6.3.4 Phenotype response .......................... 141
6.3.5 Weighted lasso phenotype inference ............... 142
6.4 Results ........................................ 144
6.4.1 Analysis of Arabidopsis germination experiment 144
6.4.2 Simulation strategies .......................... 149
6.4.3 Simulation study 1: justification of the proposed methodology .................. 152
6.4.4 Simulation study 2: comparison of the proposed methodology focusing on imputation methods 153
6.5 Conclusions ................................... 162
6.6 Acknowledgements ............................. 162
Bibliography ........................................ 163

Chapter 7: Concluding remarks ........................ 165
7.1 Summary of results ............................ 165
7.2 Why the beta distribution works and what is next ... 167
7.3 Importance of ICC as a measure of agreement, heterogeneity and heritability .............. 168
7.4 Simultaneous confidence interval for ICCs ........ 168
7.5 Alternative approaches for estimation of variance components .......................... 170
7.6 Estimation and testing of functions of variance components parameters at the boundary of the parameter space .......................... 171
Bibliography ........................................ 173

Academic summaries ................................ 175
Summary in English .................................. 175
Summary in Dutch ................................... 178
Summary in Georgian ............................... 182
Summary in Russian ............................... 186

Appendix Appendix A: Additional files of Chapter 6 191
Chapter 1
General Introduction

1.1 Motivation

Variance components models are flexible in analyzing complex structured data, and therefore are widely employed in medicine and biology. The variance components represent variability in sources of variation or indicate variability between subgroups of data. Functions of variance components can be used to quantify for instance agreement, heterogeneity, or heritability. To quantify the uncertainty of the estimates of functions of variance components, it is important to construct confidence intervals for these estimates. With the latest two decade development, variance components came to play an important role not only in linear variance components models, but also in nonlinear and generalized linear variance components models. There are challenges associated with the construction of confidence intervals for ratios of sums of variance components. This thesis is initially motivated by the lack of a generic, exact method for the construction of confidence intervals for such ratios that would work for all of these variance components models, with balanced and unbalanced designs. More generally, understanding the performance and applicability of variance components models in non-standard applications, like analyzing complex meta-analysis and family studies was a second motivation of the thesis.

1.2 Literature review

1.2.1 Linear, nonlinear and generalized linear mixed models

A mixed effects model (short a mixed model) is a statistical model that contains both, fixed and random effects. The effects are taken to be fixed if those levels of the factors are selected for inclusion in a
study in which we have an interest. The effects are taken to be random if the levels of the factors that we include in the study represent a random sample drawn from some (hypothetical) infinite population of levels. In practice though, samples are taken from (large) finite populations. A fixed effects model is a model with fixed effects only. A random effects model usually assumes an overall mean for observations, thus, it is a mixture of fixed and random effects. However, the name mixed model is reserved for models that contain fixed effects other than the overall mean. Mixed models are used for the designs where measurements are conducted on clusters or groups of related statistical units. Statistical units are frequently called subjects and may refer to individuals, animals or samples. In a broad sense clusters could represent one of the following: families, schools, litters, repeated measures under different treatments or time. Repeated measures over time are usually referred to a longitudinal data. Detailed description of clustered data is given by Aerts et al. (2002). Mixed models can also be used to model data with multiple sources of variation, heterogeneity between studies or biological variation of samples. Mixed models are applied extensively in various fields, from medical and social sciences to econometrics and these models may have different names. In social and behavioral sciences so called multilevel or hierarchical models are used to model data that vary at more than one level using (e.g. Hox and Roberts (2011)). In epidemiology, mixed models are often used to take into account the heterogeneity between studies when meta-analysis is conducted (Platt et al., 1999, Higgins and Thompson, 2002). In agriculture, mixed models can be used for growth curve modelling of traits (Jaffrézic et al., 2006, Meza et al., 2007) or for modelling the heterogeneous associations between multiple traits (Bello et al., 2012).

Suppose that for the $i$th cluster $n_i$ measurements (e.g. subjects) have been observed, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n_i$, with $I$ be the total number of clusters and $N = \sum_{i=1}^{I} n_i$ be the total number of observations. Denote the response vector $y_i = (y_{i1}, \ldots, y_{in_i})^T$ of size $n_i \times 1$ for observations in the $i$th group, with $T$ being the transpose. Note, $y_i$ are clustered measurements implying observations within the same cluster may be correlated. The general form of the mixed model in matrix form is:

$$y_i|u_i \sim F_i(\beta, u_i),$$

(1.1)
where $y_i$ conditionally on $u_i$, follows some pre-specified distribution $F$, $\beta$ is a vector of unknown parameters, common to all statistical units (i.e. subjects), through which the explanatory variables are parameterized. The vector $u_i$ of random effects coefficients associated with group $i$ is assumed to follow a so-called mixing distribution. Usually this distribution depends on a vector $\theta_u$ of unknown parameters, $u_i \sim H(\theta_u)$. This thesis focuses on conditional mixed models, in which conditionally on $u_i$, observations in $y_i$ are assumed to be independent. If $H$ is not specified, then the mixed model is nonparametric. Nonparametric models provide more flexibility, though they may or may not be able to capture the true heterogeneity between statistical units. Therefore, specific parametric form is often assumed for $H$, such as normal. Depending on the number of random effects the normal distribution can be univariate or multivariate.

