Essays on Inference in Linear Mixed Models

Dissertation

for the award of degree
"Doctor rerum naturalium" (Dr.rer.nat.)
of the Georg-August-Universität Göttingen

within the doctoral program
"Mathematical Sciences"
of the Georg-August-University School of Science (GAUSS)

submitted by

Peter Kramlinger

from Buenos Aires, Argentina

Göttingen, 2020

Thesis Committee:

Prof. Dr. Tatyana Krivobokova
Institute for Mathetmatical Stochastics, University of Göttingen

Prof. Dr. Thomas Kneib
Centre for Statistics, University of Göttingen

Prof. Dr. Stefan Sperlich
School of Economics and Management, University of Geneva

Members of the Examination Board:

Reviewer:
Prof. Dr. Tatyana Krivobokova
Institute for Mathematical Stochastics, University of Göttingen

Second Reviewer:
Prof. Dr. Stefan Sperlich
School of Economics and Management, University of Geneva

Further Members of the Examination Board:

Prof. Dr. Stephan Huckemann
Institute for Mathematical Stochastics, University of Göttingen

Prof. Dr. Ingo Witt
Mathematical Institute, University of Göttingen

Jun.-Prof. Dr. Christoph Lehrenfeld
Institute for Numerical and Applied Mathematics, University of Göttingen

Dr. Michael Habeck
Institute for Mathematical Stochastics, University of Göttingen

Date of the oral examination: April 28th, 2020

Acknowledgments

I want to thank my advisory committee for the guidance I received over the past three and a half years. During the supervision and work with Prof. Tatyana Krivobokova I have found my passion in scientific research. Further, I am thankful for the fruitful discussions with Prof. Thomas Kneib and Prof. Stefan Sperlich from which this research greatly benefited.

I am grateful to the members of the Institute of Mathematical Stochastics. The working environment, your collegial attitude and the presence of friends made this doctorate particularly enjoyable.

I thank my family, Ursula and Franz, Anna, Ruth and Michael. Your dedicated and relentless support has made this thesis possible.

I am especially grateful to Christopher and Julia. Your encouragement has motivated me to pursue and complete this doctorate.

Finally, I thankfully acknowledge the support of my project by Hanns-Seidel-Foundation and Research Training Group 1644 ‘Scaling Problems in Statistics’.

Preface

Linear mixed models (LMMs) are both mathematically intriguing and useful in practice. This dissertation aims to establish two aspects of statistical inference in such models. Those lead to confidence sets for unknown parameters that can be extended for testing statistical hypotheses in various testing scenarios.

It is based on the articles given in Addenda A and B. Devised from different ideas on how to interpret the model components in the field of 'small area estimation' (SAE), the former addresses the issue of constructing confidence sets for mixed parameters. The latter uses the particular LMM estimation methodology to adequately account for additional uncertainty induced by selecting model coefficient parameters.

This document is structured as follows. First, Chapter 1 introduces the model, its fundamental advantages and properties and motivates the aspects of inference that were investigated. Next, Chapters 2 and 3 discuss both of these separately. In each, the specific underlying problem is explained and the main results presented.

The scientific contribution to this dissertation is given in the addenda. In both articles I derived the results, designed the proofs and simulation studies. Especially the presentation, structure, phrasing, data set and motivation were the joint work of all authors.

The main body of the present text introduces and discusses the underlying problems and is merely meant to give a comprehensive overview of the topic. A thorough literature review, the rigorous model definitions, assumptions, theorems, proofs, examples, discussion and outlook is given in the respective article.

Contents

Chapter 1

Introduction

It is no surprise that mixed model methodology has been found a powerful tool in a variety of empirical sciences. Habitually, regression analysis is introduced to a layman with the assumption that observed data stems from independent and identically distributed random variables. Mathematically the implications of such conditions are well understood. Even if they may not be present in real life application, their notional premise may still serve certain tasks adequately. Yet for many research questions, in particular those that involve grouped observations, their elementary structure proves to be too restrictive. Mixed models on the other hand offer the additional flexibility to overcome those limitations, whilst preserving the mathematical simplicity of the classical approach.

Their different nature is readily understood in their simplest formulation. Consider the following model equation for $i = 1, \dots m$:

$$\begin{gathered}\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{v}_i + \mathbf{e}_i, \\ \mathbf{e}_i \overset{ind.}{\sim} \mathcal{N}_{n_i}\left(\mathbf{0}_{n_i}, \boldsymbol{\Omega}_i\right), \quad \mathbf{v}_i \overset{ind.}{\sim} \mathcal{N}_q\left(\mathbf{0}_q, \boldsymbol{\Psi}\right).\end{gathered} \tag{1.1}$$

One observes the response $\mathbf{y}_i \in \mathbb{R}^{n_i}$ and matrices $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ as well as $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$. The vector of coefficient parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and the independent random vectors $\mathbf{v}_i \in \mathbb{R}^q$ and $\mathbf{e}_i \in \mathbb{R}^{n_i}$ are unknown. Further, the covariance matrices $\boldsymbol{\Omega}_i \in \mathbb{R}^{n_i \times n_i}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{q \times q}$ as well as the sample sizes $m, n_i \in \mathbb{N}$, $i = 1, \dots, m$ and the dimensions $p, q \in \mathbb{N}$ are known.

In the classical linear model one has $\mathbf{Z}_i = \mathbf{0}_{n_i \times q}$, so that $\mathbf{y}_i$ is driven solely by the unknown, fixed coefficient parameters $\boldsymbol{\beta}$. For linear mixed models on the other hand, $\mathbf{Z}_i$ has non-zero entries, so that the fixed effects are complemented by a term of random effects.

Most importantly, these random effects model the presence of m groups in the data. All observations $\mathbf{y}_i$ from the i-th group are driven by the same realization of the random effect $\mathbf{v}_i$. In his monograph, [9] provides a suitable toy example. Consider a fictional data set of profits versus sales for certain goods visualized in Figure 1.1. Ignoring the fact that up to three pairs of observations are from the same commodity, a naïve application of the classical linear model suggests a negative relationship between sales and profit. A more convincing argument is made by evaluating the average of group-wise regression lines. Note that this does not imply that each group is treated separately. It is an inherent feature of mixed model methodology that in the course of prediction of the random effects $\mathbf{v}_1, \ldots, \mathbf{v}_m$ the overall population is used to borrow strength for each group prediction.

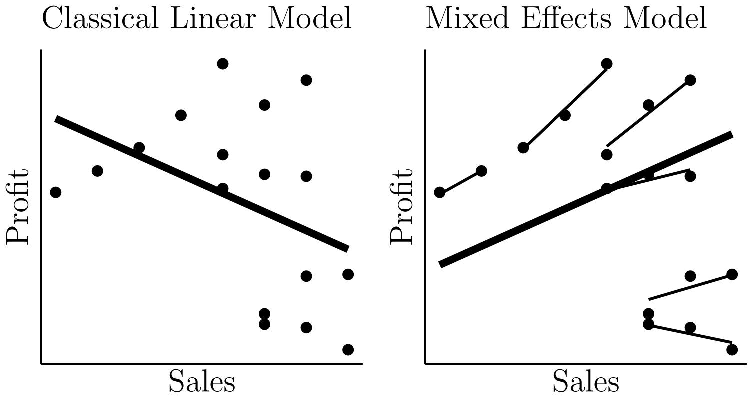

Figure 1.1: Fitting a classical linear model to grouped data may not capture their information adequately. Whereas the average regression of groups shows a positive correlation of profit and sales (right), the naïve approach indicates a negative relationship (left).

In the example above the group-wise regression lines are shrunken towards the overall mean as admissible estimators are not unbiased under quadratic loss [18, 10]. However, the amount of shrinkage remains unclear at first. Mixed models interpret the group-wise deviations from the overall mean as realizations from random variables, and hence determine the shrinkage as relative size of random effect versus error variance. Thus, the amount of shrinkage, and the interpretation of group-wise deviations as stochastic, have their own decision theoretic justification. But it is crucial to note, that, in the words of Nicholas Longford, this randomness assumption may 'merely [be] a device that enables a more natural application of a general principle that should be employed, or at least considered, universally' [25, pp. 175-176].

Even though the random effects $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are treated as stochastic, this does not imply that they are interpreted as such. But when it comes to assess the precision of their estimates, an elaboration on the true nature of the underlying random effects is crucial. If they really are seen as stochastic in practice, mixed model methodology can be applied for inference. When they are in fact understood to be fixed parameters, only treated as random to obtain shrinkage estimators in the first place, then inference has to be performed conditional on the realizations $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

This approach raises new questions on how confidence sets or testing procedures have to be constructed. Direct estimators, that do not borrow strength, suffer from large variability, which results in prohibitively large confidence sets. Borrowing strength however results in a bias, in particular for 'interesting' groups, see the example in Section 4 of Addendum A. In Chapter 2 a choice is discussed that is based on considering multiple groups simultaneously.

In the introductory example it has been made evident that even in the most basic cases, a misspecified model may cause a fallacy in interpretation. Of course, this misspecification can happen in various ways. Failing to include relevant coefficient parameters may hurt the models predictive power. Including parameters that contribute similar information on the other hand may lead to confounding. In these cases model selection is understood as selection of coefficient parameters $\boldsymbol{\beta}$. This task is often carried out prior to estimation and inference. However, the latter is disrupted if parameter selection relies on a stochastic process, such as cross validation, information criteria, or even eyeballing. All those methods are data dependent and thus stochastic. The additional uncertainty that these methods produce has to be accounted for. In Chapter 3, this problem is addressed in the context of linear mixed models.

The above described problems have made clear that the questions considered are by no means purely theoretical. Mixed models are widely applied, and their ability to treat grouped data is required in a variety of fields. In particular, these groups may represent clustered data, as in the example. In that case, sampling issues are the reason why groups emerge. This justification of mixed models is arguably the most basic one, and covered by the field of small area estimation, which is discussed in Chapter 2 as well. But, as it has been introduced above, the mixed model may be interpreted as an hierarchical, two-level setup as well. The first level models the distribution of the error terms, whereas the second level plays the role of a prior distribution on the random effects. This allows a Bayesian view on mixed model methodology, which also will be briefly considered in Chapter 2. But these motivations for the mixed model are not exclusive, nor can its usage in a specific case be uniquely attributed to one motivation only.

Consider the following example from the field of animal sciences, which was also the motivation for Charles Henderson to formulate mixed models in the first place [15, 16]. The quantity of interest is the milk production of dairy cows. It is of interest to obtain the breeding value of a bull. Obviously, although he may pass on unobserved factors that determine the milk production of his next generation, he himself does not possess any ability to produce milk. To account for his latent ability to inherit such, it may be modeled by a random effect. It is noteworthy, that this application of mixed models is not merely tool to obtain an amount of shrinkage, but a in itself justified model choice. For this particular model, it does not make sense to evaluate it conditional on the random effects. This is due to the fact that the bulls own milk producing capability is of no interest.

Another example from economics is given in Addendum A. Based on data from the Spanish survey of living conditions of 2008, interest lies in the relation between income and a panel of auxiliary variables across groups formed by a cross-section of Spain's fifty provinces and whether secondary education was completed. Now, as specifically the group deviations are of interest, inference for such a research question has to be performed conditional on the random effects.

In the former example, mixed models served to account for different sources of variation. Besides genetics, it is used in ecology to model biological heterogeneity. In the latter one, the mixed model is used to borrow strength for each group specific estimate. Problems like these are part of 'small area estimation', which is discussed in the next chapter.

The vast available literature on mixed models is rooted in its various applications. A broad overview is provided by the monographs [33, 9, 28] on that subject. After all, in the words of Eugene Demidenko: ‘Mixed model methodology brings statistics to the next level’ [9, p. 1].

Chapter 2

Marginal and Conditional Multiple Inference

2.1 Small Area Estimation

In the introductory example of the previous Chapter 1 it has been argued that mixed models 'borrow strength' from the whole population to obtain more reliable group-specific estimators. This effect is particularly prominent when the group sample sizes are very small. In 1988, George Battese, Rachel Harter and Wayne Fuller examined the soy and corn production for selected counties in north-west Iowa [1]. The groups for each crop were constituted by geographical criteria. The data set contained only up to six observations per county. Thus, a weighted average between the direct, i.e. county specific, 'regression-synthetic' estimator and 'survey regression' estimator, that considers the whole population, served as a method to obtain reliable 'composite' estimators for each county than direct estimators alone. Their novel approach sparked new research on the topic of information scarcity amongst groups, called 'small areas'. Even today, in times of electronic data processing, the gathering of larger samples is often prohibitively expensive. This is why the idea of borrowing strength remains attractive today, in particular in the framework of mixed models. Extensive reviews of current research on the subject are given in [27] or in the monographs by [31, 25].

Suppose that the vector of parameters of interest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^t$ is a linear combination of fixed and random effects $\mu_i = \mathbf{l}_i^t \boldsymbol{\beta} + \mathbf{h}_i^t \mathbf{v}_i$ for $\boldsymbol{\beta}$ and $\mathbf{v}_i$ as in (1.1), where $\mathbf{l}_i \in \mathbb{R}^p, \mathbf{h}_i \in \mathbb{R}^q$ are known, $i = 1, \ldots, m$. For example, if $\mathbf{l}_i^t = \mathbf{X}_i^t \mathbf{1}_{n_i}/n_i$ and $\mathbf{h}_i^t = \mathbf{Z}_i^t \mathbf{1}_{n_i}/n_i$, then μ_i is the conditional mean for group i. Recall that the initial motivation to turn to shrinkage estimators in the first place was to obtain more reliable estimators. In mixed model terminology, bias and variance of the target estimator have to comply with a certain optimality criterion. Under quadratic loss, the criterion is equivalent with minimizing the mean squared error (MSE) or, acknowledging its random component the mean squared prediction error (MSPE) [27]. The resulting linear estimator for $\boldsymbol{\mu}$ is called 'best linear unbiased predictor' (BLUP) $\widetilde{\boldsymbol{\mu}} = (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_m)^t$, a term coined by Charles Henderson [17]. A formal definition is deferred to Addendum A. Another estimator is given as the 'best predictor' $\mathrm{E}(\boldsymbol{\mu} \,|\, \mathbf{y})$. Under normality, both estimators coincide [31]. It is difficult to identify the distribution of the random effects, so that analytic model based small area estimation literature almost exclusively relies on normality assumptions. Distribution free approaches are manifold, but often rely on re-sampling techniques so that they lack analytic representations for inference. A comprehensive review on such methods is given by [4, 27].

For each group $i = 1, \ldots, m$, the BLUP minimizes $\mathrm{MSE}_{\mu_i}(\widetilde{\mu}_i) = \mathrm{E}(\mu_i - \widetilde{\mu}_i)^2$, and thus serves as adequate shrinkage estimator that borrows strength from other groups. Its analytic expression is readily available. However, recall from the introductory example that the amount of shrinkage is determined by the relative size of random effects versus error variance, both of which are generally unknown in practice. Denote the vector of covariance parameters as $\boldsymbol{\delta} \in \mathbb{R}^r_{>0}$, $r \in \mathbb{N}$, and let $\boldsymbol{\Omega}_i(\boldsymbol{\delta})$, $i = 1, \ldots, m$, and $\boldsymbol{\Psi}(\boldsymbol{\delta})$ from (1.1) be known matrices depending on the unknown vector $\boldsymbol{\delta}$. They may be estimated either by the method of maximum likelihood, adjusted for the loss in degrees of freedom in estimating the fixed effects $\boldsymbol{\beta}$ and then called restricted maximum likelihood (REML), or by the method of least squares, named Hendersons method III for unbalanced data sets, which do not require a distributional assumption [33]. Denote an estimator based on any of these methods as $\widehat{\boldsymbol{\delta}}$. Plugging the estimates of the covariance parameters into the BLUP then gives the empirical BLUP (EBLUP), which will be denoted as $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_m)^t$.

The EBLUP $\widehat{\boldsymbol{\mu}}$ has been developed on the basis that it is more robust than direct estimators, which suffer high variability due to small sample sizes. It is thus an adequate shrinkage estimator to treat small areas, irrespective of whether the realizations of the random effects are of interest or not. The precision of BLUP and EBLUP can be assessed by evaluation of the MSE, which in turn depends on the estimated covariance parameters. For a specific group $i = 1, \ldots, m$, it is given as $\mathrm{MSE}_{\mu_i}(\widetilde{\mu}_i) = g_{1,i}(\boldsymbol{\delta}) + g_{2,i}(\boldsymbol{\delta})$. Here, $g_{1,i}(\boldsymbol{\delta})$ is a known function quantifying the variability induced by the estimation of the random effects and $g_{2,i}(\boldsymbol{\delta})$ by the fixed effects. For the EBLUP, the estimation of $\boldsymbol{\delta}$ by $\widehat{\boldsymbol{\delta}}$ has to be taken into account. The additional variability is given by $g_{3,i}(\boldsymbol{\delta})$, so that $\mathrm{MSE}_{\mu_i}(\widehat{\mu}_i) = g_{1,i}(\boldsymbol{\delta}) + g_{2,i}(\boldsymbol{\delta}) + g_{3,i}(\boldsymbol{\delta})$. A naïve estimator for the MSE of $\widehat{\mu}_i$ is given by simply plugging in the estimated covariance parameters: $\widehat{\mathrm{MSE}}_{\mu_i}(\widehat{\mu}_i) = g_{1,i}(\widehat{\boldsymbol{\delta}}) + g_{2,i}(\widehat{\boldsymbol{\delta}}) + g_{3,i}(\widehat{\boldsymbol{\delta}})$. Problematically however, the bias of $g_{1,i}(\widehat{\boldsymbol{\delta}})$ is of the same order as $g_{2,i}(\widehat{\boldsymbol{\delta}})$ and $g_{3,i}(\widehat{\boldsymbol{\delta}})$. The explicit formulation of the functions $g_{1,i}$, $g_{2,i}$ and $g_{3,i}$ are given in [31] or Addendum A. A second order approximately unbiased estimator for the MSE, the so-called 'Prasad-Rao'-estimator was subsequently developed [30, 8, 5].

Under marginal law, that is when both errors and random effects are stochastic, and light regularity conditions [20], the EBLUP is unbiased, i.e. $\mathrm{E}(\widehat{\mu}_i) = \mu_i$. Since further $\mathrm{MSE}_{\mu_i}(\widehat{\mu}_i) = \mathrm{Var}(\widehat{\mu}_i - \mu_i)$, the Prasad-Rao estimator may serve to construct pointwise confidence intervals for single small area estimates [5, 4, 13]. Interpreting the random effect as purely stochastic, the described methods are sufficient to derive suitable shrinkage estimators, and to establish area specific inference and testing.

2.2 Conditional Inference

As in the introductory example of the previous Chapter 1 the presence of groups in the data required more elaborate estimation techniques than just direct estimators. The EBLUP minimizes the MSE under joint distribution of errors $\mathbf{e}_i$ and random effects $\mathbf{v}_i$, $i = 1, \ldots, m$, henceforth called marginal MSE, and is thus a Bayes estimator for μ_i under quadratic loss. This is a justification for the choice of estimator irrespective of its interpretation. It may be applied as a suitable estimator whether the group-wise deviations are perceived as stochastic or fixed.

Under the marginal law as stated in (1.1), the group-wise deviations follow a distribution with mean zero. The population mean is driven by the fixed effects. Therefore the BLUP, and EBLUP [14], are unbiased under marginal law. If interest lies however in the group-wise deviations, the underlying distribution must be taken as conditional on the random effects $\mathbf{v} = (\mathbf{v}_1^t, \ldots, \mathbf{v}_m^t)^t$. Under this conditional law, the EBLUP is – oxymoronicly – biased: $\mathrm{E}(\widehat{\mu}_i - \mu_i | \, \mathbf{v}) \neq 0$.

Furthermore, whereas the MSE equals the marginal variance of the estimator, it does not do so for the conditional variance, $\mathrm{MSE}_{\mu_i}(\widehat{\mu}_i) \neq \mathrm{Var}(\widehat{\mu}_i - \mu_i | \, \mathbf{v})$. The latter only depends on the variation induced by the errors, the former additionally on the variation of the random effects.

Both quantities, conditional bias and variance, are required to construct confidence intervals to perform conditional inference. But although the conditional variance can be calculated by similar means to the marginal one, see Addendum A, the bias cannot be treated with ease. Due to the small sample sizes, the direct bias estimates come with a prohibitive large variance, rendering the previous application of shrinkage estimators obsolete [19, 25, 27].

Although both issues, the conditional interpretation and its insufficient direct methods for inference were previously noted [22, 27], they have not been treated. This is even more surprising, as ignoring the misspecification results in confidence sets that do not meet nominal level [6, 7]. The effect of undercoverage is most pronounced for large deviations so that confidence intervals for groups that stand out, and for which a researcher might be particularly interested about, may be grossly misplaced. This behaviour was noted in [19]. A compact example is also provided in Addendum A.

In conclusion, group-wise confidence intervals under conditional law do not appear useful in practice, in contrast to the respective counterparts under marginal law in the previous section. However, different approaches are motivated by the phenomenon described by Grace Wahba for smoothing splines confidence intervals, namely that although they do not attain the nominal level individually they do so in average [37, 26]. Similarly, the consideration of multiple groups simultaneously under conditional law is promising, which leads to the results in the next section.

2.3 Main Results

The results obtained on the present subject are published in the article in Addendum A. It is of interest to establish multiple marginal and conditional inference for a mixed parameters vector $\boldsymbol{\mu}$, where m is the number of groups, and μ_i a linear combination of the random effect from the i-th group.

As first main result, for the marginal case, confidence sets and testing procedures that involve multiple groups are developed. In order to do so, an estimate $\widehat{\boldsymbol{\Sigma}}$ for the variance-covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{\mu})$ is derived by similar means of [30, 8]. The covariance matrix $\boldsymbol{\Sigma}$ has off-diagonal entries of order $O(m^{-1})$ and so it is crucial to verify that the second order bias corrected estimator $\widehat{\boldsymbol{\Sigma}}$ is precise enough to allow for m multiple comparisons. This is indeed confirmed in Theorem 2, which describes an m-dimensional confidence ellipsoid with coverage approaching nominal level with an error of order $O(m^{-1/2})$. As a supplementary result, it is further shown that the bias correction of $\widehat{\boldsymbol{\Sigma}}$ is actually $O(m^{-2})$ instead of $O(m^{-3/2})$, which was derived by [30]. This however does not improve the rate in Theorem 2, which also depends on the variance of the entries of $\widehat{\boldsymbol{\Sigma}}$.

For multiple conditional inference, two competing approaches are presented. First, the conditional covariance matrix $\boldsymbol{\Sigma}_c = \mathrm{Cov}(\boldsymbol{\mu} \,|\, \mathbf{v})$ is estimated by a second order bias corrected estimator $\widehat{\boldsymbol{\Sigma}}_c$, where the approach of [30] requires the treatment of additional terms. Furthermore, the bias $\lambda = \|\mathrm{E}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu} \,|\, \mathbf{v})^t \, \boldsymbol{\Sigma}_c^{-1/2} \, \|^2$, where $\| \cdot \|$ refers to the Euclidean norm, is estimated. Then, Theorem 1 describes an confidence ellipsoid with coverage approaching nominal level with an error of order $O(m^{-1/2})$, which coincides with the rate from the marginal case.

The second approach, Theorem 3, simply evaluates the marginal confidence ellipsoid under conditional law. Remarkably, the resulting coverage also attains nominal level up to an error of order $O(m^{-1/2})$. This phenomenon occurs as the misspecification of the bias for each group and the oversized variance cancel each other out in average. However, the rate in this case requires that the number of comparisons grows with m.

These results serve for the construction of confidence sets, and by inverting them also for the use of testing linear hypotheses.

Additionally, Theorem 4 lays the basis for different testing scenarios that may be helpful in practice. With Tukey's method [36, 35], all simple contrasts can be tested against, i.e. $H_0 : \mu_i = \mu_j$ for all $i, j \in S$ vs. $H_1 : \mu_i \neq \mu_j$ for at least one pair $i, j \in S$, where $S \subsetneq \mathbb{N}_{\leq m}$.

In total, all these results on marginal and conditional multiple inference are completely novel. They fill a relevant gap in the application of mixed models, as they justify a wider understanding than their narrow mathematical formulation suggests. Their usability is confirmed with an extensive simulation study. A real data set on Spanish income data gives an example how these theoretical results can be put to practice.

Chapter 3

Uniformly Valid Inference Based on the Lasso

3.1 Post Selection Inference

The motivating example from Chapter 1 discussed the association between sales and profit for certain commodities. The key message was that for an inadequately chosen model, one may fail to correctly identify the relation between response and covariates. In many real life applications, the research question is not so precisely posed. Often, many covariates are available to include within a model. Here, we only focus on this part of model selection that considers the process of deciding on a set of covariates which are to be included in the model. This understanding implies that the model equation is seen merely as a description of the association between observations and covariates. To highlight the problem of selecting the fixed coefficient parameters consider the alternative representation of model (1.1), given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_n\big\{\mathbf{0}_n, \mathbf{V}(\boldsymbol{\theta})\big\}, \tag{3.1}$$

with $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ and where $\mathbf{V}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times n}$ is a block diagonal covariance matrix with $(n_i \times n_i)$-blocks $\mathbf{Z}_i\,\boldsymbol{\Psi}(\boldsymbol{\theta})\,\mathbf{Z}_i^t + \boldsymbol{\Omega}_i(\boldsymbol{\theta})$. Deviating from the choice of notation from the previous chapter, but consistent with the notation in Addendum B, the covariance parameters are denoted as $\boldsymbol{\theta} \in \Theta \subsetneq \mathbb{R}^r$, $r \in \mathbb{N}$. This model equation is not seen as a data generating mechanism, in which the covariates exert a causal effect on the observations. Whereas in the latter understanding the inclusion of all true coefficient parameters $\boldsymbol{\beta}$ is crucial to obtain the underlying model, no such thing exists in the former case. Any model with a selected set of certain coefficient parameters may be justified from the point of a researcher. However, the classical analytic approach of estimating $\boldsymbol{\beta}$ via least squares (LS) relies on $\text{rank}(\mathbf{X}) = p \leq n$, requiring selection if more than n covariates are available. Moreover, the effect of a single coefficient within a model is expressed in terms of all other coefficients. Hence, to adequately address a single effect on the observations, a researcher might be generally interested to describe the effect with a single coefficient, to avoid confounding covariates. An extensive and insightful discussion on this problem is given in [2].

Generally, the process of model selection is performed under the principle of 'Occams razor', which postulates that amongst a set of candidate models, the simplest one is to be adopted. Superfluous complexity, in terms of coefficient covariates, is to be cut off. To find a parsimonious model, the model fit, expressed in its likelihood, is to be weighted against the number of coefficient parameters. Many such 'information criteria' have been derived on this basis. An extensive overview is provided by [3]. The fundamental problem is that those model selection techniques itself are necessarily data dependent. But since the observations are stochastic, so is any procedure that considers the model fit.

For a chosen and fitted model, one may infer about the included and underlying $\boldsymbol{\beta}$ on the basis of the estimated coefficient parameters. With classical theory one can construct confidence regions for $\boldsymbol{\beta}$, or by inverting those, derive testing procedures of interest. This is different if the model is selected by one of the established information criteria. The model is then selected based on its fit, meaning that it consists of covariates strongly related to the response. Subsequent testing for the coefficient parameters will make them appear more significant then they actually are, as the model is chosen so that they are strongly associated in the first place. For an included coefficient β_i consider the test H_0: $\beta_i = 0$ against H_1: $\beta_i \neq 0$. Then, the type-I-error $P_{\beta_i=0}(\text{reject } H_0)$ is larger compared to what the classical theory postulates [2]. See the simulation example in Section 5 in Addendum B for an visualization of this effect. Hence, classical confidence sets based on LS estimators after model selection exhibit a lower coverage than the nominal level indicates.

Recent interest has been focused on the issue of correctly quantifying the uncertainty induced by the model selection step, coined ‘post-selection inference’ (PoSI) [2]. The suggested workarounds however are either conservative by nature [2] or are conditional on the chosen model and thus not precisely what is understood to be a classical confidence set [21].

3.2 Inference Based on Penalization Methods

The problem of post-selection inference arises by the two-step nature of model fitting. The least absolute shrinkage and selection operator (Lasso) introduced by [34] is a single step procedure that selects and estimates the model coefficient parameters simultaneously. Its application thus bypasses the issue of post-selection inference. For model (3.1) and given tuning parameters $\lambda_1, \dots, \lambda_p \in \mathbb{R}$ consider the objective function

$$Q\left\{\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})\right\} = \ln|\mathbf{V}(\boldsymbol{\theta})| + \left\|\mathbf{V}(\boldsymbol{\theta})^{-1/2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right\|^2 + 2\sum_{j=1}^{p} \lambda_j \left|\beta_j\right|.$$

For the classical linear Gaussian regression model with $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $(n \times n)$-dimensional identity matrix, the Lasso for the coefficient parameters is defined as $\widehat{\boldsymbol{\beta}}_L = \operatorname{argmin}_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}, \mathbf{I}_n)$. The ℓ_1-penalization term ensures that in absolute value small coefficient parameters are shrunken to zero, and hence excluded from the model, whereas large ones are included. At the cost of this shrinkage towards zero, depending on $\lambda_1, \dots, \lambda_p$, the procedure simultaneously selects and estimates the coefficient parameters.

However, the shrinkage also results in the Lasso to be biased, see [12]. Hence the distribution of $\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}$ is shaped by the underlying coefficient parameters $\boldsymbol{\beta}$ [29]. This is in contrast to classical LS estimation. Therefore, different to inference based on LS estimators, pointwise confidence sets for fixed $\boldsymbol{\beta}$ based on the Lasso are not honest in the sense of [24]. Honest confidence sets have to achieve nominal level uniformly over the whole space of coefficient parameters [23, 29].

For a classical linear Gaussian regression model, [11] showed that limiting versions $\lim_{\boldsymbol{\beta}\to\pm\infty} Q(\boldsymbol{\beta}, \mathbf{I}_n)$ can be used to construct confidence sets based on the Lasso estimator. The resulting sets hold uniformly over the whole space of coefficient parameters.

3.3 Main Results

The contribution in Addendum B covers the construction of uniformly valid confidence sets for Lasso in LMMs. In contrast to the linear regression case, the estimation of covariance parameters has to be taken into account. The Lasso depends on the underlying covariance parameters, so the joint simultaneous estimation of both parameter vectors via

$$\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\theta}}\right) = \operatorname*{argmin}_{\boldsymbol{\beta},\boldsymbol{\theta}} Q\left\{\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\theta})\right\}$$

makes the confidence set for $\widetilde{\boldsymbol{\beta}}$ depend on $\widetilde{\boldsymbol{\theta}}$ in a complicated manner [32]. In linear regression with covariance matrix $\sigma^2\mathbf{I}_n$ with unknown variance parameter σ^2, this problem can be avoided by choosing the tuning parameters accordingly [11].

If the covariance parameters are of dimension $r > 1$, as usually considered in LMMs, one may exploit the method of restricted maximum likelihood (REML). This estimation method for the underlying covariance parameters $\boldsymbol{\theta}$ considers the loss in degrees of freedom in estimating the true coefficient parameters $\boldsymbol{\beta}$. The resulting estimator $\widehat{\boldsymbol{\theta}}$ is not only unbiased, but also based solely on transformed data $\mathbf{A}^t\mathbf{y}$ for a matrix $\mathbf{A} \in \mathbb{R}^{n\times(n-p)}$ such that $\mathbf{A}^t\mathbf{X} = \mathbf{0}_{(n-p)\times p}$. Hence, $\widehat{\boldsymbol{\theta}}$ does not depend on $\boldsymbol{\beta}$. Now, the Lasso for the LMM is defined as

$$\widehat{\boldsymbol{\beta}}_L = \operatorname*{argmin}_{\boldsymbol{\beta}} Q\left\{\boldsymbol{\beta}, \mathbf{V}(\widehat{\boldsymbol{\theta}})\right\},$$

and for this estimator similar arguments as for the case of linear regression can be applied.

