} \, |}{\| \, (\mathbf{p}_{ij}^1 + \mathbf{p}_{ij}^2 \cdot c_{\max}) \, \| \cdot \| \, \mathbf{Z}_{ij} \, \|} \right\}. \quad (8.1.3)$$

With the above results, a random realization of the random variable T can be obtained by the following algorithm:

1. Determine $\mathbf{P}_{ij}$, $\forall (i, j) \in \Lambda$.

2. Simulate independent draws $\mathbf{Z}_i \sim N(0, (\mathbf{X}_i'\mathbf{X}_i)^{-1})$, $\forall (i,j) \in \Lambda$ and $\hat{\sigma}/\sigma \sim \sqrt{\chi_v^2/v}$, where $v = \sum_i (n_i - 2)$ and χ_v^2/v denotes the Chi-square distribution with v degrees of freedom.

3. Calculate $\mathbf{Z}_{ij}$ according to (8.1.1).

4. Find Q_{ij} according to (8.1.3).

5. Compute T from the new representation (8.1.2).

Steps $2-5$ need to be repeated R times to simulate R replicates of the random variable T. The $(1-\alpha)R$th largest simulated value is the estimator of the critical value c.
To test whether the regression lines are the same, we calculate

$$\begin{aligned} LL_{ij} &= \max_{\mathbf{x}\in\mathbb{S}} \left\{ \mathbf{x}'\hat{\mathbf{b}}_i - \mathbf{x}'\hat{\mathbf{b}}_j - c\cdot\hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}} \right\} \\ UL_{ij} &= \min_{\mathbf{x}\in\mathbb{S}} \left\{ \mathbf{x}'\hat{\mathbf{b}}_i - \mathbf{x}'\hat{\mathbf{b}}_j + c\cdot\hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}} \right\}, \end{aligned}$$

which are in the case of a simple linear regression model with intercept and slope

$$\begin{aligned} LL_{ij} &= \max_{x_2\in[c_{\min}(2),c_{\max}(2)]} \left\{ \hat{b}_{0i} + \hat{b}_{1i}x_2 - \hat{b}_{0j} - \hat{b}_{1j}x_2 - c\cdot\hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}} \right\} \\ UL_{ij} &= \min_{x_2\in[c_{\min}(2),c_{\max}(2)]} \left\{ \hat{b}_{0i} + \hat{b}_{1i}x_2 - \hat{b}_{0j} - \hat{b}_{1j}x_2 + c\cdot\hat{\sigma}\sqrt{\mathbf{x}'\triangle_{ij}\mathbf{x}} \right\}. \end{aligned}$$

LL_{ij} denotes the maximum of the lower limits of the comparisons at each point $\mathbf{x} \in \mathbb{S}$, whereas UL_{ij} denotes the minimum of the upper limits of the comparisons in this region. The pair of regression lines (i, j) will be considered equal in the region $\mathbb{S}$, if the band for the pair (i, j) includes 0 over the whole range, that is equivalent to $LL_{ij} < 0 < UL_{ij}$. In the case of a simple linear regression model it is possible to do a grid search over the defined interval to assess both limits.

8.2 Comparison with a reference regression line

In method comparison studies of diagnostic assays a new regression line often needs to be compared against a reference regression line, which was not necessarily derived in the classical least-squares approach.
For example, to show the equivalence of two assays, there are specifications from clinical point of views, defining in which case the assays can be used exchangeably. One example might be that the intercept is allowed to vary between -1 and 1 and the slope between 0.98 to 1.02.
We translate these specifications in terms of a reference regression line in the following way: The best estimate of the coefficients of the reference regression line $\mathbf{b}$ is $\hat{\mathbf{b}} = (0, 1)'$, with

$$Var(\hat{\mathbf{b}}) = \mathbf{V} = \begin{pmatrix} 0.5^2 & 0 \\ 0 & 0.01^2 \end{pmatrix}.$$

To show the equivalence of a reference regression line and a new regression line the algorithm of [LJZ04] must be adapted in a few steps, but the main ideas remain the same: The comparison is based on the simultaneous confidence bands for the differences of the predicted values over a specified range.
The regression line derived in a new experiment may be written as

$$\mathbf{Y}_s = \mathbf{X}_s\mathbf{b}_s + \varepsilon_s,$$

with coefficient vector $\mathbf{b}_s$ and design matrix $\mathbf{X}_s$. The coefficients are estimated by standard least-squares regression, leading to the estimator $\hat{\mathbf{b}}_s = (\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{Y}_s$, as $(\mathbf{X}_s'\mathbf{X}_s)^{-1}$ is nonsingular. The variance-covariance matrix of the estimator is given by $\mathbf{V}_s = \sigma_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1}$. The estimator of the residual variance is given by $\hat{\sigma}_s^2 = \frac{1}{n_s-2}(\mathbf{y}_s - \mathbf{X}_s\hat{\mathbf{b}}_s)'(\mathbf{y}_s - \mathbf{X}_s\hat{\mathbf{b}}_s)$.
To test whether the reference regression line and the new regression line are equal, we construct simultaneous confidence bands for

$$\mathbf{x}'\mathbf{b} - \mathbf{x}'\mathbf{b}_s, \ \forall\, \mathbf{x} \in \mathbb{S}.$$

The variance of the estimator of this difference is given by

$$Var\left(\mathbf{x}'\hat{\mathbf{b}} - \mathbf{x}'\hat{\mathbf{b}}_s\right) = \mathbf{x}' \cdot \left(\mathbf{V} + \sigma_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1}\right) \cdot \mathbf{x} = \mathbf{x}' \triangle_s \mathbf{x}.$$

An estimator of $\triangle_s$, denoted by $\hat{\triangle}_s$, is obtained by plugging $\hat{\sigma}_s^2$ in the above equation. Therefore, we construct the following set of simultaneous confidence bands

$$(\mathbf{x}'\mathbf{b} - \mathbf{x}'\mathbf{b}_s) \in \left(\mathbf{x}'\hat{\mathbf{b}} - \mathbf{x}'\hat{\mathbf{b}}_s\right) \pm c \cdot \sqrt{\mathbf{x}'\hat{\triangle}_s\mathbf{x}}, \quad \forall \mathbf{x} \in \mathbb{S}.$$

The critical value c is determined so that the confidence level of the simultaneous confidence band is equal to $1 - \alpha$. The critical value c can be found if the distribution of T_{ms} would be known, with T_{ms} given by

$$T_{ms} = \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid \mathbf{x}' \cdot \left[\left(\hat{\mathbf{b}} - \mathbf{b}\right) - \left(\hat{\mathbf{b}}_s - \mathbf{b}_s\right)\right] \mid}{\sqrt{\mathbf{x}'\hat{\triangle}_s\mathbf{x}}}. \tag{8.2.1}$$

Now the distribution of T_{ms} is obtained via simulation, too. We apply another representation of T_{ms}, from which the derivation of a simulation algorithm becomes easy. The same ideas as in the last section are applied, with a few adjustments to account for the different variance structure of the difference estimator.
For each fixed $\sigma_s^2 > 0$, there exists a Cholesky decomposition of $\mathbf{V} + \sigma_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1}$, i.e. there exists a nonsingular matrix $\mathbf{P} \in \mathbb{R}^{p\times p}$ such that

$$\mathbf{V} + \sigma_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1} = \mathbf{P}'\mathbf{P}.$$

Let $\mathbf{Z} \in \mathbb{R}^{p\times 1}$ be a normal random vector with $\mathbf{Z} \sim N(0, \mathbf{V})$ and $\mathbf{Z}_s \in \mathbb{R}^{p\times 1}$ be a normal random vector, independent of $\mathbf{Z}$ with $\mathbf{Z}_s \sim N(0, \sigma_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1})$.
Define $\mathbf{Z}_{ms} = (\mathbf{P}')^{-1} \cdot (\mathbf{Z} - \mathbf{Z}_s)$. Then the distribution of T_{ms} is the same as the distribution of $Q_{ms} \cdot \parallel \mathbf{Z}_{ms} \parallel$, where

$$Q_{ms} = \sup_{\mathbf{x}\in\mathbb{S}} \frac{\mid (\mathbf{P}\mathbf{x})'\mathbf{Z}_{ms} \mid}{\parallel \mathbf{P}\mathbf{x} \parallel \cdot \parallel \mathbf{Z}_{ms} \parallel}.$$

The calculation of Q_{ms} in case of $p = 2$ is based on the same considerations as the calculation of Q_{ij}. But now the cone spanned by the column vectors of the matrix $\mathbf{P}$ is regarded.
Based on these considerations, the distribution of T_{ms} can be simulated by the following algorithm:

1. Simulate $\hat{\sigma}_s^2 \sim \chi_{n-2}^2$, where χ_{n-2}^2 denotes a Chi-square distribution with $n-2$ degrees of freedom and calculate $\hat{\sigma}_s^2 (\mathbf{X}_s'\mathbf{X}_s)^{-1}$.

2. Calculate the Cholesky decomposition of $\mathbf{V} + \hat{\sigma}_s^2 \cdot (\mathbf{X}_s'\mathbf{X}_s)^{-1} = \mathbf{P}'\mathbf{P}$.

3. Simulate independent random vectors

$$\begin{aligned} \mathbf{Z} &\sim N(0, \mathbf{V}) \\ \mathbf{Z}_s &\sim N(0, \hat{\sigma}_s^2 \cdot (\mathbf{X}_s' \mathbf{X}_s)^{-1}). \end{aligned}$$

4. Calculate $\mathbf{Z}_{ms} = (\mathbf{P}')^{-1} \cdot (\mathbf{Z} - \mathbf{Z}_s)$.

5. Find Q_{ms}.

6. Compute $Q_{ms} \cdot \| \mathbf{Z}_{ms} \|$.

The main difference between the two simulation algorithms is that for the first one only one Cholesky decomposition is needed for each variance-covariance matrix of differences, whereas in the second algorithm in each simulation step the Cholesky decomposition of the variance-covariance matrix of the differences has to be computed. This makes the first algorithm faster, however in regression models with a low number of parameters the differences in computation time are negligible.
Steps $1 - 6$ are repeated R time and the estimator of the critical c, is the $(1 - \alpha)R$th largest simulated value $\hat{c}$. [LJZ04] examined the standard error of $\hat{c}$ dependent on the number of simulation steps. They found that after 200.000 steps one obtains an accurate estimate of c.
To test whether the two regression lines are the same, the maximum lower limit of the comparisons LL_{ms} at each point $\mathbf{x} \in \mathbb{S}$, as well as the minimal upper limit UL_{ms} of these comparisons are calculated. The reference regression line and the new regression line are considered equal in the region $\mathbb{S}$, if the confidence band includes 0 over the whole region, being equivalent to $LL_{ms} < 0 < UL_{ms}$.
If more than one regression line needs to be compared with the reference regression line, the confidence limit α should be adjusted to the number of these comparisons, say k. This can be done for example using the Bonferroni method by setting $\alpha_{adj} = 1 - (1 - \alpha)^{1/k}$.