If the mean of responses in $y_i$ are linearly related to the parameters of fixed effects $\beta$ and random effects $u_i$ in (1.1), then one has a linear mixed model. Depending on distributional assumption of the response, linear mixed models can be classified as Gaussian (normal) and non-Gaussian linear mixed models. This thesis focuses on Gaussian linear mixed models, though non-Gaussian ones are important to study since in practice, one is never certain that normality assumption holds. Gaussian linear mixed model (LMM) assumes the response vector $y_i$ follows a multivariate normal distribution and can be written in matrix form as:

$$
\begin{align*}
    y_i | u_i &\sim N(X_i\beta + Z_iu_i, \Lambda_i(\theta_e)) \\
    u_i &\sim iid N_q(0, \Psi(\theta_u)),
\end{align*}
$$

where $X_i$ is the $n_i \times p$ matrix for the $p$ known covariates, $\beta$ is the $p \times 1$ vector of fixed population parameters, $u_i$ is the $q \times 1$ vector of random effects for group $i$ (not varying across $j$) and $u_i$’s are assumed to be independent of each other and identically distributed (iid), $Z_i$ is the $n_i \times q$ design matrix for the $q$ random effects for observations in group $i$, $\Psi$ is the $q \times q$ variance-covariance matrix of the $q$ random effects, $\epsilon_i$ is $n_i \times 1$ vector of residual errors, $\Lambda_i$ is the $n_i \times n_i$ variance-covariance matrix of residual errors, $\epsilon_i$’s are assumed to be independent of each other and $u_i$. The matrices $\Psi(\theta_u)$ and $\Lambda_i(\theta_e)$ depend on
unknown parameters $\theta_u$ and $\theta_e$, where $\theta_u$ is the vector of between-cluster variance-covariance parameters and $\theta_e$ is the vector of within-cluster variance-covariance parameters. Note, that the set of unknown parameters in $\Lambda_i$ will not depend on $i$. The landmark paper on the development of maximum likelihood (ML) estimation procedure for mixed models in general is Hartley and Rao (1967). Other contributing papers in this field during next decade were Harville (1976), Corbeil and Searle (1976), Harville (1977), Miller (1977). Major books on mixed models that has been used in the thesis are Dean and Voss (1999), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Littell et al. (2006), Jiang (2007), Demidenko (2013).

A specific subset of mixed models with particular restrictions on variance-covariance structure of random terms is the variance components model. Traditional variance components models originated from the analysis of variance (anova) introduced by Fisher (1925, Chapter 7). According to Fisher (1925), anova is the “separation of the variance ascribable to one group of causes, from the variance ascribable to other groups”. Anova is usually associated with variance decomposition and the method of moments (MM) estimation. The variance components models may use other estimation method though, for example the maximum likelihood (ML). Let use the definitions of $X_i$, $\beta$, $Z_i$, $u_i$ in (1.2) and (1.3) above. Assuming $y = (y_1, \ldots, y_N)$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ are $N \times 1$ vectors of responses and residual errors respectively, a general form of variance components model using mixed model formulation is:

$$
\mathbf{y} = \mathbf{X}_i\beta + \sum_{i=1}^{q} \mathbf{Z}_i \mathbf{u}_i + \mathbf{\epsilon},
$$

(1.4)

where each vector $\mathbf{u}_i$ corresponds to all levels of a single factor and this factor may be a main effects factor, an interaction factor or a nesting factor. In a one-way random effects model, $q = 1$ and in a two-way random effects model with interaction, $q = 3$ (Searle et al., 2006, pp. 140-141). Assumptions on random effects are usually following: (1) elements $\mathbf{u}_i$ are independent, have zero mean and variance $\text{var}[\mathbf{u}_i] = \sigma_i^2 \mathbf{I}_{n_i} \forall i$ (2) error terms in $\mathbf{\epsilon}$ are independent, have zero mean and variance $\text{var}[\mathbf{\epsilon}] = \sigma_e^2 \mathbf{I}_N$, where $\mathbf{I}_{n_i}$ and $\mathbf{I}_N$ are the identity matrices of order $n_i$ and $N$ respectively. Random terms, $\mathbf{u}_i$’s are mutually inde-
dependent with each other and with elements of $\epsilon$. Variances $\sigma_i^2$ and $\sigma_e^2$ are called *variance components* because they are the components of the variance of the response $y$, i.e. $\text{var}(y) = \sum_{i=1}^{q} \sigma_i^2 Z_i Z_i' + \sigma_e^2 I_N$. Afore-mentioned assumptions (1) and (2) on variance-covariance structures of random effects and zero covariances between all random terms are what make up the variance component model. Formulation of variance components models given in (1.4) originated from Hartley and Rao (1967). Detailed description on this formulation and many other details of variance components models is given in the book Searle et al. (2006, pp. 138-140). According to Searle et al. (2006, pp. 23-24) the first formulation of a random effects model (although not called so) seems to be given by Airy (1861). Chapters 2 and 5 of the thesis study traditional variance components models.