Then, Theorem 1 in Addendum B states that confidence sets based on $\widehat{\boldsymbol{\beta}}_L = \operatorname{argmin}_{\boldsymbol{\beta}} Q\{\boldsymbol{\beta}, \mathbf{V}(\widehat{\boldsymbol{\theta}})\}$ are uniformly valid over the space of coefficient parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\theta}$ up to an error vanishing with parametric rate. The error is induced by the estimation of the covariance parameters. To prove this result, it has been shown in Lemma 1 that the REML estimator $\widehat{\boldsymbol{\theta}}$ is uniformly consistent for $\boldsymbol{\theta}$. The results are backed up with a simulation study that visualizes the uniform nature of the resulting confidence set and its superiority to naïvely chosen ones.

Bibliography

[1] G. E. Battese, R. M. Harter, and W. A. Fuller. An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83:28–36, 1988.

[2] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid Post-Selection Inference. *Annals of Statistics*, 41(2):802–837, 2013.

[3] K. Burnham and D. Anderson. *Model Selection.* Springer, New York, NY, 2002.

[4] S. Chatterjee, P. Lahiri, and H. Li. Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models. *The Annals of Statistics*, 36(3):1221–1245, 2008.

[5] K. Das, J. Jiang, and J. N. K. Rao. Mean Squared Error of Empirical Predictor. *The Annals of Statistics*, 32(2):828–840, 2004.

[6] G. S. Datta, M. Gosh, D. D. Smith, and P. Lahiri. On the Asymptotic Theory of Conditional and Unconditional Coverage Probabilities of Empirical Bayes Confidence Intervals. *Scandinavian Journal of Statistics*, 29:139–152, 2002.

[7] G. S. Datta, T. Kubokawa, I. Molina, and J. N. K. Rao. Estimation of Mean Squared Error of Model-Based Small Area Estimators. *TEST*, 20:367–388, 2011.

[8] G. S. Datta and P. Lahiri. A Unified Measure of Uncertainty of Estimated Best Linear Predictors in Small Area Estimation Problems. *Statistica Sinica*, 10:613–627, 2000.

[9] E. Demidenko. *Mixed Models: Theory and Applications.* Wiley Series in Probability and Statistics, Hoboken, NJ, 2004.

[10] B. Efron and C. Morris. Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[11] K. Ewald and U. Schneider. Uniformly Valid Confidence Sets Based on the Lasso. *Electronic Journal of Statistics*, 12:1358–1387, 2018.

[12] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *JASA*, 96(456):1348–1360, 2001.

[13] P. Hall and T. Maiti. Nonparametric Estimation of Mean-Squared Prediction Error in Nested-Error Regression Models. *Annals of Statistics*, 34(4):1733–1750, 2006.

[14] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[15] C. R. Henderson. Estimation of Genetic Parameters. *The Annals of Mathematical Statistics*, 21:309–310, 1950.

[16] C. R. Henderson. Estimation of Variance and Covariance Components. *Biometrics*, 9(2):226–252, 1953.

[17] C. R. Henderson. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31:423–447, 1975.

[18] W. James and Charles Stein. Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium*, 1:361–379, 1961.

[19] J. Jiang and P. Lahiri. Mixed Model Prediction and Small Area Estimation. *TEST*, 15(1):1–96, 2006.

[20] R. N. Kackar and D. A. Harville. Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Assiociation*, 79(388):853–861, 1984.

[21] J. D. Lee, D. L. Sun, Y. Dun, and J. E. Taylor. Exact Post-Selection Inference, with Application to the Lasso. *Annals of Statistics*, 44(3):907–927, 2016.

[22] Y. Lee and J. A. Nelder. Conditional and marginal models: Another view. *Statist. Sci.*, 19(2):219–238, 05 2004.

[23] H. Leeb and B. Pötscher. Model Selection and Inference Facts and Fiction. *Econometric Theory*, 21:21–59, 2005.

[24] K.-C. Li. Honest Confidence Regions for Nonparametric Regression. *Annals of Statistics*, 17(3):1001–1008, 1989.

[25] N. T. Longford. *Missing Data and Small-Area Estimation.* Springer, New York, NY, 2005.

[26] D. Nychka. Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Assiociation*, 83(404):1134–1143, 1988.

[27] D. Pfefferman. New Important Developements in Small Area Estimation. *Statistical Science*, 28(1):40–68, 2013.

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

[29] B. Pötscher. Confidence Sets Based on Sparse Estimators Are Necessarily Large. *Sankhya: The Indian Journal of Statistics, Series A*, 71(1):1–18, 2009.

[30] N. G. N. Prasad and J. N. K. Rao. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

[31] J. N. K. Rao and I. Molina. *Small Area Estimation.* Wiley, Hoboken, NJ, 2nd edition, 2015.

[32] J. Schelldorfer, P. Bühlmann, and S. van de Geer. Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. *Scandinavian Journal of Statistics*, 38:197–214, 2011.

[33] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components.* Wiley, Hoboken, NJ, 1992.

[34] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *JRSS B*, 58:267–288, 1996.

[35] J. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, MA, 1977.

[36] J. W. Tukey. The Problem of Multiple Comparisons. Published in *The Collected Works of John W. Tukey: Multiple Comparisons, Volume VIII (1999). Edited by H. Braun, CRC Press, Boca Raton, Florida*, 1953.

[37] G. Wahba. Bayesian “Confidence Intervals” for the Cross-validated Smoothing Spline. *Journal of the Royal Statistical Society B*, 45(1):133–150, 1983.

Curriculum Vitae

Personal Information

Name	Kramlinger, Peter Sebastian
Origin	30th of June 1991 in Buenos Aires, Argentina

Education

2016–today	Postgraduate studies in Mathematical Statistics, University of Göttingen
2014–2016	**Master of Science** in Mathematical Finance, Technical University Munich
2013–2014	Mathematics, Complutense University of Madrid, semester abroad
2010–2013	**Bachelor of Science** in Mathematics, Technical University Munich
2003–2010	**Abitur** at Georg-Herwegh-Oberschule Berlin

Publications

Submitted	Kramlinger, Krivobokova and Sperlich (2020): Marginal and Conditional Multiple Inference in Linear Mixed Models, arXiv: 1812.09250
Manuscript in Preparation	Kramlinger, Krivobokova and Schneider (2020): Uniformly Valid Inference Based on the Lasso in Linear Mixed Models

Professional Activities

December 2018	**Invited Talk** at 10th International Calcutta Symposium on Probability and Statistics, University of Calcutta and Calcutta Statistical Association, India
June 2018	**Contributed Talk** at Small Area Estimation 2018, Shanghai, PR China
March 2018	**Poster** at Workshop YES 2018, TU Eindhoven, Netherlands
June 2017	**Summer School** Advances in Quantile Regression, Minho University, Portugal

Addendum A

Marginal and Conditional Multiple Inference in Linear Mixed Models

Marginal and Conditional Multiple Inference in Linear Mixed Models

Peter Kramlinger[1] Tatyana Krivobokova[2] Stefan Sperlich[3]

Abstract

This work introduces a general framework for multiple inference in linear mixed models. Such can be done about population parameters (marginal) and subject specific ones (conditional). For two asymptotic scenarios that adequately address settings arising in practice, consistent simultaneous confidence sets for subject specific effects are constructed. In particular, it is shown that while conditional confidence sets are feasible, remarkably, marginal confidence sets are also asymptotically valid for conditional inference. Testing linear hypotheses and multiple comparisons by Tukey's method are also considered. The asymptotic inference is based on standard quantiles and requires no re-sampling techniques. All findings are validated in a simulation study and illustrated by a real data example on Spanish income data.

MSC 2010 subject classification: Primary 62J15; secondary 62F05; 62F03; 62J10.
Keywords and phrases: marginal vs conditional confidence, simultaneous inference, multiple testing, small area estimation

1 Introduction

Linear mixed models (LMMs) were introduced by Charles Roy Henderson in 1950s [14, 15] and are applied if repeated measurements on several independent subjects of interest are available. Monographs [32], [8] and [20] give a comprehensive overview of LMMs and their

[1] peter.kramlinger@uni-goettingen.de, Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

[2] tkrivob@gwdg.de, Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

[3] stefan.sperlich@unige.ch, School of Economics and Management, Université de Genève, 40 Bd du Pont d'Arve, 1211 Genève 4, Switzerland

generalizations. The classical LMM can be written as

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\,\boldsymbol{\beta} + \mathbf{Z}_i\,\mathbf{v}_i + \mathbf{e}_i, \quad i = 1, \ldots, m \\ \mathbf{e}_i \sim \mathcal{N}_{n_i}\{\mathbf{0}_{n_i}, \mathbf{R}_i(\boldsymbol{\delta})\}, &\quad \mathbf{v}_i \sim \mathcal{N}_q\{\mathbf{0}_q, \mathbf{G}(\boldsymbol{\delta})\}, \end{aligned} \tag{1}$$

with observations $\mathbf{y}_i \in \mathbb{R}^{n_i}$, known covariates $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ and $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$, independent random effects $\mathbf{v}_i \in \mathbb{R}^q$ and error terms $\mathbf{e}_i \in \mathbb{R}^{n_i}$, such that $\mathrm{Cov}(\mathbf{e}_i, \mathbf{v}_i) = \mathbf{0}_{n_i \times q}$. Parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\delta} \in \mathbb{R}^r$ are unknown and we denote $\mathbf{V}_i(\boldsymbol{\delta}) = \mathrm{Cov}(\mathbf{y}_i) = \mathbf{R}_i(\boldsymbol{\delta}) + \mathbf{Z}_i\,\mathbf{G}(\boldsymbol{\delta})\,\mathbf{Z}_i^t$, where $\mathbf{R}_i(\boldsymbol{\delta})$ and $\mathbf{G}(\boldsymbol{\delta})$ are known up to $\boldsymbol{\delta}$.

Model (1) accommodates both settings with a fixed number of subjects m by a growing number of observations per subject n_i, as well as settings with a growing number of subjects m by few observations per subject n_i, implying two possible asymptotic scenarios for mixed models, as noted by [21]. The latter case is referred to as *small area estimation* (SAE) [34].

Depending on the research question, the focus of estimation and inference might lay either on the population parameter $\boldsymbol{\beta}$ or on subject specific effects associated with $\mathbf{v}_i$. In the former case, a LMM (1) is interpreted as a linear regression model with mean $\mathbf{X}_i\,\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_i(\boldsymbol{\delta})$ that accounts for complex dependences in the data. Inference about $\boldsymbol{\beta}$ is referred to as *marginal* and well understood. If the focus is rather on the subject specific effects, then inference should be carried out *conditional* on $\mathbf{v}_i$, which is more involved. This distinction between marginal and conditional inference is emphasized already in [13] and has attracted particular attention in the model selection context. For example, [42] argue that the conventional (i.e. marginal) Akaike information criterion (AIC) is applicable to the selection of population parameter $\boldsymbol{\beta}$ only, and suggested a conditional AIC that should be employed else. For further discussion on marginal versus conditional inference in mixed models, see [26].

Today, there is an increasing interest in studying mixed parameters, in particular linear combinations of $\boldsymbol{\beta}$ and $\mathbf{v}_i$, such as $\mu_i = \mathbf{l}_i^t\,\boldsymbol{\beta} + \mathbf{h}_i^t\mathbf{v}_i$, $i = 1, \ldots, m$ with known $\mathbf{l}_i \in \mathbb{R}^p$ and $\mathbf{h}_i \in \mathbb{R}^q$. While the SAE literature has intensively studied inference of such parameters under the marginal law for a single μ_i, little is known about conditional and/or simultaneous inference. Under two possible asymptotic scenarios we construct simultaneous confidence sets for all $\mu_1, \ldots, \mu_m$ and discuss the corresponding multiple testing problem. Thereby, we distinguish between the marginal scenario, where $\mathbf{v}_i$ are treated as proper random variables and the conditional scenario, where $\mathbf{v}_i$ are considered as pre-fixed.

There is a large body of literature on the confidence intervals for each μ_i individually under the small area asymptotic scenario. Much attention is given to the estimation of the mean squared error $\mathrm{MSE}(\hat{\mu}_i) = \mathrm{E}(\mu_i - \hat{\mu}_i)^2$, where the expectation is taken under the marginal law and $\hat{\mu}_i$ is some estimator of μ_i, which depends on unknown $\boldsymbol{\delta}$. To estimate marginal MSE, one can either plug in an appropriate estimator of $\boldsymbol{\delta}$ (e.g., restricted maximum likelihood (REML) or Hendersons method III estimator given in [36]) or use unbiased marginal MSE approximations like in [33, 7, 4]. Other distribution-free approaches to the estimation of marginal MSE comprise a diverse collection of bootstrap methods, for an extensive review consult [3].

Since inference about μ_i has often a conditional focus (under the marginal law the $\mathbf{v}_i$ are simply not available), it seems counterintuitive to base inference on the marginal MSE only. In fact, we show that the nominal coverage of the subject-wise confidence intervals for μ_i based on the marginal MSE holds under the conditional law on average (over subjects) only, see Proposition 1 in Section 4 for more details. However, $\hat{\mu}_i$ are biased under the conditional law and this bias is, in general, difficult to handle. Ignoring the bias leads to a clear under-coverage, see [5, 6], while estimating the bias leads to unacceptably wide intervals, see [22, 28, 31].

In this article we construct simultaneous confidence sets for $\mu_1, \ldots, \mu_m$ in LMMs under two possible asymptotic scenarios. To the best of our knowledge this problem remained largely untreated; only [10] points out the need for simultaneous inference and considers a related problem of inference about certain linear combinations of μ_i in the Fay-Herriot model (a special case of (1) under small area asymptotics) employing a Bayesian approach. We first consider simultaneous confidence sets for $\mu_1, \ldots, \mu_m$ under the conditional law and show that the nominal coverage is attained at the usual parametric rate. Additionally, we show that, surprisingly, the simultaneous confidence sets built under the marginal law, being also accurate at the same parametric rate, are at the same time approximately valid when conditioning on the subjects. This, however, is not true in general for the subject-wise confidence intervals, as pointed out already. We use the derived confidence sets for testing linear hypotheses. Further, we extend the scope of analysis to the special case of testing multiple comparisons by the use of Tukey's method in the context of LMMs. Eventually, the usefulness of the derived methods is demonstrated on a real data study on Spanish income data.

The main results are given in Section 2. Applications for comparative statistics and testing linear hypotheses as well as extensions are elaborated in Section 3. The fundamental problem together with our results is visualized in a simulation study in Section 4, and further exemplified on Spanish income data in Section 5. We conclude with a discussion in Section 6. Proofs are deferred to the Appendix, and some auxiliary results to the Supplement [25].

2 Simultaneous Inference

We start with introducing basic notation and assumptions. In the notation of [34], the empirical BLUP (EBLUP) as estimator of μ_i for unknown $\boldsymbol{\delta}$ reads as

$$\begin{gathered}\hat{\mu}_i = \hat{\mu}_i(\hat{\boldsymbol{\delta}}) = \mathbf{l}_i^t \hat{\boldsymbol{\beta}} + \mathbf{b}_i(\hat{\boldsymbol{\delta}})^t \big(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} \big); \\ \mathbf{b}_i(\hat{\boldsymbol{\delta}})^t = \mathbf{h}_i^t \, \mathbf{G}(\hat{\boldsymbol{\delta}}) \, \mathbf{Z}_i^t \, \mathbf{V}_i(\hat{\boldsymbol{\delta}})^{-1}, \\ \hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^m \mathbf{X}_i^t \, \mathbf{V}_i(\hat{\boldsymbol{\delta}})^{-1} \, \mathbf{X}_i \right\}^{-1} \sum_{i=1}^m \mathbf{X}_i^t \, \mathbf{V}_i(\hat{\boldsymbol{\delta}})^{-1} \, \mathbf{y}_i \, .\end{gathered} \tag{2}$$

Under the mild assumptions below $\mathrm{E}(\hat{\mu}_i) = \mu_i$, if $\mathrm{E}(\hat{\mu}_i)$ is finite [24], but $\mathrm{E}(\hat{\mu}_i | \mathbf{v}_i) \neq \mu_i$. We consider two alternative asymptotic scenarios, namely

(A1) $m \to \infty$ while $\sup_i n_i = O(1)$.

(A2) $m \to \infty$ while $n_i \to \infty \; \forall i = 1, \ldots, m$.

Condition $\sup_i n_i = O(1)$ in (A1), introduced by [12], implies $\mathrm{E}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \, | \mathbf{v}) \nrightarrow \mathbf{0}_m$. However, due to (A2), the findings below are not restricted to the SAE setting. Further, we adopt the regularity conditions from [33] and [7]:

(B1) $\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{G}(\boldsymbol{\delta}) > 0$, $\mathbf{R}_i(\boldsymbol{\delta}) > 0$, $i = 1, \ldots, m$ contain only finite values.

(B2) $\mathbf{d}_i^t = \mathbf{l}_i^t - \mathbf{b}_i(\boldsymbol{\delta})^t \, \mathbf{X}_i$ has entries $d_{ik} = O(1)$ for $k = 1, \ldots, p$.

(B3) $\left\{ \frac{\partial}{\partial \delta_j} \mathbf{b}_i(\boldsymbol{\delta})^t \, \mathbf{X}_i \right\}_k = O(1)$, for $j = 1, \ldots, r$ and $k = 1, \ldots, p$.

(B4) $\mathbf{V}_i(\boldsymbol{\delta})$ is linear in the variance components $\boldsymbol{\delta}$.

Conditions (B1) - (B3) ensure that $\boldsymbol{\mu}$ can be estimated up to a vanishing error term. Condition (B4) implies that the second derivatives of $\mathbf{R}_i$ and $\mathbf{G}$ w.r.t. $\boldsymbol{\delta}$ are zero.

The variance components $\boldsymbol{\delta}$ can be estimated using both REML and Hendersons method III. Those are unbiased, even and translation invariant, which are the conditions of [24]. Subsequently, $\hat{\boldsymbol{\delta}}$ denotes an estimator of $\boldsymbol{\delta}$ obtained with either one of these methods.

Simultaneous Confidence Sets

Now we turn to the construction of simultaneous confidence sets for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^t$. Since the inference focus in this case is conditional, we start by constructing a confidence set $\mathcal{C}_\alpha$, such that $\mathrm{P}(\boldsymbol{\mu} \in \mathcal{C}_\alpha \, | \mathbf{v}) \approx 1 - \alpha$, for a pre-specified level $\alpha \in (0, 1)$. In particular, for the conditional inference $\mathbf{v} = (\mathbf{v}_1^t, \ldots, \mathbf{v}_m^t)^t$ is treated as a fixed parameter and the assumption on normality of $\mathbf{v}$ in (1) is ignored. Thereby, all parameter estimators are still obtained under model (1).

Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_m)^t$ and $\widehat{\boldsymbol{\Sigma}}_c$ be our (approximately) second-order unbiased estimator for $\boldsymbol{\Sigma}_c = \mathrm{Cov}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\,|\mathbf{v})$, which we derive in detail in the appendix, see equation (9). It then holds:

Theorem 1. *Let model (1) hold and $\widehat{\boldsymbol{\Sigma}}_c$ be as in (9). Under (A1) or (A2), with (B1)-(B4) it holds that*

$$P\Big\{\big\|\widehat{\boldsymbol{\Sigma}}_c^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big\|^2 < \chi^2_{m,1-\alpha}(\hat{\lambda})\Big|\mathbf{v}\Big\} = 1 - \alpha + O(m^{-1/2}),$$

where $\alpha \in (0,1)$, $\chi^2_{m,\alpha}(\hat{\lambda})$ is the α-quantile of the $\chi^2_m(\hat{\lambda})$-distribution and $\hat{\lambda}$ is a least squares estimator, given in (8), for the non-centrality parameter

$$\lambda = \sum_{i=1}^{m}\left\{\sum_{k=1}^{m} E(\hat{\mu}_k - \mu_k|\mathbf{v})\big(\boldsymbol{\Sigma}_c^{-1/2}\big)_{ik}\right\}^2.$$

Since $\hat{\boldsymbol{\mu}}$ is not unbiased under the conditional law, λ has to account for the conditional bias, whereas $\widehat{\boldsymbol{\Sigma}}_c$ accounts for the correct variability under such law. Note that the result of Theorem 1 holds for any pre-fixed $\mathbf{v}$, not necessarily a realization of a normally distributed random variable.

From Theorem 1 we immediately obtain the conditional confidence set

$$\mathcal{C}_\alpha = \Big\{\boldsymbol{\mu} \in \mathbb{R}^m : \big\|\widehat{\boldsymbol{\Sigma}}_c^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big\|^2 \le \chi^2_{m,1-\alpha}(\hat{\lambda})\Big\}.$$

This defines a simultaneous confidence region over all subjects under the conditional law. The practical difficulty when constructing $\mathcal{C}_\alpha$ is the estimation of the non-centrality parameter λ which introduces additional uncertainty.

If $\mathbf{v}$ is treated as a proper random variable, this implies the following result.

Theorem 2. *Let model (1) hold and $\widehat{\boldsymbol{\Sigma}}$ be an estimator for $\boldsymbol{\Sigma} = Cov(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ given in (6). Under (A1) or (A2), with (B1)-(B4) it holds that*

$$P\Big\{\big\|\widehat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big\|^2 < \chi^2_{m,1-\alpha}\Big\} = 1 - \alpha + O(m^{-1/2}),$$

where $\alpha \in (0,1)$ and $\chi^2_{m,1-\alpha}$ is the α-quantile of the χ^2_m-distribution.

Similarly to above, one obtains the marginal confidence set

$$\mathcal{M}_\alpha = \left\{ \boldsymbol{\mu} \in \mathbb{R}^m : \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \right\|^2 \leq \chi^2_{m,1-\alpha} \right\},$$

with $P(\boldsymbol{\mu} \in \mathcal{M}_\alpha) \approx 1 - \alpha$, for $\alpha \in (0,1)$. Such marginal confidence regions have to be interpreted with care, since $\boldsymbol{\mu}$ under the marginal case remains a random parameter. However, it turns out that the marginal confidence set can be used for simultaneous inference under the conditional law. Indeed, the following theorem states that $\mathcal{M}_\alpha$, albeit derived under the marginal law, lead to the asymptotically correct coverage under the conditional law.

Theorem 3. *Let model (1) hold and $\widehat{\boldsymbol{\Sigma}}$ be as in (6). Under (A1) or (A2), with (B1)-(B4) it holds that*

$$P\left\{ \left\| \widehat{\boldsymbol{\Sigma}}^{-1/2} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \right\|^2 < \chi^2_{m,1-\alpha} \middle| \mathbf{v} \right\} = 1 - \alpha + O(m^{-1/2}).$$

From the proof one can see that the misspecification in using the marginal formulation under the conditional scenario is averaged out across the subjects under (A1) or, less surprisingly, within the subjects under (A2). Notably, the rates for the marginal formulation in the marginal versus conditional scenario coincide. The result implies $\mathrm{P}(\boldsymbol{\mu} \in \mathcal{M}_\alpha \,|\mathbf{v}) \approx 1 - \alpha$.

Note that if the quadratic form in Theorem 3 is reformulated for one subject i with $n_i < \infty$ in (A1) we get

$$\mathrm{P}\left\{ \frac{(\hat{\mu}_i - \mu_i)^2}{\hat{\sigma}_{ii}} < \chi^2_{1,1-\alpha} \middle| \mathbf{v} \right\} = 1 - \alpha + O(1).$$

In (A2) however, the bias vanishes for each subject and the nominal coverage is attained asymptotically for a single subject as well.

The results of this section suggest that simultaneous inference about $\boldsymbol{\mu}$ under the conditional law can be performed based on the confidence sets obtained under the marginal law. In particular, this allows to circumvent the problem of estimating the non-centrality parameter in practice.

Tukey's Intervals

Further interest in inferring about multiple subjects simultaneously includes the use of Tukey's method [40]. That concerns all simple contrasts $\mathbf{c}^t(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ $\forall \mathbf{c} \in \mathcal{S}_w$, $w \leq m$, where

$$\mathcal{S}_w = \Big\{ \mathbb{1}_i - \mathbb{1}_j \ \ \forall i, j \leq w, \ \text{for } \mathbb{1}_k \text{ the } k\text{-th unit vector in } \mathbb{R}^m \Big\}.$$

Conventional use of Tukey's method involves linear unbiased estimators, see e.g. [2]. This setting, however, firmly lies in the realm of the conditional law, in which $\hat{\boldsymbol{\mu}}$ are biased. Additional regularity conditions are thus required for $\forall i, k \leq m$:

(C1) $\mathbf{h}_i = \mathbf{h}_k + \{O(m^{-1/2})\}_q$ and $\mathbf{l}_i = \mathbf{l}_k + \{O(m^{-1/2})\}_p$.

(C2) $\mathbf{1}_{n_i}^t \mathbf{V}_i(\boldsymbol{\delta})^{-1} \mathbf{1}_{n_i} = \mathbf{1}_{n_k}^t \mathbf{V}_k(\boldsymbol{\delta})^{-1} \mathbf{1}_{n_k} + \{O(m^{-1/2})\}_q^t$.

These conditions ensure that the subjects' mixed parameters are sufficiently similar. A special case in which both (C1) and (C2) are fulfilled is the widely used nested error regression model (5) with a balanced panel.

Theorem 4. *Let model (1) hold and $\widehat{\boldsymbol{\Sigma}}_c$ as in (9). Under (A1) or (A2), with (B1)-(B4) and (C1), (C2) it holds for $\alpha \in (0,1)$ that*

$$P\bigg\{ \frac{|\mathbf{c}^t(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})|}{\hat{c}_+} < \eta_{\mathbf{c}} + q_{m,1-\alpha}, \ \forall \mathbf{c} \in \mathcal{S}_m \bigg| \mathbf{v} \bigg\} = 1 - \alpha + O(m^{-1/2}),$$

where $q_{m,1-\alpha}$ the α-quantile of the range distribution for m standard normal random variables, $\eta_{\mathbf{c}} = c_+^{-1} \mathbf{c}^t E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\,|\mathbf{v})$ with $c_+ = \big(\mathbf{c}^t\, \boldsymbol{\Sigma}_c^{1/2}\big)_{>0} \mathbf{1}_m$, i.e. the sum of positive entries of $\mathbf{c}^t\, \boldsymbol{\Sigma}_c^{1/2}$ and $\hat{c}_+$ analogously with $\widehat{\boldsymbol{\Sigma}}_c$.

This result establishes consistent inference for all simple contrasts and thereby forms a special case of the generalized Tukey conjecture about attaining nominal level for non-diagonal covariance matrices [2, 40]. In particular, the result states that $P(\mathbf{c}^t\, \boldsymbol{\mu} \in \mathcal{T}_{\alpha,m}(\mathbf{c}), \ \ \forall \mathbf{c} \in \mathcal{S}_m | \mathbf{v}) \approx 1 - \alpha$ for

$$\mathcal{T}_{\alpha,m}(\mathbf{c}) = \Big\{ \mathbf{c}^t\, \boldsymbol{\mu} \in \mathbb{R} : |\mathbf{c}^t(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})| \leq \hat{c}_+ \big(\eta_{\mathbf{c}} + q_{m,1-\alpha}\big) \Big\}.$$

Note that in practice $\eta_{\mathbf{c}}$ is in general unknown and the confidence interval cannot be readily constructed. However, in the next section we discuss that for relevant testing scenarios (C1) and (C2) imply that $\eta_{\mathbf{c}}$ vanishes quickly enough, so that $\mathcal{T}_{\alpha,w}(\mathbf{c})$ can serve for pairwise testing for equality of $\mu_i, \ldots, \mu_w$, $w < m$.

3 Testing

It is appealing to use the derived results to test either linear hypotheses or multiple comparisons of μ_i, $i = 1, \dots, m$, under conditional law. The former is concerned about testing whether $\boldsymbol{\mu}$ lies in a given subspace of $\mathbb{R}^m$. It can, for example, be applied to examine if subject specific effects are present within subsets, as done in Section 5. In case of rejection, one may want to know which subjects are the cause for it. Tukey's method controls the family-wise error rate whilst simultaneously testing multiple comparisons for all pairwise differences $\mu_i - \mu_j$, $i, j = 1, \dots, w < m$.

Linear Hypotheses

Let us assume it is of interest to test

$$H_0 : \ \mathbf{L}(\boldsymbol{\mu} - \mathbf{a}) = \mathbf{0}_u \quad \text{vs.} \quad H_1 : \quad \mathbf{L}(\boldsymbol{\mu} - \mathbf{a}) \neq \mathbf{0}_u, \tag{3}$$

where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{L}$ is a given $(u \times m)$-matrix with $u \leq m$ and $\operatorname{rank}(\mathbf{L}) = u$. The dimension u of the linear subspace of $\mathbb{R}^m$ corresponds to the number of simultaneous tests of linear combinations, whereas each linear combination of interest is specified in the rows of $\mathbf{L}$. For example, for $\mathbf{L} = \mathbf{I}_m$ and $\mathbf{a} = (a_1, \dots, a_m)^t$, $a_i \neq a_j$, $i, j \leq m$, implies testing whether the mixed parameters take on some ex-ante assumed value. For conditional inference in (1) about $\boldsymbol{\mu}$, Theorem 1 gives the α-level test for (3), that rejects H_0 if $\mathbf{a} \notin \mathcal{C}_{\alpha,\mathbf{L}}$, where

$$\mathcal{C}_{\alpha,\mathbf{L}} = \left\{ \mathbf{a} \in \mathbb{R}^m : \left\| \left(\mathbf{L} \widehat{\boldsymbol{\Sigma}}_c \mathbf{L}^t \right)^{-1/2} \mathbf{L} (\hat{\boldsymbol{\mu}} - \mathbf{a}) \right\|^2 \leq \chi^2_{u,1-\alpha}(\hat{\lambda}_{\mathbf{L}}) \right\}.$$

This test is consistent with an error $O(m^{-1/2})$. Parameter $\hat{\lambda}_{\mathbf{L}}$ is the non-centrality parameter that depends on the modified covariance $\mathbf{L}\widehat{\boldsymbol{\Sigma}}_c\mathbf{L}^t$.

Furthermore, Theorem 3 allows to employ the confidence set $\mathcal{M}_\alpha$ as well. An α-level test rejects H_0 if $\mathbf{a} \notin \mathcal{M}_{\alpha,\mathbf{L}}$, where

$$\mathcal{M}_{\alpha,\mathbf{L}} = \left\{ \mathbf{a} \in \mathbb{R}^m : \left\| \left(\mathbf{L} \widehat{\boldsymbol{\Sigma}} \mathbf{L}^t \right)^{-1/2} \mathbf{L} (\hat{\boldsymbol{\mu}} - \mathbf{a}) \right\|^2 \leq \chi^2_{u,1-\alpha} \right\}.$$

This test is again consistent with rate $O(m^{-1/2})$ under (A2), while under (A1) the rate is $O(u^{-1/2})$ for $u = m^{\xi_1}$, where $\xi_1 \in (0, 1]$ bounded away from zero. This affirms that individual confidence intervals ($u = 1$) can not be constructed using $\mathcal{M}_{\alpha,\mathbf{L}}$ under (A1), the standard SAE assumption.

It is often of interest to test if some or all μ_i are equal, which implies equality of random effects. If $w < m$ random effects are tested to be equal, then model (1) is altered in that only $m' = m - w + 1$ different subjects remain under H_0. In that case, above tests are consistent with m replaced by m': $O[\{\min(m', u)\}^{-1/2}]$. For $w = m$, the underlying model (1) of H_0 reduces to a linear model, for which conventional tests are readily available. Details are given in the appendix.

Tukey's Method

Multiple comparisons, as $\mu_i - \mu_j$, $i, j = 1, \ldots, w < m$, allow for multiple testing against w equal random effects. Formally,

$$H_0: \ \mathbf{c}^t \boldsymbol{\mu} = 0 \ \ \forall \mathbf{c} \in \mathcal{S}_w \quad \text{vs.} \quad H_1: \ \mathbf{c}^t \boldsymbol{\mu} \neq 0 \ \text{ for some } \mathbf{c} \in \mathcal{S}_w, \tag{4}$$

where $w = m^{\xi_2}$, with $\xi_2 \in (0, 1)$. Under (C1) and (C2), $\eta_{\mathbf{c}}$, $\mathbf{c} \in \mathcal{S}_w$ vanishes under H_0. See (13) in the appendix for details. It follows that all $\binom{w}{2}$ simple contrasts in (4) can be tested by Tukey's method [40, 27] with Theorem 4. The test rejects H_0 if $\exists \mathbf{c} \in \mathcal{S}_w$ such that $0 \notin \mathcal{T}_{\alpha,w}(\mathbf{c})$ and is consistent with $O(m^{-1/2})$ under (A1) and (A2), where m is replaced by m'.