8.3 Application to examples

For both presented algorithms, there are applications in the field of standardization of diagnostic assays. As an example for the first one, we discuss how to compare method comparison studies that are repeated after fixed time intervals. For the second algorithm we present a problem from reagent-lot comparability studies.

8.3.1 Annual comparison of standardization networks

The scope of the IFCC network for standardization of HbA1c is the development of a reference material for HbA1c testing, such that all HbA1c assays worldwide available are standardized to this material. However, up to now, there exist national standardization networks with measurement methods that differ from the IFCC method. This implies that reported HbA1c values are not comparable among the networks. For example, a HbA1c value of 3.5% based on IFCC standardization is about 5.35% according to the US standardization network, called NGSP* [HWJ+04]. Recognizing that HbA1c changes of 1% will cause patients to change their treatment, these differences are not acceptable in a clinical sense. Therefore, the relationship between these values needs to be established, to transform values from other networks to IFCC values and vice versa. Twice a year, method comparison studies are launched between the IFCC network and networks in the US, Japan, Sweden and Australia, the so-called "Designated Reference Methods" (DCMs). Based on the assumption that the differences between these methods are systematic and/or proportional, a linear relationship is assumed. Hence, a regression line between each national standardization network and the IFCC network is determined twice a year. To check the stability of the relationships, these regression lines have to be compared.

As an example we regard the comparison of the Barcelona 2 regression line between the IFCC network and the NGSP network with all available studies up to this point, as well as the regression line between the IFCC network and the Australian network. In Table 8.1, the estimated parameters of the regression lines between these two networks and the IFCC network are given.

As the coefficient of multiple determination R^2 ([DS98]) is near to one for all regression lines, the assumptions of linear relationship holds for the IFCC - NGSP and the IFCC - Australian relationship. Note that the slope and the intercept vary much more from study to study for the IFCC - Australian relationship than for the IFCC - NGSP relationship.

To compare the Barcelona 2 regression line with the other regression lines, we apply the algorithm of Section 8.1. There are $k = 5$ comparisons, the set Λ is $\Lambda = \{(6, j) \mid j = 1, .., 5\}$. The comparison should be based on the measurement range of the HbA1c assays, which goes from 0% up to 20%. The derivation of the critical value is based on 200.000 simulations. A grid search over the measurement range is performed for the estimation of LL_{ij} and UL_{ij}.

In Table 8.2, the minimal upper limit as well as the maximal lower limit for each com-

*National Glycohemoglobin Standardization Program

parison are given. For the IFCC - NGSP relationship, all differences of the predicted values stay within the simultaneous confidence bands, hence there are no differences between the regression lines of the Barcelona 2 study with the other studies.
Regarding the IFCC - Australian relationship, the Barcelona 2 regression line differs from the regression lines of the Kyoto 2, Chicago and Marrakech study. In these cases the minimum upper limit is smaller than zero. Note especially UL_{ij} of the comparison with the Marrakech study.
The different behaviour of the two relationships is explainable: The NGSP network consists of 10 laboratories whereas the Australian network only of one laboratory. Differences from study to study of the single laboratory are directly reflected in the mean of the measured values, whereas this kind of differences are mostly averaged in the NGSP network.
We should keep this in mind, as in the next chapter we deal with the question how multiple regression lines should be combined to an average regression line.

Table 8.1: **Estimated slope, intercept and R^2 for the IFCC - NGSP and IFCC - Australia relationship, for the studies Marrakech to Barcelona 2.**

Relationship	Study	Intercept	Slope	R^2
IFCC - NGSP	Barcelona 2	2.21	0.90	0.9999
	Barcelona 1	2.23	0.91	0.9999
	Kyoto 2	2.18	0.91	0.9998
	Kyoto 1	2.22	0.91	0.9993
	Chicago	2.04	0.93	0.9991
	Marrakech	2.14	0.92	0.9989
IFCC - Australia	Barcelona 2	2.3	0.89	0.9995
	Barcelona 1	2.07	0.94	0.9977
	Kyoto 2	2.6	0.89	0.9992
	Kyoto 1	1.94	0.95	0.9993
	Chicago	1.79	1.0	0.9968
	Marrakech	1.79	1.01	0.9989

Table 8.2: **Comparison statistics of the IFCC - NGSP and IFCC - Australian relationship for the comparison of the Barcelona 2 study with previous studies, within the range from 0% HbA1c to 20 % HbA1c.**

		NGSP		Australia	
Study(i)	Study(j)	LL_{ij}	UL_{ij}	LL_{ij}	UL_{ij}
Barcelona 2	Barcelona 1	-.222	0.094	-.303	0.136
Barcelona 2	Kyoto 2	-.193	0.102	-.549	-.073
Barcelona 2	Kyoto 1	-.215	0.097	-.192	0.234
Barcelona 2	Chicago	-.130	0.120	-.216	-.086
Barcelona 2	Marrakech	-.219	0.033	-.183	-.238

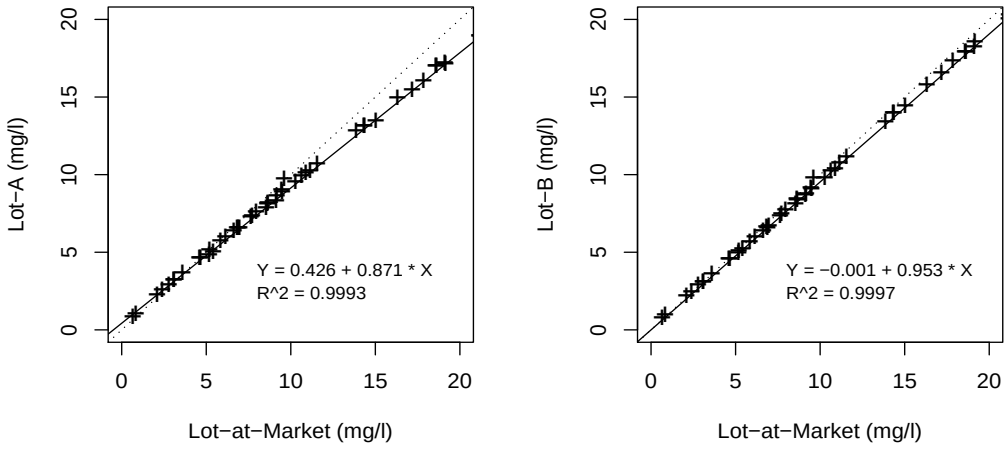
8.3.2 Reagent-lot comparability

As an example for the application of the second algorithm, we consider the approval process of new reagent-lots for a diagnostic assay. Errors in the production process of reagent-lots may cause a shift in the values of the measured samples, hence new reagent-lots need to be compared with the reagent-lots at market.
The approval experiment for a new reagent-lot consists in a method comparison study, where 50 samples are measured with the reagent-lot at market and with the new one. At the moment the approval rule is based on a global criteria of intercept and slope. For example, if the intercept of the fitted regression line falls within the range of 0 ± 0.3 and the slope within 1 ± 0.1, the reagent-lot will be approved. If either the intercept or the slope lies outside these limits the reagent-lot will not be approved. In our example the measurement range of the assay is 0 mg/l – 20 mg/l.
In Figure 8.2, the data of two comparisons together with the fitted regression lines are shown. Note that for reagent-lot A, the global approval rule would cause the reagent lot not to be approved. However, in the region around 5 mg/l the regression line intersects the bisection line, so that at least in this region the new lot could be used interchangeable to the lot-at-market. To answer, whether this holds for the whole measurement range of the assay, the algorithm of Section 8.2 is used.
The regression line drawn in Figure 8.2 is derived by simple least-squares regression. However, as the values of both axes are measured by the same methods, both axes are subject to error and a orthogonal regression method would be more appropriate. But as the coefficient of determination (R^2) is in both cases very close to 1, both regression method result approximately in the same estimates [CVN99]. Therefore, the error made by the usage of the less appropriate least-squares regression method is negligible. It rests for further research to accommodate the algorithms of Section 8.1 and 8.2 to the orthogonal regression case.
To apply the algorithm of Section 8.2 to the presented problem, we translate the acceptance rules of this experiment in the following way: The best estimate of the reference regression line is $\hat{\mathbf{b}} = (0, 1)'$, with

$$Var(\hat{\mathbf{b}}) = \begin{pmatrix} 0.15^2 & 0 \\ 0 & 0.05^2 \end{pmatrix}.$$

The region of interest for differences between the new reagent-lots and the lot-at-market is the measurement range 0 mg/l – 20 mg/l. We are interested in the construction of 95% simultaneous confidence bands for the differences of predicted values within this range. To obtain the critical value c, we repeat the simulation of the random variable T, 200.000

Figure 8.2: **Plot of the method comparison data for two new reagent-lots versus the lot-at-market.**



times. The minimum upper and maximum lower differences are obtained by a grid search in the interval, at 10.000 equally distant points.
In Table 8.3, the intercept and the slope of both regression lines are shown, as well as the critical value c, the maximum lower difference LL_{ms} and minimum upper difference UL_{ms} over the specified range. For reagent-lot A the minimum upper limit is negative,

Table 8.3: **Intercept, slope and comparison statistics for the comparisons between the new reagent-lots A and B with the lot-at-market, within the range** 0 **mg/l –** 20 **mg/l**, **with** $\alpha = 0.95$.

Reagent - Lot	Intercept	Slope	T	LL_{ms}	UL_{ms}
Lot A	0.426	0.871	2.33	-0.21	-0.06
Lot B	-0.001	0.953	2.33	-0.33	0.36

hence this regression line can not be considered equal to the reference regression line over the region of interest. In Figure 8.3, the differences over the measurement range

are shown together with the simultaneous confidence band for the differences. There we note that the differences for reagent-lot A fall outside the confidence band at very low concentration levels. The differences for reagent-lot B comparison are within the confidence bands over the whole measurement range.

The presented approach for the approval of reagent-lots has some considerable advantages over the global approval rule, which is only based on the comparison of the coefficients of the new regression line with the specified limits.