**Nonlinear mixed models** (NLMM) are an extension of model (1.2) such that the responses in $y_i$ are nonlinearly related to fixed and/or random effects parameters:

$$y_i | u_i \sim N (g(X_i, \beta, Z_i, u_i), \Lambda_i(\theta_e)) \quad (1.5)$$

$$u_i \overset{iid}{\sim} N_q(0, \Psi(\theta_u)), \quad (1.6)$$

for some known nonlinear function $g(\cdot)$. The definitions of $X_i, \beta, Z_i, \Lambda_i(\theta_e)$ and $\Psi(\theta_u)$ remain the same as in (1.2). For nonlinear mixed models our major reference is Vonesh (2012). A special type of nonlinear mixed model (1.5) is *mixed effects nonlinear regression model* (Vonesh and Carter, 1992) and it is employed in Chapter 3 of the thesis. This specific model is nonlinear in fixed effects and linear in random effects: $y_i = g(X_i, \beta) + Z_i u_i + \epsilon_i$ with some nonlinear function $g(\cdot)$ and a design matrix $Z_i$ linked to the linear random effects $u_i$.

**Generalized linear mixed models** (GzLMM) are an extension of linear mixed models (1.2) that allow response variables from an exponential family of distributions, such as the Gaussian, binomial, Poisson, gamma, beta, etc. These models are mostly used for discrete or categorical outcomes. GzLMM were introduced by Breslow and Clayton (1993) and McCulloch (1997). The basis of the GzLMM are the generalized linear models (GLM). GLM were introduced by Nelder and Wedderburn (1972) and described in details by McCullagh and Nelder (1989). To formulate the parametric GzLMM the first key assumption
used in GLM is adjusted to random effects incorporation, namely: (1) conditional distribution of \( y_i | u_i \) is from an exponential family. Other two elements of GLM are automatically inherited by GzLMM and are: (2) the mean of the response relates the explanatory variables through a link function, (3) the variance of the response is a function of the mean of the response. The forth assumption is specific for GzLMM and it is frequently as follows: (4) the elements of \( y_i \) given a set of normally distributed random effects are conditionally independent, \( u_i \sim \text{iid} N_q(0, \Psi(\theta)) \). Then, the conditional probability density function in canonical form is:

\[
    f(y_i | u_i) = \exp\left\{ \left( y_i \eta_i - b(\eta_i) \right)/a(\phi) + c(y_i, \phi) \right\} \tag{1.7}
\]

where \( a(\cdot), b(\cdot), c(\cdot) \) are known functions, \( \phi \) is a known dispersion parameter (with unknown \( \phi \) the function \( f \) may or may not belong to an exponential family). Further the model is specified through a differentiable and monotonic link function \( g(\cdot) \) that relates the expected value of response \( \mu_i = E(y_i | u_i) \) to a linear predictor \( \eta_i \) containing a set of predictors with fixed effects and random effects. That is \( g(\mu_i) = \eta_i = X_i \beta + Z_i u_i \). Respectively, the inverse link function \( g^{-1} \) is \( \mu_i = g^{-1}(\eta_i) = g^{-1}(X_i \beta + Z_i u_i) \). Specification of the link function, mean and variance of the response depends on the distribution of the response. In Chapter 4 of this thesis for binary outcomes we use a logit link function \( g(\mu_i) = \logit(\mu_i) = (\mu_i/(1 - \mu_i)) \) where \( \ln \) denotes the natural logarithm. When the link function is the identity, \( g(\mu_i) = \mu_i \), one has the specification of means and variances for the normal distribution, which is typical for linear mixed models shown in (1.2).

### 1.2.2 Estimation of variance components

Let the density functions \( f_i(y_i) \) and \( h(u_i) \) correspond to distribution functions \( F_i \) and \( H \) in (1.1). Considering all models above and the notations introduced in (1.2-1.7), the marginal density of general mixed model can be written:

\[
    f_i(y_i | \gamma) = \int_{-\infty}^{\infty} f_i(y_i | x_i; \beta, u_i, \theta_e) h(u_i; \theta_u) du_i. \tag{1.8}
\]

The density \( f_i \) in (1.8) depends on unknown parameters \( \gamma = (\beta, \theta_e, \theta_u) \). A likelihood of (1.8), based on joint density of all observations \( y \) and
can be written:

\[ L(\gamma|y) = \prod_{i=1}^{I} \int_{-\infty}^{\infty} f_i(y_i|x_i; \beta, u_i, \theta_e) h(u_i; \theta_u) du_i. \]  \hspace{1cm} (1.9)