Again, for $w = m$, the classical versions of Tukey's method can be applied, see the discussion in the appendix.

4 Simulation Study

Consider a special case of (1), the nested error regression model [1] with $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$, $v_i \sim \mathcal{N}(0, \sigma_v^2)$, and

$$y_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta} + v_i + e_{ij}, \ \ i = 1, \ldots, m, \ j = 1, \ldots, n_i. \tag{5}$$

The data are simulated as follows. For each given set of the parameters m, n_i, σ_e^2, σ_v^2, the value of the subject effect v_i is obtained as a realization of a $\mathcal{N}(0, \sigma_v^2)$-distributed random variable and remains fixed in all Monte Carlo samples. Parameters $\boldsymbol{\beta} \in \mathbb{R}^2$ were drawn once from a standard normal distribution, whereas $\mathbf{X}_i \in \mathbb{R}^{n_i \times 2}$ consists of a column of 1's and a column of entries drawn once from the uniform distribution. The parameter of interest is $\mu_i = \overline{\mathbf{X}}_i \boldsymbol{\beta} + v_i$, where $\overline{\mathbf{X}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$.

Before we report simulation results for simultaneous inference, we visualize consequences of using marginal law for subject-wise inference about single μ_i. We set $(\sigma_v^2, \sigma_e^2) = (4, 4)$, $m = 100$, $n_i = 5$ under (A1) and $n_i = 50$ under (A2). The results are based on 1.000

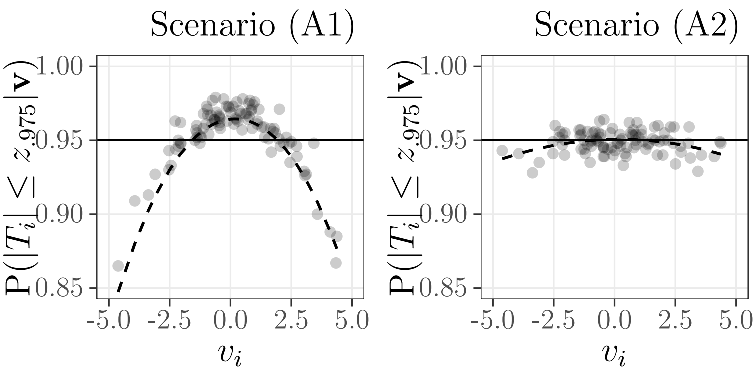

Figure 1: Empirical coverage of marginal 95% subject-wise confidence intervals for μ_i under conditional law under (A1) (left) and (A2) (right). The dashed lines give the theoretical coverage.

Monte Carlo samples. Figure 1 shows the subject-wise coverage of confidence intervals for μ_i built under the marginal law. The left hand side of Figure 1 corresponds to the small area asymptotics (A1). Subjects which comprise a large $|v_i|$, being those with most prominent subject effect, exhibit a severe undercoverage. This is particularly annoying, since such subjects are arguably those that a practitioner might be most interested in, see [22]. On average (over all subjects), however, over- and undercoverage cancel each other out. Under (A2), this problem is less pronounced, since the bias for every subject vanishes asymptotically and so does the difference between conditional and marginal variance, as visible on the right hand side of Figure 1. These observations are formalized in

Proposition 1. *Let model (1) hold, $\boldsymbol{\delta}$ known, $T_i = (\hat{\mu}_i - \mu_i)\, Var(\hat{\mu}_i - \mu_i)^{-1/2}$ and $z_{1-\alpha/2}$ the two-sided α-quantile of $\mathcal{N}(0,1)$. Then,*

(a) for $Z \sim \mathcal{N}(0,1)$ and c_1 as well as $c_2(\mathbf{v})$ as given in (14) it holds

$$P\big(|T_i| \leq z_{1-\alpha/2}|\mathbf{v}\big) = P\big\{|Z| \leq z_{1-\alpha/2} + c_1 \pm c_2(\mathbf{v})|\mathbf{v}\big\}.$$

(b) under (A1) or (A2) with (B1) and (B2) it holds

$$\frac{1}{m}\sum_{i=1}^{m} P(|T_i| \leq z_{1-\alpha/2}|\mathbf{v}) = 1 - \alpha + O\big(m^{-1/2}\big).$$

That is, although $c_1 \pm c_2(\mathbf{v})$ is almost surely nonzero under marginal law, the coverage probability of marginal confidence intervals under the conditional law still attains its nominal level on average over all subjects. For the simulated data in Figure 1 the average coverage under (A1) is 95.4%, while under (A2) it is 94.9%.

Table 1: Coverage of 95%-confidence ellipsoids in model (5) under conditional law.

				Marginal		Conditional	
$\boldsymbol{\delta}$	m	n_i	n_k	known $\boldsymbol{\delta}$	REML	known $\boldsymbol{\delta}$	REML
	10	5	5	.96 (.98)	.92 (1)	.95 (.92)	.88 (.79)
	100	5	5	.95 (.97)	.93 (1)	.95 (.89)	.93 (1.6)
$\sigma_v^2 = 8$	10	10	10	.95 (.96)	.93 (1)	.95 (.99)	.93 (1.0)
$\sigma_e^2 = 2$	100	10	10	.95 (.96)	.93 (1)	.95 (1.0)	.94 (1.5)
	10	5	10	.96 (.96)	.92 (1)	.95 (.83)	.90 (.98)
	10	10	100	.95 (.95)	.95 (1)	.95 (.93)	.95 (1.1)
	10	5	5	.96 (.70)	.92 (1)	.94 (.66)	.96 (1.6)
	100	5	5	.95 (.81)	.93 (1)	.94 (.25)	.94 (3.8)
$\sigma_v^2 = 4$	10	10	10	.95 (.79)	.94 (1)	.95 (.80)	.96 (1.8)
$\sigma_e^2 = 4$	100	10	10	.94 (.72)	.94 (1)	.94 (.59)	.95 (5.1)
	10	5	10	.96 (.73)	.94 (1)	.94 (.39)	.97 (2.3)
	10	10	100	.96 (.80)	.96 (1)	.95 (.73)	.97 (1.3)
	10	5	5	.97 (.24)	.96 (1)	.85 (.32)	.82 (1.0)
	100	5	5	.97 (.32)	.88 (1)	.86 (.01)	.88 (1.5)
$\sigma_v^2 = 2$	10	10	10	.94 (.42)	.93 (1)	.91 (.24)	.97 (.94)
$\sigma_e^2 = 8$	100	10	10	.94 (.32)	.92 (1)	.92 (.01)	.94 (2.8)
	10	5	10	.98 (.28)	.97 (1)	.90 (.05)	.92 (3.3)
	10	10	100	.97 (.44)	.97 (1)	.93 (.25)	.99 (1.8)

We now turn to simultaneous inference: Table 1 contains results based on $10,000$ Monte Carlo samples. For each sample the estimates $\hat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$, as well as $\widehat{\boldsymbol{\Sigma}}_c$ and $\hat{\lambda}$ are calculated, and it is checked whether $\boldsymbol{\mu}$ lies within the 95%−confidence set. The resulting coverage probability is reported together with the one of the oracle confidence sets for known $\boldsymbol{\delta} = (\sigma_v^2, \sigma_e^2)^t$. The relative volume of the confidence sets to the volume of the REML-based marginal set is given in brackets.

Under (A1), the asymptotic behavior relies on m, and is therefore studied for $m = 10$ and $m = 100$. One case is carried out for $n_i = n_k = 5$ observations for all subjects i, k, relating to the study of [1]; the other for $n_i = n_k = 10$. Under (A2), 80% percent of subjects had n_i observations, while the remaining 20% had n_k. As it is well known that the relation of σ_v^2 and σ_e^2, the so-called intraclass correlation coefficient (ICC), plays a key role in the reliability of the estimators, different ICC are considered.

The columns "Marginal" and "Conditional" of Table 1 give the simulated coverage of the confidence sets for the nominal coverage of 0.95. The differences between each of the two marginal and conditional coverages display the impact of the REML estimation. The estimation of the variance components is indeed influential, in accuracy as well as in size. Further, a comparison between marginal REML and conditional REML coverages reveals the performance of $\mathcal{M}_\alpha$ and $\mathcal{C}_\alpha$. The marginal sets are generally smaller. This is due to

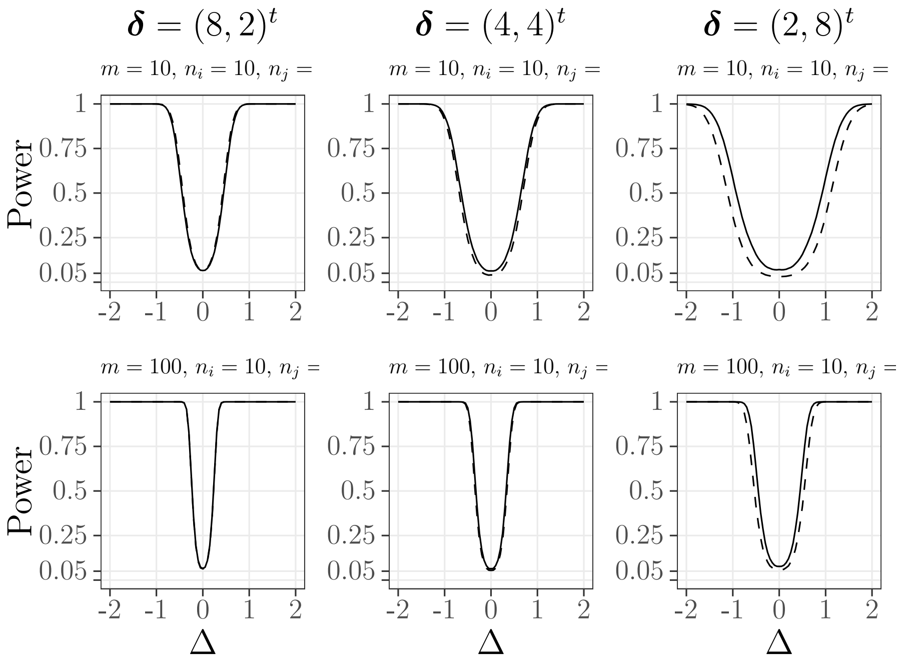

Figure 2: Power of tests based on confidence ellipsoids $\mathcal{M}_\alpha$ (solid line) and $\mathcal{C}_\alpha$ (dashed) for model (5) in the conditional setting with $H_1 : \boldsymbol{\mu} = \mathbf{a} + \mathbf{1}_m\Delta$.

the conditional sets being amplified by the non-central quantile to meet the nominal level, but not stretched in the direction where the multivariate distribution of $\hat{\boldsymbol{\mu}}$ has most of the mass.

The first two rows of each configuration of $\boldsymbol{\delta}$ show the asymptotic behavior for (A1) with $n_i = n_k = 5$ observations only, whereas the less extreme case for (A1) is given in lines three and four. Clearly, larger m produce better results. However, the reported coverage seems to be stronger influenced by the number of observations in each subject. This is the realm of case (A2). Convergence for that scenario seems to be more sensitive, although this is likely due to the smaller sample size.

The ICC (and the signal-noise ratio) proves to be quite influential, with coverage being closest to the nominal level for large σ_v^2. This is not unexpected as these parameters determine the validity of the REML-estimates which has already been observed for individual confidence intervals, see [4]. However, even for known $\boldsymbol{\delta}$, $\mathcal{C}_\alpha$ can exhibit undercoverage if the ICC is too small and/or too few data is available, whereas $\mathcal{M}_\alpha$ does not.

Let us turn to the test $H_0 : \boldsymbol{\mu} = \mathbf{a}$ vs. $H_1 : \boldsymbol{\mu} = \mathbf{a} + \mathbf{1}_m\Delta$, $\mathbf{a} \in \mathbb{R}^m$ with $\Delta \in \mathbb{R}$. Power functions studying the error of the second kind for different parameters m and n_i, cf. Table 1, are given in Figure 2 with different ICC. Unsurprisingly, the power growths steeper for larger m and n_i, but again is sensitive to the relative size of σ_v^2 to σ_e^2. The power of the tests based on the marginal set (solid line) is notably steeper than the slope of the power based on the conditional set (dashed). Although being of less importance if

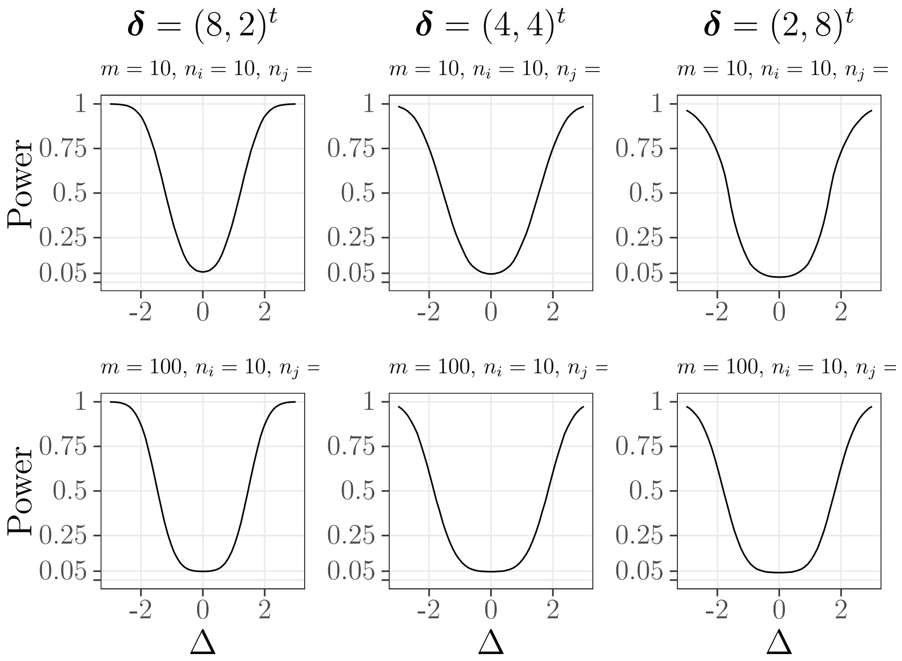

Figure 3: Power of tests based on Tukey's method with $i, j = 1, \ldots, m^*$ for $H_0 : \mu_i = \mu_j$ for against $H_1 : \exists! i : \mu_i = \mu_j + \Delta$.

n_i is large, the plots favor the use of the marginal confidence sets for testing.

A similar visualization for Tukey's method is obtained by testing $H_0 : \mu_i = \mu_j$ for $\forall i, j = 1, \ldots, m/2$ vs. $H_1 : \exists! i : \mu_i = \mu_j + \Delta$ for $\forall i, j = 1, \ldots, m/2$, where $\Delta \in \mathbb{R}$. That is, all but one $\mu_1, \ldots, \mu_{m/2}$ are equal. Figure 3 shows that, similar to the case for confidence ellipsoids, the ICC is influential. In difference to Figure 2, more subjects greatly increase the number of tests, and the power function undesirably flattens around zero.

5 Study on Spanish Income

The discussed methods are now applied to a case study on log-transformed yearly income for working population in Spain obtained from the survey of living conditions in 2008 [17]. As income varies non-linearly with age, we restrict the study on people of age 50 and older. The subjects, henceforth small areas, are formed by cross-section of all 50 provinces of Spain and whether secondary school was completed. In total, $n = 3,335$ observations are available for $m = 100$ small areas, with a median of 20 per area. Available explanatory variables are gender, municipality size and nationality. The observations are assumed to follow a nested error regression model (5). The variance components are estimated via REML as $\hat{\sigma}_v^2 \approx .55 \times 10^{-2}$ and $\hat{\sigma}_e^2 \approx 5.17 \times 10^{-2}$.

Interest lies in determining whether the hypothesis that no area specific effect is present in those 16 small areas that lie in the autonomous community of Andalucía, can be rejected.

Formally, let $\mathbf{L}_1 = (\mathbf{0}_{15}, \ldots, \mathbf{L}_1^*, \mathbf{0}_{15}, \ldots)$ be (15×100), with $\mathbf{L}_1^* = \big(\mathbf{I}_{15}, \mathbf{0}_{15}\big) - \mathbf{1}_{15}\mathbf{1}_{16}^t/16$ corresponding to all small areas in Andalucía. The test then checks the linear hypothesis $H_0 : \mathbf{L}_1\,\boldsymbol{\mu} = \mathbf{0}_{15}$ against $H_1 : \mathbf{L}_1\,\boldsymbol{\mu} \neq \mathbf{0}_{15}$. For $\alpha = 0.05$, 86.7% of individual tests do not reject H_0, and neither does the conservative Bonferroni correction. Both tests based on marginal and conditional ellipsoids however do reject, as

$$\big\|(\mathbf{L}_1\widehat{\boldsymbol{\Sigma}}\mathbf{L}_1^t)^{-1/2}\mathbf{L}_1\hat{\boldsymbol{\mu}}\big\|^2 \approx 25.7 > 25.0 \approx \chi^2_{15,.95},$$
$$\big\|(\mathbf{L}_1\widehat{\boldsymbol{\Sigma}}_c\mathbf{L}_1^t)^{-1/2}\mathbf{L}_1\hat{\boldsymbol{\mu}}\big\|^2 \approx 37.2 > 25.0 \approx \chi^2_{15,.95}(0).$$

Both sets have the same nominal coverage, and here they both yield the same result, although the conditional fails to produce a positive estimate of $\lambda_{\mathbf{L}_1}$ for this data set. Moreover, if there was interest in investigating other regions, this approach would require to re-estimate the non-centrality parameter on the new subset of interest. This aspect makes the application of the marginal set more appealing.

Although both accurate tests reject H_0, it remains unknown by which areas this is caused. Tukey's method allows for those kind of multiple comparisons. Let $\mathcal{S}_{\text{And}} = \{\mathbb{1}_i - \mathbb{1}_j,\ \ \forall i, j \text{ in Andalucía}, \mathbb{1}_k \text{ the } k\text{-th unit vector}\}$, $|\mathcal{S}_{\text{And}}| = 120$, and $H_0 :\ \mathbf{c}^t\,\boldsymbol{\mu} = 0\ \ \forall \mathbf{c} \in \mathcal{S}_{\text{And}}$. One can verify that the bias $\eta_{\mathbf{c}}$ is of a negligible order for this test, so that Theorem 4 can be applied. Then, H_0 can be rejected by two contrasts, namely

$$\frac{|\hat{\mu}_{\text{Cádiz, school}} - \hat{\mu}_{\text{Granada, no school}}|}{\hat{c}_{+,\text{Cádiz, school; Granada, no school}}} \approx 5.20 > 4.85 \approx q_{16,.95},$$
$$\frac{|\hat{\mu}_{\text{Cádiz, school}} - \hat{\mu}_{\text{Córdoba, no school}}|}{\hat{c}_{+,\text{Cádiz, school; Córdoba, no school}}} \approx 4.89 > 4.85 \approx q_{16,.95}.$$

If interest does not concern all pairwise differences, but only those within a single province, the test $H_0 : \mu_{i,\text{school}} - \mu_{i,\text{no school}} = 0$ for $\forall i$ in Andalucía is appropriate. Similarly as above, let $\mathbf{L}_2 = (\mathbf{0}_8, \ldots, \mathbf{L}_2^*, \mathbf{0}_8, \ldots)$ be (8×100), with i-th row corresponding to the i-th province as $\mathbf{L}_{2,i}^* = \big(0, \ldots, 1, -1, 0, \ldots\big)$. We test the linear hypothesis $H_0 : \mathbf{L}_2\,\boldsymbol{\mu} = \mathbf{0}_8$ against $H_1 : \mathbf{L}_2\,\boldsymbol{\mu} \neq \mathbf{0}_8$.

At $\alpha = 0.05$, 75% of individual tests do not reject H_0, and, again, neither does the Bonferroni correction, whereas both ellipsoid-based methods reject:

$$\big\|(\mathbf{L}_2\widehat{\boldsymbol{\Sigma}}\mathbf{L}_2^t)^{-1/2}\mathbf{L}_2\hat{\boldsymbol{\mu}}\big\|^2 \approx 17.8 > 15.5 \approx \chi^2_{16,.95},$$
$$\big\|(\mathbf{L}_2\widehat{\boldsymbol{\Sigma}}_c\mathbf{L}_2^t)^{-1/2}\mathbf{L}_2\hat{\boldsymbol{\mu}}\big\|^2 \approx 25.9 > 15.5 \approx \chi^2_{16,.95}(0).$$

Note that as only pairwise differences within a single province are tested against, Tukey's method cannot be applied here, as it does not extend to test against all pairwise differences

Table 2: Means of income in Euro for school graduates.
If $\hat{\mu}_{i,\text{no school}}$ were equal $\mu^*_{i,\text{no school}}$, $H_0 : \mathbf{L}\,\boldsymbol{\mu} = \mathbf{0}_8$ could not be rejected.

Province	$\hat{\mu}_{i,\text{school}}$	$\hat{\mu}_{i,\text{no school}}$	$\mu^*_{i,\text{no school}}$	$\mu^*_{i,\text{no school}} - \hat{\mu}_{i,\text{no school}}$
Córdoba	16,394	11,177	11,720	543
Granada	16,149	10,900	11,101	201
Sevilla	15,742	12,099	12,516	417
Total				1,161

of pairwise differences.

However, the method based on confidence ellipsoids allows to project $\mathbf{L}_2\hat{\boldsymbol{\mu}}$ onto $\mathcal{M}_{\alpha,\mathbf{L}_2}$ in order to obtain $\boldsymbol{\mu}^*_{\text{no school}}$ for which H_0 could not have been rejected. Results are given in Table 2. This procedure indicates how much effort, and in which province is to be made to attain statistically insignificant differences. Such findings could not have been obtained from so far existing tools. They exemplify the wide range of applications of multiple inference in SAE, with others easily conceivable.

6 Discussion

We derive simultaneous confidence sets for mixed parameters $\mu_1, \ldots, \mu_m$, namely linear combinations of fixed and random effects of LMMs. This is done under the two scenarios, $m \to \infty$ or $n_i \to \infty$. These simultaneous confidence sets are derived under conditional law and require the estimation of a non-centrality parameter of the respective $\chi^2(\lambda)$-distribution. We can show that with its estimate, the wanted nominal coverage is still attained at the usual parametric rate. Further, we extend the theory for marginal law, for which no such parameter is required.

We find that, surprisingly, the simultaneous confidence sets built under marginal law are approximately valid (at the same parametric rate) when conditioning on the subjects. This, however, is not true in general for the subject–wise confidence intervals. We use the confidence sets for multiple testing, and demonstrate its usefulness in practice.

Our results hold for all kind of linear combinations of mixed (fixed and random) parameters μ_i of a subject i. The results show that the problem of bias estimation in the conditional case can be overcome either by straight-forward estimation of λ or by directly applying marginal sets as the bias is averaged out over multiple subjects. A simulation study confirms this effect already for samples of small and moderate size.

For the special case when it is of interest to test whether the specific effects within a subset of subjects are equal, we extended the testing procedures to cover multiple comparisons by Tukey's method. However, the application of this method is limited to special cases of LMMs where the corresponding bias can be shown to be negligible.

Most uncertainty is induced by the estimation of $\boldsymbol{\delta}$. If the normality of errors and random effects is not met, it has been shown that the estimates for hierarchical [35] and non-hierarchical LMM [19] or by Hendersons Method III [11] are still consistent and asymptotically normal. For the latter, we obtain asymptotically the same results [25]. However, it is to be expected that depending on the deviations from normality larger samples are needed to reach the nominal coverage probability.

A popular strategy is to transform the data in order to achieve normality for errors and random effects, see [41] for a recent review. Software provides checks for the distributions of residuals and predictors $\hat{\mathbf{v}}_i$ [18]. Alternatively, bootstrap methods for LMM can account for non-Gaussian data, see [9].

We expect that our results can be extended to other predictors of LMMs, such as the best predictor of [23].

Acknowledgements

The authors thank Domingo Morales, Carmen Cadarso-Suárez, J.N.K. Rao, Gauri Datta, Jiming Jiang, Maria-Jose Lombardia, and Wenceslao Gonzalez-Manteiga for helpful discussion. They also gratefully acknowledge the funding by the German Research Association (DFG) via Research Training Group 1644 “Scaling Problems in Statistics”. Peter Kramlinger also gratefully acknowledges the support by Hanns-Seidel-Foundation, funded by the Federal Ministry of Education and Research (BMBF).

Appendix

Notation

Throughout the appendix the following notation is used. The (i,j)-th entry of matrix $\mathbf{A}$ is denoted as a_{ij} or $(\mathbf{A})_{ij}$. The $(\mathbf{A})_i$ denotes the i-th column vector of matrix $\mathbf{A}$. Other ways to display a vector or matrix is by e.g. $\{O_p(1)\}_{n\times n}$, a $(n \times n)$ stochastic matrix with each entry being of probabilistic order $O_p(1)$, or $\mathbf{A} = (a_{ij})_{i,j}$, if it is obvious that $i,j = 1,\dots,m$. Furthermore, for a matrix $\mathbf{A}$, $\|\mathbf{A}\|^2 = \mathrm{tr}\{\mathbf{A}^t\mathbf{A}\}$ is the Frobenius norm. The square-root of a symmetric positive-definite matrix $\mathbf{A}^{1/2}$ is defined as the unique symmetric matrix such that $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$. For easier readability, the dependence on the $\boldsymbol{\delta}$ is suppressed for various quantities. It should be clear from the context if e.g. $\mathbf{G}$ or $\mathbf{V}$ depend on $\boldsymbol{\delta}$ or $\hat{\boldsymbol{\delta}}$. Further, if not otherwise noted, we adapt the notation of [34] and denote $\tilde{\mu}_i = \hat{\mu}_i(\boldsymbol{\delta})$ as given in (2) and $\tilde{\boldsymbol{\beta}}$ analogously. For convenience, dropping the subject index $i = 1, \ldots, m$ labels the respective quantities over all observations, e.g. $\mathbf{y} = (\mathbf{y}_1^t, \ldots, \mathbf{y}_m^t)^t$, $\mathbf{V} = \mathrm{diag}\{\mathbf{V}_i(\boldsymbol{\delta})\}_{i=1,\ldots,m}$ and $\mathbf{X} = (\mathbf{X}_1^t, \ldots, \mathbf{X}_m^t)^t$, etc. If the range of the index is clear from the context, it will not be dropped as well. For the proofs, $i, k = 1, \ldots, m$ denote the subject and $j = 1, \ldots, n_i$ the respective observation for the i-th subject. Eventually, $e, f, g, d = 1, \ldots, r$ are indices referring the entries in $\boldsymbol{\delta}$.

Proof and Definitions for Theorem 2

The estimator for the across-area generalization of the Prasad-Rao MSE estimator from [33] is defined below. Let $\overline{\mathbf{V}} = \mathrm{Cov}(\hat{\boldsymbol{\delta}})$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\delta}}$. Then,

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}) &= \mathbf{K}_1(\hat{\boldsymbol{\delta}}) + \mathbf{K}_2(\hat{\boldsymbol{\delta}}) + 2\hat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}}); \\
\mathbf{K}_1(\boldsymbol{\delta}) &= \mathrm{diag}\big\{\mathbf{h}_i^t\big(\,\mathbf{G} - \mathbf{G}\,\mathbf{Z}_i^t\,\mathbf{V}_i^{-1}\,\mathbf{Z}_i\,\mathbf{G}\,\big)\mathbf{h}_i\big\}_{i=1,\ldots,m}, \\
\mathbf{K}_2(\boldsymbol{\delta}) &= \left\{\mathbf{d}_i^t\left(\sum_{l=1}^{m}\mathbf{X}_l^t\,\mathbf{V}_l^{-1}\,\mathbf{X}_l\right)^{-1}\mathbf{d}_k\right\}_{i,k=1,\ldots,m}, \qquad (6) \\
\widehat{\mathbf{K}}_3(\boldsymbol{\delta}) &= \mathrm{diag}\left\{\mathrm{tr}\left(\frac{\partial\mathbf{b}_i^t}{\partial\,\boldsymbol{\delta}}\,\mathbf{V}_i\,\frac{\partial\mathbf{b}_i}{\partial\,\boldsymbol{\delta}^t}\overline{\mathbf{V}}\right)\right\}_{i=1,\ldots,m}.
\end{aligned}$$

$\widehat{\boldsymbol{\Sigma}}$ is a second-order unbiased estimator of $\boldsymbol{\Sigma} = \mathbf{K}_1(\boldsymbol{\delta}) + \mathbf{K}_2(\boldsymbol{\delta}) + \mathbf{K}_3$. The leading term $\mathbf{K}_1(\boldsymbol{\delta})$ is an estimator for the variability induced in the prediction of the random effect, whereas $\mathbf{K}_2(\boldsymbol{\delta})$ describes the variability induced by the estimation of $\boldsymbol{\beta}$ such that $\mathbf{K}_1(\boldsymbol{\delta}) + \mathbf{K}_2(\boldsymbol{\delta}) = \{\mathrm{E}(\mu_i - \tilde{\mu}_i)(\mu_k - \tilde{\mu}_k)\}_{i,j=1,\ldots,m}$. Finally, $(\mathbf{K}_3)_{i,j} = \mathrm{E}(\tilde{\mu}_i - \hat{\mu}_i)(\tilde{\mu}_k - \hat{\mu}_k)$ the variability of the estimation of $\boldsymbol{\delta}$.