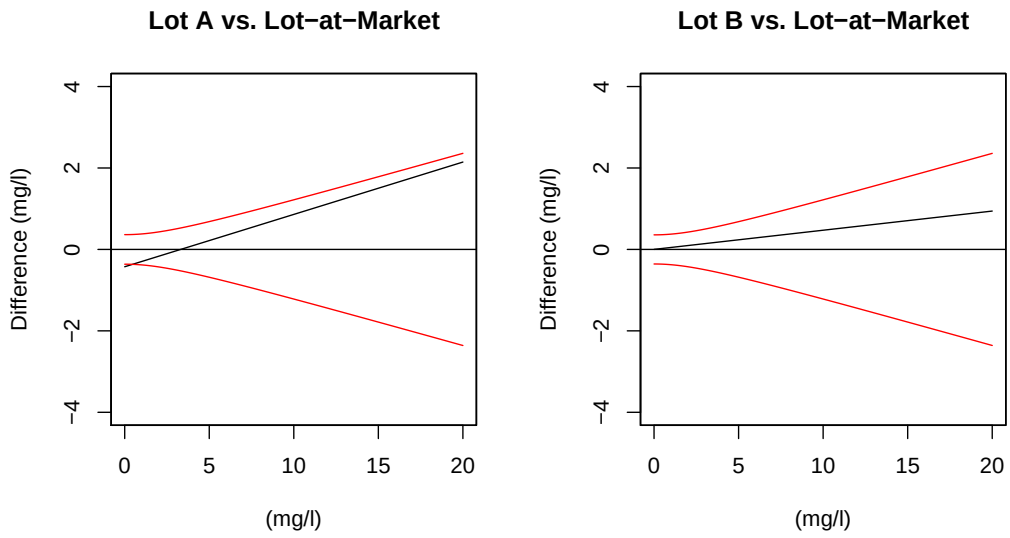
The most obvious advantage is that the approach focuses on the region of interest, which is in most cases the measurement range of the assay.

The second one is that the two-dimensional problem of comparing intercept and slope is reduced to a one-dimensional one, as the difference between predicted values is regarded. Hence it would be easy to incorporate also the correlation structure between the intercept and the slope in the specifications.

Further on, in addition to the variation of the specifications, the variation of the new regression line is taken into account, by this enhanced method.

However, it rests for further research to accommodate this approach to errors-in-variables problems, e.g. to adopt it to orthogonal regression.

Figure 8.3: **Plot of the differences for reagent-lot A and B versus the lot-at-market. The simultaneous confidence bands** ($\alpha = 0.95$) **are printed in red.**



Chapter 9

Meta-analysis of regression lines

In this chapter we consider the problem of the combination of several regression lines, being repetitions of the same method comparison experiment, to obtain an average regression line. This problem appears usually, if the method comparison experiment is repeated after a certain time. For instance the IFCC network for standardization of HbA1c needs to compare its values with other national standardization networks for HbA1c. So twice a year a method comparison study is launched to derive the relationship between the IFCC-HbA1c values and the HbA1c values of another standardization network. In each method comparison study another set of samples is used. After a certain number of studies an average regression line should be derived, which describes the transformation rule of IFCC-HbA1c values to the other HbA1c values and vice versa.
Hierarchical linear models are appropriate for this situation, but as both methods are subject to error, these models need to be enhanced to incorporate the measurement errors in both axes. We consider now the situation, that there are replicates of each measurement in both methods. This is different to the errors-in-variables model regarded in Chapter 3.1. The goal of the analysis is the derivation of the averaged regression line and its uncertainty.
In the first section, we repeat the main steps of fitting a hierarchical linear model in a Bayesian approach and derive the main Gibbs sampling algorithm. Afterwards, we extend this model and the algorithm to incorporate errors in both axes. For choosing the appropriate prior distribution, we discuss an approach based on posterior predictive checks. We finish with the analysis of the IFCC - Sweden relationship.

9.1 Hierarchical linear models

From a Bayesian point of view, linear mixed models are named hierarchical linear models, as in Bayesian analysis no distinction is made between fixed and random effects - all parameters of a model are regarded as random variables. The difference between these two effects is that a hyper-distribution is assigned to the so-called random effects. This introduces a further hierarchy in the model. Hence, the linear mixed model (5.1.1) may be written in terms of a hierarchical model as

$$\begin{aligned} Y_{ij}|x_{ij}, \mathbf{b}, \beta_i, \sigma_\varepsilon^2 &\sim N(b_0 + \beta_{0i} + b_1 x_{ij} + \beta_{1i} x_{ij}, \sigma_\varepsilon^2),\ i = 1, .., I,\ j = 1, .., J_i, \quad (9.1.1) \\ \beta_i|\mathbf{D} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D}\right),\ i = 1, .., I. \end{aligned}$$

As no closed forms of the posterior distributions of the parameters are available, their derivation is based on Markov Chain Monte Carlo techniques. If the prior distributions of the parameters are chosen appropriately a Gibbs sampling algorithm (see Section 2.1.2) can be set up. Gibbs samplers are most efficient when the parametrization is done in terms of independent components. If highly dependent components are used the convergence to the posteriors is slow. Centering and reparametrization can be used to improve the converge of the Gibbs sampler.
In linear regression models, the estimators of intercept and slope are correlated as long as the mean of the predictor values is not zero. This applies also to hierarchical linear models. Hence, centering the predictor values around their mean breaks the correlation between the coefficients; see e.g. [ZGF02] and [GSC95].
The parameters of Model (9.1.1) are $(\mathbf{b}, \sigma_\varepsilon^2, \mathbf{D})$, for which prior distributions need to be specified. Priors of parameters are specified either based on historical data, or non-informative priors are assigned.
Most authors working on hierarchical linear models specify conjugate prior distributions, to achieve proper full conditional distributions, such that the Gibbs sampler is easily applicable; see for example [LLV04], [ZGF02], [Car96]. The parameters of the distributions are chosen such that these prior distributions are almost non-informative. Nevertheless, the sensitivity of the results on the prior settings must be analyzed. Especially in cases of small sample sizes it is difficult to assess the vagueness of such a prior; see for example [LSB$^+$05] for a sensitivity analysis of different prior settings in hierarchical linear models. We will discuss this issue further in Section 9.3.

We will work with conjugate prior distributions for hierarchical linear models, given by

$$\begin{aligned} p(\sigma_\varepsilon^2) &= InvGamma(u_y, v_y), \\ p\left(\left(\begin{array}{c} b_0 \\ b_1 \end{array}\right)\right) &= N_2(\mathbf{M}, \mathbf{\Phi}), \\ p(\mathbf{D}) &= InvWhishart(\mathbf{\Omega}, \lambda). \end{aligned} \tag{9.1.2}$$

To derive the Gibbs sampling algorithm, the full conditional posterior distributions of the parameters need to be derived. In the following we will denote the conditional posterior distribution of a parameter Θ, by $p(\Theta|\cdot)$. That is the distribution of Θ given all the other parameters and the data. For Model (9.1.1) they can be derived by considering the following proportional relationships (see [Gil96]):

$$\begin{aligned} p(\sigma_\varepsilon^2|\cdot) &\propto p(\mathbf{Y}|\cdot) \cdot p(\sigma_\varepsilon^2), \\ p\left(\left(\begin{array}{c} \beta_{0i} \\ \beta_{1i} \end{array}\right)|\cdot\right) &\propto p(\mathbf{Y}_i|\cdot) \cdot p\left(\left(\begin{array}{c} \beta_{0i} \\ \beta_{1i} \end{array}\right)|\mathbf{D}\right), \ \forall\, i = 1, .., I, \\ p\left(\left(\begin{array}{c} b_0 \\ b_1 \end{array}\right)|\cdot\right) &\propto p(\mathbf{Y}|\cdot) \cdot p\left(\left(\begin{array}{c} b_0 \\ b_1 \end{array}\right)\right), \\ p(\mathbf{D}|\cdot) &\propto p\left(\left(\begin{array}{c} \beta_{0i} \\ \beta_{1i} \end{array}\right)|\mathbf{D}\right) \cdot p(\mathbf{D}). \end{aligned} \tag{9.1.3}$$

According to Model (9.1.1) the likelihood functions are given by

$$\begin{aligned} p(\mathbf{Y}_i|\cdot) &= (2p)^{-J_i/2} \cdot |\mathbf{\Psi}_i|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\beta_i)'\mathbf{\Psi}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\beta_i)\right), \\ p(\mathbf{Y}|\cdot) &= \prod_{i=1}^{I} p(\mathbf{Y}_i|\cdot), \end{aligned}$$

where $\mathbf{\Psi}_i = \sigma_\varepsilon^2 \cdot \mathbf{I}_{J_i}$ and $\mathbf{X}_i = \mathbf{Z}_i$, defined as

$$\mathbf{X}_i = \left(\begin{array}{cc} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{iJ_i} \end{array}\right).$$

Based on the prior settings and the proportional relationships given in (9.1.3), the conditional posteriors can be deduced, leading to the following Gibbs sampling algorithm:

Step 1 Generate initial values,

$$\sigma^2_{\varepsilon^0},\ \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}^0,\ \forall i = 1,..,I, \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}^0, \mathbf{D}^0.$$

Step 2 Let $t > 0$. Given

$$\sigma^2_{\varepsilon^t},\ \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}^t,\ \forall i = 1,..,I, \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}^t, \mathbf{D}^t,$$

draw:

(a)

$$\sigma^2_{\varepsilon^{t+1}} \sim InvGamma\left(u_y + \frac{1}{2}n, \frac{1}{2}\sum_i \sum_j (y_{ij} - E.y^t_{ij})^2 + v_y\right), \quad (9.1.4)$$

where $E.y^t_{ij} = b^t_0 + \beta^t_{0i} + b^t_1 x_{ij} + \beta^t_{1i} x_{ij}$ and $n = \sum_i J_i$.

(b) $\forall\ i = 1,..,I$

$$\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}^{t+1} \sim N_2\left((\mathbf{\Lambda}^{t+1}_i)^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}^t)(\mathbf{\Psi}^{t+1}_i)^{-1}\mathbf{Z}_i, (\mathbf{\Lambda}^{t+1}_i)^{-1}\right), \quad (9.1.5)$$

where $\mathbf{\Lambda}^{t+1}_i = \mathbf{Z}'_i(\mathbf{\Psi}^{t+1}_i)^{-1}\mathbf{Z}_i + (\mathbf{D}^t)^{-1}$.

(c)

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}^{t+1} \sim N_2\left((\mathbf{\Lambda}^{t+1})^{-1} \cdot \left(\sum_i ((\mathbf{Y}_i - \mathbf{Z}_i\beta^{t+1}_i)'(\mathbf{\Psi}^{t+1}_i)^{-1}\mathbf{X}_i) + \mathbf{\Phi}^{-1}\mathbf{M}\right), (\mathbf{\Lambda}^{t+1})^{-1}\right), \quad (9.1.6)$$

where $\mathbf{\Lambda}^{t+1} = \sum_i \mathbf{X}'_i(\mathbf{\Psi}^{t+1}_i)^{-1}\mathbf{X}_i + \mathbf{\Phi}^{-1}$.