This likelihood requires maximization of the integral with respect to all parameters \( \gamma \). An integral in (1.9) has analytical solution when the random effects are assumed to follow multivariate normal, as in (1.3), and when the conditional distribution of \( y_i \) given \( u_i \) is also normal, as in (1.2). The ML is one of three most widely used estimation methods for LMM which has following advantages. ML estimators of variance components are 1) consistent, asymptotically normal, and efficient according to Miller (1977), 2) nonnegative, 3) with obtainable information matrix (Harville, 1977). The information matrix or expected Fisher information matrix for ML is used to obtain the standard errors of ML estimators. Note, ML does not take into account the degrees of freedom lost by estimating the fixed effects parameters \( \beta \) and consequently ML estimates are biased. For that reason ML is frequently substituted by restricted maximum likelihood (REML). REML estimators take into account the loss of degrees of freedom due to fixed effects. REML provides unbiased estimates of variance components for special cases and less biased estimates of variance components than ML in general. REML also guarantees nonnegativity of variance components. The third estimation method, the MM is typically used in traditional variance components models (as outlined in previous section). For balanced data, REML and MM estimators are the same when the variance components are positive. MM are known to be unbiased with minimal variance. Despite this attractive property, MM estimators have the important disadvantage of generating negative variance components. Overview on estimation methods for variance components models is given in Chapter 2 of the thesis.

For nonlinear and generalized linear mixed models analytical solutions of (1.9) are mostly not possible due to non-normal likelihood, implied by (1.5) and (1.7). Therefore, literature contains different approximate estimation methods. The most cited two estimation techniques are based on: 1) model linearization and 2) numerical integration. The idea of linearization is to approximate the model by the one
that is linear in the random effects $u_i$. Two strategies are used to linearize the likelihood function of NLMM using first-order Taylor expansion. One involves linearization of the model function about the mean of the random effects, $u_i = 0$ (Sheiner and Beal, 1980) and the other at the conditional modes of the random effects, $u_i = \tilde{u}_i$ (Lindstrom and Bates, 1990). The first strategy is usually described as the population-averaged (marginal or mean) expansion and the second as the cluster-specific expansion. Using linearization techniques the parameters are not estimated jointly as they would be when the true likelihood is maximized directly, but rather sequentially in two steps. Consequently, a so-called pseudo-likelihood (PL) or restricted pseudo-likelihood (REPL) is maximized instead of the true likelihood. For GzLMM, a linearization-based PL was presented by Breslow and Clayton (1993). The advantage of linearization estimation technique is its simplicity. The disadvantages are: 1) absence of a true likelihood that would allow testing using likelihood ratio test, 2) asymptotic properties of estimators such as consistency are unclear (at least to our knowledge) and 3) bias in estimates, particularly for variance components (see simulation results of Pinheiro and Bates (1995)). The work of Claassen (2014) provides methods to reduce the bias in variance components in GzLMM.

Laplace’s method (Tierney and Kadane, 1986) is usually used for integral approximation. It is based on a second-order Taylor expansion. Laplace’s method was the first estimation technique used for GzLMM to approximate the penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) (Breslow and Clayton, 1993). PQL and MQL are based on linearization ideas, similar to the ones proposed by Sheiner and Beal (1980) and Lindstrom and Bates (1990) in the context of NLMM. Details on Laplacian approximation for GzLMM and NLMM are given by Pinheiro and Bates (1995) and Pinheiro and Chao (2006). To achieve better accuracy in estimates, higher order Laplace’s approximations were proposed for GzLMM with nested random effects by Raudenbush et al. (2000), though the same authors and others (Pinheiro and Chao, 2006) have shown that the ML estimates of variance components using Laplace’s approximation are severely biased under the model and data configurations assumed in simulations. Furthermore, the fixed effects estimates are also biased under investigated scenarios, but relatively
1.2 Literature review

Two strategies of numerical integration are common for GzLMM and NLMM, deterministic and stochastic. The deterministic method approximates the value of an integral by a weighted average of the integrand which are evaluated at points conveniently chosen. In the literature this approach is known as Gauss-Hermite quadrature (originated from Abramowitz and Stegun (1972), Davis and Rabinowitz (1975)). For NLMM, Davidian and Gallant (1992) employed Gauss-Hermite quadrature to compute the approximated likelihood for general setting on random effects. They made no parametric assumption about the form of the random effects distribution. Moreover, Davidian and Gallant (1992) and Pinheiro and Bates (1995) demonstrated how to transform $q$-dimensional integrals in (1.9) into a series of one-dimensional integrals to simplify the computation. To achieve higher accuracy in estimates, an adaptive version of Gauss-Hermite quadrature was proposed by Pinheiro and Bates (1995). Original Gauss-Hermite quadrature uses fixed abscissas and weights for the integrand while an adaptive Gauss-Hermite quadrature allows a pseudo-random mechanisms for them. As usual, gain comes at a price and the price is computational time.