Proof. (of Theorem 2). Consider first (A1). We first show that

$$\|\widehat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 = \|\,\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 + O_p(m^{1/2}).$$

It has been shown for both Hendersons method III [33] and REML [7] that $\mathrm{E}(\hat{\sigma}_{ik}) = \sigma_{ik} + O_p(m^{-3/2})$, as well as $\tilde{\sigma}_{ik} = \sigma_{ik} + O_p(m^{-3/2})$. Note that $\hat{\boldsymbol{\delta}}_e - \boldsymbol{\delta}_e = O_p(m^{-1/2})$. Further, $\tilde{\sigma}_{ii} = O(1)$ as well as $\tilde{\sigma}_{ik} = O(m^{-1})$ for $i \neq k$ and this order is preserved for its derivatives with respect to $\boldsymbol{\delta}$. Thus,

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_{ik}) &= \mathrm{E}\big[\{\hat{\sigma}_{ik} - \tilde{\sigma}_{ik} + O(m^{-3/2})\}^2\big] \\
&= \mathrm{E}\left[\left\{(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t\frac{\partial\tilde{\sigma}_{ik}}{\partial\,\boldsymbol{\delta}} + (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t\frac{\partial^2\tilde{\sigma}_{ik}}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O_p(m^{-3/2})\right\}^2\right]
\end{aligned}$$

$$= \mathbb{1}_{i=k} O(m^{-1}) + O(m^{-3}).$$

Using that for a random variable X with finite variance $X = \mathrm{E}(X) + O_p\{\sqrt{\mathrm{Var}(X)}\}$, it follows that $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} - \mathbf{C}$ where

$$\begin{aligned}\mathbf{\Sigma} &= \mathrm{diag}[\{O(1)\}_m] + \{O(m^{-1})\}_{m\times m},\\ \mathbf{C} &= \mathrm{diag}[\{O_p(m^{-1/2})\}_m] + \{O_p(m^{-3/2})\}_{m\times m}.\end{aligned}$$

It is now shown that inverting preserves the error. Note that $(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1} = \{O_p(m^{-1})\}_{p\times p}$ and let $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_m)$ for $\mathbf{d}_i$ as in (B2) as well as $\mathbf{K}_1 = \mathbf{K}_1(\boldsymbol{\delta})$. With (6), the matrix inversion formula yields

$$\begin{aligned}\mathbf{\Sigma}^{-1} &= \left\{\mathbf{K}_1 + \mathbf{D}^t\left(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X}\right)^{-1}\mathbf{D}\right\}^{-1}\\ &= \mathbf{K}_1^{-1} - \mathbf{K}_1^{-1}\,\mathbf{D}^t\left(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X} + \mathbf{D}\,\mathbf{K}_1^{-1}\,\mathbf{D}^t\right)^{-1}\mathbf{D}\,\mathbf{K}_1^{-1}\\ &= \mathbf{K}_1^{-1} + \{O(m^{-1})\}_{m\times m}.\end{aligned}$$

Thus, $\mathbf{C}\,\mathbf{\Sigma}^{-1} = \mathrm{diag}[\{O_p(m^{-1/2})\}_m] + \{O_p(m^{-3/2})\}_{m\times m}$. Denote $\lambda_{\mathbf{C}\,\mathbf{\Sigma}^{-1}}$ as largest eigenvalue of $\mathbf{C}\,\mathbf{\Sigma}^{-1}$. With the column-sum norm, $\lambda_{\mathbf{C}\,\mathbf{\Sigma}^{-1}} \leq \max_{k=1,\ldots,m}\sum_{i=1}^m |\{\mathbf{C}\,\mathbf{\Sigma}^{-1}\}_{ik}| = O(m^{-1/2}) < 1$ for large m. Writing the inverse as Neumann-series, $(\mathbf{I}_m - \mathbf{C}\,\mathbf{\Sigma}^{-1})^{-1} = \mathbf{I}_m + \mathrm{diag}[\{O_p(m^{-1/2})\}_m] + \{O_p(m^{-3/2})\}_{m\times m}$. Now

$$\begin{aligned}\widehat{\mathbf{\Sigma}}^{-1} &= \mathbf{\Sigma}^{-1}\left(\mathbf{I}_m - \mathbf{C}\,\mathbf{\Sigma}^{-1}\right)^{-1}\\ &= \mathbf{\Sigma}^{-1} + \mathrm{diag}[\{O_p(m^{-1/2})\}_m] + \{O_p(m^{-3/2})\}_{m\times m}.\end{aligned}$$

Eventually, since $m^{-1/2}\sum_{i=1}^m(\hat{\mu}_i - \mu_i)^2 = O_p(m^{1/2})$, it holds first that

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 = \|\,\mathbf{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 + O_p(m^{1/2}),$$

and second $Q = m^{-1}\|\,\mathbf{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 = O_p(1)$. Further, $U = O_p(m^{-1/2})$ with probability density function f_U and $z = m^{-1}\chi^2_{m,1-\alpha} = O(1)$, such that

$$\begin{aligned}&\mathrm{P}\left\{\|\widehat{\mathbf{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 < \chi^2_{m,1-\alpha}\right\} = \mathrm{P}\big(Q + U < z\big)\\ &= \int_{\mathbb{R}} \mathrm{P}\big(Q < z - u\big) f_U(u)du = \int_{\mathbb{R}}\left\{\mathrm{P}\big(Q < z\big) + O(m^{-1/2})\right\} f_U(u)du\\ &= 1 - \alpha + O(m^{-1/2}),\end{aligned}$$

which concludes the proof for (A1). For (A2), the reasoning with $m \to \infty$ goes analogously, so that it suffices to consider case $m = O(1)$. Analogous results to [33] and [7] follow directly as the diagonal entries of $\mathbf{\Sigma}$ are of the same order as off-diagonal entries, as

$$\begin{aligned}\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii} &= \mathbf{h}_i^t\big(\,\mathbf{G} - \mathbf{G}\,\mathbf{Z}_i^t\,\mathbf{V}_i^{-1}\,\mathbf{Z}_i\,\mathbf{G}\,\big)\mathbf{h}_i \\ &= \mathbf{h}_i^t\big(\,\mathbf{G}^{-1} - \mathbf{Z}_i^t\,\mathbf{R}_i\,\mathbf{Z}_i\,\big)^{-1}\mathbf{h}_i = O(n_i^{-1}),\end{aligned}$$

as $\mathbf{V}_i = \mathbf{Z}_i\,\mathbf{G}\,\mathbf{Z}_i^t + \mathbf{R}_i$. The same holds for derivatives with respect to $\boldsymbol{\delta}$. As $\hat{\boldsymbol{\delta}}_e - \boldsymbol{\delta}_e = O_p(m^{-1/2})$, a Taylor expansion yields

$$n_i\big[\{\mathbf{K}_1(\hat{\boldsymbol{\delta}})\}_{ii} - \{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}\big] = n_i(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t\frac{\partial\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}}{\partial\,\boldsymbol{\delta}} + O(m^{-1}) = O(m^{-1/2}).$$

Now, $s\widehat{\mathbf{\Sigma}}^{-1} = s\,\mathbf{\Sigma}^{-1} + \{O_p(m^{-1/2})\}_{m\times m}$ by analogous reasoning as in (A1). As for $m = O(1)$ the number of parameters does not grow,

$$\|\widehat{\mathbf{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 = \|\,\mathbf{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 + O_p(m^{-1/2}).$$

However, as neither does the quantile, $\chi^2_{m,1-\alpha} = O(1)$. This gives

$$\mathrm{P}\bigg\{\big\|\widehat{\mathbf{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\big\|^2 < \chi^2_{m,1-\alpha}\bigg\} = 1 - \alpha + O_p(m^{-1/2}),$$

which concludes the proof. □

Proof and Definitions for Theorem 1

First, note that $\boldsymbol{\delta}$ is not well-defined in the conditional model as parts of this vector that only describe the variability of the now-fixed random effects are meaningless. Below, $\boldsymbol{\delta}^c$ is interpreted as the solution of the respective expected minimization problem when either estimating with REML of Hendersons method III. Now, for the conditional scenario, let

$$\begin{aligned}\widehat{\mathbf{A}} = \mathbf{A}(\widehat{\mathbf{\Sigma}}_{\mathrm{c}}, \hat{\boldsymbol{\delta}}^c) = \big\{\big(\widehat{\mathbf{\Sigma}}_{\mathrm{c}}^{-1/2}\big)_i(\mathbf{b}_i^t\mathbf{Z}_i - \mathbf{h}_i^t)\big\}_{i=1,\ldots,m}(\mathbf{Z}^t\,\mathbf{Z})^{-1}\,\mathbf{Z}^t \\ + \sum_{i=1}^m(\widehat{\mathbf{\Sigma}}_{\mathrm{c}}^{-1/2})_i\mathbf{d}_i^t(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}^t\,\mathbf{V}^{-1}\,.\end{aligned} \tag{7}$$

With $\mathbf{S} = \mathbf{X}(\mathbf{X}^t\,\mathbf{H}\,\mathbf{X})^{-1}\,\mathbf{X}^t\,\mathbf{H}$ and $\mathbf{H} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\,\mathbf{Z}(\mathbf{Z}^t\,\mathbf{R}^{-1}\,\mathbf{Z})^{-1}\,\mathbf{Z}^t\,\mathbf{R}^{-1}$,

$$\hat{\lambda} = \tilde{\lambda}(\widehat{\mathbf{\Sigma}}_c, \hat{\boldsymbol{\delta}}^c) = \big\|\widehat{\mathbf{A}}\big(\mathbf{I}_n - \mathbf{S}\big)\,\mathbf{y}\,\big\|^2 - \big\|\widehat{\mathbf{A}}\big(\mathbf{I}_n - \mathbf{S}\big)\mathbf{R}^{1/2}\big\|^2. \tag{8}$$

Further, denote $\mathbf{A} = \mathbf{A}(\boldsymbol{\Sigma}_c, \boldsymbol{\delta}^c)$ and $\tilde{\lambda} = \tilde{\lambda}(\boldsymbol{\Sigma}_c, \boldsymbol{\delta}^c)$ if the variance components are known. Now, for $\widehat{\boldsymbol{\Sigma}}_c$ as an estimator for $\boldsymbol{\Sigma}_c$ reads as

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_c &= \widehat{\boldsymbol{\Sigma}}_c(\hat{\boldsymbol{\delta}}^c) = \mathbf{L}_1(\hat{\boldsymbol{\delta}}^c) + \mathbf{L}_2(\hat{\boldsymbol{\delta}}^c) + \widehat{\mathbf{L}}_3(\hat{\boldsymbol{\delta}}^c) + \widehat{\mathbf{L}}_4(\hat{\boldsymbol{\delta}}^c) - \ \widehat{\mathbf{L}}_5(\hat{\boldsymbol{\delta}}^c); \qquad (9)\\
\mathbf{L}_1(\boldsymbol{\delta}^c) &= \operatorname{diag}\big(\mathbf{b}_i^t\,\mathbf{R}_i\,\mathbf{b}_i\big)_{i=1,\dots,m},\\
\mathbf{L}_2(\boldsymbol{\delta}^c) &= \Big\{\mathbf{d}_i^t(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{R}\,\mathbf{V}^{-1}\,\mathbf{X}(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\mathbf{d}_k\\
&\qquad\qquad +\mathbf{b}_i^t\mathbf{R}_i\,\mathbf{V}_i^{-1}\,\mathbf{X}_i(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\mathbf{d}_k\\
&\qquad\qquad\quad +\mathbf{b}_k^t\mathbf{R}_k\,\mathbf{V}_k^{-1}\,\mathbf{X}_k(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\mathbf{d}_i\Big\}_{i,k=1,\dots,m},\\
\widehat{\mathbf{L}}_4(\boldsymbol{\delta}^c) &= \operatorname{diag}\left\{\operatorname{tr}\left(\frac{\partial \mathbf{b}_i^t}{\partial\,\boldsymbol{\delta}^c}\,\mathbf{R}_i\,\frac{\partial \mathbf{b}_i}{\partial(\boldsymbol{\delta}^c)^t}\overline{\mathbf{V}}\right)\right\}_{i=1,\dots,m},\\
\widehat{\mathbf{L}}_5(\boldsymbol{\delta}^c) &= \frac{1}{2}\operatorname{diag}\left\{\operatorname{tr}\left[\frac{\partial^2\{\mathbf{L}_1(\boldsymbol{\delta}^c)\}_{ii}}{\partial\,\boldsymbol{\delta}^c\,\partial(\boldsymbol{\delta}^c)^t}\overline{\mathbf{V}}\right]\right\}_{i=1,\dots,m}.
\end{aligned}
$$

As in the marginal case $\boldsymbol{\Sigma}_c = \mathbf{L}_1(\boldsymbol{\delta}^c) + \mathbf{L}_2(\boldsymbol{\delta}^c) + \mathbf{L}_3 + \mathbf{L}_4$. $\widehat{\mathbf{L}}_5$ serves as a estimator for the bias of the leading term $\mathbf{L}_1(\hat{\boldsymbol{\delta}}^c)$. The fourth term accounts for the estimation of the random effects, i.e. $\mathbf{L}_4 = \big[\operatorname{Cov}\{\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}} - \mathrm{E}(\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}|\mathbf{v})\}\big|\,\mathbf{v}\,\big]_{ik}$, whereas the third term does so for the cross-terms, that do not vanish in the conditional model as $\hat{\boldsymbol{\mu}}$ is biased. The term $\widehat{\mathbf{L}}_3 = \widehat{\mathbf{L}}_3(\boldsymbol{\delta}^c)$ differs for $\hat{\boldsymbol{\delta}}^c$ being a REML- or Hendersons method III-based. It can be split into $\widehat{\mathbf{L}}_3 = \widehat{\mathbf{L}}_3^* + (\widehat{\mathbf{L}}_3^*)^t$, $\widehat{\mathbf{L}}_3^* = \mathrm{E}\big[\{\tilde{\mu}_i - \mathrm{E}(\tilde{\mu}_i|\mathbf{v})\}\{\hat{\mu}_k - \tilde{\mu}_k - \mathrm{E}(\hat{\mu}_k - \tilde{\mu}_k|\mathbf{v})\}\big|\,\mathbf{v}\,\big]$. Define $\mathbf{w}_i \in \mathbb{R}^n$ such that $\hat{\mu}_i(\boldsymbol{\delta}^c) - \mathrm{E}\{\hat{\mu}_i(\boldsymbol{\delta}^c)|\mathbf{v}\} = \mathbf{w}_i^t\mathbf{e}$. Now, for Hendersons method III, $\hat{\delta}_e^c = \mathbf{y}^t\,\mathbf{C}_e\,\mathbf{y}$ for $e = 1,\dots,r$ and $\mathbf{C}_e \in \mathbb{R}^{n\times n}$ as given in [36]. It is an estimator for $\delta_e^c = \mathrm{E}(\mathbf{y}^t\,\mathbf{C}_e\,\mathbf{y}\,|\mathbf{v})$. Then, for $i, k = 1,\dots,m$,

$$
\big\{\widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c)\big\}_{ik} = \sum_{e=1}^{r}\left\{2\operatorname{tr}\left(\mathbf{w}_i\frac{\partial \mathbf{w}_k^t}{\partial \delta_e^c}\,\mathbf{R}\,\mathbf{C}_e\,\mathbf{R}\right) + \sum_{g=1}^{r}\operatorname{tr}\left(\mathbf{w}_i\frac{\partial^2 \mathbf{w}_k^t}{\partial\delta_e^c\partial\delta_g^c}\,\mathbf{R}\right)\overline{\mathbf{V}}_{eg}\right\}.
$$

For REML, let $\boldsymbol{\delta}^c$ such that $\frac{\partial}{\partial\,\boldsymbol{\delta}^c}\mathrm{E}\{\ell_{\mathrm{RE}}(\boldsymbol{\delta}^c)|\mathbf{v}\} = \mathbf{0}_r$, where ℓ_{RE} is the marginal restricted log-likelihood as spelled out in (10). Now let

$$
\begin{aligned}
\mathbf{D}_{ik}(e,d) &= \sum_{f=1}^{r}(\overline{\mathbf{V}})_{ef}\mathbf{w}_i(\overline{\mathbf{V}})_d^t\frac{\partial\overline{\mathbf{V}}^{-1}}{\partial\delta_e^c}\overline{\mathbf{V}}\frac{\partial \mathbf{w}_k^t}{\partial\,\boldsymbol{\delta}^c},\\
\mathbf{F}_{ik}(e,d) &= \sum_{f,g=1}^{r}(\overline{\mathbf{V}})_{ef}(\overline{\mathbf{V}})_{fg}\mathbf{w}_i\frac{\partial^2\mathbf{w}_k^t}{\partial\delta_e^c\partial\delta_d^c}.
\end{aligned}
$$

Further, with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}\,\mathbf{V}^{-1}\mathbf{X}^t)^{-1}\mathbf{X}^t\,\mathbf{V}^{-1}$, an estimator $\widehat{\boldsymbol{\Sigma}}_c = \widehat{\boldsymbol{\Sigma}}_c(\hat{\boldsymbol{\delta}}^c)$ for $\boldsymbol{\Sigma}_c$ is given by

$$\begin{aligned}\big\{\widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c)\big\}_{ik} = {}& 2\sum_{e=1}^{r}\mathrm{tr}\Big\{\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta_e^c}\mathbf{P}\,\mathbf{R}\,\mathbf{w}_i(\overline{\mathbf{V}})_e^t\frac{\partial\mathbf{w}_k^t}{\partial\,\boldsymbol{\delta}^c}\,\mathbf{R}\,\Big\} \\ &+2\sum_{e,d=1}^{r}\mathrm{tr}\Big[\big\{2\mathbf{F}_{ik}(e,d)-\mathbf{D}_{ik}(e,d)\big\}\,\mathbf{R}\,\Big]\big(\overline{\mathbf{V}}^{-1}\big)_{ed} \\ &+\sum_{e,d,g=1}^{r}\mathrm{tr}\Big\{\mathbf{w}_i(\overline{\mathbf{V}})_e^t\frac{\partial\mathbf{w}_k^t}{\partial\,\boldsymbol{\delta}^c}\,\mathbf{R}\,\Big\}\frac{\partial(\overline{\mathbf{V}}^{-1})_{ef}}{\partial\delta_g^c}(\overline{\mathbf{V}})_{ed}.\end{aligned}$$

For the proof of Theorem 1, two preliminary results are required.

Lemma 1. *Let $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ be nonstochastic and $\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{V})$. Then,*

$$\text{(i)}\ \ \mathrm{E}\Big(\prod_{i=1}^{2}\mathbf{u}^t\mathbf{A}_i\mathbf{u}\Big) = 2\mathrm{tr}\big(\mathbf{A}_1\,\mathbf{V}\,\mathbf{A}_2\,\mathbf{V}\,\big)+\mathrm{tr}\big(\mathbf{A}_1\,\mathbf{V}\,\big)\mathrm{tr}\big(\mathbf{A}_2\,\mathbf{V}\,\big),$$

$$\begin{aligned}\text{(ii)}\ \ \mathrm{E}\Big(\prod_{i=1}^{3}\mathbf{u}^t\mathbf{A}_i\mathbf{u}\Big) = {}& \prod_{i=1}^{3}\mathrm{tr}\big(\mathbf{A}_i\,\mathbf{V}\,\big)+2\mathrm{tr}\big(\mathbf{A}_1\,\mathbf{V}\,\big)\mathrm{tr}\big(\mathbf{A}_2\,\mathbf{V}\,\mathbf{A}_3\,\mathbf{V}\,\big) \\ &+2\mathrm{tr}\big(\mathbf{A}_2\,\mathbf{V}\,\big)\mathrm{tr}\big(\mathbf{A}_1\,\mathbf{V}\,\mathbf{A}_3\,\mathbf{V}\,\big)+4\mathrm{tr}\big(\mathbf{A}_2\,\mathbf{V}\,\mathbf{A}_1\,\mathbf{V}\,\mathbf{A}_3\,\mathbf{V}\,\big) \\ &+2\mathrm{tr}\big(\mathbf{A}_3\,\mathbf{V}\,\big)\mathrm{tr}\big(\mathbf{A}_2\,\mathbf{V}\,\mathbf{A}_1\,\mathbf{V}\,\big)+4\mathrm{tr}\big(\mathbf{A}_1\,\mathbf{V}\,\mathbf{A}_2\,\mathbf{V}\,\mathbf{A}_3\,\mathbf{V}\,\big).\end{aligned}$$

This Lemma follows by direct application of Theorem 1 of [38].

Lemma 2. *Let model (1) hold. Under (A1) with regularity conditions conditions (B1) and (B4) and $\hat{\boldsymbol{\delta}}^c$ being a REML estimate, let $\mathbf{s}$ be the score vector of $\hat{\boldsymbol{\delta}}^c$ and $\overline{\mathbf{V}}^{-1}$ its information matrix, and $\boldsymbol{\Lambda}$ as in (11). Then,*

$$\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c = \mathbf{g}_1 + \mathbf{g}_2 - \mathbf{g}_3 + \{O_p(m^{-3/2})\}_r;$$

$$\begin{aligned}\mathbf{g}_1 &= \overline{\mathbf{V}}\mathbf{s} = \{O_p(m^{-1/2})\}_r, \\ \mathbf{g}_2 &= \overline{\mathbf{V}}\,\boldsymbol{\Lambda}\,\overline{\mathbf{V}}\mathbf{s} = \{O_p(m^{-1})\}_r, \\ \mathbf{g}_3 &= \frac{1}{2}\sum_{g=1}^{r}(\overline{\mathbf{V}}\mathbf{s})_g\overline{\mathbf{V}}\frac{\partial\overline{\mathbf{V}}^{-1}}{\partial\delta_g^c}\overline{\mathbf{V}}\mathbf{s} = \{O_p(m^{-1})\}_r.\end{aligned}$$

Proof. Denote $\ell_{\mathrm{RE}}(\boldsymbol{\delta}^c)$ as the restricted log-likelihood function such that

$$\ell_{\mathrm{RE}}(\boldsymbol{\delta}^c) \propto -\frac{1}{2}\mathrm{log}|\,\mathbf{V}\,| - \frac{1}{2}\mathrm{log}|\,\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X}\,| - \frac{1}{2}\,\mathbf{y}^t\,\mathbf{P}\,\mathbf{y} \tag{10}$$

with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\,\mathbf{X}(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}^t\,\mathbf{V}^{-1}$. As $\frac{\partial}{\partial\delta^c_f}\mathbf{P} = -\mathbf{P}\frac{\partial}{\partial\delta^c_f}\mathbf{V}\mathbf{P}$ and $\mathbf{P}\mathbf{V}\mathbf{P} = \mathbf{P}$, score vector and matrix of second derivatives read as

$$\begin{aligned}
\mathbf{s}(\boldsymbol{\delta}^c) = \frac{\partial \ell_{\mathrm{RE}}}{\partial\,\boldsymbol{\delta}^c}(\boldsymbol{\delta}^c) &= \left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_d}\right) + \frac{1}{2}\,\mathbf{y}^t\,\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_d}\mathbf{P}\,\mathbf{y}\right\}_d;\\
\frac{\partial^2\ell_{\mathrm{RE}}}{\partial\,\boldsymbol{\delta}^c\,\partial(\boldsymbol{\delta}^c)^t}(\boldsymbol{\delta}^c) &= -\overline{\mathbf{V}}^{-1} + \boldsymbol{\Lambda},\\
\left(\overline{\mathbf{V}}^{-1}\right)_{ef} &= \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_f}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_e}\right),\\
(\boldsymbol{\Lambda})_{ef} &= \mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_f}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_e}\right) - \mathbf{y}^t\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_f}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta^c_e}\mathbf{P}\mathbf{y}.
\end{aligned} \tag{11}$$

The information matrix of $\hat{\boldsymbol{\delta}}^c$ is $\overline{\mathbf{V}}^{-1} = \{O(m)\}_{r\times r}$. Further, as $\mathrm{E}\{s(\boldsymbol{\delta}^c)|\mathbf{v}\} = \mathbf{0}_r$, it follows that

$$\mathrm{E}\big\{(\boldsymbol{\Lambda})_{fe}|\mathbf{v}\big\} = \mathrm{tr}\left\{\mathbf{P}\,\frac{\partial\,\mathbf{V}}{\partial\delta^c_e}\,\mathbf{P}\,\frac{\partial\,\mathbf{V}}{\partial\delta^c_f}\,\mathbf{P}\,\mathbf{Z}(\mathbf{v}\,\mathbf{v}^t - \mathbf{G})\,\mathbf{Z}^t\right\} = 0,$$

too. Further, by Lemma 1 (ii) this gives $\mathrm{E}\{(\boldsymbol{\Lambda})_{ef}|\,\mathbf{v}\} + O_p[\sqrt{\mathrm{Var}\{(\boldsymbol{\Lambda})_{ef}|\,\mathbf{v}\}}] = O_p(m^{1/2})$. Together with $\mathbf{s}(\boldsymbol{\delta}^c) = \{O_p(m^{1/2})\}_r$, this gives

$$\left\{\frac{\partial^3\ell_{\mathrm{RE}}}{\partial\delta^c_e\partial\delta^c_f\partial\delta^c_g}(\boldsymbol{\delta}^c)\right\}_{e,f} = -\frac{\partial}{\partial\delta^c_g}\overline{\mathbf{V}}^{-1} + \{O_p(m^{1/2})\}_{r\times r}.$$

We continue with a Taylor expansion for $\mathbf{s}(\hat{\boldsymbol{\delta}}^c) = \mathbf{0}_r$ around the score vector $\mathbf{s}(\boldsymbol{\delta}^c)$. Next, suppress the argument of the score vector, e.g. $\mathbf{s}$ refers to the score vector and $\mathbf{s}_d$ to its d-th entry. Then,

$$\begin{aligned}
\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c &= \overline{\mathbf{V}}\mathbf{s} + \overline{\mathbf{V}}\,\boldsymbol{\Lambda}(\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c)\\
&\qquad - \frac{1}{2}\sum_{g=1}^{r}(\hat{\delta}^c_g - \delta^c_g)\overline{\mathbf{V}}\frac{\partial\overline{\mathbf{V}}^{-1}}{\partial\delta^c_g}(\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c) + \{O_p(m^{-3/2})\}_r\\
&= \overline{\mathbf{V}}\mathbf{s} + \overline{\mathbf{V}}\,\boldsymbol{\Lambda}\,\overline{\mathbf{V}}\mathbf{s} - \frac{1}{2}\sum_{g=1}^{r}(\overline{\mathbf{V}}\mathbf{s})_g\overline{\mathbf{V}}\frac{\partial\overline{\mathbf{V}}^{-1}}{\partial\delta^c_g}\overline{\mathbf{V}}\mathbf{s} + \{O_p(m^{-3/2})\}_r.
\end{aligned}$$

This gives the claim. □

Lemma 3. *Let model (1) hold with definitions above and let* $\hat{\boldsymbol{\delta}}^c$ *be being a REML estimator. Under (A1) or (A2), with (B1)-(B4) it holds*

(i) $\mathbf{L}_1(\boldsymbol{\delta}^c) = \mathrm{E}\big\{\mathbf{L}_1(\hat{\boldsymbol{\delta}}^c) - \widehat{\mathbf{L}}_5(\hat{\boldsymbol{\delta}}^c)\big|\,\mathbf{v}\big\} + \{O(m^{-3/2})\}_{m\times m}$,

(ii) $\mathbf{L}_2(\boldsymbol{\delta}^c) = \mathrm{E}\{\mathbf{L}_2(\hat{\boldsymbol{\delta}}^c) |\, \mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$,

(iii) $\quad \mathbf{L}_3 = \mathrm{E}\{\widehat{\mathbf{L}}_3(\hat{\boldsymbol{\delta}}^c) |\, \mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$,

(iv) $\quad \mathbf{L}_4 = \mathrm{E}\{\widehat{\mathbf{L}}_4(\hat{\boldsymbol{\delta}}^c) |\, \mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$.

Proof. Consider (A1) only, as (A2) goes analogously to the considerations in the proof of Theorem 2 as $\mathbf{L}_1(\boldsymbol{\delta}^c) = \{O_p(n_i^{-1})\}_{m\times m}$ for $m = O(1)$. Also, (ii) and (iv) are obtained analogously to [7]. For (iii) we show first that

$$\mathbf{L}_3 = \widehat{\mathbf{L}}_3(\boldsymbol{\delta}^c) + \{O_p(m^{-3/2})\}_{m\times m}. \tag{12}$$

Consider the Taylor expansion of $\hat{\mu}_k - E[\hat{\mu}_k|\mathbf{v}]$ around $\boldsymbol{\delta}^c$, multiply with $\mathbf{w}_i^t\mathbf{e}$ and take expectation. Then, $\mathbf{L}_3^* = \mathrm{E}\big\{\mathbf{w}_i^t\mathbf{e}(\widehat{\mathbf{w}}_k - \mathbf{w}_k)^t\mathbf{e}\big\}$. With $\mathbf{g}_j$, $j = 1, 2, 3$ as given in Lemma 2, this yields

$$(\widehat{\mathbf{w}}_k - \mathbf{w}_k)^t\mathbf{e} = (\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c)^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c} + (\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c)^t \frac{\partial^2 \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c\, \partial(\boldsymbol{\delta}^c)^t}(\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c) + O(m^{-3/2}).$$

Multiplying with $\mathbf{w}_i^t\mathbf{e}$ and taking expectations then gives

$$\begin{aligned}\mathbf{L}_3^* = \mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_1^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c}\bigg|\, \mathbf{v}\right) + \mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_2^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c}\bigg|\, \mathbf{v}\right) - \mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_3^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c}\bigg|\, \mathbf{v}\right)\\ + \mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_1^t \frac{\partial^2 \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c\, \partial(\boldsymbol{\delta}^c)^t}\mathbf{g}_1\bigg|\, \mathbf{v}\right) + O(m^{-3/2}),\end{aligned}$$

which we will show to lead to $\widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c)$. Each expectation above is evaluated one by one, using Lemma 1. First, Lemma 1 (i) yields

$$\begin{aligned}\mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_1^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c}\bigg|\, \mathbf{v}\right) &= \sum_{e=1}^r \mathrm{E}\left\{\mathbf{s}_e\mathbf{e}^t\mathbf{w}_i(\overline{\mathbf{V}})_e^t \frac{\partial \mathbf{w}_k^t}{\partial\, \boldsymbol{\delta}^c}\mathbf{e}\bigg|\, \mathbf{v}\right\}\\ &= 2\sum_{e=1}^r \mathrm{tr}\left\{\mathbf{P}\frac{\partial \mathbf{V}}{\partial \delta_e^c}\mathbf{P}\,\mathbf{R}\,\mathbf{w}_i(\overline{\mathbf{V}})_e^t \frac{\partial \mathbf{w}_k^t}{\partial\, \boldsymbol{\delta}^c}\,\mathbf{R}\right\}.\end{aligned}$$

Similarly, with Lemma 1 (ii) the next term gives

$$\begin{aligned}\mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\; \mathbf{g}_2^t \frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial\, \boldsymbol{\delta}^c}\bigg|\, \mathbf{v}\right) &= \sum_{e,g,f=1}^r \mathrm{E}\left\{\mathbf{s}_e(\boldsymbol{\Lambda})_{fg}\mathbf{e}^t\mathbf{w}_i(\overline{\mathbf{V}})_e^t \frac{\partial \mathbf{w}_k^t}{\partial\, \boldsymbol{\delta}^c}\mathbf{e}(\overline{\mathbf{V}})_{ef}\bigg|\, \mathbf{v}\right\}\\ = -2\sum_{e,g,f=1}^r \mathrm{tr}\left\{\mathbf{w}_i(\overline{\mathbf{V}})_e^t \frac{\partial \mathbf{w}_k^t}{\partial\, \boldsymbol{\delta}^c}\,\mathbf{R}\right\}&\mathrm{tr}\left(\frac{\partial\, \mathbf{V}}{\partial \delta_e^c}\mathbf{P}\frac{\partial\, \mathbf{V}}{\partial \delta_f^c}\mathbf{P}\frac{\partial\, \mathbf{V}}{\partial \delta_g^c}\mathbf{P}\right)(\overline{\mathbf{V}})_{ef} + O(m^{-2}),\end{aligned}$$

and note that $-2\partial(\overline{\mathbf{V}}^{-1})_{ef}/\partial\delta_g^c = \mathrm{tr}(\frac{\partial\mathbf{V}}{\partial\delta_e^c}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta_g^c}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta_e^c}\mathbf{P}) + \mathrm{tr}(\frac{\partial\mathbf{V}}{\partial\delta_e^c}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta_g^c}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\delta_f^c}\mathbf{P})$. For the next term, note that $\mathbf{D}_{ik}(e,d)$ has only entries of order $O(m^{-3})$ except on the submatrix $\{O(m^{-2})\}_{n_i\times n_k}$ corresponding to the respective subjects. Hence,

$$\begin{aligned}\mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\;\mathbf{g}_3^t\frac{\partial\mathbf{w}_k^t\mathbf{e}}{\partial\,\boldsymbol{\delta}^c}\middle|\,\mathbf{v}\right) &= \frac{1}{2}\sum_{e,d=1}^{r}\mathrm{E}\left\{\mathbf{s}_e\mathbf{s}_d\;\mathbf{e}^t\mathbf{D}_{ik}(e,d)\mathbf{e}\middle|\,\mathbf{v}\right\}\\ &= 2\sum_{e,d=1}^{r}\mathrm{tr}\{\mathbf{D}_{ik}(e,d)\,\mathbf{R}\,\}(\overline{\mathbf{V}}^{-1})_{ed} + O(m^{-2}),\end{aligned}$$

by Lemma 1 (ii). The last term eventually gives also by Lemma 1 (ii) that

$$\begin{aligned}\mathrm{E}\left(\mathbf{w}_i^t\mathbf{e}\;\mathbf{g}_1^t\frac{\partial^2\mathbf{w}_k^t\mathbf{e}}{\partial\,\boldsymbol{\delta}^c\,\partial(\boldsymbol{\delta}^c)^t}\mathbf{g}_1\middle|\,\mathbf{v}\right) &= \sum_{e,d=1}^{r}\mathrm{E}\left\{\mathbf{s}_e\mathbf{s}_d\;\mathbf{e}^t\mathbf{F}_{ik}(e,d)\mathbf{e}\middle|\,\mathbf{v}\right\}\\ &= 4\sum_{e,d=1}^{r}\mathrm{tr}\{\mathbf{F}_{ik}(e,d)\,\mathbf{R}\,\}(\overline{\mathbf{V}}^{-1})_{ed} + O(m^{-2}).\end{aligned}$$

Putting all terms together eventually gives $\mathbf{L}_3^* = \widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c) + \{O_p(m^{-3/2})\}_{m\times m}$ and thus (12). Note that $\widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c) = \{O_p(m^{-1})\}_{m\times m}$. Taking derivatives preserves the order and a Taylor expansion and taking expectations yields

$$\mathrm{E}\big[\{\widehat{\mathbf{L}}_3(\hat{\boldsymbol{\delta}}^c)\}_{ik}|\,\mathbf{v}\,\big] = \{\widehat{\mathbf{L}}_3(\boldsymbol{\delta}^c)\}_{ik} + O(m^{-2}) = (\mathbf{L}_3)_{ik} + O(m^{-2}),$$

as $\mathrm{E}(\hat{\boldsymbol{\delta}}^c - \boldsymbol{\delta}^c\,|\,\mathbf{v}) = \mathbf{0}_r$. The last equation follows by (12). This gives (iii).