(d)

$$\mathbf{D}^{t+1} \sim InvWishart\left(\lambda + I, \mathbf{\Omega} + \sum_i \beta'^{t+1}_i \beta^{t+1}_i\right). \quad (9.1.7)$$

This conditional posterior distribution is obtained, due to the property of the

trace of a matrix, given in (A.0.3).

Step 3 Repeat Step 2 R times until the chains have converged to the posterior distribution. Criteria for assessing convergence are discussed in Section 2.2.

The OpenBUGS2.2.0 software [STBL05] provides an easy interface for the specification of hierarchical linear models and the use of the Gibbs sampler. This software in combination with the R package **BRugs** [Lig06] provides a comfortable way for the derivation of the posterior distribution of the parameters.

9.2 Hierarchical linear models with errors in both axes

As already stated in the introduction of this chapter, both methods are subject to error in method comparison studies. So these errors need to be incorporated in the hierarchical model. In contrast to the model in Chapter 3, multiple measurements of each sample are available for each method.
Let $X_{ijk},\ i = 1,..,I, j = 1,..,J_i, k = 1,..,K_{ij}$ be the kth measurement of sample j and experiment i of the first method and let $Y_{ijl},\ i = 1,..,I, j = 1,..,J_i, l = 1,..,L_{ij}$ be the lth measurement of sample j in experiment i of the second method. Expanding Model (9.1.1) to errors in both axes leads to

$$\begin{aligned} X_{ijk} &= \eta_{ij} + \delta_{ijk},\ i = 1,..,I, j = 1,..,J_i, k = 1,..,K_{ij}, \\ Y_{ijl} &= \xi_{ij} + \varepsilon_{ijl},\ i = 1,..,I, j = 1,..,J_i, l = 1,..,L_{ij}, \\ \xi_{ij} &= b_0 + \beta_{0i} + b_1 \cdot \eta_{ij} + \beta_{1i} \cdot \eta_{ij},\ i = 1,..,I, j = 1,..,n_i, \end{aligned} \tag{9.2.1}$$

where the measurement errors δ_{ijk} and ε_{ijk} are independent normally distributed with $N(0, \sigma^2_{\delta_{ij}})$ and $N(0, \sigma^2_{\varepsilon_{ij}})$ respectively.
Just as for Model (9.1.1) we assume that the intercept-slope vectors of the experiments are random draws from a multivariate normal distribution:

$$\beta_i \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D}\right),\ i = 1,..,I, \tag{9.2.2}$$

independent of δ_{ijk} and ε_{ijl}.
In terms of a hierarchical linear model (9.2.1) and (9.2.2) may be written as

$$\begin{aligned}
\mathbf{X}_{ij}|\eta_{ij}, \sigma^2_{\delta_{ij}} &\sim N_{K_{ij}}(\eta_{ij}\mathbf{1}_{K_{ij}}, \sigma^2_{\delta_{ij}}\mathbf{I}_{K_{ij}}), \\
\mathbf{Y}_{ij}|\xi_{ij}, \sigma^2_{\varepsilon_{ij}} &\sim N_{L_{ij}}(\xi_{ij}\mathbf{1}_{L_{ij}}, \sigma^2_{\varepsilon_{ij}}\mathbf{I}_{L_{ij}}), \\
\xi_i &= \begin{pmatrix} \mathbf{1}_{J_i} & \eta_i \end{pmatrix} \cdot \mathbf{b} + \begin{pmatrix} \mathbf{1}_{J_i} & \eta_i \end{pmatrix} \cdot \beta_i, \\
\beta_i|\mathbf{D} &\sim N_2((0,0)', \mathbf{D}),
\end{aligned} \tag{9.2.3}$$

with prior distributions on the parameters considered below.
Additional to the parameters of the simple hierarchical model, we have the parameters η_{ij} - the true values of the samples in the X-method and ξ_{ij} - the true values of the samples in the Y-method and the respective measurement error variances $\sigma^2_{\delta_{ij}}, \sigma^2_{\varepsilon_{ij}}$. Hence, the Gibbs sampling algorithm expands to drawing from

$$\begin{aligned}
p(\eta_{ij}|\cdot) &\propto p(\mathbf{X}_{ij}|\eta_{ij}, \sigma^2_{\delta_{ij}}) \cdot p(\eta_{ij}), \ \forall i = 1,..,I, \ j = 1,..,J_i, \\
p(\sigma^2_{\delta_{ij}}|\cdot) &\propto p(\mathbf{X}_{ij}|\eta_{ij}, \sigma^2_{\delta_{ij}}) \cdot p(\sigma^2_{\delta_{ij}}), \ \forall i = 1,..,I, \ j = 1,..,J_i, \\
p(\sigma^2_{\varepsilon_{ij}}|\cdot) &\propto p(\mathbf{Y}_{ij}|\cdot) \cdot p(\sigma^2_{\varepsilon_{ij}}), \ \forall i = 1,..,I, \ j = 1,..,J_i, \\
p\left(\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}|\cdot\right) &\propto p(\mathbf{Y}_i|\cdot) \cdot p\left(\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}|\mathbf{D}\right), \ \forall \ i = 1,..,I, \\
p\left(\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}|\cdot\right) &\propto p(\mathbf{Y}|\cdot) \cdot p\left(\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}\right), \\
p(\mathbf{D}|\cdot) &\propto p\left(\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}|\mathbf{D}\right) \cdot p(\mathbf{D}).
\end{aligned}$$

The conjugate prior distributions of the parameters are given by

$$\begin{aligned}
p(\sigma^2_{\delta_{ij}}) &= InvGamma(u_x, v_x), \ \forall i = 1,..,I, \ j = 1,..,J_i \\
p(\sigma^2_{\varepsilon_{ij}}) &= InvGamma(u_y, v_y), \ \forall i = 1,..,I, \ j = 1,..,J_i \\
p\left(\begin{pmatrix} b_0 \\ b_1 \end{pmatrix}\right) &= N_2(\mathbf{M}, \mathbf{\Phi}) \\
p(\mathbf{D}) &= InvWishart(\mathbf{\Omega}, \lambda).
\end{aligned} \tag{9.2.4}$$

For η_{ij} we specify a uniform prior, for example uniform on the measurement range of the assay. Based on these prior settings and the proportional relationships given above, the conditional posteriors can be deduced, leading to a Gibbs sampling algorithm similar to

the one given in Section 9.1. However, this time $\eta_{ij}, \sigma^2_{\delta_{ij}}$ are drawn, too, and the design and variance-covariance matrices are adapted to the situation of repeated measurements and heteroscedastic variances.

9.3 Model checking

For the analysis of hierarchical linear models, different authors pointed out that with the specification of proper prior distributions, especially the inverse Gamma or inverse Whishart distribution, it is hard to assess non-informativeness (see e.g. [LSB+05], [NM98], [Gel06].)
Hence, it is necessary to analyze the sensitivity of the results to the specified prior distributions and to choose the most appropriate one in cases where inferences are different. We will do this by posterior predictive checks, as already described in Section 2.3.
In hierarchical linear models posterior predictive checks can be made on different levels of the model. On the first level, the data of the observed experiments can be compared to the posterior predictive distribution given the observed experiments. On the second level, one can derive the posterior predictive distribution for a new experiment, and compare it with an experiment which will be observed in the future.
The posterior predictive checks on the first level are adequate for choosing the best prior distribution among different priors or to assess the best model, given the data at hand. Posterior predictive checks of the second level may be used to assess whether a new experiment still fits to the same conditions as the other experiments. For experiments that are repeated over time, it might be that the experimental conditions change unperceived. To check if these changes cause a change in the parameters of the regression model, the posterior predictive checks of the second level may be used. As test quantities we compare the study-wise coefficients of the observed studies to the study-wise coefficients based on the replicated data.

Let us regard this in more detail by first considering Model (9.1.1).
The algorithm for the posterior predictive distribution given the observed experiments is the following:

Step 1 For each experiment $i = 1, .., I$ and each simulated value of the posterior distribution $n = 1, .., N$ calculate

$$\begin{aligned} E.Y_{ij}^{rn} &= b_0^n + \beta_{0i}^n + b_1^n \cdot x_{ij} + \beta_{1i}^n \cdot x_{ij}. \\ Y_{ij}^{rn} &\sim N(E.Y_{ij}^{rn}, \sigma^2_{\varepsilon^n}). \end{aligned}$$

Step 2 For each simulated value $n = 1, .., N$, calculate the study-wise coefficients β_i^n based on $Y_{ij}^{rn}\ \forall\ i = 1, .., I, j = 1, .., J_i$ and the design-matrix $\mathbf{X}_i$ of experiment i based on least-squares regression.

Step 3 Calculate for the intercept and the slope the p-values $p_i(\beta_{0i})$ and $p_i(\beta_{1i})$ as the proportion of the N calculated intercepts and slopes, respectively, which are smaller than $\hat{\beta}_{0i}, \hat{\beta}_{1i}$, which are the intercept and slope derived via least-squares regression from the data of experiment i.

Step 4 Calculate for each experiment i, the mean squares error, given by

$$MSE_i = \frac{1}{n}\sum_n ||\beta_i^n - \hat{\beta}_i||^2.$$

Having calculated the p-values and MSE of different models and/or prior settings, the model with still acceptable p-values and low MSE may be chosen as the most appropriate one. In the example at the end of this chapter we will discuss this further.

Having data from a new experiment, we can check how well the chosen model predicts the outcome of the new experiment, or regarding things the other way around, how well the new experiment fits to the previous data situation.
Denoting with $\mathbf{X}_{i'}$ the design matrix of the new experiment, the algorithm for posterior predictive checks is

Step 1 For each simulated value $n = 1, .., N$, calculate

$$\begin{aligned}
\begin{pmatrix} \beta_{0i'}^{rn} \\ \beta_{1i'}^{rn} \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D}^n\right) \\
E.Y_{i'j}^{rn} &= b_0^n + \beta_{0i'}^{rn} + b_1^n \cdot x_{i'j} + \beta_{1i'}^{rn} \cdot x_{i'j} \\
Y_{i'j}^{rn} &\sim N(E.Y_{i'j}^{rn}, \sigma^2_{\varepsilon^n}),\ \forall\ j = 1, .., J_{i'}.
\end{aligned}$$

Step 2 For each simulated value the study-wise coefficients are calculated based on $Y_{i'j}^{rn}$ and the design matrix of the new experiment.