The second strategy is stochastic integration or the so called Monte Carlo integration. Monte Carlo methods conduct integration via simulation: by simulating (or sampling) from the target distribution (i.e. the desired distribution) over many trials we can learn properties of a random variable (mean, variance, etc.). When the target distribution is unknown, importance sampling can be used. The idea of importance sampling is to draw the sample from an importance distribution and re-weight the integral using importance weights so that the correct distribution is targeted. According to Pinheiro and Bates (1995), importance sampling provides reliable estimation results, comparable to those by Laplacian and adaptive Gaussian approximations. Several challenges are associated with importance sampling however, namely: 1) choice of importance distribution that would ideally correspond to the density of the integrand, 2) choice of proposal weights, 3) computational efficiency, 4) stochastic convergence is not always guaranteed. The main advantage of importance sampling is that it can handle dis-
tributions other than the normal for both random effects and residual errors. Stochastic approximation based on expectation-maximization (EM) algorithm are also common for NLMM, see for example Delyon et al. (1999). Furthermore, stochastic EM has been coupled with the Markov Chain Monte Carlo (MCMC) (Kuhn and Lavielle, 2004, 2005). MCMC by itself can also be used to approximate the density in (1.8) (e.g. Subhash et al. (2010)), but we will not outline these stochastic methods as they fall outside the scope of the thesis.

Generalized estimating equations (GEE) is a non-likelihood estimation approach to account for correlated responses in settings with repeated or longitudinal measures (Liang and Zeger (1986)). To be specific, GEE is a method of moments approach. It originated from GzLM where the later assumes independence of observations. GEE a semi-parametric approach since no full specification is made about the joint distribution of responses $y_i$ in the cluster $i$ and it requires only correct specification of the marginal mean of response, $E[y_i] = \mu_i(\beta)$. Prentice and Zhao (1991) utilized the results of Gourieroux et al. (1984) and extended GEE (GEE1) to a second-order generalized estimating equations (GEE2). GEE2 requires correct specification of both, the mean and the variance-covariance of response, $\text{var}[y_i] = \Sigma_i(\gamma)$. Further details on GEE and GEE2 can be found for example in Vonesh (2012, pp. 110-115, 151-158). Chapter 3 of the thesis employs GEE2 for NLMM and describes this estimation method.

1.2.3 Ratios of sums of variance components: intraclass correlation coefficient, heritability, heterogeneity

Sums and ratios of sums of variance components have been studied abundantly after the introduction of anova by Fisher (1925). Some of these ratios are of particular interest since they result in interpretable notions, such as intraclass correlation coefficient (ICC), heritability, heterogeneity. In general, the ICC is a correlation coefficient on clusters or groups of related statistical units (e.g. individuals, animals, samples). As explained in section 1.2.1, clusters may contain repeated or longitudinal measures. The concept of ICC was introduced and formulated by Fisher (1925, chapter 7). The work of Fisher (1925) on ICC was impacted by the work of Galton (1886) and Pearson (1896) on interclass correlation (Pearson’s correlation coefficient). ICC has its origin from genetics to calculate the resemblance between brothers in the same
“fraternity” or class, or to study the resemblance between leaves on the same tree (Fisher, 1925). The key idea of ICC is that the observations on individuals in the same class tend to be more alike than the observations on individuals from different classes. Later, in genetics the concept of heritability was formalized (Lush, 1940) which took its origin from Fisher (1919) and Wright (1920). Thereafter, the term heritability became common in genetics (biology) rather than the ICC. Extensive review on the concept and misconceptions of heritability is given by Visscher et al. (2008). Here, we briefly define both the broad and narrow sense heritabilities. The variance of the observable phenotypes \( \sigma^2_p \) are decomposed into unobserved underlying genetic \( \sigma^2_g \) and environmental \( \sigma^2_{env} \) variances, that is \( \sigma^2_p = \sigma^2_g + \sigma^2_{env} \). The genetic variance \( \sigma^2_g \) consists of the following three variance components: \( \sigma^2_a \) due to additive genetic effects, \( \sigma^2_d \) due to dominance genetic effects (interaction between alleles in the same locus), and \( \sigma^2_l \) due to epistatic genetic effects (interaction between alleles at different loci). Broad heritability is the proportion of variance due to genetic components in the total phenotypic variance, \( H^2 = \sigma^2_g / \sigma^2_p \) (Sham, 1998, p.212). Narrow heritability is the proportion of variance due to additive genetic component in the total phenotypic variance, \( h^2 = \sigma^2_a / \sigma^2_p \). Chapter 5 of the thesis studies the heritability.

The concept of ICC was extended to psychology, social and medical sciences (Cohen, 1960, Bartko, 1966, Fleiss, 1971, Fleiss and Shrout, 1978, Fleiss and Cuzick, 1979), and also to engineering (Lin et al., 2002, Burdick et al., 2005). New instruments, methods, tests, assays, manufacturing processes require to ensure that the measurements are reliable and accurate before accepting them in practice (Lin et al., 2002, Barnhart et al., 2007). Thus, it is important to assess closeness (agreement) of observations. ICC is frequently used to assess agreement of measurements made by multiple observers (e.g. doctors) on the same statistical unit (e.g. patients). Chapter 2 of the thesis employs ICC for agreement studies and provides further details on this topic.