(i) A similar approach for $\{\mathbf{L}_1(\hat{\boldsymbol{\delta}}^c)\}_{ii}$ around $\{\mathbf{L}_1(\boldsymbol{\delta}^c)\}_{ii}$ gives

$$\mathrm{E}\big[\{\mathbf{L}_1(\hat{\boldsymbol{\delta}}^c)\}_{ii}|\,\mathbf{v}\,\big] = \{\mathbf{L}_1(\boldsymbol{\delta}^c)\}_{ii} + \frac{1}{2}\mathrm{diag}\left\{\mathrm{tr}\left[\frac{\partial^2\{\mathbf{L}_1(\boldsymbol{\delta}^c)\}_{ii}}{\partial\,\boldsymbol{\delta}^c\,\partial(\boldsymbol{\delta}^c)^t}\overline{\mathbf{V}}\right]\right\}_i + O(m^2).$$

This concludes the proof for Lemma 3. □

Lemma 4. *Let model (1) hold with definitions above and let $\hat{\boldsymbol{\delta}}^c$ given by Hendersons method III. Under (A1) or (A2), with (B1)-(B4) it holds*

(i) $\mathbf{L}_1(\boldsymbol{\delta}^c) = \mathrm{E}\{\mathbf{L}_1(\hat{\boldsymbol{\delta}}^c) - \widehat{\mathbf{L}}_5(\hat{\boldsymbol{\delta}}^c)|\,\mathbf{v}\,\} + \{O(m^{-3/2})\}_{m\times m}$,

(ii) $\mathbf{L}_2(\boldsymbol{\delta}^c) = \mathrm{E}\{\mathbf{L}_2(\hat{\boldsymbol{\delta}}^c)|\,\mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$,

(iii) $\mathbf{L}_3 = \mathrm{E}\{\widehat{\mathbf{L}}_3(\hat{\boldsymbol{\delta}}^c)|\,\mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$,

(iv) $\mathbf{L}_4 = \mathrm{E}\{\widehat{\mathbf{L}}_4(\hat{\boldsymbol{\delta}}^c)|\,\mathbf{v}\} + \{O_p(m^{-3/2})\}_{m\times m}$.

Proof. (of Lemma 4). Consider (A1) only, as (A2) analogously unless $m = O(1)$, in which case the leading term is $O(m^{-1})$. Recall that $\hat{\delta}_e = \mathbf{y}^t\,\mathbf{C}_e\,\mathbf{y}$ where $\mathbf{C}_e = \mathrm{diag}[\{O(m^{-1})\}_{n_i\times n_i}]_{i=1,\dots,m} + \{O(m^{-2})\}_{n\times n}$. Further, (i) and (ii) hold as in Lemma 3 as they do not depend on the different nature of $\hat{\boldsymbol{\delta}}^c$.

(iii) The only part that remains to be treated is $\widehat{\mathbf{L}}_3^*$. As all entries are of order $O(m^{-1})$, it suffices to show

$$\mathbf{L}_3^* = \widehat{\mathbf{L}}_3^*(\boldsymbol{\delta}^c) + \{O(m^{-3/2})\}_{m\times m}$$

To show this, rewrite $\hat{\boldsymbol{\delta}}^c$ terms of $\mathbf{e}$, namely

$$\hat{\delta}_e^c - \delta_e^c = \mathbf{e}^t\mathbf{C}_e\mathbf{e} - \mathrm{tr}\{\mathbf{C}_e\,\mathbf{R}\} + 2\mathbf{e}^t\mathbf{C}_e(\mathbf{Z}\,\mathbf{v} + \mathbf{X}\,\boldsymbol{\beta}).$$

Adapting to the abbreviations of before, $\mathbf{L}_3^* = \mathrm{E}\{\mathbf{w}_i^t\mathbf{e}(\widehat{\mathbf{w}}_k - \mathbf{w}_k)^t\mathbf{e}\} = a_{ik} + b_{ik} + O(m^{-3/2})$, where

$$\begin{aligned} a_{ik} &= \sum_{e=1}^{r} \mathrm{E}\left\{ \mathbf{w}_i^t\mathbf{e}(\hat{\delta}_e^c - \delta_e^c)\frac{\partial \mathbf{w}_k^t\mathbf{e}}{\partial \delta_e^c}\right| \mathbf{v} \left.\right\}, \\ b_{ik} &= \sum_{e=1}^{r}\sum_{g=1}^{r} \mathrm{E}\left\{ \mathbf{w}_i^t\mathbf{e}(\hat{\delta}_e^c - \delta_e^c)(\hat{\delta}_g^c - \delta_g^c)\frac{\partial^2 \mathbf{w}_k^t\mathbf{e}}{\partial \delta_e^c\partial\delta_g^c}\right| \mathbf{v} \left.\right\}. \end{aligned}$$

Both terms are treated in turn. First, Lemma 1 (i) gives

$$a_{ik} = \sum_{e=1}^{m} 2\mathrm{tr}\left\{ \mathbf{w}_i \frac{\partial \mathbf{w}_k^t}{\partial \delta_e^c}\,\mathbf{R}\,\mathbf{C}_e\,\mathbf{R} \right\} .$$

For the next term, Lemma 1 (ii) yields

$$b_{ik} = \sum_{e=1}^{r}\sum_{g=1}^{r} \mathrm{tr}\left\{ \mathbf{w}_i \frac{\partial^2 \mathbf{w}_k^t}{\partial \delta_e^c\partial\delta_f^c}\,\mathbf{R} \right\}\overline{\mathbf{V}}_{eg} + O(m^{-2}).$$

(iv) The proof goes similar to Lemma 3. As all entries of $\widehat{\mathbf{L}}_4$ are of order $O(m^{-1})$, it suffices to show

$$\mathbf{L}_4 = \widehat{\mathbf{L}}_4(\boldsymbol{\delta}^c) + \{O(m^{-3/2})\}_{m\times m}.$$

Using that $\mathbf{C}_f\,\mathbf{R}\,\mathbf{C}_e = \mathrm{diag}[\{O(m^{-2})\}_{n_i\times n_i}] + \{O(m^{-3})\}_{n\times n}$ and $\frac{\partial \mathbf{w}_k}{\partial \delta_f^c}\frac{\partial \mathbf{w}_i^t}{\partial \delta_e^c}$ has entries

$O(m^{-1})$ except on the submatrix $\{O(1)\}_{n_k\times n_i}$, it follows that

$$
\begin{aligned}
(\mathbf{L}_4)_{ik} &= \sum_{e=1}^{r}\sum_{f=1}^{r}\mathrm{E}\left\{(\hat{\delta}^c_e-\delta^c_e)(\hat{\delta}^c_f-\delta^c_f)\mathbf{e}^t\frac{\partial \mathbf{w}_k}{\partial \delta^c_f}\frac{\partial \mathbf{w}_i^t}{\partial \delta^c_e}\mathbf{e}\,\middle|\,\mathbf{v}\right\}+O(m^{-3/2})\\
&=\left\{\sum_{e=1}^{r}\sum_{f=1}^{r}2\mathrm{tr}(\mathbf{C}_e\,\mathbf{R}\,\mathbf{C}_f\,\mathbf{R})+4(\mathbf{X}\,\boldsymbol{\beta}-\mathbf{Z}\,\mathbf{v})^t\mathbf{C}_e\,\mathbf{R}\,\mathbf{C}_f(\mathbf{X}\,\boldsymbol{\beta}-\mathbf{Z}\,\mathbf{v})\right\}\\
&\qquad\qquad\qquad\qquad\cdot\,\mathrm{tr}\left\{\frac{\partial \mathbf{b}}{\partial \delta^c_e}\frac{\partial \mathbf{b}^t}{\partial \delta^c_f}\mathbf{R}\right\}+O(m^{-3/2})\\
&=\widehat{\mathbf{L}}_4(\boldsymbol{\delta}^c)+O(m^{-3/2})
\end{aligned}
$$

as the first factor equals $\mathrm{Cov}\{(\mathbf{y}^t\mathbf{C}_e\mathbf{y})_{e=1,\dots,r}|\,\mathbf{v}\}=\overline{\mathbf{V}}$. Since $\hat{\boldsymbol{\delta}}^c$ is unbiased and $\boldsymbol{\delta}^c=\hat{\boldsymbol{\delta}}^c+O(m^{-1/2})$, the remaining part of the proof follows analogously to that one of Lemma 3. □

Proof. (of Theorem 1). With Lemma 3 and Lemma 4 the proof for Theorem 2 can be replicated, which gives

$$
\mathrm{P}\left\{\|\widehat{\boldsymbol{\Sigma}}_c^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\|^2<\chi^2_{m,1-\alpha}(\lambda)\,\middle|\,\mathbf{v}\right\}=1-\alpha+O(m^{-1/2}).
$$

Thus, for (A1) it remains to show that $\chi^2_{m,1-\alpha}(\lambda)=\chi^2_{m,1-\alpha}(\hat{\lambda})+O_p(m^{1/2})$. First, define $\tilde{\lambda}$ for $\hat{\lambda}$ from (8) with $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\delta}^c$ instead of $\widehat{\boldsymbol{\Sigma}}_c$ and $\hat{\boldsymbol{\delta}}^c$. Similarly, define $\mathbf{A}$ for $\widehat{\mathbf{A}}$ as in (7) and see that

$$
\mathbf{A}=\mathrm{diag}[\{O(1)\}_{m\times n_i}]_{i=1,\dots,m}+\{O(m^{-1})\}_{m\times n}.
$$

Analogously to the marginal scenario it holds that $\widehat{\boldsymbol{\Sigma}}_c^{-1/2}=\boldsymbol{\Sigma}_c^{-1/2}+\mathbf{B}$, for $\mathbf{B}=\mathrm{diag}[\{O_p(m^{-1/2})\}_m]+\{O_p(m^{-1})\}_{m\times m}$. As $\hat{\boldsymbol{\delta}}^c=\boldsymbol{\delta}^c+\{O_p(m^{-1/2})\}_r$ by Lemma 2 and thus it holds

$$
\begin{aligned}
\widehat{\mathbf{A}}&=\mathbf{A}+\left\{\mathbf{B}_i(\mathbf{b}_i^t\mathbf{Z}_i-\mathbf{h}_i^t)\right\}_i(\mathbf{Z}^t\,\mathbf{Z})^{-1}\,\mathbf{Z}^t+\sum_{i=1}^{m}\mathbf{B}_i\mathbf{d}_i^t(\mathbf{X}\,\mathbf{V}^{-1}\,\mathbf{X}^t)^{-1}\,\mathbf{X}\,\mathbf{V}^{-1}\\
&=\mathbf{A}+\mathrm{diag}[\{O(m^{-1/2})\}_{m\times n_i}]_i+\{O(m^{-1})\}_{m\times n}\;=\;\mathbf{A}+\mathbf{C}\,.
\end{aligned}
$$

Note that $\mathbf{A}^t\mathbf{C}=\mathrm{diag}[\{O_p(m^{-1/2})\}_{n_i\times n_i}]_i+\{O_p(m^{-1})\}_{n\times n}$ and further $n^{-1}\sum_{k=1}^{n}\{(\mathbf{I}_n-\mathbf{S})\,\mathbf{y}\}_k=O_p(m^{-1/2})$. As $\hat{\boldsymbol{\delta}}^c$ only occurs in $\mathbf{R}(\hat{\boldsymbol{\delta}}^c)$ in $\hat{\lambda}=\tilde{\lambda}(\widehat{\boldsymbol{\Sigma}}_c,\hat{\boldsymbol{\delta}}^c)$ by (8), multiplying out and a Taylor expansion for $\hat{\boldsymbol{\delta}}^c$ around $\boldsymbol{\delta}^c$ with $\hat{\boldsymbol{\delta}}^c-\boldsymbol{\delta}^c=\{O_p(m^{-1/2})\}_r$ leads to

$$
\hat{\lambda}\;=\tilde{\lambda}\;\;+2\,\mathbf{y}^t(\mathbf{I}_n-\mathbf{S})\mathbf{A}^t\mathbf{C}(\mathbf{I}_n-\mathbf{S})\,\mathbf{y}+\left\|\mathbf{C}(\mathbf{I}_n-\mathbf{S})\,\mathbf{y}\,\right\|^2
$$

$$-2\mathrm{tr}\big\{(\mathbf{I}_n-\mathbf{S})\mathbf{A}^t\mathbf{C}(\mathbf{I}_n-\mathbf{S})\,\mathbf{R}(\hat{\boldsymbol{\delta}}^c)\big\}-\big\|\mathbf{C}(\mathbf{I}_n-\mathbf{S})\{\mathbf{R}(\hat{\boldsymbol{\delta}}^c)\}^{1/2}\big\|^2$$
$$+O_p(m^{-1/2})\big\|\mathbf{A}(\mathbf{I}_n-\mathbf{S})\{\mathbf{R}(\boldsymbol{\delta}^c)\}^{1/2}\big\|^2+O_p(1)=\tilde{\lambda}+O_p(m^{1/2}).$$

By construction $\mathrm{E}\{\tilde{\lambda}|\mathbf{v}\}=\lambda$. For its variance, we get

$$\begin{aligned}\mathrm{Var}\{\tilde{\lambda}|\mathbf{v}\} &= 6\mathrm{Var}\{\|\mathbf{A}(\mathbf{I}_n-\mathbf{S})\,\mathbf{y}\,\|^2|\mathbf{v}\}=\mathrm{Var}\{\|\mathbf{A}\,\mathbf{Z}\,\mathbf{v}+\mathbf{A}(\mathbf{I}_n-\mathbf{S})\mathbf{e}\|^2|\mathbf{v}\}\\ &= 2\mathrm{tr}\big\{(\mathbf{I}_n-\mathbf{S})\mathbf{A}^t\mathbf{A}(\mathbf{I}_n-\mathbf{S})\mathbf{R}(\mathbf{I}_n-\mathbf{S})\mathbf{A}^t\mathbf{A}(\mathbf{I}_n-\mathbf{S})\mathbf{R}\big\}\\ &\qquad +\mathbf{v}^t\,\mathbf{Z}^t\,\mathbf{A}^t\mathbf{A}(\mathbf{I}_n-\mathbf{S})\,\mathbf{R}(\mathbf{I}_n-\mathbf{S})\mathbf{A}^t\mathbf{A}\,\mathbf{Z}\,\mathbf{v}=O(m)\end{aligned}$$

by Lemma 1 (i). Hence, $\hat{\lambda}=\lambda+O_p(m^{1/2})$. Eventually,

$$\chi^2_{m,1-\alpha}\big(\hat{\lambda}\big)=\chi^2_{m,1-\alpha}\big\{\lambda+O_p(m^{1/2})\big\}=\chi^2_{m,1-\alpha}(\lambda)+O_p(m^{1/2}).$$

This concludes the proof for (A1). For (A2) the reasoning is similar. If moreover $m=O(1)$, the respective quantities are smaller, namely

$$\mathbf{A}=\mathrm{diag}[\{O(m^{-1})\}_{m\times n_i}]_{i=1,\ldots,m}+\{O(m^{-2})\}_{m\times n},$$
$$\mathbf{C}=\mathrm{diag}[\{O(m^{-3/2})\}_{m\times n_i}]_{i=1,\ldots,m}+\{O(m^{-2})\}_{m\times n},$$

which gives that $\hat{\lambda}=\tilde{\lambda}(\boldsymbol{\Sigma}_c,\boldsymbol{\delta}^c)+O_p(m^{-1/2})$, $\mathrm{Var}\{\tilde{\lambda}(\boldsymbol{\Sigma}_c,\boldsymbol{\delta}^c)|\mathbf{v}\}=O(m^{-1})$ and thus $\chi^2_{m,1-\alpha}\big(\hat{\lambda}\big)=\chi^2_{m,1-\alpha}(\lambda)+O_p(m^{-1/2})$. This proves Theorem 1. □

Proof and Definitions for Theorem 3

Another way to obtain a pivotal for simultaneous inference is to evaluate the distribution of the quadratic form $Q=\|\,\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\|^2$ under conditional law. It is distributed as generalized non-central χ^2, and thus has no analytically tractable probability density function. However, due to the linearity of $\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}$ in $\mathbf{v}$, the quadratic form Q can be suitably split up in treatable terms.

In the conditional scenario, $\mathbf{v}$ is seen as pre-fixed. Hereafter it is treated as being generated by the underlying marginal model $\mathbf{v}\sim\mathcal{N}(\mathbf{0}_m,\mathbf{G})$. Generally, it is merely required that $\mathbf{v}$ does not depart too much from $\mathbf{G}$, namely,

$$\frac{1}{\sqrt{q}}\sum_{i=1}^{q}\sum_{j=1}^{q}(\mathbf{v})_i(\mathbf{v})_j-(\mathbf{G})_{ij}=O(1).$$

Proof. (of Theorem 3). Due to linearity of $\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}$, it holds that $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_c+\boldsymbol{\Sigma}_b$, where

$\boldsymbol{\Sigma}_b = \mathrm{Cov}(\boldsymbol{\mu}_b)$ for $\boldsymbol{\mu}_b = \mathrm{E}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\,|\mathbf{v})$ by the law of total variance. Moreover,

$$\boldsymbol{\Sigma}^{-1} = \big(\boldsymbol{\Sigma}_c + \boldsymbol{\Sigma}_b\big)^{-1} = \boldsymbol{\Sigma}_c^{-1} - \boldsymbol{\Sigma}_c^{-1}\big(\boldsymbol{\Sigma}_c^{-1} + \boldsymbol{\Sigma}_b^{-1}\big)^{-1}\boldsymbol{\Sigma}_c^{-1} = \boldsymbol{\Sigma}_c^{-1} - \mathbf{T}_c^{-1},$$

where $\mathbf{T}_c^{-1}$ fulfills $\boldsymbol{\Sigma}_c\,\mathbf{T}_c^{-1} = \boldsymbol{\Sigma}_b\,\boldsymbol{\Sigma}^{-1}$. Now consider $Q = S + R$ with

$$\begin{aligned} S &= \|\,\boldsymbol{\Sigma}_c^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \boldsymbol{\mu}_b)\|^2, \\ R &= \|\,\boldsymbol{\Sigma}^{-1/2}\,\boldsymbol{\mu}_b\,\|^2 + 2\,\boldsymbol{\mu}_b^t\,\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \boldsymbol{\mu}_b) - \|\mathbf{T}_c^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \boldsymbol{\mu}_b)\|^2. \end{aligned}$$

First consider the marginal law. Clearly, $Q \sim \chi^2_m$ and $S \sim \chi^2_m$. Thus, $\mathrm{E}(R) = 0$ and $\mathrm{Var}(R) = 4\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\Sigma}_b) - 2\mathrm{tr}\{(\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\Sigma}_b)^2\}$, which can be verified by calculating all quantities directly. Under conditional law, $Q|\mathbf{v} \not\sim \chi^2_m$ a.s., but as in the marginal case, $S|\mathbf{v} \sim \chi^2_m$. Thus, $\mathrm{E}(R|\mathbf{v}) \neq 0$ a.s. Note that *almost surely* (a.s.) refers to the joint, marginal distribution. In order to evaluate Q under conditional law, R is replaced by its marginal expectation and variance $R = \mathrm{E}(R) + O_p\{\sqrt{\mathrm{Var}(R)}\}$. For $m \to \infty$ this gives $R = O_p(m^{1/2})$ as $\mathrm{Var}(R) = O(m)$ by $\mathrm{diag}(\boldsymbol{\Sigma}) = \{O(1)\}_m$. For $m = O(1)$ we have $\mathrm{Var}(R) = \{O(m^{-1})\}_m$ and thus $R = O_p(m^{-1/2})$. This is a natural procedure, insofar $R|\mathbf{v}$ is interpreted as random variable that depends on the realization of $\mathbf{v}$, and those can be wrapped up by their marginal expectation and square-rooted variance. Now for $m \to \infty$, using that $S = O_p(m)$,

$$\begin{aligned} \mathrm{P}\big(Q < \chi^2_{m,1-\alpha}\big|\mathbf{v}\big) &= \mathrm{P}\left\{\frac{S}{m} + O_p(m^{-1/2}) < \frac{\chi^2_{m,1-\alpha}}{m}\middle|\,\mathbf{v}\right\} \\ &= \mathrm{P}\big(S < \chi^2_{m,1-\alpha}\big|\mathbf{v}\big) + O(m^{-1/2}) = 1 - \alpha + O_p(m^{-1/2}). \end{aligned}$$

Replacing $\boldsymbol{\Sigma}$ in Q by $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \{O_p(m^{-1/2})\}_{m\times m}$ gives $\|\widehat{\boldsymbol{\Sigma}}^{-1/2}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\|^2 = Q + O_p(m^{1/2})$ as in the proof of Theorem 2. The order of the error coincides with $\sqrt{\mathrm{Var}(R)} = O_p(m^{1/2})$ and above equation still holds. Analogously, for $m = O(1)$, $\chi^2_{m,1-\alpha} = O(1)$, which gives the stated result. □

Proof of Theorem 4

The proof transforms the simple contrast upon the unstandardized pivot to a general contrast upon a standardized pivot.

Proof. First we show that $\tilde{\mathbf{c}}^t\mathbf{1}_m = \mathbf{c}^t\,\boldsymbol{\Sigma}_c^{1/2}\,\mathbf{1}_m = O(m^{-1/2})$. Note that (C1), (C2) imply $\mathbf{b}_i^t\mathbf{R}_i\mathbf{b}_i = \mathbf{b}_k^t\mathbf{R}_k\mathbf{b}_k$, $i, k \leq w$, as

$$\mathbf{b}_i^t\mathbf{R}_i\mathbf{b}_i = \mathbf{h}_i^t\mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\big(\mathbf{I}_q + \mathbf{G}\mathbf{Z}_i^t\mathbf{V}_i^{-1}\mathbf{Z}_i\big)\mathbf{G}\mathbf{h}_i.$$

Further, $\mathbf{\Sigma}_c = \mathbf{b}_i^t \mathbf{R}_i \mathbf{b}_i \mathbf{I}_m + \mathbf{C}$, for $\mathbf{C} = \{O(m^{-1})\}_{m\times m}$ and as both matrices commute, they are simultaneously diagonalizable and the eigenvalues of $\mathbf{\Sigma}_c$ the sum of eigenvalues of its components above. It thus remains to evaluate $\mathbf{c}^t\mathbf{C}$ by (9). Some calculations finally yield with (C1), (C2) that $\tilde{\mathbf{c}}^t \mathbf{1}_m = O(m^{-1/2})$.Now let $\mathbf{c} \in \mathcal{S}_m$ and $Z \sim \mathcal{N}_m(\mathbf{0}_m, \mathbf{I}_m)$. Then,

$$\tilde{\mathbf{c}}^t Z = \tilde{\mathbf{c}}^t\, \mathbf{\Sigma}_c^{-1/2} \left\{ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - \mathrm{E}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\,|\mathbf{v}) \right\} = \mathbf{c}^t(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) - c_+ \eta_{\mathbf{c}}.$$

The claim follows for $\tilde{c}_+ = (\tilde{\mathbf{c}}^t)_{>0}\mathbf{1}_m$ by [37, Theorem 7.11]. Replacing $\tilde{\mathbf{c}}$ by $\hat{\mathbf{c}}$ as in the proof of Theorem 2 gives

$$\begin{aligned} \mathbf{c}^t \widehat{\mathbf{\Sigma}}_c^{1/2} &= \mathbf{c}^t\, \mathbf{\Sigma}_c^{1/2} + \mathbf{c}^t \mathrm{diag}\big[\big\{O_p(m^{-1/2})\big\}_m\big] + \mathbf{c}^t \big\{O_p(m^{-3/2})\big\}_{m\times m} \\ &= \mathbf{c}^t\, \mathbf{\Sigma}_c^{1/2} + O_p(m^{-1/2}), \end{aligned}$$

so $\hat{c}_+ = \tilde{c}_+ + O_p(m^{-1/2})$, which shows Theorem 4. □

Tukey's Method

For subjects with $\mathbf{v}_i = \mathbf{v}_k$ for all $i, k \le w < m$, (C1), (C2) imply that $\mathbf{c}^t \mathrm{E}\big(\hat{\mu}_i - \mu_i|\mathbf{v}\big) = O(m^{-1/2})$, and thus $\eta_{\mathbf{c}} = O(m^{-1/2})$:

$$\begin{aligned} \mathrm{E}\big(\hat{\mu}_i - \mu_i|\mathbf{v}\big) &= (\mathbf{l}_i^t - \mathbf{b}_i^t \mathbf{X}_i)(\mathbf{X}^t \mathbf{V}_i^{-1} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{V}_i^{-1} \mathbf{Z}\mathbf{v} + (\mathbf{l}_i^t - \mathbf{b}_i^t\, \mathbf{X}_i)\mathbf{v}_i \\ &= \mathbf{h}_i^t\, \mathbf{G}\, \mathbf{Z}_i^t \mathbf{V}_i^{-1} \mathbf{Z}_i \mathbf{v}_i - \mathbf{h}_i^t \mathbf{v}_i + O(m^{-1/2}). \end{aligned} \tag{13}$$

Testing for Equality of all Random Effects

For $w = m$, all random effects are zero under H_0 and the underlying model reduces to a linear model. For linear hypotheses, this allows for the application of F-tests, see [16]. For testing equality of all pairwise differences, the standard version of Tukey's method for balanced, or Tukey-Kramer for inbalanced sets, have to be applied, see [39, 30].

References

[1] G. E. Battese, R. M. Harter, and W. A. Fuller. An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83:28–36, 1988.

[2] Y. Benjamini and H. Braun. John W. Tukey's Contributions to Multiple Comparisons. *Annals of Statistics*, 30(6):1576–1594, 2002.

[3] S. Chatterjee, P. Lahiri, and H. Li. Parametric Bootstrap Approximation to the Distribution of EBLUP and Related Prediction Intervals in Linear Mixed Models. *The Annals of Statistics*, 36(3):1221–1245, 2008.

[4] K. Das, J. Jiang, and J. N. K. Rao. Mean Squared Error of Empirical Predictor. *The Annals of Statistics*, 32(2):828–840, 2004.

[5] G. S. Datta, M. Gosh, D. D. Smith, and P. Lahiri. On the Asymptotic Theory of Conditional and Unconditional Coverage Probabilities of Empirical Bayes Confidence Intervals. *Scandinavian Journal of Statistics*, 29:139–152, 2002.

[6] G. S. Datta, T. Kubokawa, I. Molina, and J. N. K. Rao. Estimation of Mean Squared Error of Model-Based Small Area Estimators. *TEST*, 20:367–388, 2011.

[7] G. S. Datta and P. Lahiri. A Unified Measure of Uncertainty of Estimated Best Linear Predictors in Small Area Estimation Problems. *Statistica Sinica*, 10:613–627, 2000.

[8] E. Demidenko. *Mixed Models: Theory and Applications.* Wiley Series in Probability and Statistics, Hoboken, NJ, 2004.

[9] D. Flores Agreda. *On the inference of random effects in Generalized Linear Mixed Models.* PhD thesis at University of Geneva, Geneva, 2017.

[10] N. Ganesh. Simultaneous Credible Intervals for Small Area Estimation Problems. *Journal of Multivariate Analysis*, 100(8):1610–1621, 2009.

[11] W. González-Manteiga, M.-J. Lombardía, I. Molina, D. Morales, and L. Santamaría. Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5):443–462, 2008.

[12] H.O. Hartley and J. N. K. Rao. Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika,*, 54(1/2):93–108, 1967.

[13] D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[14] C. R. Henderson. Estimation of Genetic Parameters. *The Annals of Mathematical Statistics*, 21:309–310, 1950.

[15] C. R. Henderson. Estimation of Variance and Covariance Components. *Biometrics*, 9(2):226–252, 1953.

[16] F. K. C. Hui, Samuel Müller, and A. H. Welsh. Testing random effects in linear mixed models: another lookat the F-test (with discussion). *Aust. N. Z. J. Stat.*, 61(1):61–84, 2019.

[17] Instituto Nacional de Estadística. Living Conditions Survey, 2008. Spain.

[18] H. Jiang, S. Sibillot, C. Proust, J.-M. Molina, and R. Thiébaut. Robustness of the linear mixed model to misspecified error distribution. *Computaional Statistics & Data Analysis*, 51:5142–5154, 2007.

[19] J. Jiang. Asymptotic Properties of the Empirical BLUP and BLUE in Mixed Linear Models. *Statistica Sinica*, 8:861–885, 1998.

[20] J. Jiang. *Linear and Generalized Linear Mixed Models and Their Applications.* Springer Series in Statistics, New York, NY, 2007.

[21] J. Jiang. *Asymptotic Analysis of Mixed Effects Models.* CRC Press, Hoboken, NJ, 2017.

[22] J. Jiang and P. Lahiri. Mixed Model Prediction and Small Area Estimation. *TEST*, 15(1):1–96, 2006.

[23] J. Jiang, T. Nguyen, and J.S. Rao. Best Predictive Small Area Estimation. *Journal of the American Statistical Association*, 106(494):732–745, 2011.

[24] R. N. Kackar and D. A. Harville. Unbiasedness of Two-Stage Estimation and Prediction Procedures for Mixed Linear Models. *Communication in Statistics*, 10:1249–1261, 1981.

[25] P. Kramlinger, T. Krivobokova, and S. Sperlich. Supplement to "Marginal and Conditional Multiple Inference in Linear Mixed Models", 2020.

[26] Y. Lee and J. A. Nelder. Conditional and marginal models: Another view. *Statist. Sci.*, 19(2):219–238, 05 2004.

[27] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses.* Springer, New York, NY, 2005.

[28] N. T. Longford. *Missing Data and Small-Area Estimation.* Springer, New York, NY, 2005.

[29] K. S. Miller. On the Inverse of the Sum of Matrices. *Mathematical Association of America*, 54(2):67–72, 1981.

[30] G. A. Milliken and D. E. Johnson. *Analysis of Messy Data, Volume 1.* Chapman & Hall, London, 1992.

[31] D. Pfeffermann. New important developments in small area estimation. *Statist. Sci.*, 28(1):40–68, 02 2013.

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

[33] N. G. N. Prasad and J. N. K. Rao. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

[34] J. N. K. Rao and I. Molina. *Small Area Estimation.* Wiley, Hoboken, NJ, 2nd edition, 2015.

[35] A. Richardson and A. H. Welsh. Asymptotic properties of restricted maximum likelihood (reml) estimates for hierarchical mixed linear models. *Australian & New Zealand Journal of Statistics*, 36:31–43, 1994.

[36] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components.* Wiley, Hoboken, NJ, 1992.

[37] J. Shao. *Mathematical Statistics.* Springer, New York, NY, 2010.

[38] V. K. Srivastava and R. Tiwari. Evaluation of Expectations of Products of Stochastic Matrices. *Scandinavian Journal of Statistics*, 3(3):135–138, 1976.

[39] J. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Reading, MA, 1977.

[40] J. W. Tukey. The Problem of Multiple Comparisons. Published in *The Collected Works of John W. Tukey: Multiple Comparisons, Volume VIII (1999). Edited by H. Braun, CRC Press, Boca Raton, Florida*, 1953.

[41] N. Tzavidis, L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla. From start to finish: a framework for the production of small area official statistics. *Journal of the Royal Statistical Society, Series A*, 181:927–979, 2018.

[42] F. Vaida and S. Blanchard. Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, 92(2):351–370, 2005.

Supplement to “Marginal and Conditional Multiple Inference in Linear Mixed Models”

This document includes the proof of Proposition 1 in [4], simulation studies evaluating Tukey’s multiple comparison tests as well as performance of marginal ellipsoids and as an extension to the results of [2] and [6] it is shown that the bias of the estimators actually vanishes with rate $O(m^{-2})$ instead of $O(m^{-3/2})$. The proof for this claim is related to the findings of [4].

Simulation Study.

Naturally, the set $\mathcal{M}_\alpha$ is especially suitable to use for marginal models. Table 1 shows results of a simulation for such a marginal scenario, i.e. in each iteration the random effects are drawn again. All other quantities are as described in [4]. As stated in Theorem 2 the nominal coverage is achieved asymptotically, though for finite samples undercoverage is induced due to uncertainty caused by the REML estimation.

Table 1: Coverage of 95%-confidence ellipsoids in model (5) under marginal law.