Step 3 p-values and MSE are obtained by the comparison of the calculated coefficients based on the replicated data and the coefficients derived from the observed data of the new experiment.

If the p-values of intercept or slope are further away from 0.5 than a certain level α, $(0 < \alpha < 0.5)$ we may conclude that the experimental conditions have changed, since

the actual model can only poorly explain the data from the new experiment.

Regarding the hierarchical linear model with errors in both axes, the simulation algorithm for the posterior predictive checks is extended by drawing the repeated measurements of the first method, too. Thus Step 1 is extended to

Step 1 For each experiment $i = 1, .., I$ and each simulated value $n = 1, .., N$, calculate

$$\begin{aligned} X_{ijk}^{rn} &\sim N(\eta_{ij}^n, \sigma_{\delta_{ij}}^{2n}),\ i = 1, .., I, j = 1, .., J_i, k = i, .., K_{ij} \\ \xi_{ij}^{rn} &= b_0^n + \beta_{0i}^{rn} + b_1^n \cdot \eta_{ij}^n + \beta_{1i}^{rn} \cdot \eta_{ij}^n\ i = 1, .., I, j = 1, .., J_i \\ Y_{ijl}^{rn} &\sim N(\xi_{ij}^{rn}, \sigma_{\varepsilon_{ij}}^{2n}). \end{aligned}$$

P-values can be calculated for the means of the samples of the X-axis, as well as for the study-wise coefficients.
In the next section, we present the application of these ideas to the derivation of the average IFCC - Sweden relationship.

9.4 IFCC - Sweden relationship

We return to the method comparison studies of the IFCC network for standardization of HbA1c, which we introduced already in Section 8.3.1. This time we regard the IFCC - Sweden relationship in more detail, for which 10 comparisons are performed up to now. Based on these data an average regression line, the so-called master equation shall be derived, such that in the future HbA1c values based on the Swedish standardization can be transformed into IFCC values and vice versa.
We will proceed in three steps to derive the master equation appropriately. In a first step, we pool the data from all studies and apply a simple regression model to it. [HWJ+04] already derived a master equation based on the first four studies by pooling all the data. At that time this was perhaps the best way to do, due to the small number of studies. We will show that now, having more studies, a hierarchical model is more appropriate. It explains better the variation within the data and provides therefore a better estimate for the relationship and its uncertainty.
In a second step, we turn to the hierarchical linear model, by ignoring the measurement-error in the IFCC values. Instead of regarding the individual measurements, we fit the model to the means of the Sweden and IFCC samples. This model is faster to fit and we can already derive some settings for the prior distributions. We use the Gibbs sampling algorithm given in Section 9.1.

In the third step, the model based on the individual observations is fitted by the Gibbs sampling algorithm of Section 9.2.
As the first two studies had a different design from the other ones, they are not included in the derivation of the master equation. However, we can use them for the comparison of the prediction of observations from new studies.
In each study 5 samples, distributed over the measurement range, are measured within the IFCC and Swedish standardization network. As the IFCC network for standardization of HbA1c consists of up to 10 laboratories, up to 40 repeated measurements for one sample are available. For the Swedish network there are up to four measurements per sample.
In Figure 9.1, a plot of the intercept/slope of the individual studies is shown. Studies indicated in red are used for the model fit, the black ones are those not included in the analysis.

9.4.1 Linear Bayesian model

By pooling the data over all studies, we obtain a set of 40 sample pairs $(x_i, y_i), i = 1, .., 40$, where x_i denotes the IFCC - HbA1c value of the ith sample and y_i the Sweden - HbA1c value, respectively. Both values are the means of the repeated measurements per sample. The likelihood function and the prior distributions of the linear Bayesian model are given by:

$$\begin{aligned} Y_i &= b_0 + b_1(X_i - \bar{X}) + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ p((b_0, b_1)') &= N((E.b_0, E.b_1)', \mathbf{\Phi}) \\ p(\sigma_\varepsilon^2) &= InvGamma(u, v). \end{aligned} \tag{9.4.1}$$

We analyze different prior distributions: For the prior distribution of the regression coefficients, we set $(E.b_0, E.b_1) = (5.54, 0.989)$, these are the coefficients of the master equation derived by [HWJ+04], recalculated to the centered model. To achieve that the normal distribution becomes non-informative, $\mathbf{\Phi}$ was set to 10^3 and 10^6, respectively.
Regarding the inverse gamma distribution, three different pairs for (u, v) are analyzed:

$$(0.0001, 0.0001),\ (0.1, 0.1),\ (2, 0.018).$$

The first pair is the "classical" set of parameters, it is used for example in almost all OpenBUGS2.2.0 examples ([STBL05]). The third one is derived by setting $E(\sigma_\varepsilon^2) =$

0.0064, the value obtained by [HWJ+04] and $Var(\sigma_\varepsilon^2) = 10^6$. The second pair is seen as an intermediate one.
The analysis is done with OpenBUGS2.2.0 [STBL05], running three MCMC chains, with starting points for σ_ε^2 being $1, 0.1, 0.01$ and for (b_0, b_1) $((0, 10), (10, 10), (10, 0))$. After a burn-in of 100.000 values, every 50th value is saved until 5000 values are obtained from each chain. Convergence of the three chains was assessed by regarding the potential scale reduction factor.

The different settings for $\mathbf{\Phi}$ have no influence on the posterior distribution of the parameters. On the other side, the different parameter settings of the inverse gamma prior lead to different posteriors of the coefficients and the error variance. In Table 9.1, the 2.5%, 50% and 97.5% quantiles of the posterior distributions of the parameters are given for these different settings. The posterior medians of the coefficients are the same

Table 9.1: **2.5%, 50% and 97.5% quantiles of the posterior distributions of the parameters for different parameters of the prior of σ_ε^2, based on the linear Bayesian model.**

Parameters	(u, v)	2.5% quantile	50% quantile	97.5% quantile
b_0	$(0.0001, 0.0001)$	0.915	1.008	1.099
	$(0.1, 0.1)$	0.897	1.008	1.120
	$(2, 0.018)$	0.915	1.007	1.102
b_1	$(0.0001, 0.0001)$	0.955	0.968	0.981
	$(0.1, 0.1)$	0.952	0.968	0.983
	$(2, 0.018)$	0.955	0.968	0.981
σ_ε^2	$(0.0001, 0.0001)$	0.090	0.111	0.142
	$(0.1, 0.1)$	0.108	0.133	0.171
	$(2, 0.0018)$	0.090	0.110	0.138

for all priors, but the range of the posterior distributions differ. The InvGamma(0.1,0.1) prior leads to a wider range of the posterior distribution of the coefficients. The posterior of σ_ε^2 based on the InvGamma(0.1,0.1) prior is shifted to the right, compared to the other two prior distributions.
The adequacy of the likelihood and the prior settings is analyzed by posterior predictive checks, according to the first algorithm presented in Section 9.3. In Table 9.2, the p-values and MSE obtained from the comparison of the study-wise coefficients for the Kyoto 1, Kyoto 2 and Orlando 1 study are shown. Regarding Figure 9.1, we see that intercept and slope of the Kyoto 1 study are away from the center of these parameters

averaged over all studies. For the Kyoto 2 study, the intercept is to the right of this center, whereas the slope is quite near to the average slope. Both parameters of the Orlando 1 study are near to the center. Regarding the p-values we see, that the simple

Table 9.2: **P-values and MSE from the posterior predictive checks for studies Kyoto 1, Kyoto 2, Orlando 1 for different parameters of the prior of σ_ε^2, based on the linear Bayesian model.**

Study	(u, v)	p-value Intercept	p-value Slope	MSE
Kyoto 1	$(0.0001, 0.0001)$	0.07	0.96	0.065
	$(0.1, 0.1)$	0.11	0.93	0.074
	$(2, 0.018)$	0.06	0.96	0.065
Kyoto 2	$(0.0001, 0.0001)$	0.79	0.52	0.049
	$(0.1, 0.1)$	0.74	0.52	0.061
	$(2, 0.018)$	0.78	0.53	0.046
Orlando 1	$(0.0001, 0.0001)$	0.43	0.59	0.021
	$(0.1, 0.1)$	0.44	0.58	0.030
	$(2, 0.018)$	0.43	0.59	0.021

probability model does not explain the variation of the observations in an adequate form. Especially for the Kyoto 1 study the p-values of intercept and slope are either too low or too near at 1. The second prior seems to better explain the variability, as all p-values are closer to 0.5 for this prior. However, this is due to the increased range of the posterior, seen by the increased MSE.
In summary, we can conclude that the linear Bayesian model is not adequate to explain the variation within the data. In the next section we extend this model to a hierarchical linear model.

9.4.2 Hierarchical linear model

Modelling the data under consideration as hierarchical linear model, according to (9.1.1), leads to $I = 8$ studies and $J_i = 5, i = 1, .., 8$ samples per study.
As observations Y_{ij} we take the Sweden - HbA1c values of the respective sample and X_{ij} the IFCC - HbA1c values.
The prior distribution of (b_0, b_1) is defined as $N((5.54, 0.989), 10^6 \cdot \mathbf{I}_2)$ and the prior of σ_ε^2 as InvGamma(2,0.018), according to the arguments for the linear Bayesian model.
Different settings for the inverse Whishart distribution are analyzed, with $\lambda = 3$ and $\mathbf{\Omega} = 10^{-2} \cdot \mathbf{I}, 5 \cdot 10^{-3} \cdot \mathbf{I}, 10^{-3} \cdot \mathbf{I}, 5 \cdot 10^{-4} \cdot \mathbf{I}, 10^{-4} \cdot \mathbf{I}$.