The ICC is usually calculated as the proportion of variance unrelated to statistical units in the total amount of variation. Using model
(1.4), the general formulation of ICC is:

\[
\text{ICC} = \frac{\sum_{i=1}^{q} \sigma_i^2}{\sum_{i=1}^{q} \sigma_i^2 + \sigma_e^2},
\]

where \(\sigma_e^2\) is the variance component due to residual error and \(\sigma_i^2\)'s are the variance components due to main effects, interactions or nestings of certain factors influencing the outcomes.

Assume, the model contains only one random effects factor, \(q = 1\) in addition to random residual error (i.e. one-way random effects model). The correlation between two individuals \(j\) and \(l\) in cluster \(i\) is then equal to 

\[
\text{cor}(Y_{ij}, Y_{il}) = \frac{\text{cov}(Y_{ij}, Y_{il})}{\sqrt{\text{var}(Y_{ij})\text{var}(Y_{il})}} = \frac{\sigma_1^2}{(\sigma_1^2 + \sigma_e^2)} = \text{ICC}.
\]

The ICC can be interpreted as the correlation coefficient between any two members within the same cluster. The other interpretation of ICC is that it is the proportion of variance that is due to clustering in the total amount of variation. This interpretation of the ICC fits with the definition of \(I^2\) used to quantify heterogeneity across medical studies when meta-analysis is conducted (Takkouche et al., 1999, Higgins and Thompson, 2002). Heterogeneity in a meta-analysis for nonlinear and generalized mixed models are studied in Chapters 3 and 4 of the thesis, where more insight is given into the concept of heterogeneity.

### 1.2.4 Construction of confidence intervals for ratios of sums of variance components

Suppose \(\mathbf{Y} = (Y_1, \ldots, Y_n)\) has a joint distribution which depends on a parameter \(\theta\). Let \(L(\mathbf{Y})\) and \(U(\mathbf{Y})\) be two statistics where \(L(\mathbf{Y}) < U(\mathbf{Y})\). Assume \(\theta\) is the parameter of interest. A confidence interval is a random interval whose lower confidence bound \(L(\mathbf{Y})\) and upper confidence bound \(U(\mathbf{Y})\) are functions of the sample values such that:

\[
P[L(\mathbf{Y}) \leq \theta \leq U(\mathbf{Y})] = 1 - \alpha.
\]

The term \(1 - \alpha\) is the confidence coefficient to denote two-sided confidence interval. The confidence coefficient is typically selected prior to data sampling. In this thesis we study equal-tailed two-sided confidence interval with confidence coefficient 0.95. When exact confidence interval is not possible to find, an approximate 95% confidence
interval is used:

\[ P[L(Y) \leq \theta \leq U(Y)] \approx 0.95. \quad (1.12) \]

Knight (2000, Chapter 7) and Casella and Berger (2002, Chapter 9) provide more details on confidence intervals and their construction in general. Burdick and Graybill (1992) are focused on construction of confidence intervals for variance components models.

**Pivotal** and **delta** methods are two major approaches to construct confidence intervals. Pivotal methods may provide exact or approximate confidence intervals. Delta method which uses the Taylor series expansion is based on a large sample normal approximation. Thus, delta method is an approximate method. Detailed description of the delta method is given in Chapter 3 of the thesis where we derive approximate confidence interval for ICC.

The pivotal method is one of the basic approaches to find confidence intervals. The idea of the pivotal method is to find a random variable \( g(Y; \theta) \) whose distribution does not depend on \( \theta \), thus the distribution function \( P[g(Y; \theta) \leq y] = G(Y) \) is independent of \( \theta \). The random variable \( g(Y; \theta) \) is called an exact pivot for the parameter \( \theta \). We have to find constants \( a \) and \( b \) such that

\[ P[a \leq g(Y; \theta) \leq b] = 0.95. \quad (1.13) \]

After some mathematical operations the confidence interval is usually obtained. Since exact pivots are difficult to obtain, particularly for functions of variance components, therefore approximate pivots \( g(Y; \theta) \) are sometimes used such that

\[ P[a \leq g(Y; \theta) \leq b] \approx 0.95. \quad (1.14) \]

This means that the asymptotic distribution of the pivot is independent of the parameter \( \theta \). Approximate pivots are usually justified via asymptotic arguments. Below we provide two examples to demonstrate pivotal method.

**Example from Knight (2000, p. 343):** Assume \( Y_1, \ldots, Y_n \) are iid normal variables with unknown mean \( \mu \) and variance \( \sigma^2 \). To find a confidence interval for \( \mu \), define \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) and the ran-
dom variable
\[ \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \sim t_{n-1} \]  
(1.15)

has t distribution with \( n - 1 \) degrees of freedom; this distribution is independent of both parameters, \( \mu \) and \( \sigma^2 \) and therefore \( \sqrt{n}(\bar{Y} - \mu)/S \) is an exact pivot for \( \mu \). Afterwards we can obtain confidence interval for \( \mu \). If we divide both sides of equality with \( S^2 \) on \( \sigma^2 \), then we have:

\[ \frac{(n - 1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \chi^2_{n-1} \]  
(1.16)

and thus \( (1/\sigma^2) \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) is an exact pivot for \( \sigma^2 \). Afterwards we can obtain confidence interval for \( \sigma^2 \).