			$\boldsymbol{\delta} = (8, 2)$		$\boldsymbol{\delta} = (4, 4)$		$\boldsymbol{\delta} = (2, 8)$	
m	n_i	n_j	$\boldsymbol{\delta}$	REML	$\boldsymbol{\delta}$	REML	$\boldsymbol{\delta}$	REML
10	5	5	.95	.92	.95	.92	.95	.95
100	5	5	.95	.92	.95	.93	.95	.89
10	10	10	.95	.93	.95	.94	.95	.93
100	10	10	.95	.94	.95	.94	.95	.93
10	5	10	.95	.93	.96	.94	.96	.96
10	5	100	.96	.95	.96	.97	.98	.98

Table 2: Accuracy in % of Tukey’s multiple comparisons test at 5% for model (5) under conditional law.

			$\boldsymbol{\delta} = (8, 2)$		$\boldsymbol{\delta} = (4, 4)$		$\boldsymbol{\delta} = (2, 8)$	
m	n_i	n_j	$\boldsymbol{\delta}$	REML	$\boldsymbol{\delta}$	REML	$\boldsymbol{\delta}$	REML
10	5	5	4.93	6.96	5.08	5.18	5.06	3.47
100	5	5	4.71	5.49	4.72	5.08	4.53	3.85
10	10	10	5.24	5.63	5.19	4.59	5.23	2.86
100	10	10	4.81	4.86	4.83	4.70	4.81	4.12
10	5	10	4.95	7.49	4.93	5.79	4.90	4.18
10	10	100	5.21	8.47	5.21	7.60	5.21	5.27

Similarly, Table 2 assesses the accuracy of Tukey’s test for multiple comparisons. The simulations parameters are chosen to match those in Section 4 in [4]. That is, for half the subjects ($m^* = m/2$) being generated with equal random effect, the hypothesis $H_0 : \mu_i = \mu_j$ for $\forall i, j = 1, \dots, m^*$ was tested against a two-sided alternative. For all $\binom{m^*}{2}$ simple contrasts it was then checked whether the Tukey interval included zero, and the resulting rejection of H_0.

The simulation reassures that Tukey's method works very well within the generated data set. The nominal level is readily being achieved, even for estimated variance components, both for 10 ($m^* = 5$) and 1125 ($m^* = 50$) tests carried out in total.

Proof of Proposition 1

For result (a) it suffices to identify all quantities solely depending on $\mathbf{v}$ and marginal law, namely

$$c(\mathbf{v}) = z_{1-\frac{\alpha}{2}} \sum_{k=1}^{\infty} \binom{1/2}{k} \left\{ \frac{\operatorname{Var}(\hat{\mu}_i) - \operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})} \right\}^k - \operatorname{sign}(Z) \frac{\operatorname{E}(\hat{\mu}_i - \mu_i|\,\mathbf{v})}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}}. \tag{14}$$

For the second result, the the described phenomenon in scenario (A1) has been previously found by [9] and [5] for nonparametric regression. We borrow the ansatz of the latter for the proof.

Proof. (a) The result follows immediately from

$$\begin{aligned}
\operatorname{P}\left\{ \left| \frac{\hat{\mu}_i - \mu_i}{\sqrt{\operatorname{Var}(\hat{\mu}_i)}} \right| \le z_{1-\frac{\alpha}{2}} \,\middle|\, \mathbf{v} \right\} &= \operatorname{P}\left\{ \left| \frac{\hat{\mu}_i - \mu_i}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} \right| \le z_{1-\frac{\alpha}{2}} \frac{\sqrt{\operatorname{Var}(\hat{\mu}_i)}}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} \,\middle|\, \mathbf{v} \right\} \\
&= \operatorname{P}\left\{ |Z| \le z_{1-\frac{\alpha}{2}} \frac{\sqrt{\operatorname{Var}(\hat{\mu}_i)}}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} - \operatorname{sign}(Z) \frac{\operatorname{E}(\hat{\mu}_i - \mu_i|\,\mathbf{v})}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} \,\middle|\, \mathbf{v} \right\} \\
&= \operatorname{P}\left\{ |Z| \le z_{1-\frac{\alpha}{2}} \sqrt{1 + \frac{\operatorname{Var}(\hat{\mu}_i) - \operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} - \operatorname{sign}(Z) \frac{\operatorname{E}(\hat{\mu}_i - \mu_i|\,\mathbf{v})}{\sqrt{\operatorname{Var}(\hat{\mu}_i|\,\mathbf{v})}} \,\middle|\, \mathbf{v} \right\} \\
&= \operatorname{P}\left\{ |Z| \le z_{1-\frac{\alpha}{2}} + c(\mathbf{v}) \,\middle|\, \mathbf{v} \right\}.
\end{aligned}$$

(b) First consider (A1). Let ξ be a random variable independent to $\mathbf{e}$ with distribution putting equal weight on the points in $\{1, \dots, m\}$. Then,

$$\frac{1}{m} \sum_{i=1}^{m} \operatorname{P}\left(|T_i| \le z_{1-\frac{\alpha}{2}} \,|\, \mathbf{v}\right) = \operatorname{E}\left\{ \operatorname{E}\left(\mathbb{1}_{|T_\xi| \le z_{1-\frac{\alpha}{2}}} \,\middle|\, \mathbf{v}, \xi \right) \right\} = P\left(|T_\xi| \le z_{1-\frac{\alpha}{2}} \,|\, \mathbf{v}\right).$$

Now study the distribution of T_ξ under the joint law $(\xi, \mathbf{e})$. In particular, due to $\tilde{\boldsymbol{\beta}} =$

$\boldsymbol{\beta} + \{O_p(m^{-1/2})\}_p$,

$$T_\xi = \frac{\hat{\mu}_\xi - \mu_\xi}{\sqrt{\mathrm{Var}(\hat{\mu}_\xi - \mu_\xi)}} = \frac{(\mathbf{b}_\xi^t \mathbf{Z}_\xi - \mathbf{h}_\xi^t)\mathbf{v}_\xi + \mathbf{b}_\xi^t \mathbf{e}_\xi}{\sqrt{\mathrm{Var}(\hat{\mu}_\xi - \mu_\xi)}} + O_p(m^{-1/2}),$$

where $\mathbf{b}_\xi$ was defined in (2). Now, first and second moments are expressed in terms of the joint expectation. Due to independence of ξ and $\mathbf{e}$, the expectation with respect to the former can be treated as the average again, while the order of the remaining part is assessed in terms of the marginal case. Since $\mathrm{Var}(\hat{\mu}_i - \mu_i) = \mathbf{h}_i^t \mathbf{G}_i \mathbf{h}_i - \mathbf{b}_i^t\, \mathbf{V}_i\, \mathbf{b}_i + O(m^{-1})$, this amounts to

$$\mathrm{E}(T_\xi |\, \mathbf{v}) = \frac{1}{m} \sum_{i=1}^{m} \frac{(\mathbf{b}_i^t \mathbf{Z}_i - \mathbf{h}_i^t)\mathbf{v}_i}{\sqrt{\mathbf{h}_i^t \mathbf{G}_i \mathbf{h}_i - \mathbf{b}_i^t\, \mathbf{V}_i\, \mathbf{b}_i}} + O(m^{-1/2}) = O(m^{-1/2}),$$

by Lindeberg's central limit. The same approach gives for the variance

$$\begin{aligned}\mathrm{Var}\big(T_\xi |\, \mathbf{v}\,\big) &= \mathrm{E}(T_\xi^2 |\, \mathbf{v}) - \mathrm{E}(T_\xi |\, \mathbf{v})^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} \frac{(\mathbf{b}_i^t \mathbf{Z}_i - \mathbf{h}_i^t)\mathbf{v}_i \mathbf{v}_i^t (\mathbf{b}_i^t \mathbf{Z}_i - \mathbf{h}_i^t)^t + \mathbf{b}_i^t \mathbf{R}_i \mathbf{b}_i}{\mathbf{h}_i^t \mathbf{G}_i \mathbf{h}_i - \mathbf{b}_i^t\, \mathbf{V}_i\, \mathbf{b}_i} + O(m^{-1}) \\ &= 1 + O(m^{-1}),\end{aligned}$$

again by central limit and since $\mathbf{V}_i = \mathbf{Z}_i^t\, \mathbf{G}_i\, \mathbf{Z}_i + \mathbf{R}_i$. It follows that $T_\xi = Z + O(m^{-1/2})$ for $Z \sim \mathcal{N}(0,1)$. The claim follows.

For (A2), the claim follows for each subject as matrix inversion gives

$$\begin{aligned}(\mathbf{b}_i^t \mathbf{Z}_i - \mathbf{h}_i^t)\mathbf{v}_i &= \mathbf{h}_i^t \big\{\, \mathbf{G}_i\, \mathbf{Z}_i^t \,\big(\, \mathbf{Z}_i\, \mathbf{G}_i\, \mathbf{Z}_i^t + \mathbf{R}_i \,\big)^{-1} \mathbf{Z}_i\, \mathbf{G}_i - \mathbf{G}_i \,\big\}\, \mathbf{G}_i^{-1}\, \mathbf{v}_i \\ &= -\mathbf{h}_i^t \big(\, \mathbf{G}_i^{-1} + \mathbf{Z}_i^t\, \mathbf{R}_i^{-1}\, \mathbf{Z}_i \,\big)^{-1} \mathbf{G}_i^{-1}\, \mathbf{v}_i = O(n_i^{-1}).\end{aligned}$$

Thus, $\mathrm{E}(T_i |\, \mathbf{v}) = O(n_i^{-1})$. Similarly, for $\mathrm{Var}(T_i\, \mathbf{v})$, we find that the denominator is $O(n_i^{-1})$, as well as $\mathbf{b}_i^t \mathbf{R}_i \mathbf{b}_i = O(n_i^{-1})$ in the nominator. The remaining part of the nominator, by the same reasoning as above, is $O(n_i^{-2})$. It follows that $\mathrm{Var}(T_i) = O(n_i^{-1})$. This gives the claim. □

Auxiliary Results.

We only consider the marginal case, as the conditional case follows from analogous considerations for results in the appendix of [4]. These findings do not improve the error rate obtained for simultaneous comparisons however, as the error rate in Theorem 1 and Theorem 2 is induced by the variability of the estimators $\widehat{\boldsymbol{\Sigma}}_c$ and $\widehat{\boldsymbol{\Sigma}}$, respectively. Some preliminary results are required.

Lemma 1. *Let* $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ *be symmetric and nonstochastic for* $i \in \{1,2,3,4\}$, *and* $\mathbf{u} \sim \mathcal{N}_n(\mathbf{0}_n, \mathbf{V})$. *For* $\mathcal{R} = \{(1,2,3,4),(1,3,2,4),(1,4,2,3)\}$ *and* $\mathcal{Q} = \{(1,2,3,4),(2,1,3,4),(3,1,2,4),(4,1,2,3)\}$ *it holds*

(i) $\mathrm{E}\Big(\prod_{i=1}^{2} \mathbf{u}^t\mathbf{A}_i\mathbf{u}\Big) = 2\mathrm{tr}\big(\mathbf{A}_1\mathbf{V}\mathbf{A}_2\mathbf{V}\big) + \mathrm{tr}\big(\mathbf{A}_1\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_2\mathbf{V}\big)$,

(ii) $$\begin{aligned}\mathrm{E}\Big(\prod_{i=1}^{3} \mathbf{u}^t\mathbf{A}_i\mathbf{u}\Big) = {}& \prod_{i=1}^{3}\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\big) + 2\mathrm{tr}\big(\mathbf{A}_1\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_2\mathbf{V}\mathbf{A}_3\mathbf{V}\big)\\ &+ 2\mathrm{tr}\big(\mathbf{A}_2\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_1\mathbf{V}\mathbf{A}_3\mathbf{V}\big) + 4\mathrm{tr}\big(\mathbf{A}_2\mathbf{V}\mathbf{A}_1\mathbf{V}\mathbf{A}_3\mathbf{V}\big)\\ &+ 2\mathrm{tr}\big(\mathbf{A}_3\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_2\mathbf{V}\mathbf{A}_1\mathbf{V}\big) + 4\mathrm{tr}\big(\mathbf{A}_1\mathbf{V}\mathbf{A}_2\mathbf{V}\mathbf{A}_3\mathbf{V}\big),\end{aligned}$$

(iii) $$\begin{aligned}\mathrm{E}\Big(\prod_{i=1}^{4} \mathbf{u}^t\mathbf{A}_i\mathbf{u}\Big) = {}& \prod_{i=1}^{4}\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\big)\\ &+ \sum_{(i,j,k,l)\in\mathcal{R}} 2\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_j\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_k\mathbf{V}\mathbf{A}_l\mathbf{V}\big)\\ &+ \sum_{(k,l,i,j)\in\mathcal{R}} 2\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_j\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_k\mathbf{V}\mathbf{A}_l\mathbf{V}\big)\\ &+ \sum_{(i,j,k,l)\in\mathcal{Q}} 4\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\big)\big\{\mathrm{tr}\big(\mathbf{A}_j\mathbf{V}\mathbf{A}_k\mathbf{V}\mathbf{A}_l\mathbf{V}\big)\\ &\qquad\qquad + \mathrm{tr}\big(\mathbf{A}_k\mathbf{V}\mathbf{A}_j\mathbf{V}\mathbf{A}_l\mathbf{V}\big)\big\}\\ &+ \sum_{(i,j,k,l)\in\mathcal{R}} 4\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\mathbf{A}_j\mathbf{V}\big)\mathrm{tr}\big(\mathbf{A}_k\mathbf{V}\mathbf{A}_l\mathbf{V}\big)\\ &\qquad\qquad + 16\mathrm{tr}\big(\mathbf{A}_i\mathbf{V}\mathbf{A}_j\mathbf{V}\mathbf{A}_k\mathbf{V}\mathbf{A}_l\mathbf{V}\big).\end{aligned}$$

This results is an extension of Lemma 1 from [4], a result that was derived by direct application of Theorem 1 of [8].

Lemma 2. *Let model (1) hold with (6) and let* $\hat{\boldsymbol{\delta}}$ *be being a REML estimator. Under (B1) - (B4) it holds*

(i) $\mathbf{K}_1(\boldsymbol{\delta}) = \mathrm{E}\big\{\mathbf{K}_1(\hat{\boldsymbol{\delta}}) + \widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$,

(ii) $\mathbf{K}_2(\boldsymbol{\delta}) = \mathrm{E}\big\{\mathbf{K}_2(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$,

(iii) $\mathbf{K}_3 = \mathrm{E}\big\{\widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$.

Proof. (of Lemma 2) First, (A1) is considered, and part (ii) is proved. Adapt the notation of the proof of Lemma 2 for $\mathbf{X}$ and $\mathbf{V}$. Recall that $\{\mathbf{K}_2(\boldsymbol{\delta})\}_{ik} = \mathbf{d}_i^t(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{V})^{-1}\mathbf{d}_k =$

$O(m^{-1})$, and all derivatives preserve the order. Thus, for $i, k = 1, \dots, m$, a Taylor expansion around $\boldsymbol{\delta}$ and taking expectations yields

$$\begin{aligned}\mathrm{E}\big[\{\mathbf{K}_2(\hat{\boldsymbol{\delta}})\}_{ik}\big] &= \{\mathbf{K}_2(\boldsymbol{\delta})\}_{ik} + \frac{1}{2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t \frac{\partial^2 \{\mathbf{K}_2(\boldsymbol{\delta})\}_{ik}}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O(m^{-2}) \\ &= \{\mathbf{K}_2(\boldsymbol{\delta})\}_{ik} + O(m^{-2}),\end{aligned}$$

noting that the REML estimates fulfill $\boldsymbol{\delta} - \hat{\boldsymbol{\delta}} = \{O_p(m^{-1/2})\}_r$ and are unbiased, hence having that the second term of the expansion is zero under expectation.

(iii) For the next statement, we start showing that

$$\widehat{\mathbf{K}}_3(\boldsymbol{\delta}) + \{O(m^{-2})\}_{m\times m} = \mathbf{K}_3. \tag{15}$$

The proof is similar to [2], but in contrast to these authors the Taylor expansion is performed including the second order term. Again using that the REML estimates fulfill $\boldsymbol{\delta} - \hat{\boldsymbol{\delta}} = \{O_p(m^{-1/2})\}_r$,

$$\hat{\mu}_i - \tilde{\mu}_i = (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t \frac{\partial \tilde{\mu}_i}{\partial\,\boldsymbol{\delta}} + \frac{1}{2}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^t \frac{\partial^2 \tilde{\mu}_i}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}) + O_p\big(m^{-3/2}\big). \tag{16}$$

Since further $\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial\,\boldsymbol{\delta}^t} = \{O_p(m^{-1/2})\}_{p\times r}$ as shown in [1],

$$\begin{aligned}\frac{\partial \tilde{\mu}_i}{\partial\,\boldsymbol{\delta}} &= \frac{\partial \tilde{\mu}_i|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}}{\partial\,\boldsymbol{\delta}} + \frac{\partial \tilde{\boldsymbol{\beta}}^t}{\partial\,\boldsymbol{\delta}} \frac{\partial \tilde{\mu}_i|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}}{\partial\,\boldsymbol{\beta}} = \mathbf{f}_{1,i} + \mathbf{f}_{2,i}, \\ \frac{\partial^2 \tilde{\mu}_i}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t} &= \frac{\partial^2 \tilde{\mu}_i|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}}}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t} + \{O_p(m^{-1/2})\}_{r\times r} = 2\mathbf{F}_3 + \{O_p(m^{-1/2})\}_{r\times r}.\end{aligned}$$

With the notation from Lemma 2, the explicit forms read as

$$\begin{aligned}\mathbf{f}_{1,i} &= \frac{\partial \mathbf{b}_i^t}{\partial\,\boldsymbol{\delta}}(\mathbf{Z}_i\,\mathbf{v}_i + \mathbf{e}_i) = \{O_p(1)\}_r, \\ \mathbf{f}_{2,i} &= -\left\{\mathbf{d}_i^t(\mathbf{X}^t\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}\,\mathbf{V}^{-1}\,\frac{\partial\,\mathbf{V}}{\partial \delta_d}\mathbf{P}(\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta})\right\}_d = \{O_p(m^{-1/2})\}_r, \\ 2\mathbf{F}_{3,i} &= \left\{\frac{\partial^2 \mathbf{b}_i^t}{\partial \delta_d \partial \delta_e}(\mathbf{Z}_i\,\mathbf{v}_i + \mathbf{e}_i)\right\}_{d,e} = \{O_p(1)\}_{r\times r}.\end{aligned}$$

Using Lemma 2, it follows that (16) can be rewritten as

$$\hat{\mu}_i - \tilde{\mu}_i = \underbrace{\mathbf{g}_1^t\mathbf{f}_{1,i}}_{O_p(m^{-1/2})} + \underbrace{\mathbf{g}_1^t\mathbf{f}_{2,i}}_{O_p(m^{-1})} + \underbrace{\mathbf{g}_2^t\mathbf{f}_{1,i}}_{O_p(m^{-1})} - \underbrace{\mathbf{g}_3^t\mathbf{f}_{1,i}}_{O_p(m^{-1})} + \underbrace{\mathbf{g}_1^t\mathbf{F}_{3,i}\mathbf{g}_1}_{O_p(m^{-1})} + O_p(m^{-3/2}).$$

Now, (15) is shown by splitting $\mathbf{K}_3$ into nine terms. Five terms are considered separately, the other four yield the same result by symmetry. In particular,

$$\begin{aligned}(\mathbf{K}_3)_{ik} = \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{1,k}) &+ \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_2^t\mathbf{f}_{1,k}) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_2^t\mathbf{f}_{1,i})\\ &- \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_3^t\mathbf{f}_{1,k}) - \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_3^t\mathbf{f}_{1,i})\\ &+ \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{2,k}) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_1^t\mathbf{f}_{2,i})\\ &+ \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_1^t\mathbf{F}_{3,i}\mathbf{g}_1) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{F}_{3,k}\mathbf{g}_1) + O(m^{-2}).\end{aligned}$$

In the following it will be shown that $\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{1,k}) = \{\widehat{\mathbf{K}}_3(\boldsymbol{\delta})\}_{ik} + O(m^{-2})$ and all other terms are of order $O(m^{-2})$, which is sufficient to show (15). Repeating the calculations of [2], the leading term gives

$$\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{1,k}) = \mathbb{1}_{i=k}\,\mathrm{tr}\left(\frac{\partial\mathbf{b}_i^t}{\partial\,\boldsymbol{\delta}}\,\mathbf{V}_i\,\frac{\partial\mathbf{b}_i}{\partial\,\boldsymbol{\delta}^t}\overline{\mathbf{V}}\right) + O(m^{-2}),$$

using Lemma 1 (i) and exploiting that $\mathbf{V}$ being of block-diagonal form and $\mathbf{P} = \mathrm{diag}[\{O(1)\}_{n_i\times n_i}]_{i=1,\dots,m} + \{O(m^{-1})\}_{n\times n}$. For the next term consider the matrix $\mathbf{M}_{ik}(e,d,g,f)$ of dimension $(n\times n)$ with only non-zero entries being $(\overline{\mathbf{V}})_{ef}\frac{\partial\mathbf{b}_k}{\partial\,\boldsymbol{\delta}^t}(\overline{\mathbf{V}})_d(\overline{\mathbf{V}})_g^t\frac{\partial\mathbf{b}_k^t}{\partial\,\boldsymbol{\delta}} = \{O(m^{-3})\}_{n_i\times n_k}$ at the $(n_i\times n_k)$-submatrix, corresponding to the respective subjects. Further, by construction, $\mathbf{P}\,\mathbf{y} = \mathbf{P}(\mathbf{Z}\,\mathbf{v}+\mathbf{e})$. Tedious calculations yield with Lemma 1 (iii) that

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_2^t\mathbf{f}_{1,k}) &= \sum_{e,g,d,f=1}^{r}\mathrm{E}\Big\{\mathbf{s}_g\mathbf{s}_d(\boldsymbol{\Lambda})_{ef}(\mathbf{Z}\,\mathbf{v}+\mathbf{e})^t\mathbf{M}_{ik}(e,g,d,f)(\mathbf{Z}\,\mathbf{v}+\mathbf{e})\Big\}\\ &= O(m^{-2}).\end{aligned}$$

Now define the matrix $\mathbf{O}_{ik}(e,d,g,f)$ of dimension $(n\times n)$ with only non-zero entries being $(\overline{\mathbf{V}})_{ef}\frac{\partial\mathbf{b}_k}{\partial\,\boldsymbol{\delta}^t}(\overline{\mathbf{V}})_d(\overline{\mathbf{V}}\frac{\partial\overline{\mathbf{V}}^{-1}}{\partial\delta_e}\overline{\mathbf{V}})_g^t\frac{\partial\mathbf{b}_k^t}{\partial\,\boldsymbol{\delta}} = \{O(m^{-3})\}_{n_i\times n_k}$ at the $(n_i\times n_k)$-submatrix, corresponding to the respective subjects. Further, by construction, $\mathbf{P}\,\mathbf{y} = \mathbf{P}(\mathbf{Z}\,\mathbf{v}+\mathbf{e})$. Now, Lemma 1 yields

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_3^t\mathbf{f}_{1,k}) &= \frac{1}{2}\sum_{e,d,f,g=1}^{r}\mathrm{E}\Big\{\mathbf{s}_g\mathbf{s}_d\mathbf{s}_f(\mathbf{Z}\,\mathbf{v}+\mathbf{e})^t\mathbf{O}_{ik}(e,d,g,f)(\mathbf{Z}\,\mathbf{v}+\mathbf{e})\Big\}\\ &= O(m^{-2}).\end{aligned}$$

Similarly, by Lemma 1 (ii), and matrix $\mathbf{Q}_{ik}(e,d)$ with only non-zero entries on the n_i-columns corresponding to the respective i-th subject with entries of order $\{O(m^{-3})\}_{n_i\times n}$

it holds that

$$\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{2,k}) = \sum_{e,d=1}^{r} \mathrm{E}\Big\{\mathbf{s}_e\mathbf{s}_d(\mathbf{Z}\,\mathbf{v}+\mathbf{e})^t\mathbf{Q}_{ik}(e,d)(\mathbf{Z}\,\mathbf{v}+\mathbf{e})\Big\} = O(m^{-2}).$$

It remains to treat the last term. As before, for a matrix $\mathbf{U}_{i,k}(e,d,f)$ with zero entries except on the $(n_i \times n_k)$-submatrix $\{O(m^{-3})\}_{n_i\times n_k}$ corresponding to the respective subjects it holds that

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_1^t\mathbf{F}_{3,i}\mathbf{g}_1) &= \frac{1}{2}\sum_{e,d,f=1}^{r} \mathrm{E}\Big\{\mathbf{s}_e\mathbf{s}_d\mathbf{s}_f(\mathbf{Z}\,\mathbf{v}+\mathbf{e})^t\mathbf{U}_{ik}(e,d,f)(\mathbf{Z}\,\mathbf{v}+\mathbf{e})\Big\}\\ &= O(m^{-2}).\end{aligned}$$

The other terms are $O(m^{-2})$ by symmetry when replacing i and k. Hence $\mathbf{K}_3 = \widehat{\mathbf{K}}_3(\boldsymbol{\delta}) + \{O(m^{-2})\}_{m\times m}$, which was the claim in (15). The remaining proof is now similar to (ii). Note that $\{\widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\}_{ik} = O(m^{-1})$. As above, taking derivatives preserves the order and a Taylor expansion and taking expectations yields

$$\begin{aligned}\mathrm{E}\big[\{\widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\}_{ik}\big] &= \{\widehat{\mathbf{K}}_3(\boldsymbol{\delta})\}_{ik} + \frac{1}{2}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^t\frac{\partial^2\{\widehat{\mathbf{K}}_3(\boldsymbol{\delta})\}_{ik}}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}) + O(m^{-2})\\ &= \{\widehat{\mathbf{K}}_3(\boldsymbol{\delta})\}_{ik} + O(m^{-2}) = (\mathbf{K}_3)_{ik} + O(m^{-2}),\end{aligned}$$

where the last equation follows by (15). This gives (iii).

(i) Finally, as before, a Taylor expansion of $\{\mathbf{K}_1(\hat{\boldsymbol{\delta}})\}_{ii}$ around $\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}$ and taking expectation yields

$$\begin{aligned}\mathrm{E}\big[\{\mathbf{K}_1(\hat{\boldsymbol{\delta}})\}_{ii}\big] &= \{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii} + \mathrm{E}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^t\frac{\partial\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}}{\partial\,\boldsymbol{\delta}} + \frac{1}{2}\mathrm{tr}\left[\frac{\partial^2\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}\overline{\mathbf{V}}\right]\\ &\quad + \frac{1}{6}\mathrm{E}\left[\sum_{e=1}^{r}(\hat{\delta}_e-\delta_e)(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^t\frac{\partial^3\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii}}{\partial\delta_e\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\right] + O(m^{-2}).\end{aligned}$$

By Lemma 1 (ii), the fourth term is of order $O(m^{-2})$ as

$$\begin{aligned}\sum_{e,d,f=1}^{r}\mathrm{E}\big(\mathbf{s}_e\mathbf{s}_d\mathbf{s}_f\big) &= \sum_{e,d,f=1}^{r}\mathrm{tr}\left(\frac{\partial\,\mathbf{V}}{\partial\delta_e}\mathbf{P}\frac{\partial\,\mathbf{V}}{\partial\delta_f}\mathbf{P}\frac{\partial\,\mathbf{V}}{\partial\delta_d}\mathbf{P} + \frac{\partial\,\mathbf{V}}{\partial\delta_f}\mathbf{P}\frac{\partial\,\mathbf{V}}{\partial\delta_e}\mathbf{P}\frac{\partial\,\mathbf{V}}{\partial\delta_d}\mathbf{P}\right)\\ &= O(m),\end{aligned}$$

exploiting again the block diagonal structure of $\mathbf{V}$ and detailed structure of $\mathbf{P}$, with the same reasoning as for the proof of (15). Further, some calculations yield $\frac{\partial^2}{\partial\,\boldsymbol{\delta}\,\partial\,\boldsymbol{\delta}^t}\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii} =$

$-2\frac{\partial \mathbf{b}_i^t}{\partial \boldsymbol{\delta}} \mathbf{V}_i \frac{\partial \mathbf{b}_i}{\partial \boldsymbol{\delta}^t}$ [2, p. 624-625]. Together with the proof in (iii), which implies that $\mathrm{E}\{\widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\} = \widehat{\mathbf{K}}_3(\boldsymbol{\delta}) + \{O(m^{-2})\}_m$, it follows that $\mathrm{E}\big\{\mathbf{K}_1(\hat{\boldsymbol{\delta}}) + \widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\big\} + \mathrm{diag}\big[\{O(m^{-2})\}_m\big] = \mathbf{K}_1(\boldsymbol{\delta})$. Altogether, this gives (i) and proves Lemma 2 for (A1).

For (A2), the leading term itself is of lower order and it holds

$$\begin{aligned}\{\mathbf{K}_1(\boldsymbol{\delta})\}_{ii} &= \mathbf{h}_i^t\big(\mathbf{G}_i - \mathbf{G}_i\,\mathbf{Z}_i^t\,\mathbf{V}_i^{-1}\,\mathbf{Z}_i\,\mathbf{G}_i\big)\mathbf{h}_i \\ &= \mathbf{h}_i^t\big(\mathbf{G}_i^{-1} - \mathbf{Z}_i^t\,\mathbf{R}_i\,\mathbf{Z}_i\big)^{-1}\mathbf{h}_i = O(n_{\mathcal{I}}^{-1}),\end{aligned}$$

as $\mathbf{V}_i = \mathbf{Z}_i\,\mathbf{G}_i\,\mathbf{Z}_i^t + \mathbf{R}_i$. Further, $\{\mathbf{K}_2(\boldsymbol{\delta})\}_{ii}$, $\{\mathbf{K}_3(\boldsymbol{\delta})\}_{ii}$ as well as subject crossterms are of lower order. Thus, statements (i)-(iii) for (A2) follow analogously from the reasoning in (ii) and (iii) above. This proves Lemma 2. □

First, [6] derived a second-order unbiased estimator for the MSE, but they considered estimation of the variance components by Hendersons Method III [7]. They can be written as $\hat{\delta}_e = \mathbf{y}^t\,\mathbf{C}_e\,\mathbf{y}$ where it further holds that $\mathbf{C}_e = \mathrm{diag}[\{O(m^{-1})\}_{n_i\times n_i}]_{i=1,\ldots,m} + \{O(m^{-2})\}_{n\times n}$. An explicit formulation for the nested error regression model (5) is e.g. given in [3]. The analogous result from Lemma 2 holds true.

Lemma 3. *Let model (1) hold with (6) and let $\hat{\boldsymbol{\delta}}$ be being estimate obtained via Hendersons Method III. Under (B1) - (B4) it holds*

(i) $\mathbf{K}_1(\boldsymbol{\delta}) = \mathrm{E}\big\{\mathbf{K}_1(\hat{\boldsymbol{\delta}}) + \widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$,

(ii) $\mathbf{K}_2(\boldsymbol{\delta}) = \mathrm{E}\big\{\mathbf{K}_2(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$,

(iii) $\mathbf{K}_3 = \mathrm{E}\big\{\widehat{\mathbf{K}}_3(\hat{\boldsymbol{\delta}})\big\} + \{O(m^{-2})\}_{m\times m}$.