Setting $\lambda = 3$ lets the Whishart distribution have a still finite density. For $|\mathbf{\Omega}| \to 0$, Jeffrey's prior is obtained. However in this limiting case, the posterior would no longer be proper, see e.g. [STH01] and [Sta06] for a further discussion of this issue.
Three chains are run with starting values for $(b_0, b_1) = (0, 10), (10, 10), (10, 0)$. Convergence for all priors could be assessed, based on the potential scale reduction factor, after 600.000 simulations. After the burn-in, every 50th value is saved until 5000 values are obtained from each chain.
In Table 9.3, a summary of the posterior distribution of the mean coefficients for the different prior settings is given. The smaller the absolute value of $\mathbf{\Omega}$, the smaller be-

Table 9.3: **2.5%, 50% and 97.5% quantiles of the posterior distributions of** (b_0, b_1) **for different parameters of the prior of D, based on the hierarchical linear model.**

Parameters	$\mathbf{\Omega}$	2.5% quantile	50% quantile	97.5% quantile
b_0	$10^{-2} \cdot \mathbf{I}$	0.760	0.976	1.191
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	0.811	0.980	1.149
	$10^{-3} \cdot \mathbf{I}$	0.864	0.983	1.106
	$5 \cdot 10^{-4} \cdot \mathbf{I}$	0.877	0.980	1.100
	$10^{-4} \cdot \mathbf{I}$	0.897	0.994	1.096
b_1	$10^{-2} \cdot \mathbf{I}$	0.941	0.973	1.006
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	0.947	0.973	0.999
	$10^{-3} \cdot \mathbf{I}$	0.953	0.972	0.990
	$5 \cdot 10^{-4} \cdot \mathbf{I}$	0.955	0.972	0.988
	$10^{-4} \cdot \mathbf{I}$	0.955	0.970	0.985

comes the range of the posterior distribution of the coefficients. This is mostly due to the fact that the non-informativeness of the inverse Whishart distribution is not given by its flatness, but because of the lower probability of matrices with higher absolute value. For these prior settings we make posterior predictive checks for the observed studies. In Table 9.4, the p-values for the slope and the intercept as well as the MSE for the Kyoto 1, Kyoto 2 and Orlando 1 study are given. Based on these we discuss the choice of the most adequate prior distribution from the ones considered here.
The p-values for slope and intercept of all studies have improved compared to the linear Bayesian model. With decreasing absolute values of $\mathbf{\Omega}$ they become worse, however the first two priors give quite the same p-values. The minimum of the mean-square error is given by the first prior for the Kyoto 1 study, by the second for the Kyoto 2 study and by the third for the Orlando 1 study. Regarding all studies included in the analysis, we have two studies for which the minimum mean-square error is reached for the first prior,

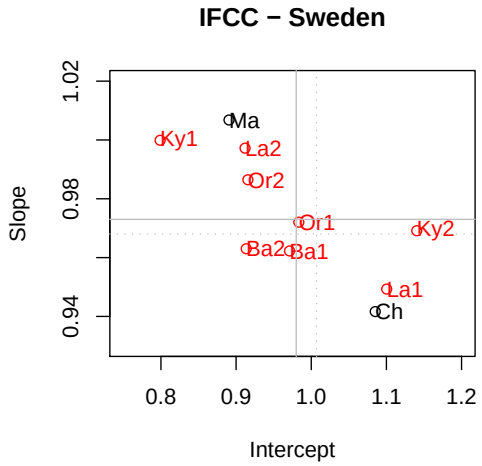
Table 9.4: **P-values and MSE from the posterior predictive checks for studies Kyoto 1, Kyoto 2, Orlando 1 for different parameters of the prior of D, based on the hierarchical linear model.**

Study	$\boldsymbol{\Omega}$	p-value Intercept	p-value Slope	MSE
Kyoto 1	$10^{-2} \cdot \mathbf{I}$	0.43	0.57	0.018
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	0.39	0.62	0.019
	$10^{-3} \cdot \mathbf{I}$	0.27	0.73	0.023
	$5 \cdot 10^{-4} \cdot \mathbf{I}$	0.22	0.79	0.027
	$10^{-4} \cdot \mathbf{I}$	0.13	0.89	0.041
Kyoto 2	$10^{-2} \cdot \mathbf{I}$	0.52	0.55	0.027
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	0.55	0.55	0.027
	$10^{-3} \cdot \mathbf{I}$	0.62	0.51	0.029
	$5 \cdot 10^{-4} \cdot \mathbf{I}$	0.65	0.49	0.031
	$10^{-4} \cdot \mathbf{I}$	0.73	0.47	0.039
Orlando 1	$10^{-2} \cdot \mathbf{I}$	0.50	0.50	0.018
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	0.50	0.50	0.017
	$10^{-3} \cdot \mathbf{I}$	0.50	0.50	0.017
	$5 \cdot 10^{-4} \cdot \mathbf{I}$	0.50	0.51	0.017
	$10^{-4} \cdot \mathbf{I}$	0.48	0.53	0.018

and three studies for which the minimum mean-square error is reached for the second or third prior, respectively. As for the two studies with minimum MSE in the first prior, the change in the p-values as well as in the MSE is not large when going to the second prior, we conclude that the second prior is the most adequate for this data.
Comparing the posterior distribution of the mean coefficients based on the hierarchical linear model, with their posterior distribution from the linear Bayesian model, we see that the range of the posterior distribution is wider. The location of the posterior is slightly shifted (see also Figure 9.1).

Posterior predictive checks for new studies are exemplarily made for the data of the Marrakech and the Chicago study. Based on the posterior distributions of the parameters, derived with the second prior ($\boldsymbol{\Omega} = 5 \cdot 10^{-3} \cdot \mathbf{I}$), p-values of the intercept and the slope for the Marrakech study are (0.36, 0.82) and for the Chicago study (0.67, 0.20). As all p-values are greater than 0.1 or less than 0.9 we can conclude that both studies fit to the assumed model. No change in the experimental conditions has occurred.

Figure 9.1: **Plot of the intercept and slope for the method comparison studies between the IFCC network and the Swedish network for standardization of HbA1c. Red indicates the studies used for the derivation of the master equation, black the first two studies which are not included in the analysis. The solid line represents the location of the posterior of intercept and slope based on the hierarchical linear model, the dashed line based on the pooled model.**



9.4.3 Hierarchical linear model with errors in both axes

To account for the measurement error in both methods, we regard the repeated measurements in both methods. For each sample measured within the IFCC network for standardization of HbA1c about 40 values and for each sample measured in the Swedish network four repeated values are available.
For the analysis four different prior settings are analyzed: The parameters of the inverse Gamma distributions of the measurement errors of both networks (u_x, v_x), (u_y, v_y) are set to $(0.1, 0.1)$ or $(2, 0.018)$, the parameter $\mathbf{\Omega}$ of the inverse Whishart distribution is set to $5 \cdot 10^{-3}\mathbf{I}$, $10^{-3}\mathbf{I}$. The other priors are the same as for the hierarchical linear model.

In Table 9.5, a summary of the posterior distribution of the mean coefficients for the different prior settings is given. The range of the posterior distributions of the coefficients is much wider for the last two priors, than for the first two. For the first prior setting it is comparable with the range from the hierarchical linear model. The last two prior settings result in flatter posterior distributions. The same prior distribution is assigned to both measurement error variances. However, the InvGamma(2, 0.018) is derived based only on information on the Swedish network and does not well describe the information of the measurement error variance of the IFCC network, the InvGamma(0.1, 0.1) is less informative.

Table 9.5: **2.5%, 50% and 97.5% quantiles of the posterior distributions of** (b_0, b_1) **for different priors, based on the hierarchical linear model with errors in both axes.**

Parameters	$\mathbf{\Omega}$	(u,v)	2.5% quantile	50% quantile	97.5% quantile
b_0	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.792	0.995	1.201
	$10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.844	1.001	1.158
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.497	1.006	1.507
	$10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.553	1.004	1.456
b_1	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.939	0.970	1.001
	$10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.946	0.969	0.993
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.894	0.968	1.042
	$10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.903	0.968	1.034

In Table 9.6, the p-values of the intercept and the slope, as well as the MSE is shown

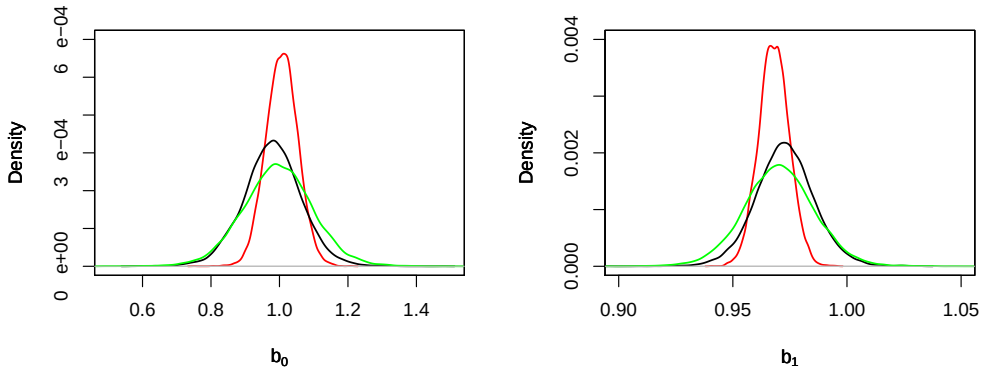
for the different prior settings. The p-values of the priors are comparable to the p-values of the hierarchical linear model, however there are huge differences in the MSE for the different choices of the parameters of the inverse Gamma distribution. The last two priors give no adequate results.

Table 9.6: **P-values and MSE from the posterior predictive checks for studies Kyoto 1, Kyoto 2, Orlando 1 for different priors, based on the hierarchical linear model with errors in both axes.**

Study	$\boldsymbol{\Omega}$	(u,v)	p-value Intercept	p-value Slope	MSE
Kyoto 1	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.39	0.60	0.069
	$10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.32	0.69	0.072
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.42	0.58	0.942
	$10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.40	0.61	0.910
Kyoto 2	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.53	0.59	0.109
	$10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.59	0.55	0.105
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.52	0.53	1.335
	$10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.53	0.53	1.294
Orlando 1	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.48	0.52	0.063
	$10^{-3} \cdot \mathbf{I}$	$(0.1, 0.1)$	0.48	0.53	0.056
	$5 \cdot 10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.48	0.52	0.938
	$10^{-3} \cdot \mathbf{I}$	$(2, 0.018)$	0.48	0.52	0.842

In Figure 9.2, the posterior densities of the coefficients of the IFCC - Sweden relationship are shown. The red line indicates the posterior density of the linear Bayesian model, the black line of the hierarchical linear model, the green one of the hierarchical linear model with errors in both axes. We see very clearly that the posterior density becomes flatter the more variation sources are included in the model. The inclusion of the study hierarchy is more important than the additional inclusion of the measurement error in both axes.

Figure 9.2: **Plot of the posterior densities of b_0 and b_1 for the IFCC - Sweden master equation derived from the different models. The red line indicates the posterior density of the linear Bayesian model, the black line of the hierarchical linear model, the green one of the hierarchical linear model with errors in both axes.**



Chapter 10

Discussion

A stable standardization concept for diagnostic assays is mandatory for the long-term quality of these medical devices. It guarantees that the measured values are stable over space and time, such that confidence is provided to physicians to support their medical decisions. Statistical methods assist in achieving these goals. In this thesis, three major standardization issues are analyzed from a statistical point of view:

(i) The derivation of assigned values for calibrators via sample reading and the calculation of their uncertainty by taking into account the uncertainties of metrologically higher calibrators is discussed.