**Example from Searle et al. (2006, pp. 65-66):** Assume a one-way random effects model \( y_{ij} = \mu + u_i + \epsilon_{ij} \), applies to a set of observations \( y_{ij} (j = 1, 2, \ldots, n_i) \) from different clusters \( i = 1, 2, \ldots, I \) with \( I \) be the total number of clusters and \( N = \sum_{i=1}^{I} n_i \) be the total number of observations. We assume that the data are balanced, \( n_i = n \ \forall i \). Normality assumptions on random effects are required, for instance with mean \( E[u_i] = 0, E[\epsilon_{ij}] = 0 \) and variances \( \text{var}[u_i] = \sigma_1^2 \ \forall i, \) \( \text{var}[\epsilon_{ij}] = \sigma_e^2 \ \forall i \) and \( j \), in order to obtain the confidence interval for \( \sigma_1^2/\sigma_e^2 \) and then for ICC using pivotal method. Random terms are mutually independent. See further details in Searle et al. (2006, pp. 44-46).

Let us employ the method of moments. Under normality assumption, the independence of sums of squares for clusters \( SSU \) and residuals \( SSE \) have been established and these terms are proportional to a central \( \chi^2 \) distribution with degrees of freedom \( I - 1 \) and \( I(n - 1) \) respectively. Assumption of independence and following two properties,

\[ \frac{SSU}{(n\sigma_1^2 + \sigma_e^2)} \sim \chi^2_{I-1} \quad \text{and} \quad \frac{SSE}{\sigma_e^2} \sim \chi^2_{I(n-1)} \]

lead ratio \( \frac{MSU/(n\sigma_1^2 + \sigma_e^2)}{MSE/\sigma_e^2} \sim F_{I-1}^{I(n-1)} \) to \( F \) distribution with degrees of freedom \( I - 1 \) and \( I(n - 1) \). Extensive description of \( F \)-distribution is given by Johnson et al. (1995, p. 322).

Denoting the ratio of mean squares by F-statistic, \( F = MSU/MSE \), after mathematical operations the random variable:

\[ \frac{F \sigma_e^2}{n\sigma_1^2 + \sigma_e^2} \sim F_{I-1}^{I(n-1)} \]  
(1.17)

has a central \( F \)-distribution with degrees of freedom \( I - 1 \) and \( I(n - \)
1.2 Literature review

1); this distribution is independent of both parameters, $\sigma^2_1$ and $\sigma^2$ and therefore $F\sigma^2_e/(n\sigma^2_1 + \sigma^2_e)$ is an exact pivot for F-statistic. If we define upper and lower bound of $\mathcal{F}$-distribution by $F_U$ and $F_L$, then we have confidence interval for $F$-statistic:

$$P[F_L \leq \mathcal{F}_{I(n-1)} \leq F_U] = 0.95.$$  

(1.18)

Rearranging (1.17) leads to

$$\frac{\sigma^2_1}{\sigma^2_e} \sim \frac{F/\mathcal{F} - 1}{n},$$  

(1.19)

and thus using notation of lower $F_U$ and upper $F_L$ bounds, we have exact confidence interval $[F/F_U - n^{-1} F/F_L - n^{-1}]$ for $\sigma^2_1/\sigma^2_e$. Since ICC $= \sigma^2_1/(\sigma^2_1 + \sigma^2_e)$ can be rewritten as $(\sigma^2_1/\sigma^2_e)/(1 + \sigma^2_1/\sigma^2_e)$, we can also obtain an exact confidence interval on ICC, as follows:

$$\left[ \frac{F/F_U - 1}{F/F_U + n - 1}, \frac{F/F_L - 1}{F/F_L + n - 1} \right].$$  

(1.20)

Obviously, we reluctantly obtained confidence interval for ICC in a one-way random effects model with balanced data. Exact pivots are mostly available in simple situations with one or two unknown parameters, as shown in examples above. In unbalanced data, assumptions used in balanced data on independence of the mean squares are not valid any more. Since ratios of sums of variance components are complex functions, therefore it becomes even more complicated to obtain confidence intervals for such functions. Chapter 2 of the thesis studies this topic in depth. Major books on construction of confidence intervals for functions of variance components used in the thesis are Searle et al. (2006) and Burdick and Graybill (1992).

Bootstrap method introduced by Efron (1979) may be considered as alternative approximate pivotal method. The idea of bootstrapping is to assume that given study data plays a role of population. We have to draw random samples (resamples) repeatedly from the study data set to create bootstrap data sets of the same size as the study data. Many bootstrap data sets are generated on which the estimates are calculated. Often, percentile approach is used to obtain confidence intervals from these estimates. For dependent data the bootstrap method
is not particularly recommended (Chapter 3 of the thesis covers this topic in details).

1.3 Goal and specific objectives

A major goal of the thesis is to construct confidence intervals on ratios of sums of variance components that would work for linear, nonlinear and generalized linear mixed models in non-standard applications and complex structured data. Below we briefly outline the specific motivation and objectives of five chapters which are given in full text hereafter.