Proof. (of Lemma 3). We treat (A1) only as (A2) follows by analogous considerations as in Lemma 2. In fact the proof is very similar to the proof of Lemma 2, but we have to account for the different nature of the estimator $\hat{\boldsymbol{\delta}}$. Replicating the calculations of Lemma 2, and adapting the notation of $\mathbf{g}_1 = \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}$ for simplicity, the terms to consider for $\mathbf{K}_3$ are

$$\begin{aligned}(\mathbf{K}_3)_{ik} = \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{1,k}) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{f}_{2,k}) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_1^t\mathbf{F}_{3,i}\mathbf{g}_1)& \\ + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,k}\mathbf{g}_1^t\mathbf{f}_{2,i}) + \mathrm{E}(\mathbf{g}_1^t\mathbf{f}_{1,i}\mathbf{g}_1^t\mathbf{F}_{3,k}\mathbf{g}_1) + O(m^{-2})&\ .\end{aligned}$$

For the first term, verify that $\mathbf{C}_f\,\mathbf{V} = \mathrm{diag}[\{O(m^{-1})\}_{n_i\times n_i}] + \{O(m^{-2})\}_{n\times n}$, $\mathbf{C}_f\,\mathbf{V}\,\mathbf{C}_e\,\mathbf{V} = \mathrm{diag}[\{O(m^{-2})\}_{n_i\times n_i}] + \{O(m^{-3})\}_{n\times n}$ and further $\mathbf{V}\,\mathbf{C}_e\,\mathbf{X}\,\boldsymbol{\beta} = \{O(m^{-1})\}_n$. Further adapt the notation $\mathbf{u} = \mathbf{Z}\,\mathbf{v} + \mathbf{e}$ and respectively $\mathbf{u}_i = \mathbf{Z}_i\,\mathbf{v}_i + \mathbf{e}_i$ for $i = 1, \ldots, m$. Now, let

$\boldsymbol{\Omega} = \boldsymbol{\Omega}_{i,j}(e,f)$ with only entries $\frac{\partial \mathbf{b}_i}{\partial \delta_e} \frac{\partial \mathbf{b}_k^t}{\partial \delta_f}$ on the $(n_i \times n_k)$-submatrix, corresponding to the respective subjects. Then it follows that

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1^t \mathbf{f}_{1,i} \mathbf{g}_1^t \mathbf{f}_{1,k}) &= \sum_{e=1}^{r} \sum_{f=1}^{r} \mathrm{E}\big\{(\hat{\delta}_e - \delta_e)(\hat{\delta}_f - \delta_f)\mathbf{u}^t \boldsymbol{\Omega} \mathbf{u}\big\} \\ &= \mathbb{1}_{i=k} \mathrm{tr}\Bigg[\mathrm{Cov}\big\{(\mathbf{y}^t \mathbf{C}_e \mathbf{y})_e\big\} \mathrm{Cov}\Bigg\{\bigg(\frac{\partial \mathbf{b}_i^t}{\partial \delta_e} \mathbf{u}_i\bigg)_e\Bigg\}\Bigg] + O(m^{-2}).\end{aligned}$$

As before, the remaining parts are of lower order. Let $\boldsymbol{\Xi} = \boldsymbol{\Xi}_{i,k}(e,f)$ with only non-zero entries being the n_i rows $\frac{\partial \mathbf{b}_i}{\partial \delta_e} \mathbf{d}_i^t (\mathbf{X}^t \, \mathbf{V}^{-1} \, \mathbf{X})^{-1} \, \mathbf{X} \, \mathbf{V}^{-1} \, \frac{\partial \, \mathbf{V}}{\partial \delta_f} \mathbf{P} = \{O(m^{-1})\}_{n_i \times n}$ corresponding to the i-th subject. Then, as above it holds

$$\mathrm{E}(\mathbf{g}_1^t \mathbf{f}_{1,i} \mathbf{g}_1^t \mathbf{f}_{2,k}) = \sum_{e=1}^{r} \sum_{f=1}^{r} \mathrm{E}\{(\hat{\delta}_e - \delta_e)(\hat{\delta}_f - \delta_f)\mathbf{u}^t \boldsymbol{\Xi} \mathbf{u}\} = O(m^{-2}) = O(m^{-2}).$$

Eventually, let $\boldsymbol{\Lambda}_{i,k}(e,f,g) = \boldsymbol{\Lambda}$ with only non-zero entries $\frac{\partial \mathbf{b}_i}{\partial \delta_e} \frac{\partial^2 \mathbf{b}_k^t}{\partial \delta_f \partial \delta_g}$ on the $(n_i \times n_k)$-submatrix, corresponding to the respective subjects. Moreover noting that it holds that $\mathrm{tr}(\mathbf{C}_f \, \mathbf{V} \, \mathbf{C}_e \, \mathbf{V} \, \mathbf{C}_g \, \mathbf{V}) = O(m^{-2})$, $\mathrm{tr}(\mathbf{C}_f \, \mathbf{V} \, \mathbf{C}_e \, \mathbf{V} \, \boldsymbol{\Lambda} \, \mathbf{V}) = O(m^{-2})$, $\mathrm{tr}(\mathbf{C}_f \, \mathbf{V} \, \boldsymbol{\Lambda} \, \mathbf{V}) = O(m^{-1})$ and $\mathrm{tr}(\mathbf{C}_f \, \mathbf{V} \, \mathbf{C}_e \, \mathbf{V} \, \mathbf{C}_g \, \mathbf{V} \, \boldsymbol{\Lambda} \, \mathbf{V}) = O(m^{-3})$. Now, similarly to the considerations above,

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1^t \mathbf{f}_{1,k} \mathbf{g}_1^t \mathbf{F}_{3,i} \mathbf{g}_1) &= \frac{1}{2} \sum_{e=1}^{r} \sum_{f=1}^{r} \sum_{g=1}^{r} \mathrm{E}\{(\hat{\delta}_e - \delta_e)(\hat{\delta}_f - \delta_f)(\hat{\delta}_g - \delta_g)\mathbf{u}^t \boldsymbol{\Lambda} \mathbf{u}\} \\ &= O(m^{-2}).\end{aligned}$$

Hence, $\widehat{\mathbf{K}}_3(\boldsymbol{\delta}) = \mathbf{K}_3 + \{O(m^{-2})\}$ as $\mathrm{Cov}\{(\mathbf{y}^t \mathbf{C}_e \mathbf{y})_{e=1,\ldots,r}\} = \mathrm{Cov}(\hat{\boldsymbol{\delta}}) = \overline{\mathbf{V}}$. Since $\hat{\boldsymbol{\delta}}$ is unbiased and still $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}} + O(m^{-1/2})$, the remaining part of the proof follows analogously to that one of Lemma 2. □

References

[1] D. R. Cox and N. Reid. Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society B*, 49(1):1–39, 1987.

[2] G. S. Datta and P. Lahiri. A Unified Measure of Uncertainty of Estimated Best Linear Predictors in Small Area Estimation Problems. *Statistica Sinica*, 10:613–627, 2000.

[3] P. Hall and T. Maiti. Nonparametric Estimation of Mean-Squared Prediction Error in Nested-Error Regression Models. *Annals of Statistics*, 34(4):1733–1750, 2006.

[4] P. Kramlinger, T. Krivobokova, and S. Sperlich. Marginal and Conditional Multiple Inference in Linear Mixed Models. *Submitted*, 2020.

[5] D. Nychka. Bayesian Confidence Intervals for Smoothing Splines. *Journal of the American Statistical Assiociation*, 83(404):1134–1143, 1988.

[6] N. G. N. Prasad and J. N. K. Rao. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

[7] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. Wiley, Hoboken, NJ, 1992.

[8] V. K. Srivastava and R. Tiwari. Evaluation of Expectations of Products of Stochastic Matrices. *Scandinavian Journal of Statistics*, 3(3):135–138, 1976.

[9] G. Wahba. Bayesian “Confidence Intervals” for the Cross-validated Smoothing Spline. *Journal of the Royal Statistical Society B*, 45(1):133–150, 1983.

Addendum B

Uniformly Valid Inference Based on the Lasso in Linear Mixed Models

Uniformly Valid Inference Based on the Lasso in Linear Mixed Models

Peter Kramlinger[1] Tatyana Krivobokova[2] Ulrike Schneider[3]

Abstract

In a Gaussian linear mixed model we construct confidence sets for fixed effects that are estimated via a Lasso-type penalization. It is shown that those are uniformly valid over the space of coefficient and covariance parameters. They adequately quantify the joint uncertainty of model selection and estimation. Their superiority to naïve LS confidence sets is demonstrated in a simulation example.

MSC 2010 subject classification: Primary 62F25; secondary 62J10; 62J07.
Keywords and phrases: Sparsity, confidence sets, REML, variance components.

1 Introduction

Linear mixed models (LMMs) are regression models that allow for more elaborate dependency structures within the observed data. They are widely applied in many empirical sciences ranging from genetics [Henderson, 1950] to survey statistics [Pfefferman, 2013] and more. Comprehensive reviews are given by [Demidenko, 2004, Pinheiro and Bates, 2000]. The classical LMM can be written as

$$\begin{gathered} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_0 + \mathbf{Z}_i\mathbf{v}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, m; \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}\{\mathbf{0}_{n_i}, \boldsymbol{\Omega}_i(\boldsymbol{\theta}_0)\}, \quad \mathbf{v}_i \sim \mathcal{N}_q\{\mathbf{0}_q, \boldsymbol{\Psi}(\boldsymbol{\theta}_0)\}, \end{gathered} \tag{1}$$

with observations $\mathbf{y}_i \in \mathbb{R}^n$, known covariates $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ and $\mathbf{Z}_i \in \mathbb{R}^{n_i \times q}$ and $\mathbf{v}_i \in \mathbb{R}^q$ and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{n_i}$ independent and each independently distributed for all $i = 1, \ldots, m$. The term $\mathbf{X}_i\boldsymbol{\beta}_0$ is referred to as ‘fixed effects’, and $\mathbf{Z}_i\mathbf{v}_i$ as ‘random effects’. The coefficient $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and covariance parameters $\boldsymbol{\theta}_0 \in \mathbb{R}^q$ are unknown and to be estimated.

Two-stage methods for fitting regression models involve model selection first, e.g. based on information criteria, and thereafter estimation of the parameters in the obtained model.

[1] peter.kramlinger@uni-goettingen.de, Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

[2] tkrivob@gwdg.de, Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

[3] ulrike.schneider@tuwien.ac.at, Institute of Statistics and Mathematical Methods in Economics, Technische Universität Wien, Wiedner Hauptstr. 8, 1040 Vienna, Austria

However, classical inferential theory does not consider the selection procedure as stochastic. Accounting for the additional uncertainty of model selection is a difficult task, that received attention as post-selection inference in the recent past [Berk et al., 2013].

A single stage approach is given by the least absolute shrinkage and selection operator (Lasso) [Tibshirani, 1996], which performs selection and estimation jointly.

In the context of mixed models the Lasso can be applied on both the fixed and random effects. For example, Ibrahim et al. [2011] and Bondell et al. [2010] penalize both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, whereas Peng and Lu [2012] penalize the random effects directly. A review on these methods is given by Müller et al. [2013]. Even when only the fixed effects are penalized, the joint estimation of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ in LMMs raises difficulties from computational aspects [Schelldorfer et al., 2011, Juming and Shang, 2019].

Its usefulness and wide application sparked interest in how to construct confidence intervals based on the Lasso. The general difficulty is that its asymptotic distribution is crucially shaped by the unknown coefficient parameters $\boldsymbol{\beta}_0$ [Pötscher and Leeb, 2009]. Therefore, honest confidence sets in the sense of Li [1989] are necessarily obtaining the nominal coverage over the whole parameter space. For a low dimensional framework ('$p < n$'), Ewald and Schneider [2018] propose limiting versions of the objective function in order to obtain uniformly valid confidence sets.

This contribution builds on these obtained results and extends them to LMMs. It is exploited that the classical estimation of $\boldsymbol{\theta}_0$ via restricted maximum likelihood (REML) in LMMs is separated from $\boldsymbol{\beta}_0$, see Section 3. Thus, it is shown that the results on Lasso-type penalized estimators for the fixed effects carry over to LMMs. Eventually, the main result establishes confidence sets that are uniformly valid over the space of coefficient and variance parameters together.

The rest of the article is structured as follows. First, we specify the settings and regularity conditions and state the estimation procedure that avoids the need of a non-convex optimization problem for both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ in Section 2. Next, in Section 3, the estimation of covariance parameters $\boldsymbol{\theta}_0$ is discussed and their uniform consistency established. In Section 4 the main results are presented. Their usefulness and limitations, and in particular their superiority to a naïve approach based on least squares is demonstrated in a simulation study in Section 5.

2 Setting and Regularity Conditions

We rewrite model (1) to a linear model with n dependent observations and model equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}_n\{\mathbf{0}_n, \mathbf{V}(\boldsymbol{\theta}_0)\}. \tag{2}$$

The covariance matrix $\mathbf{V}(\cdot) \in \mathbb{R}^{n\times n}$ is known and models the dependency amongst the observations. The vectors $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ as well as $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^r_{>0}$ are unknown and remain to be estimated. The objective function for Lasso-penalized fixed effects and given tuning parameters $\lambda_1, \dots, \lambda_p$ is given by

$$Q(\boldsymbol{\beta}, \boldsymbol{\theta}) = \ln|\mathbf{V}(\boldsymbol{\theta})| + \left\|\mathbf{V}(\boldsymbol{\theta})^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\|^2 + 2\sum_{j=1}^{p} \lambda_j |\beta_j|. \tag{3}$$

The joint minimization over both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is a non-convex optimization problem [Schelldorfer et al., 2011]. In order to establish uniform coverage for both parameter spaces by adapting the approach of Ewald and Schneider [2018], we consider the Lasso estimator for $\boldsymbol{\beta}_0$ for a given estimator $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ to be

$$\widehat{\boldsymbol{\beta}}_L = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}). \tag{4}$$

We aim to find a set $M \subset \mathbb{R}^p$ such that $\inf_{\boldsymbol{\beta}_0, \boldsymbol{\theta}_0} \mathrm{P}\{\sqrt{n}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) \in M\} = 1 - \alpha + O(n^{-1/2})$ for some nominal level $\alpha \in (0,1)$. Here and throughout the rest of this article, the order of the remainder term is understood uniformly over $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and $\boldsymbol{\theta}_0 \in \Theta$.
The reasoning closely follows Ewald and Schneider [2018], by which the notation is borrowed and slightly adapted. Key idea is that instead of treating $\widehat{\boldsymbol{\beta}}_L$ with its intractable distribution directly, consider

$$\widehat{\mathbf{u}} = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^p} \ \mathbf{u}^t \widehat{\mathbf{C}} \mathbf{u} - 2\mathbf{u}^t \widehat{\mathbf{w}} + 2\mathbf{u}^t \boldsymbol{\Lambda} \mathbf{d} = \widehat{\mathbf{C}}^{-1}\Big(\widehat{\mathbf{w}} - \boldsymbol{\Lambda}\mathbf{d}\Big). \tag{5}$$

where $\widehat{\mathbf{C}} = n^{-1}\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}\mathbf{X}$, $\widehat{\mathbf{w}} = n^{-1/2}\mathbf{X}^t\mathbf{V}(\widehat{\boldsymbol{\theta}})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$, $\boldsymbol{\Lambda} = n^{-1/2}\mathrm{diag}(\lambda_1, \dots, \lambda_p)$ and $\mathbf{d} \in \{-1, 1\}^p$. The hindmost plays the role of adjusting sign of the coefficients, such that, for fixed $\widehat{\mathbf{w}}$, $\widehat{\mathbf{u}} \to \sqrt{n}\big(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\big)$ for $d_i\beta_{0,i} \to \infty$ for all $i = 1, \dots, p$. Further, denote $\mathbf{u}$, $\mathbf{C}$ and $\mathbf{w}$ analogously with $\widehat{\boldsymbol{\theta}}$ replaced by $\boldsymbol{\theta}_0$. As the distribution of $\widehat{\mathbf{u}}$ is not analytically available, the proofs exploit that $\mathbf{u} \sim \mathcal{N}_p(-\mathbf{C}^{-1}\boldsymbol{\Lambda}\mathbf{d}, \mathbf{C}^{-1})$.
We only consider confidence sets of ellipsoidal shape, i.e. $E(\widehat{\mathbf{C}}, k) = \{\mathbf{z} \in \mathbb{R}^p \mid \mathbf{z}^t\widehat{\mathbf{C}}\mathbf{z} \le k\}$, as those are required in order to attain the infimum over both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$. For details on non-ellipsoidal confidence sets see Ewald and Schneider [2018]. Regularity conditions below are further imposed for all $n \in \mathbb{N}$.

(A) $\Theta \subseteq \mathbb{R}^r_{>0}$.

(B) $\mathrm{rank}(\mathbf{X}) = p < n$.

(C) $\mathbf{V}(\boldsymbol{\theta}_0) = \sum_{k=1}^r \theta_{0,k}\mathbf{H}_k$ positive definite with positive semi-definite, symmetric and linear independent $\mathbf{H}_k \in \mathbb{R}^{n \times n}$, $k = 1 \dots, r$.

(D) For $i = 1 \dots, n$, $j = 1 \dots, p$ and $k = 1 \dots, r$, there exist non-zero $c_1, c_2, c_3 \in \mathbb{R}$ constant with respect to n, such that $\sum_{s=1}^n x_{sj} \asymp nc_1$, $\sum_{s=1}^p x_{is} \asymp c_2$, $\sum_{s=1}^n h_{is}^k \asymp c_3$, where x_{ij} and h_{ij}^k are the (i,j)-th entry of $\mathbf{X}$ and $\mathbf{H}_k$, respectively.

Condition (A) with $\theta_{0,i} > 0$, $i = 1, \dots, r$ fulfills the conditions of positive covariance parameters and non-degenerativity as introduced by Jiang [1996]. Condition (B) allows that $\mathbf{C} = n^{-1}\mathbf{X}^t\mathbf{V}(\boldsymbol{\theta}_0)^{-1}\mathbf{X}$ can be inverted, which is essential in the reasoning below. Further, (C) ensures that $\mathbf{V}(\boldsymbol{\theta}_0)$ can be inverted and that random variables depending on $\widehat{\boldsymbol{\theta}}$ can be approximated by ones depending on $\boldsymbol{\theta}_0$. It corresponds to the condition of identifiablility of covariance parameters condition from Jiang [1996]. Eventually, Condition (D) is required in order assure consistent estimation of both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, and is a similar to the condition of Prasad and Rao [1990]. The condition ensures that all entries in $\mathbf{X}$ are bounded and do not vanish for $n \to \infty$. The condition on $\mathbf{H}_k$ allows for block-diagonal matrices with constant entries and with blocks of size $n_i = O(1)$, corresponding to the small area setting [Rao and Molina, 2015]. Also, for blocks of size $n_i \neq O(1)$, a constant number of entries in each row has to consist of constant entries, with other entries being $O(n^{-1})$. Also allowed are Toeplitz matrices with a constant number of non-zero off-diagonals.

Example. The linear mixed model with (1) with $n_i = O(1)$, $i = 1, \dots, m$ is of type (2). For convenience, drop the class index to i label the respective quantities over all observations. Then, $\boldsymbol{\epsilon} = \mathbf{Z}\mathbf{v} + \boldsymbol{\varepsilon}$ with block diagonal covariance matrix $\mathbf{V}(\boldsymbol{\theta}_0) = \mathbf{Z}\boldsymbol{\Psi}(\boldsymbol{\theta}_0)\mathbf{Z}^t + \boldsymbol{\Omega}(\boldsymbol{\theta}_0)$.

The usual representation of LMMs is given in the form of above example. However, the reasoning in this contribution allows for different dependency structures in $\mathbf{V}(\boldsymbol{\theta}_0)$ from (2) as long as $\boldsymbol{\theta}_0$ can be estimated with restricted maximum likelihood.

3 Estimation of the Covariance Parameters

The key idea in deriving uniformly valid confidence sets in LMMs is that the estimation of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ is separated. By using restricted maximum likelihood (REML), the covariance parameters are estimated while accounting for the loss of degrees of freedom when estimating $\boldsymbol{\beta}_0$ [Searle et al., 1992, Demidenko, 2004]. Let $\mathbf{A} \in \mathbb{R}^{n\times(n-p)}$ such that $\mathbf{A}^t\mathbf{X} = \mathbf{0}_{(n-p)\times p}$ and consider only transformed data $\mathbf{A}^t\mathbf{y}$. Then, the REML estimate $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ is the minimizer of

$$l_R(\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{V}(\boldsymbol{\theta})| - \frac{1}{2}\ln|\mathbf{X}^t\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}^t\mathbf{P}(\boldsymbol{\theta})\mathbf{y}, \tag{6}$$

where $\mathbf{P}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta})^{-1} - \mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\{\mathbf{X}^t\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^t\mathbf{V}(\boldsymbol{\theta})^{-1}$. By construction, $\widehat{\boldsymbol{\theta}}$ does not depend on $\boldsymbol{\beta}_0$ as $\mathbf{P}(\boldsymbol{\theta})\mathbf{X}\boldsymbol{\beta}_0 = \mathbf{0}_{n\times n}$.
As interest lies in valid inference, uniform consistency of the REML estimator is required. Consistency for ML estimators was first shown by Wald [1949], and uniform ML consistency by Moran [1970]. Both required the parameter space to be compact, as well as independently drawn observations. A consistency result that omits those assumptions was first given by Weiss [1971, 1973]. Miller [1977] applied this on ML in LMMs, which describe a dependency structure. Similarly, Jiang [1996] did so for REML estimators. However, none of the latter three explicitly considered uniform consistency for $\widehat{\boldsymbol{\theta}}$. This is given by the following Lemma.

Lemma 1. *Let model (2) and (A) - (D) hold. Then, $\nu_i(\boldsymbol{\theta}_0)|\widehat{\theta}_i - \theta_{0,i}| = O_P(1)$ for all $\boldsymbol{\theta}_0 \in \Theta$ and $i = 1, \dots, r$ for $\nu_i(\boldsymbol{\theta}_0) = \sqrt{-E\{\partial^2\ell_R(\boldsymbol{\theta})/\partial\theta_i^2|_{\boldsymbol{\theta}_0}\}}$, that is*

$$\nu_i(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{2}}\mathrm{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial\mathbf{V}}{\partial\theta_i}\right\}^{1/2}.$$

Note that $\nu_i(\boldsymbol{\theta}_0)^2$ is the asymptotic variance of $\widehat{\theta}_i$. Also, by conditions (C) and (D), $\partial\mathbf{V}/\partial\theta_i$ is a bounded $(n \times n)$-matrix with independent of $\boldsymbol{\theta}_0$. The proof of Lemma 1 is similar to Moran [1970]. It is merely required to check if their conditions hold uniformly. The second part of the proof mimics Weiss [1971].

4 Main Results

As main result of Ewald and Schneider [2018], it is known how the minimum coverage over the whole parameter space can be expressed in terms of the limiting distribution, depending only on the sign of the entries of $\boldsymbol{\beta}_0$. This result directly carries over to minimum coverage over the space of coefficient and covariance parameters, as the latter, when estimated with REML, does not depend on the fixed effects.

Lemma 2. *Let model (2) and (A)-(D) hold, $\boldsymbol{\theta}_0$ estimated by REML and $k > 0$. Then,*

$$\inf_{\boldsymbol{\beta}_0, \boldsymbol{\theta}_0} P\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\right) \in E\left(\widehat{\mathbf{C}}, k\right)\right\} = \inf_{\boldsymbol{\theta}_0, \mathbf{d}} P\left\{\widehat{\mathbf{u}} \in E\left(\widehat{\mathbf{C}}, k\right)\right\}.$$

The proof follows from Ewald and Schneider [2018, Theorem 1], as $\widehat{\boldsymbol{\theta}}$ is independent from $\boldsymbol{\beta}_0$. Using properties of the optimization problem, is proven by showing that both sets are equal, not by evaluating the probabilities. This can be done by minimizing over a discrete set, as $\mathbf{d} \in \{-1, 1\}^p$, and obviates the need to treat the underlying parameter $\boldsymbol{\beta}_0$ directly. It is clear that the coverage of the resulting confidence set heavily depends on the signs of the specific value of $\boldsymbol{\beta}_0$. See section 5 for similar occurrences and Ewald and Schneider [2018] for an extensive discussion.

To apply this result for LMMs, the next theorem additionally treats the minimization over Θ. The additional variability induced by the estimation of $\boldsymbol{\theta}_0$ is incorporated within a term of known stochastic order. The result shows that due to the uniform consistency of $\widehat{\boldsymbol{\theta}}$, the infimal coverage meets nominal level up to a term of vanishing order.

Theorem 1. *Let model (2) and (A)-(D) hold, $\boldsymbol{\theta}_0$ estimated by REML and*

$$\widehat{\kappa} = \max_{\mathbf{d} \in \{-1,1\}^p} \chi^2_{p,1-\alpha}\left(\left\|\widehat{\mathbf{C}}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\right\|^2\right)$$

the corresponding quantile of the non-central χ^2_p-distribution. Then,

$$\inf_{\boldsymbol{\beta}_0, \boldsymbol{\theta}_0} P\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\right) \in E\left(\widehat{\mathbf{C}}, \widehat{\kappa}\right)\right\} = 1 - \alpha + O\left(\frac{1}{\sqrt{n}}\right).$$

The result differs from Lemma 2 in two aspects. First, the minimization over $\mathbf{d} \in \{-1, 1\}^p$ has been shifted to the choice of non-centrality parameter of the χ^2-distribution, based on Propositions 4 and 5 from Ewald and Schneider [2018]. Note the maximization is invariant with respect to sign, in that if $\mathbf{d}^*$ is the maximizer, so is $-\mathbf{d}^*$. Second, the additional variability induced by $\widehat{\boldsymbol{\theta}}$ has been treated with an error term of vanishing, usual parametric order $n^{-1/2}$.

From Theorem 1 we obtain the confidence set

$$M = \left\{\boldsymbol{\beta} \in \mathbb{R}^p\colon \quad n\left\|\widehat{\mathbf{C}}^{1/2}\left(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}\right)\right\|^2 \leq \widehat{\kappa}\right\} \tag{7}$$

uniformly attains nominal coverage up to an error of parametric rate. Although this implies that the resulting testing procedure is not of nominal level $1 - \alpha$ as discussed in Leeb and Pötscher [2017], it is shown in the simulations in Section 5 that this error term seems to have little influence in finite samples.

5 Simulations

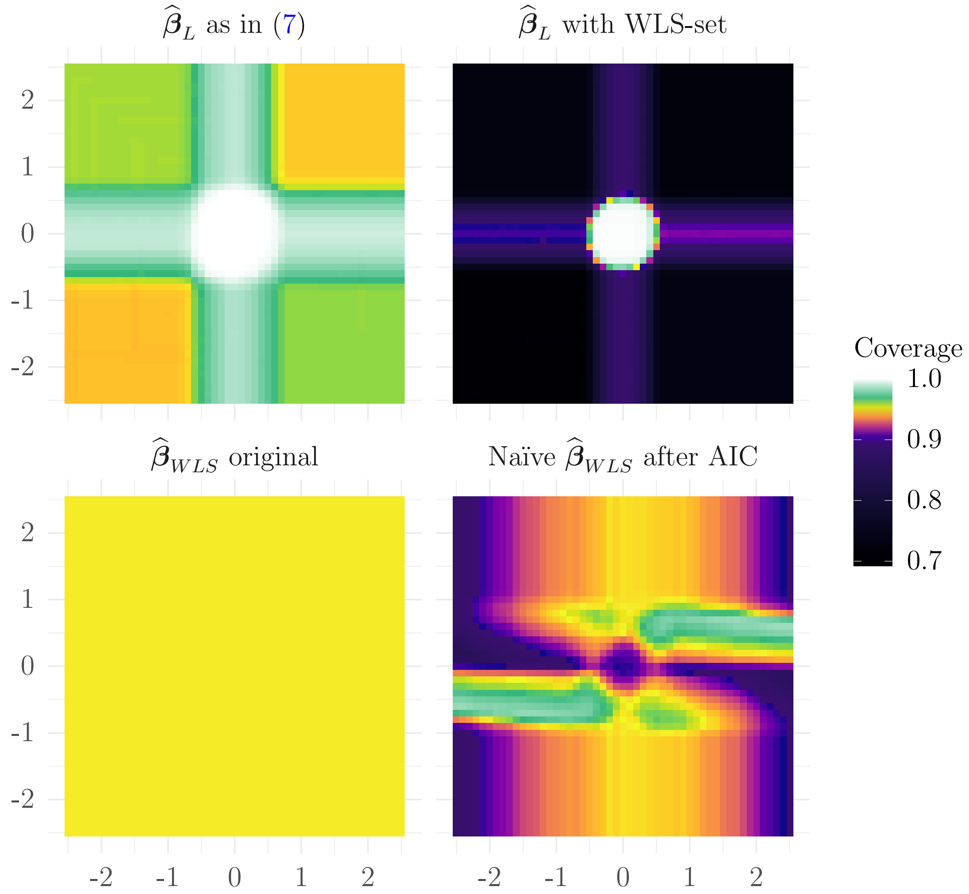

Figure 1: Only confidence sets based on the Lasso as in (7) or WLS estimator (left) achieve nominal level over the whole parameter space (yellow), the former is conservative (green and white) around the origin and axes. Naïvely applying WLS sets to the Lasso or for the WLS estimator after AIC model selection (right) yields undercoverage (dark).

The derived confidence sets are uniformly valid over the whole parameter space. They are constructed from the limiting distribution based on the sign of the parameters. This implies that they attain nominal coverage in two orthants only (two, as they are invariant to sign), whereas they exhibit overcoverage in all other orthants.

In order to visualize this effect, we use the following simulation design for two coefficient parameters. Consider the 'random intercept model', a special case of model (1):

$$\begin{gathered} y_{ij} = \mathbf{x}_{ij}^t \boldsymbol{\beta}_0 + v_i + u_{ij}, \quad i = 1, \ldots, m, \; j = 1, \ldots, n_i; \\ u_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_u^2), \quad v_i \sim \mathcal{N}(0, \sigma_v^2). \end{gathered} \tag{8}$$

We fix the parameters to $m = 20$, $n = 400$, $n_i = 20$, $\sigma_u = \sigma_v = 4$, resembling to a similar scenario as in Kramlinger et al. [2020]. For visualization purposes, we restrict the simulation to $p = 2$ parameters. The tuning parameters are chosen to be $\lambda_i = n^{1/2}/2$ for $i = 1, 2$, resembling a conservative tuning regime. The entries of the matrix of covariables are independently drawn from $\mathcal{N}(0, 4)$, so that the fixed and random effects are of a comparable magnitude. Hence, for $\beta_0 \in [-2, 2]^2$ the empirical coverage probability is computed by checking if $\boldsymbol{\beta}_0 \in M$, for M with $\alpha = .05$ from (7). For each $\boldsymbol{\beta}_0$, 3.000 simulations were carried out. Figure 1 shows the results. The probabilities shown are average empirical coverages for all configurations of $\boldsymbol{\beta}_0$.

First note the classical confidence sets based on the WLS estimator $\widehat{\boldsymbol{\beta}}_{WLS}$ (bottom left). As the distribution of $\widehat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}_0$ is independent of $\boldsymbol{\beta}_0$, the coverage based on its confidence set is attained uniformly over the coefficient parameters space. One finds that no deviations from the the nominal level of 95% (yellow) can be observed in the simulation.

Next, consider the confidence set based on the Lasso as given in (7) (top left). Nominal coverage is only attained up to a small error is attained in two orthants, namely for $\text{sign}(\beta_1) = \text{sign}(\beta_2)$. Note that due to Lemma 6, those orthants can be determined in advance, up to the uncertainty induced by the estimation of $\boldsymbol{\theta}_0$. The other orthants exhibit a slight overcoverage (green), whereas a significant overcoverage (white) occurs at the axes and around the origin. The latter effects are due in the event of variable selection, when a component in $\widehat{\boldsymbol{\beta}}_L$ being zero. Hence, at the axes, a coverage close to 1 is achieved, and the 95%-confidence sets prove to be too wide. These findings are in line with the example of the linear regression model from Ewald and Schneider [2018, Fig. 4]. Although additional uncertainty is present due to the estimation of random effects, the additional error term does not appear too influential. Hence, although Theorem 1 postulates that nominal coverage is only achieved with an additional term of vanishing order, the experiment indicates that the confidence sets still prove to be adequately close to the nominal level.

In contrast to the methods on the left, which meet the nominal level thoroughly, but without model selection (bottom left) or conservatively, but with selection (upper left), two additional naïve approaches are displayed on the right column.

First, one observes that naïvely applying classical WLS confidence sets around a WLS estimator $\widehat{\boldsymbol{\beta}}_{WLS}$ after performing model selection with AIC (bottom right) yields inconsistent coverages over the parameter space. Type-I-error inflation [Berk et al., 2013] occurs in regions with undercoverage (purple, dark) in event of variable selection at the axes. Another approach is applying a WLS confidence set around the estimator $\widehat{\boldsymbol{\beta}}_L$ (top right). Note that a generalization of this approach in the case of $n > p$ has been proposed by van der Geer et al. [2014], although the authors note that their resulting sets hold uniformly. Again, these sets are not theoretically justified and indeed, an overcoverage occurs at the origin, whereas a severe undercoverage over the rest of the coefficient parameter space. Both naïve approaches thus yield misleading confidence sets, and their use is inadvisable.