(ii) Techniques for the identification of outliers within data, coming from multiple laboratories, are shown. These includes the possibility to define rules for the approval of laboratories as members of standardization networks.

(iii) Procedures for the comparison and the combination of method comparison studies are established.

The derivation of the assigned value of a calibrator is usually embedded in a whole standardization cascade, i.e. metrologically higher calibrators are needed for the derivation of the assigned value and the particular calibrator might be used in the value assignment of a metrological lower calibrator. The assigned values of calibrators carry an uncertainty, its calculation is demanded by different institutions ([IVD98], [GUM93]). However, up to now these uncertainty values are not used as valuable information in the downstream standardization cascade.
We show that the incorporation of the uncertainty of the calibrators is necessary for the establishment of a meaningful calibration curve, as well as for the reading of an unknown sample from this calibration curve and for the derivation of the uncertainty of

this reading.
In a simulation of the sample reading from a linear calibration curve, we analyze different situations for the error structure of the assigned values of the calibrators (low vs. high uncertainty, no correlation vs. high correlation, flat calibration curve vs. steep calibration curve). In all cases, the incorporation of the knowledge on the uncertainty of the calibrators produces much better results in terms of point estimates of the value of the read sample and coverage of the confidence intervals.
Bayesian modelling of the sample reading and MCMC algorithm are a good tool for this situation. However, these calculations are still too slow to be incorporated in automatic analyzers. Therefore it would be worth to analyze how approximations of these models, as e.g. proposed by [CFH04], match the results of the Bayesian modelling.
Knowing the uncertainty of a single measurement, it is easy to combine multiple measurements coming from different laboratories by a one-way random effects model and known individual error variance. The combination of such structured data is a well known issue for the value assignment of calibrators. However, most authors dealing with it do not incorporate the uncertainty of the individual measurement in their models. We show that the incorporation provides better estimates of the assigned values and narrower confidence intervals.

In the analysis of data from standardization networks the outlier detection plays an important role. In the structure of the data different types of outliers can occur. A whole laboratory is an outlier if its measured values are extreme in comparison to the measured values of the other laboratories on the same samples. Or a laboratory is an outlier, if the variation of its repeated measured values is different to the variation of the other laboratories. Further a single measurement within a laboratory can be an outlier, if this value is extreme, compared to the other values within this laboratory. We provide rules to detect these three types of outliers within data of standardization networks, based on linear mixed models.
Two data situations are regarded: If data of a single sample is regarded, the one-way random effects model is appropriate for modelling this situation. Based on the sum of squares of the laboratory effects and the sum of squares of the residuals, outlier identification rules are defined. If data from multiple samples is regarded the random coefficients model is used to model laboratory effects. In this case a regression line is estimated for each laboratory, such that for each laboratory a kind of systematic and proportional effect is estimated.
For both situations limits are defined for the sum of squares of the laboratory effects and residuals. The limits are given by Chi-square quantiles, as under the assumption

of normally distributed laboratory effects and residuals the respective sums of squares follow this distribution. However, it rests for further research to define better limits, based on the distribution of the estimates of the laboratory effects and residuals.
Based on actual data, we show that the estimation of the parameters of the linear mixed models with normally distributed laboratory effects and residuals leads to masking and swamping effects for the outlier identification. Therefore we propose to use a more robust estimation method, based on the t-distribution with appropriate degrees of freedom. Reanalyzing the data with this estimation method leads to better results concerning the outlier identification.
Finally we discuss how these rules can be generalized to be used for multiple standardization studies. This means that the limits for outlier identification are not derived from the actual data, but set in advance. This is a kind of quality control procedure for data of standardization networks. On this procedure rules can be defined to approve laboratories as members of a standardization network.
On one hand side we propose some ideas for the derivation of such limits, based on historical data. On the other side we discuss how these limits can be interpreted in terms of allowable deviations.

When new standardization principles for a diagnostic assay are introduced the old and established values have to be linked to the new values, to ensure the continuity of diagnostic and therapeutic decisions. The IFCC network for standardization of HbA1c is a recent example for this problem. A new standardization approach is established worldwide and replaces national standardizations. To compare the values based on these national standardizations with the IFCC values, twice a year method comparison studies are launched. This means that a set of samples is measured based on the national standardization and based on the IFCC standardization. A linear regression line is afterwards fitted to the data.
We deal with two issues in the context of repeated method comparison studies. The first one is the comparison of two or more of these regression lines. It is important to know if there are differences between the outcome of these experiments. For this comparison we use the algorithm proposed by [LJZ04], which bases the comparison on the construction of simultaneous confidence bands for the difference of predicted values. Further we show how this algorithm can be changed to compare a new regression line with a reference regression line.
Both algorithms are based on least-squares estimation of the respective regression line. This is not always appropriate for the analysis of method comparison studies, as both methods are subject to error. Hence, estimation techniques for errors-in-variables mod-

els would be better. It would be interesting to adjust the algorithms to these estimation methods.
The second issue deals with the combination of multiple regression lines to obtain an averaged regression line and its uncertainty. Bayesian hierarchical models are used for modelling the data of multiple method comparison studies, to obtain the posterior distribution of the coefficients of the averaged regression line. We extend the simple Bayesian hierarchical model to incorporate the errors in both variables.
As for Bayesian models prior distributions need to be assigned, the sensitivity of the priors on the results must be checked. We present a method, based on posterior predictive checks, to analyze different prior settings. Finally we discuss based on data of the IFCC network for standardization of HbA1c the importance for the hierarchical modelling and the incorporation of the errors in both variables.

Bibliography

[Ber69] J. Berkson. Estimation of a linear function for a calibration line; Consideration of a recent proposal. *Technometrics*, 11:649–660, 1969.

[BG92] R.K. Burdick and F.A. Graybill. *Confidence Intervals on Variance Components*. Marcel Dekker Inc., New York, 1992.

[BG98] S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[Bil98] J.A. Bilmes. *A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models*. International Computer Science Institute, Berkeley CA, USA, 1998.

[Car96] B.P. Carlin. Hierarchical longitudinal modelling. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *In Markov Chain Monte Carlo in Practice*, pages 303–320. Chapmann & Hall, 1996.

[CFH04] M.G. Cox, A.B. Forbes, and P.M. Harris. *The classification and solution of regression problems in calibration*. National Physics Laboratory, Teddington, UK, 2004.

[CG92] G. Casella and E.I. George. Explaining the Gibbs sampler. *The American Statistician*, 46:167–174, 1992.

[CG94] S. Chib and E. Greenberg. Bayes inference in regression models with ARMA(p,q) errors. *Journal of Econometrics*, 64:183–206, 1994.

[CG95] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–334, 1995.

[Coc37] W.G. Cochran. Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, 4(Suppl):102–118, 1937.

[Coc54] W.G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10:101–129, 1954.

[CVN99] C.-L. Cheng and J.W. Van Ness. *Statistical Regression with Measurement Error*. Oxford University Press Inc., New York, 1999.

[DCC93] DCCTRG - The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of longterm complications in insulin-dependent diabetes mellitus. *New England Journal of Medicine*, 329:977–986, 1993.

[Die91] C.F. Dietrich. *Uncertainty, Calibration and Probability: The statistics of scientific and industrial measurement*. Adam Hilger, New York, 1991.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–22, 1977.

[DS95] P. Dellaportas and D.A. Stephens. Bayesian analysis of errors-in-variables regression models. *Biometrics*, 51:1085 – 1095, 1995.

[DS98] R.D. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons Inc., New York, 1998.

[Eis39] C. Eisenhart. The interpretation of certain regression methods and their use in biological and industrial research. *The Annals of Mathematical Statistics*, 10:162 – 186, 1939.

[Eno99] D.R. Eno. *Noninformative prior bayesian analysis for statistical calibration problems, PhD Thesis*. Virginia Polytechnic Institute, Blacksburg, Virginia, USA, 1999.

[FC60] R.A. Fisher and E.A. Cornish. The percentile points of distributions having unknown cumulants. *Technometrics*, 2:209–226, 1960.

[FHT96] L. Fahrmeier, A. Hamerle, and G. Tutz, editors. *Multivariate statistische Verfahren*. Walter de Gruyter & Co., Berlin, 1996.

[Ful87] A.W. Fuller. *Measurement Error Models*. John Wiley & Sons Inc., New York, 1987.

[GCS04] A. Gelman, G.L. Chew, and M. Shnaidman. Bayesian analysis of serial dilution assays. *Biometrics*, 60:407–417, 2004.

[GCSR04] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, 2004.

[GD93] U. Gather and L. Davies. The identification of multiple outliers. *American Statistical Association*, 88:782–801, 1993.

[Gel06] A. Gelman. Prior distributions for variance parameters in hierarchical linear models. *Bayesian Analysis*, 1:515:533, 2006.

[Gil96] W.R. Gilks. Full conditional distributions. In W.R. Gilks, Richardson S., and Spiegelhalter D.J., editors, *Markov Chain Monte Carlo in Practice*, pages 75–88. Chapman & Hall, 1996.

[GR92] A. Gelman and D.R. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

[GRS96] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[GSC95] A.E. Gelfand, S.K. Sahu, and B.P. Carlin. Efficient parametrisations for normal linear mixed models. *Biometrika*, 82:479–488, 1995.

[GUM93] International Organization for Standardization, Geneva. *Guide to the Expression of Uncertainty in Measurement*, 1993. ISO.

[Gut67] I. Guttman. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B*, 29:83 – 100, 1967.

[Har74] D.A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61:383–385, 1974.

[Hil81] B.M. Hill. Discussion of Hunter-Lamboy's paper. *Technometrics*, 23:335–338, 1981.

[HL81] W.G. Hunter and W.F Lamboy. A bayesian analysis of the linear calibration problem. *Technometrics*, 23:323–328, 1981.

[HM96] W. Hoelzel and K. Miedema. Development of a reference system for the international standardization of HbA1c/glycohemoglobin determinations. *Journal of the International Federation of Clinical Chemistry*, 9:62–67, 1996.

[Hoa70] B. Hoadley. A bayesian look at inverse linear regression. *Journal of the American Statistical Association*, 65:356–369, 1970.

[How03] P. Howarth. *Metrology - in short, 2nd edition*. EUROMET, 2003.

[HWJ+04] W. Hoelzel, C. Weykamp, J.-O. Jeppson, K. Miedema, J.R. Barr, I. Goodall, et al. IFCC reference system for the measurement of hemoglobin A1c in human blood and the national standardization schemes in the United States, Japan and Sweden: A method comparison study. *Clinical Chemistry*, 50:166–174, 2004.

[IVD98] European Parliament, Strasbourg. *Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic mediacal devices*, 1998. IVD.