1. Confidence intervals for intraclass correlation coefficients in variance components models

Motivation and Objectives: Two major challenges are associated with the construction of confidence intervals for the ratios of sums of variance components in linear mixed effects models. One challenge is appropriate estimation of variance components within the non-negative parameter space. For instance, the method of moments allows negative estimates of variance components. Likelihood based estimates are non-negative, but can be biased. The second issue is the lack of an exact (closed-form) method for the construction of confidence intervals that could be generic for “any” variance components model. Satterthwaite’s approximation for construction of confidence intervals is a general approach, but it is not always applicable for unbalanced designs and it requires tedious calculations every time another model is required. The objective is to construct generic, closed-form methods for ICCs of the form \[ \frac{\sum_{q=1}^{Q} \sigma_{q}^{2}}{(\sum_{q=1}^{Q} \sigma_{q}^{2} + \sum_{p=Q+1}^{P} \sigma_{p}^{2})} \] for balanced and unbalanced linear, random and mixed effects models.

2. Confidence intervals for intraclass correlation coefficients in a nonlinear dose-response meta-analysis

Motivation and Objectives: To our knowledge, no generic approach exists to construct the confidence intervals for ICCs in nonlinear mixed effects models. There are only two suitable methods that can be given consideration. These are the bootstrap and large sample normal approximation (delta method). The bootstrap method is not straightforward for correlated data and the delta method does not apply to small samples. Estimation of variance components remains a challenge for nonlinear mixed effects models, since maximum likelihood underestimates the variance components and restricted maximum likelihood
1.3 Goal and specific objectives

does not exist. The objective is to determine how well the proposed
generic beta-approach would perform for construction of confidence
intervals on ICCs in nonlinear mixed effects models when different
methods of estimation are applied.

3. A generalized linear mixed model for meta-analysis of test-
negative design case-control studies

Motivation and Objectives: Test-negative design (TND) case-control
studies are a special type of observational studies for investigating in-
fluenza vaccine effectiveness. These studies are conducted yearly at
different places, but have never been combined in a meta-analysis.
Commonly used simple method of DerSimonian and Laird for esti-
mating the log odds ratios in case-control designs can be applied to
TND, but this method is biased when it is applied to two-by-two con-
tingency tables, especially in case of substantial heterogeneity across
studies. More sophisticated approaches could be applied, but none of
these approaches would analyze the proportion of cases across studies,
which is expected to change from study to study. The first objective
is to propose a generalized linear mixed model for a meta-analysis of
multiple TND studies that addresses all heterogeneity aspects. The sec-
ond objective is to apply the beta-approach to construct confidence
intervals on ICCs in a generalized linear mixed model to quantify the
uncertainty in heterogeneity.

4. The consequences of family structures in LifeLines cohort
study for BMI mediated health related quality of life scores

Motivation and Objectives: LifeLines is a three generation cohort
study. It is important to get insight to what extent the family structure
(correlation between family members) influences the associations be-
tween determinants and health outcomes. One objective is to study the
impact of relatedness in a family on the association of body mass index
(BMI) mediated health related quality of life (HRQoL) scores. Another
objective is to estimate the contribution of variances in outcomes due
to heredity and shared environment.

5. Probability genotype imputation method and integrated
weighted lasso for QTL identification

Motivation and Objectives: The phenotype of a quantitative trait is
the cumulative result of several genes, their interactions and the en-
vIRONMENT. Genome regions that contain genes associated with a par-
ticular quantitative trait are known as quantitative trait loci (QTL). The primary biological goal is to identify the QTL associated with variation in traits and the eventual goal is to improve the quality of seed production in Arabidopsis thaliana. Many QTL studies have two common challenges. One challenge is that there is missing marker information and another one is that among many markers involved in the biological process only a few are causal. To increase the accuracy of QTL identification, the objective is to develop a methodology that makes use of the information from immediate neighbouring markers of a marker with missing genotype and maps QTL by incorporating this information.

1.4 Structure of the thesis

In addition to introduction and discussion, this thesis includes five chapters. In Chapter 2 we give an extensive overview on challenges of estimation and construction of confidence intervals for ratios of sums of variance components in linear mixed models. In this chapter we propose two generic methodologies (F-approach and beta-approach) for construction of confidence intervals on the example of ICC that overcome existing challenges. In Chapter 3 we show that our introduced beta-approach can be successfully employed to construct confidence intervals for heterogeneity in nonlinear mixed models. In Chapter 4 we introduce a generalized linear mixed model for the meta-analysis of TND case-control studies and demonstrate that the beta-approach works well on ICC again (with categorical outcome). In Chapter 5 we employ several linear mixed models to study correlated data within families. We assessed the impact of heritabilities on outcomes. In this chapter we again used the beta-approach to construct confidence intervals for heritabilities. In Chapter 6 we propose likelihood-based imputation and estimation methods for sparse variable (genetic marker) selection in the context of plant breeding.
Bibliography