[JKB+02] J-O. Jeppsson, U. Kobold, J.R. Barr, A. Finke, W. Hoelzel, T. Hoshino, et al. Approved IFCC reference method for the measurement of HbA1c in human blood. *Clinical Chemistry and Laboratory Medicine*, 40:78–89, 2002.

[KAS+06] A. Konnert, S. Arends, S. Schubert, C. Berding, C. Weykamp, and C. Siebelder. Uncertainty calculation for calibrators of the IFCC HbA1c standardization network. *Journal of Accrediation and Quality Assurance*, 11:319 – 328, 2006.

[KBA+06] A. Konnert, C. Berding, S. Arends, C. Parvin, C.L. Rohlfing, H.-M. Wiedmeyer, et al. Statistical rules for laboratory networks. *Journal of Testing and Evaluation*, 34:128–134, 2006.

[Kru67] R.G. Krutchkoff. Classical and inverse regression methods of calibration. *Technometrics*, 9:425–439, 1967.

[Law81] J.F. Lawless. Discussion of Hunter-Lamboy's paper. *Technometrics*, 23:334–335, 1981.

[Lig06] U. Ligges. *The BRugs Package*, 2006. R package version 0.2-5.

[LJZ04] W. Liu, M. Jamshidian, and Y. Zhang. Multiple comparison of several linear regression models. *Journal of the American Statistical Association*, 99:395–403, 2004.

[LLV04] R. Landes, P. Loutzenhiser, and S. Vardeman. *Hierarchical Bayes Statistical Analyses for a Calibration Experiment*. Iowa State University, Iowa, USA, 2004.

[LSB+05] P.C. Lambert, A.J. Sutton, P.R. Burton, K.R. Abrams, and D.R. Jones. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24:2401 – 2428, 2005.

[LW82] N.M. Laird and J.H. Ware. Random effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[Lwi81] T. Lwin. Discussion of Hunter-Lamboy's paper. *Technometrics*, 23:339–341, 1981.

[Man95] J. Mandel. Structure and outliers in interlaboratory studies. *Journal of Testing and Evaluation*, 23:364–369, 1995.

[MKW+00] G.L. Myers, M.M. Kimberly, P.P. Waymacks, S.J. Smith, G.R. Cooper, and E.J. Sampson. A reference method laboratory network: A model for standardization and improvement of clinical laboratory measurements. *Clinical Chemistry*, 46:11:1762–1772, 2000.

[MP70] J. Mandel and R.C. Paule. Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry*, 42:1194–1197, 1970.

[MQ06] A.D. Martin and K. M. Quinn. *MCMCpack: Markov chain Monte Carlo (MCMC) Package*, 2006. R package version 0.7-2.

[MR93] X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993.

[MRR02] A. Martinez, J. Riu, and F.X. Rius. Evaluating bias in method comparison studies using linear regression with errors in both axes. *The Analyst*, 16:41–53, 2002.

[MSLW96] G.A. Milliken, W.W. Stroup, R.C. Littell, and R.D. Wolfinger. *The SAS System for Mixed Models*. SAS Publishing, Carry, 1996.

[NM98] R. Natarajan and C.E. McCulloch. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7:267–277, 1998.

[Orb81] J.E. Orban. Discussion of Hunter-Lamboy's paper. *Technometrics*, 23:342–343, 1981.

[PB83] H. Passing and W. Bablock. A new biometrical procedure for testing the equality of measurements from two different analytical methods. *Clin. Chem. Clin. Biochem.*, 21:709–720, 1983.

[PB00] J.C. Pinheiro and D.M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer, New York, 2000.

[PBCV06] M. Plummer, N. Best, K. Cowles, and K. Vines. *coda: Output analysis and diagnostics for MCMC*, 2006. R package version 0.10-5.

[PBD+06] Jose Pinheiro, Douglas Bates, Saikat DebRoy, , and Deepayan Sarkar. *nlme: Linear and nonlinear mixed effects models*, 2006. R package version 3.1-73.

[PLW01] J.C. Pinheiro, C. Liu, and Y.N. Wu. Efficient alorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, 10:249–276, 2001.

[PM82] R.C. Paule and J. Mandel. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87:377–385, 1982.

[R D06] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[RBV00] A.L. Rukhin, B.J. Biggerstaff, and M.G. Vangel. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. *Journal of Statistical Planning and Inference*, 83:319–330, 2000.

[RKC81] P.S. Rao, J. Kaplan, and W.G. Cochran. Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76:89–97, 1981.

[RL96] A.E. Raftery and S.M. Lewis. Implementing MCMC. In W.R. Gilks, Richardson S., and Spiegelhalter D.J., editors, *Markov Chain Monte Carlo in Practice*, pages 115–130. Chapman & Hall, 1996.

[Rob91] G.K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6:15–51, 1991.

[RPWS91] A. Racine-Poon, C. Weihs, and A.F.M. Smith. Estimation of relative potency with sequential dilution errors in radioimmunoassays. *Biometrics*, 47:1235–1246, 1991.

[RRR01] F.J. Rio, J. Riu, and F.X. Rius. Robust linear regression taking into account errors in the predictor and response variable. *The Analyst*, 126:1113–1117, 2001.

[Rub81] D.B. Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6:377–401, 1981.

[Rub84] D.B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistic*, 12:1151–1172, 1984.

[RV98] A.L. Rukhin and M.G. Vangel. Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association*, 93:303–308, 1998.

[SAS06] SAS Institute, Inc., Cary, NC, USA. *SAS 9.1.3*, 2006.

[SCM92] S. Searle, G. Casella, and C. McCulloch. *Variance Components*. John Wiley & Sons Inc., New York, 1992.

[Sea71] S. Searle. *Linear Models*. John Wiley & Sons Inc., New York, 1971.

[SG92] A.F.M. Smith and A.E. Gelfand. Bayesian statistics without tears: a sampling-resampling approach. *The American Statistician*, 46:84–88, 1992.

[Spu99] J.D. Spurrier. Exact confidence bounds for all contrasts of three or more regression lines. *Journal of the American Statistical Association*, 94:483–488, 1999.

[Sta06] S. Stampf. Bayesianische Strukturierte Additive Regression: Schätzung der Glättungsparameter. Master's thesis, LMU München, München, Germany, 2006.

[STBL05] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, Version 2.10*. University of Helsinki, Dep. of Mathematics and Statistics, Helsinki, Finnland, 2005.

[STH01] D. Sun, R.K. Tsutakawa, and Z. He. Propriety of posteriors with improper priors in hierarchical linear models. *Statistica Sinica*, 11:77–95, 2001.

[Tie94] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762, 1994.

[TUM+05] L. Thienpont, K. van Uytfanghe, J. Marriot, L. Siekmann, A. Kessler, D. Bunk, et al. Metrologic traceability of total tyroxine measurements in human serum: Efforts to establish a network of reference measurement laboratories. *Clinical Chemistry*, 51:161–168, 2005.

[VIM93] International Organization for Standardization, Geneva. *International Vocabulary of basic and general standard terms in Metrology*, 1993. ISO.

[VM00] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.

[WG03] J. Wellmann and U. Gather. Identification of outliers in a one-way random effects model. *Statistical Papers*, 44:335–348, 2003.

[YC54] F. Yates and W. G. Cochran. The analysis of groups of experiments. *Journal of Agricultural Science*, 28:556–580, 1954.

[ZGF02] G. Zuur, P.H. Garthwaite, and R.J. Fryer. Practical use of MCMC methods: Lessons from a case study. *Biometrical Journal*, 44:433 – 455, 2002.

Appendix A

Matrix Notations

The following matrix identities are used throughout the thesis.

The trace of a square matrix $trace(\mathbf{A})$ is defined as the sum of the diagonal elements of $\mathbf{A}$. Given two square matrices of the same dimension $\mathbf{A}, \mathbf{B}$ we have

$$trace(\mathbf{A} + \mathbf{B}) = trace(\mathbf{A}) + trace(\mathbf{B}) \tag{A.0.1}$$

$$trace(\mathbf{AB}) = trace(\mathbf{BA}) \tag{A.0.2}$$

$$\sum_i \mathbf{x}_i' \mathbf{A} \mathbf{x}_i = trace\left(A \sum_i \mathbf{x}_i \mathbf{x}_i' \right). \tag{A.0.3}$$

Also remember, that $|\mathbf{A}|$ denotes the determinant of a $n \times n$ matrix $\mathbf{A}$, defined as

$$n = 1 \qquad |\mathbf{A}| = a, \text{ where } \mathbf{A} = (a)$$

$$n \geq 2 \qquad |\mathbf{A}| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}|, \ \forall 1 \leq i \leq n,$$

where $\mathbf{A}_{ij}$ denotes the $(n-1) \times (n-1)$ matrix, which is obtained, by discarding the ith row and jth column of $\mathbf{A}$. $(-1)^{i+j} |\mathbf{A}_{ij}|$ is called the cofactor of the element a_{ij} of $\mathbf{A}$.

For the determination of the values that maximize the log-likelihood function, we have to take derivatives of a function of a matrix $f(\mathbf{A})$. Hence, we define $\frac{\partial f(\mathbf{A})}{\mathbf{A}}$, as the

matrix, with i, jth entry

$$\left(\frac{\partial f(\mathbf{A})}{\mathbf{A}}\right)_{ij} = \frac{\partial f(\mathbf{A})}{\partial a_{ij}}.$$

One can show, that the following identities hold (see e.g. [FHT96])

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = (\mathbf{A} + \mathbf{A}')\mathbf{x} \tag{A.0.4}$$

$$\frac{\partial |\mathbf{A}|}{\partial a_{ij}} = \begin{cases} (-1)^{i+j}|\mathbf{A}_{ij}|, & \text{if i=j} \\ 2(-1)^{i+j}|\mathbf{A}_{ij}|, & \text{if } i \neq j \end{cases} \quad \forall \mathbf{A} = \mathbf{A}' \tag{A.0.5}$$

$$\begin{aligned} \frac{\partial ln|\mathbf{A}|}{\partial \mathbf{A}} &= \begin{cases} (-1)^{i+j}|\mathbf{A}_{ij}|/|\mathbf{A}|, & \text{if i=j} \\ 2(-1)^{i+j}|\mathbf{A}_{ij}|/|\mathbf{A}|, & \text{if } i \neq j \end{cases} \\ &= 2\mathbf{A}^{-1} - diag(\mathbf{A}^{-1}) \; \forall \mathbf{A} = \mathbf{A}' \end{aligned} \tag{A.0.6}$$

$$\frac{\partial trace(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B} + \mathbf{B}' - diag(\mathbf{B}). \tag{A.0.7}$$