6 Discussion

This contributions presents a solution to estimate both coefficient and covariance parameters in a low-dimensional LMM in which the fixed effects only are estimated with a

Lasso-penalization. Our aim is to construct uniformly consistent confidence sets for the fixed effects. We suggest that a two-stage estimation procedure, where the covariance parameters are estimated via REML-estimators $\widehat{\boldsymbol{\theta}}$ first, and the parameters with $\widehat{\boldsymbol{\theta}}$ plugged in second. In doing so, the REML estimators do not depend on $\boldsymbol{\beta}_0$, and hence previous results of Ewald and Schneider [2018] can be employed. Eventually, we prove that the resulting confidence sets are uniformly valid under both the coefficient and covariance parameters.

To the best of our knowledge, this work is the first that considers inference specifically for the Lasso in LMMs. We expect that this approach can serve as a basis for proper inference for an estimation procedure that penalizes fixed and random effects in the future.

7 Proofs

In order to prove Lemma 1 is proved, the following preliminary result is helpful.

Lemma 3. *Let model (2) and (A) - (D) hold. Denote for* $i, j, k = 1, \dots, r$

$$\begin{aligned} \mathbf{Q}_{ij}(\boldsymbol{\theta}_0) &= \mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_j}, \\ \mathbf{Q}_{ijk}(\boldsymbol{\theta}_0) &= \mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_i}\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_k}. \end{aligned} \tag{9}$$

Then, there exist constants $c_{ij}, c_{ijk} \in \mathbb{R}$ *with respect to* n *and* $\boldsymbol{\theta}_0$ *such that*

$$\begin{aligned} \operatorname{tr}\left\{\mathbf{Q}_{ij}(\boldsymbol{\theta}_0)\right\} &\asymp \frac{c_{ij}}{n}\operatorname{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_i}\right\}\operatorname{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_j}\right\}, \\ \operatorname{tr}\left\{\mathbf{Q}_{ijk}(\boldsymbol{\theta}_0)\right\} &\asymp \frac{c_{ijk}}{n^2}\operatorname{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_i}\right\}\operatorname{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_j}\right\}\operatorname{tr}\left\{\mathbf{P}(\boldsymbol{\theta}_0)\frac{\partial \mathbf{V}}{\partial \theta_k}\right\}. \end{aligned} \tag{10}$$

Proof of Lemma 3. As notational convenience denote the dependency of quantities involved for this proof only by $\mathbf{P}(\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}$. First, we show that $\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}_{\boldsymbol{\theta}}^{-1}$ is independent of $\boldsymbol{\theta}$, by considering its derivative. Note that $\mathbf{S}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S}_{\boldsymbol{\theta}}$ and thus

$$\frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i} = \mathbf{S}_{\boldsymbol{\theta}}\mathbf{H}_i\mathbf{P}_{\boldsymbol{\theta}} = \frac{\partial \mathbf{S}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i} = \frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i}\mathbf{S}_{\boldsymbol{\theta}} + \mathbf{S}_{\boldsymbol{\theta}}\frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i} = \mathbf{S}_{\boldsymbol{\theta}}\frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i},$$

where the last equality holds as $(\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i)\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S}_{\boldsymbol{\theta}}\mathbf{H}_i\mathbf{P}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{0}_{n\times n}$ by the construction of $\mathbf{P}_{\boldsymbol{\theta}}$. This gives that $\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i = \mathbf{S}_{\boldsymbol{\theta}}\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i$, which implies either that $\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i$ is the null element, i.e. $\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i = \mathbf{0}_{n\times n}$, or, as $\mathbf{S}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S}_{\boldsymbol{\theta}}$, $\partial \mathbf{S}_{\boldsymbol{\theta}}/\partial \theta_i = \mathbf{S}_{\boldsymbol{\theta}}$. However, the latter as well implies

$$\frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i} = \mathbf{S}_{\boldsymbol{\theta}} = \mathbf{S}_{\boldsymbol{\theta}}^2 = \left(\frac{\partial \mathbf{S}_{\boldsymbol{\theta}}}{\partial \theta_i}\right)^2 = \mathbf{S}_{\boldsymbol{\theta}}\mathbf{H}_i\mathbf{P}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}}\mathbf{H}_i\mathbf{P}_{\boldsymbol{\theta}} = \mathbf{0}_{n\times n},$$

again using $\mathbf{P}_{\boldsymbol{\theta}}\mathbf{S}_{\boldsymbol{\theta}} = \mathbf{0}_{n\times n}$. Thus, since the derivative of $\mathbf{S}_{\boldsymbol{\theta}}$ is zero, $\mathbf{S}_{\boldsymbol{\theta}}$ is constant and hence independent of $\boldsymbol{\theta}$. Now, (10) can be shown. Let $\mathbf{R}_i \in \mathbb{R}^{n\times n}$ such that

$$\mathbf{P}_{\boldsymbol{\theta}_0}\frac{\partial \mathbf{V}}{\partial \theta_i} = \mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\left[\mathbf{I}_n - \mathbf{X}\left\{\mathbf{X}^t\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{X}\right\}^{-1}\mathbf{X}^t\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\right]\frac{\partial \mathbf{V}}{\partial \theta_i} = \mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i.$$

By condition (C), $\mathbf{R}_i = (\mathbf{I}_n - \mathbf{S}_{\boldsymbol{\theta}_0})\partial \mathbf{V}/\partial \theta_i$ is independent of $\boldsymbol{\theta}_0$. Let $c_0, \dots, c_9 \in \mathbb{R}$ be constants with respect to n and $\boldsymbol{\theta}_0$. Further, let $I_s = \left\{t \in \{1, \dots, n\} : (\mathbf{V}_{\boldsymbol{\theta}_0}^{-1})_{st} \neq 0\right\}$ being the set of indices of non-zero entries in the s-th column of $\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}$ and similarly $I_s^i = \{t \in \{1, \dots, n\} : (\mathbf{R}_i)_{st} \not\twoheadrightarrow 0\}$ the set of constant entries. Also, as the entries of $\mathbf{X}$ do not vanish by condition (D), $t \notin I_s^i$: $(\mathbf{R}_i)_{st} \asymp c_0/n$. Now,

$$\left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{st} \asymp \sum_{u \in I_s \cap I_t^i} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\right)_{su} (\mathbf{R}_i)_{ut} + \frac{c_1}{n} \sum_{u \in I_s \setminus I_t^i} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\right)_{su}, \tag{11}$$

for any $s, t = 1, \dots, n$. If t is such that $I_s \cap I_t^i \neq \emptyset$ this implies that $(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i)_{st} \asymp c_2(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i)_{ss}$. Finally, note that $I_s \cap I_s^i \neq \emptyset$ for all s. Now consider the left hand side of the the first line of (10):

$$\begin{aligned}
\operatorname{tr}\{\mathbf{Q}_{ij}(\boldsymbol{\theta}_0)\} &= \sum_{s=1}^{n}\sum_{t=1}^{n} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{st} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} \\
&= \sum_{s=1}^{n} \sum_{t : I_s \cap I_t^i \neq \emptyset} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{st} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} + \sum_{s=1}^{n} \sum_{t : I_s \cap I_t^i = \emptyset} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{st} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} \\
&= A + B.
\end{aligned}$$

For A, note that as $|\{t : I_s \cap I_t^i \neq \emptyset\}| = O(1)$, it follows that

$$\begin{aligned}
A \asymp c_3 \sum_{s=1}^{n} \sum_{t : I_s \cap I_t^i \neq \emptyset} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{ss} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} &\asymp c_4 \sum_{s=1}^{n} \sum_{t : I_s \cap I_t^i \neq \emptyset} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{ss} \frac{1}{n} \sum_{u=1}^{n} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{tu} \\
&\asymp \frac{c_5}{n} \operatorname{tr}\left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right) \operatorname{tr}\left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right).
\end{aligned}$$

Similarly, by (11) and as $|\{t : I_s \cap I_t^i = \emptyset\}| \asymp c_6 n$ it holds for B that

$$\begin{aligned}
B \asymp c_7 \sum_{s=1}^{n} \sum_{t : I_s \cap I_t^i = \emptyset} \frac{1}{n} \sum_{u \in I_s \setminus I_t^i} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\right)_{su} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} &\asymp \frac{c_8}{n} \sum_{s=1}^{n} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right)_{ss} \sum_{t : I_s \cap I_t^i = \emptyset} \left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right)_{ts} \\
&\asymp \frac{c_9}{n} \operatorname{tr}\left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_i\right) \operatorname{tr}\left(\mathbf{V}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_j\right).
\end{aligned}$$

Altogether, this gives the first line of (10). The second line is shown analogous. □

The result is helpful as $\nu_i(\boldsymbol{\theta}_0) = \sqrt{\operatorname{tr}\{\mathbf{Q}_{ii}(\boldsymbol{\theta}_0)\}/2} \asymp c_{10}\operatorname{tr}\{\mathbf{Q}_{ii}(\boldsymbol{\theta}_0)\}$. Now Lemma 1 is shown.

Proof of Lemma 1. This proof is adapted to account for uniformity w.r.t. $\boldsymbol{\theta}_0$ from Lemma 7.2 (i) from Jiang [1996] and closely follows the lines of Weiss [1971]. The four boundedness conditions below on the derivatives of the log-likelihood are surrogates of the boundedness conditions imposed on the log-likelihood directly by Wald [1949] and Moran [1970]. By a Taylor expansion the conditions below include boundedness of the log-likelihood as well. They are further required to omit the compactness condition of Θ, by constructing a compact set Θ'_ϵ, see details below. For all $i, j = 1 \dots, r$, it holds for any $\boldsymbol{\theta}_0$ that

(i) $\mathrm{E}\left\{\frac{\partial}{\partial\theta_i} l_R(\boldsymbol{\theta})|_{\boldsymbol{\theta}_0}\right\} = 0$,

(ii) $\frac{1}{\nu_i(\boldsymbol{\theta}_0)}\frac{\partial}{\partial\theta_i} l_R(\boldsymbol{\theta})|_{\boldsymbol{\theta}_0} = O_P(1)$,

(iii) $J_{ij}(\boldsymbol{\theta}_0) = -\frac{1}{\nu_i(\boldsymbol{\theta}_0)\nu_j(\boldsymbol{\theta}_0)}\mathrm{E}\left\{\frac{\partial^2}{\partial\theta_i\theta_j}\ell_R(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}_0}\right\} = c$ for some constant $c \in \mathbb{R}$,

(iv) $\epsilon_{ij}(\boldsymbol{\theta}) = \frac{1}{\nu_i(\boldsymbol{\theta})\nu_j(\boldsymbol{\theta})}\frac{\partial^2}{\partial\theta_i\partial\theta_j} l_R(\boldsymbol{\theta}) + J_{ij}(\boldsymbol{\theta}_0) = o_P(1)$ for $\boldsymbol{\theta} \in \Theta_q = \{\boldsymbol{\theta} : \nu_i(\boldsymbol{\theta}_0)|\theta_i - \theta_{0,i}| \leq q\}$,

where $q \asymp n^{1/(2+\epsilon)}$ for some $\epsilon > 0$. For readability, suppress the dependency from $\boldsymbol{\theta}$ when the argument is clear from the context. Now, (i)-(iv) are shown.

(i) Note that $\partial\mathbf{P}/\partial\theta_i = -\mathbf{P}\partial\mathbf{V}/\partial\theta_i\mathbf{P}$ and $\mathbf{PVP} = \mathbf{P}$, as well as

$$\frac{\partial l_R}{\partial\theta_i}\bigg|_{\boldsymbol{\theta}_0} = \frac{1}{2}\mathbf{y}^t\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{P}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_i}\right).$$

The claim now follows after taking expectation.

(ii) As $\mathrm{E}(\partial l_R/\partial\theta_i|_{\boldsymbol{\theta}_0}) = 0$ by (i) and $\mathrm{Var}(\partial l_R/\partial\theta_i|_{\boldsymbol{\theta}_0}) = \nu_i^2$, Chebychevs inequality can be applied, it holds uniformly, and gives that for any $\epsilon > 0$ there exists $k > 0$, such that

$$\sup_{\boldsymbol{\theta}_0} P\left(\frac{1}{\nu_i}\left|\frac{\partial l_R}{\partial\theta_i}\bigg|_{\boldsymbol{\theta}_0}\right| \geq k\right) \leq \frac{1}{k^2} < \epsilon,$$

or, equivalently, $\nu_i^{-1}\partial l_R/\partial\theta_i|_{\boldsymbol{\theta}_0} = O_P(1)$.

(iii) The same reasoning as for (i) with the help of Lemma 3 gives for a constant c:

$$J_{ij}(\boldsymbol{\theta}_0) = -\frac{1}{\nu_i\nu_j}\mathrm{E}\left\{\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\right) - \mathbf{y}^t\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{P}\mathbf{y}\right\} = \frac{\mathrm{tr}(\mathbf{Q}_{ij})}{\sqrt{\mathrm{tr}(\mathbf{Q}_{jj})\mathrm{tr}(\mathbf{Q}_{jj})}} = c.$$

(iv) First, for any $\boldsymbol{\theta} \in \Theta_q$, consider expectation and variance of the random term in (iv) w.r.t. $\boldsymbol{\theta}$. For that note that for any $k = 1, \ldots, r$,

$$\sum_{i=1}^{r}(\theta_i - \theta_{0,i})\frac{\partial\mathbf{V}}{\partial\theta_i} \leq \sum_{i=1}^{r}\frac{q}{\nu_i}\frac{\partial\mathbf{V}}{\partial\theta_i} = O\left[\left\{\frac{q}{\sqrt{\mathrm{tr}(\mathbf{Q}_{kk})}}\right\}_{n\times n}\right].$$

This implies $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\theta}_0) + [O\{q\,\mathrm{tr}(\mathbf{Q}_{kk})^{-1/2}\}]_{n\times n}$ for any $k = 1, \ldots, r$, by a Taylor expansion. With Lemma 3, this gives

$$\begin{aligned}\mathrm{E}\left\{-\frac{1}{\nu_i\nu_j}\frac{\partial^2\ell_R}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta})\right\} &= -\frac{1}{\nu_i\nu_j}\left[\frac{1}{2}\mathrm{tr}(\mathbf{Q}_{ij}) - \mathrm{tr}\left\{\mathbf{Q}_{ij}\mathbf{P}\mathbf{V}(\boldsymbol{\theta}_0)\right\}\right] \\ &= J_{ij}(\boldsymbol{\theta}) + O\left\{q\frac{\mathrm{tr}(\mathbf{Q}_{ijk})}{\mathrm{tr}(\mathbf{Q}_{ii})^{3/2}}\right\} = J_{ij}(\boldsymbol{\theta}) + O\left(\frac{q}{\sqrt{n}}\right),\end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta_q$ on which all quantities except $\mathbf{V}(\boldsymbol{\theta}_0)$ depend upon. Similarly, for any $k = 1, \ldots, r$,

$$\begin{aligned}
\mathrm{Var}\left\{-\frac{1}{\nu_i\nu_j}\frac{\partial^2 \ell_R}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta})\right\} &= -\frac{1}{\nu_i^2\nu_j^2}\mathrm{tr}\left\{\mathbf{Q}_{ij}\mathbf{P}\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{Q}_{ij}\mathbf{P}\mathbf{V}(\boldsymbol{\theta}_0)\right\} \\
&= O\left\{\frac{\mathrm{tr}(\mathbf{Q}_{ij}\mathbf{Q}_{ij})}{\mathrm{tr}(\mathbf{Q}_{kk})^2} + q\frac{\mathrm{tr}(\mathbf{Q}_{ij}\mathbf{Q}_{ijk})}{\mathrm{tr}(\mathbf{Q}_{kk})^{5/2}} + q^2\frac{\mathrm{tr}(\mathbf{Q}_{ijk}\mathbf{Q}_{ijk})}{\mathrm{tr}(\mathbf{Q}_{kk})^3}\right\} \\
&= O\left(\frac{q^2}{n}\right)
\end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta_q$. Putting the previous two results together, Chebychev gives that for any $\epsilon > 0$ there exists $k > 0$ such that for $\boldsymbol{\theta} \in \Theta_q$, where Θ_q depends on $\boldsymbol{\theta}_0$,

$$\sup_{\boldsymbol{\theta}_0} P_{\boldsymbol{\theta}}\left\{\frac{\sqrt{n}}{q}\left|\frac{1}{\nu_i\nu_j}\frac{\partial^2 \ell_R}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta}) + J_{ij}(\boldsymbol{\theta})\right| \geq k\right\} \leq \frac{1}{k^2} < \epsilon, \tag{12}$$

or $-\frac{1}{\nu_i\nu_j}\frac{\partial^2\ell_R}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta}) = J_{ij}(\boldsymbol{\theta}) + O_P(qn^{-1/2})$ for any $\boldsymbol{\theta}_0$. In order to prove (iv), this must hold for $\boldsymbol{\theta}_0$ in the right hand side. Taking derivatives gives with Lemma 3 for $\boldsymbol{\theta} \in \Theta_q$ and $i, j, k = 1, \ldots, r$ that

$$\begin{aligned}
(\theta_k - \theta_{0,k})\frac{\partial J_{ij}}{\partial\theta_k}\bigg|_{\boldsymbol{\theta}_0} &\leq \frac{q}{\nu_k}\frac{\partial J_{ij}}{\partial\theta_k}\bigg|_{\boldsymbol{\theta}_0} \\
&= \sum_{\substack{a,b\in\{i,j\}\\ a\neq b}} q\frac{\mathrm{tr}(\mathbf{Q}_{aa})\left\{\mathrm{tr}(\mathbf{Q}_{ab})\mathrm{tr}(\mathbf{Q}_{bbk}) + \mathrm{tr}(\mathbf{Q}_{bb})\mathrm{tr}(\mathbf{Q}_{abk})\right\}}{\mathrm{tr}(\mathbf{Q}_{kk})^{1/2}\mathrm{tr}(\mathbf{Q}_{ii})^{3/2}\mathrm{tr}(\mathbf{Q}_{jj})^{3/2}} = O\left(\frac{q}{\sqrt{n}}\right)
\end{aligned}$$

Finally, a Taylor expansion for $J_{ij}(\boldsymbol{\theta})$ around $J_{ij}(\boldsymbol{\theta}_0)$ gives the second equality below, while the first is due to (12), and it follows that

$$\frac{1}{\nu_i\nu_j}\frac{\partial^2 l_R}{\partial\theta_i\partial\theta_j}(\boldsymbol{\theta}) = -J_{ij}(\boldsymbol{\theta}) + O_P\left(\frac{q}{\sqrt{n}}\right) = -J_{ij}(\boldsymbol{\theta}_0) + O_P\left(\frac{q}{\sqrt{n}}\right),$$

so that (iv) holds as well, as $qn^{-1/2} \to 0$.

The second part of the proof mimics the reasoning of Weiss [1971]. Let $J(\boldsymbol{\theta}_0) = \{J_{ij}(\boldsymbol{\theta}_0)\}_{ij}$ and $\mathbf{s}(\boldsymbol{\theta}_0) = \{s_1(\boldsymbol{\theta}_0), s_2(\boldsymbol{\theta}_0), \ldots, s_r(\boldsymbol{\theta}_0)\}$ with $s_i(\boldsymbol{\theta}_0) = \nu_i(\boldsymbol{\theta}_0)^{-1}\partial l_R/\partial\theta_i|_{\boldsymbol{\theta}_0}$ and define $\boldsymbol{\theta}'$ such that

$$\nu(\boldsymbol{\theta}_0) \otimes (\boldsymbol{\theta}' - \boldsymbol{\theta}_0) = \mathbf{s}(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0)^{-1}.$$

Note that $\mathbf{J}(\boldsymbol{\theta}_0)$ is non-singular as its i-th and j-th row are linearly independent by condition (C). By (i) and (iii), $\boldsymbol{\theta}' \in \Theta_q$ for some q large enough. Now, for any $\boldsymbol{\theta} \in \Theta_q$,

$$\begin{aligned}
\ell_R(\boldsymbol{\theta}) = \ell_R(\boldsymbol{\theta}_0) &+ \sum_{i=1}^{r} \nu_i(\boldsymbol{\theta}_0)(\theta_i - \theta_{0,i}) s_i(\boldsymbol{\theta}_0) \\
&- \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r} \nu_i(\boldsymbol{\theta}_0)(\theta_i - \theta_{0,i})\nu_j(\boldsymbol{\theta}_0)(\theta_j - \theta_{0,j}) J_{ij}(\boldsymbol{\theta}_0) + R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \\
= \ell_R(\boldsymbol{\theta}_0) &+ \frac{1}{2}\mathbf{s}(\boldsymbol{\theta}_0)^t I(\boldsymbol{\theta}_0)^{-1}\mathbf{s}(\boldsymbol{\theta}_0) \\
&- \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r} \nu_i(\boldsymbol{\theta}_0)(\theta'_i - \theta_i)\nu_j(\boldsymbol{\theta}_0)(\theta'_j - \theta_j) J_{ij}(\boldsymbol{\theta}_0) + R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}),
\end{aligned} \tag{13}$$

where $R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r} \nu_i(\boldsymbol{\theta}_0)(\theta_i - \theta_{0,i})\nu_j(\boldsymbol{\theta}_0)(\theta_j - \theta_{0,j})\epsilon(\widetilde{\boldsymbol{\theta}})$ for some $\widetilde{\theta}_i = \theta_{0,i} + t(\theta_i - \theta_{0,i})$ where $t \in [-1, 1]$. By (iv) we have $\sup_{\boldsymbol{\theta} \in \Theta_q} |R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})| = o_P(1)$. Now consider the set $\Theta'_\epsilon = \{\boldsymbol{\theta} : \nu_i(\boldsymbol{\theta}_0)|\theta_i - \theta'_i| < \epsilon\}$ and its boundary $\overline{\Theta}'_\epsilon$. By (i) and (iii), there exists a sequence $\epsilon_n \to 0$ such that $\inf_{\boldsymbol{\theta}_0} P(\overline{\Theta}'_{\epsilon_n} \subset \Theta_q) \to 1$ for q large enough. Hence, $\sup_{\boldsymbol{\theta} \in \overline{\Theta}'_{\epsilon_n}} |R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})| = o_P(1)$. Second-to-last, consider

$$\delta(\epsilon_n) = \min_{\boldsymbol{\theta} \in \overline{\Theta}'_{\epsilon_n}} \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{r} \nu_i(\boldsymbol{\theta}_0)(\theta'_i - \theta_i)\nu_j(\boldsymbol{\theta}_0)(\theta'_j - \theta_j) I_{ij}(\boldsymbol{\theta}_0),$$

and note that $\delta(\epsilon)$ is not stochastic, increasing and $\delta(0) = 0$. Finally, let ϵ_n such that $\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}_0} P\{2\sup_{\boldsymbol{\theta} \in \overline{\Theta}'_{\epsilon_n}} |R(\boldsymbol{\theta})| < \delta(\epsilon_n)\} = 1$, for which the infimum holds due to (iii) and $\boldsymbol{\theta}' \in \Theta_q$ for q large enough. Then, using (13) for $\boldsymbol{\theta}' \in \Theta_q$ for q large enough, gives

$$\begin{aligned}
&\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}_0} P\left\{ \min_{\boldsymbol{\theta} \in \overline{\Theta}'_{\epsilon_n}} \ell_R(\boldsymbol{\theta}') - \ell_R(\boldsymbol{\theta}) > 0 \right\} \\
&\qquad = \liminf_{n\to\infty} \inf_{\boldsymbol{\theta}_0} P\left\{ R(\boldsymbol{\theta}', \widetilde{\boldsymbol{\theta}}') + \delta(\epsilon_n) + \min_{\boldsymbol{\theta} \in \overline{\Theta}'_{\epsilon_n}} -R(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) > 0 \right\} = 1.
\end{aligned}$$

Thus, there is a maximum $\widehat{\boldsymbol{\theta}} \in \Theta'_{\epsilon_n}$ such that $\nu_i|\widehat{\theta}_i - \theta'_i| < \epsilon_n$, and thus $\nu_i(\boldsymbol{\theta}_0)(\widehat{\theta}_i - \theta_{0,i}) - \nu_i(\boldsymbol{\theta}_0)(\theta_{0,i} - \theta'_i) = o_P(1)$ for any $\boldsymbol{\theta}_0$. The claim follows. □

In order to address the infimum over Θ, we use the following result.

Lemma 4. *Let X_n and Y_n be random variables where $X_n = O_P(a_n)$ and $Y_n = O_P(a_n b_n)$ with $b_n = o(1)$. Then,*

$$P(X_n + Y_n \leq a_n) = P(X_n \leq a_n) + O(b_n).$$

The asymptotic result is clear as convergence in probability implies convergence in distribution. Above result further specifies the rate of convergence.

Proof of Lemma 4. First, let $\phi(s,t) = P(X_n + s \leq a_n | Y_n = t)$ and consider a Taylor expansion for s/a_n around zero, which gives $\phi(s,t) = \phi(0,t) + O(s/a_n)$. This implies $\int_{a_n - t}^{a_n} p_{X_n|Y_n=t}(z)\, dz = \phi(0,t) - \phi(t,t) = O(t/a_n)$. Using convolution formula we obtain

$$\begin{aligned}
\mathrm{P}\big(X_n + Y_n \leq a_n\big) &= \int_{-\infty}^{a_n} p_{X_n+Y_n}(z)\, dz \\
&= \int_{-\infty}^{a_n} \int_{-\infty}^{\infty} p_{X_n,Y_n}(z-t,t)\, dt\, dz \\
&= \int_{-\infty}^{a_n} \int_{-\infty}^{\infty} p_{X_n|Y_n=t}(z-t) p_{Y_n}(t)\, dt\, dz \\
&= \int_{-\infty}^{\infty} p_{Y_n}(t) \int_{-\infty}^{a_n - t} p_{X_n|Y_n=t}(z)\, dz\, dt \\
&= \int_{-\infty}^{\infty} p_{Y_n}(t) \left\{ \int_{-\infty}^{a_n} p_{X_n|Y_n=t}(z)\, dz + \int_{a_n - t}^{a_n} p_{X_n|Y_n=t}(z)\, dz \right\} dt \qquad (14) \\
&= \int_{-\infty}^{\infty} p_{Y_n}(t) \left\{ \int_{-\infty}^{a_n} p_{X_n|Y_n=t}(z)\, dz + O\left(\frac{t}{a_n}\right) \right\} dt \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{a_n} p_{X_n,Y_n}(z,t)\, dz\, dt + O\left\{ \mathrm{E}\left(\frac{Y_n}{a_n}\right) \right\} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{a_n} p_{X_n}(z) p_{Y_n|X_n=z}(t)\, dz\, dt + O(b_n) \\
&= \mathrm{P}(X_n \leq a_n) + O(b_n),
\end{aligned}$$

which gives the claim. □

The next two result are helpful to prove Theorem 1.

Lemma 5. *Let model (2) and (A)-(D) hold,* $\boldsymbol{\theta}_0$ *estimated by REML and* $\widehat{k} = k + c > 0$ *with*

$$k = O\left(p + \left\|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\right\|^2\right), \qquad c = O_P\left(\frac{1}{\sqrt{n}} \left\|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\right\|^2\right).$$

Then,

$$\inf_{\boldsymbol{\theta}_0, \mathbf{d}} P\left\{\widehat{\mathbf{u}} \in E\left(\widehat{\mathbf{C}}, \widehat{k}\right)\right\} = \min_{\mathbf{d}} P\left\{\mathbf{u} \in E\left(\mathbf{C}, k\right)\right\} + O\left(\frac{1}{\sqrt{n}}\right).$$

Proof of Lemma 5. First, let $\xi = \|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2$ and observe that $\|\mathbf{C}^{1/2}\mathbf{u}\|^2 \sim \chi^2_p(\xi)$. This implies that $\|\mathbf{C}^{1/2}\mathbf{u}\|^2 = O_P(1+\xi)$ for all $\boldsymbol{\theta}_0$. Now, consider its derivative with respect to θ_i, $i = 1, \ldots, r$.

$$\frac{\partial}{\partial \theta_i} \|\mathbf{C}^{1/2}\mathbf{u}\|^2 = \|\boldsymbol{\Omega}_i^{1/2}(\mathbf{w} - \boldsymbol{\Lambda}\mathbf{d})\|^2 + 2(\mathbf{w} - \boldsymbol{\Lambda}\mathbf{d})^t \mathbf{C}^{-1} \frac{\partial \mathbf{w}}{\partial \theta_i},$$

where $\boldsymbol{\Omega}_i = \mathbf{C}^{-1}\mathbf{X}^t\mathbf{V}^{-1}\partial\mathbf{V}/\partial\theta_i \mathbf{V}^{-1}\mathbf{X}\mathbf{C}^{-1}/n$ as $\partial\mathbf{C}^{-1}/\partial\theta_i = \boldsymbol{\Omega}_i$. Now,

$$\mathrm{E}\left(\frac{\partial}{\partial \theta_i} \|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) = \|\boldsymbol{\Omega}_i^{1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2 - \mathrm{tr}(\mathbf{C}\boldsymbol{\Omega}_i).$$

For two symmetric positive semi-definite matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ it holds $\mathrm{tr}(\mathbf{AB}) = \|\mathbf{A}^{1/2}\mathbf{B}^{1/2}\|^2 \leq \|\mathbf{A}^{1/2}\|^2\|\mathbf{B}^{1/2}\|^2 = \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B})$. Hence, for $\mathbf{A} = \mathbf{V}^{-1/2}\partial\mathbf{V}/\partial\theta_i\mathbf{V}^{-1/2}$ and $\mathbf{B} = \mathbf{V}^{-1/2}\mathbf{X}\mathbf{C}^{-1}(\boldsymbol{\Lambda}\mathbf{d}\mathbf{d}^t\boldsymbol{\Lambda} + \mathbf{C})\mathbf{C}^{-1}\mathbf{X}^t\mathbf{V}^{-1/2}/n$, it follows

$$\mathrm{E}\left(\frac{\partial}{\partial\theta_i}\|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) \leq \mathrm{tr}(\mathbf{AB}) \leq \mathrm{tr}(\mathbf{A})\mathrm{tr}(\mathbf{B}) = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\right)(1+\xi),$$

and thus $\mathrm{E}\left(\partial\|\mathbf{C}^{1/2}\mathbf{u}\|^2/\partial\theta_i\right) = O\{\mathrm{tr}(\mathbf{V}^{-1}\partial\mathbf{V}/\partial\theta_i)(1+\xi)\}$ for all $\boldsymbol{\theta}_0$. Proceeding analogously for $\nu_i(\boldsymbol{\theta}_0)$ as defined in Lemma 1 gives $\nu_i(\boldsymbol{\theta}_0) = O\{\sqrt{n-p}\,\mathrm{tr}(\mathbf{V}^{-1}\partial\mathbf{V}/\partial\theta_i)\}$ for all $\boldsymbol{\theta}_0$. Altogether,

$$\mathrm{E}\left(\frac{\partial}{\partial\theta_i}\|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) = O\left\{\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\right)(1+\xi)\right\} = O\left\{\frac{\nu_i(\boldsymbol{\theta}_0)}{\sqrt{n}}(1+\xi)\right\}$$

for all $\boldsymbol{\theta}_0$. Similarly, lengthy calculations give that

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta_i}\|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) = 2\mathrm{tr}(\boldsymbol{\Omega}_i\mathbf{C}\boldsymbol{\Omega}_i\mathbf{C}) - 4\left\|\mathbf{C}^{1/2}\boldsymbol{\Omega}_i\boldsymbol{\Lambda}\mathbf{d}\right\|^2 + \frac{4}{n}\left\|\mathbf{V}^{-1/2}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{V}^{-1}\mathbf{X}\mathbf{C}^{-1}\boldsymbol{\Lambda}\mathbf{d}\right\|^2.$$

Proceeding as above,

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta_i}\|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) = O\left\{\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\right)^2(1+\xi)\right\} = O\left\{\frac{\nu_i(\boldsymbol{\theta}_0)^2}{n}(1+\xi)\right\}$$

for all $\boldsymbol{\theta}_0$. With Chebychev, this gives the representation

$$\frac{\partial}{\partial\theta_i}\left\|\mathbf{C}^{1/2}\mathbf{u}\right\|^2 = O_P\left(\frac{\nu_i(\boldsymbol{\theta}_0)}{\sqrt{n}}\|\mathbf{C}^{1/2}\mathbf{u}\|^2\right) \tag{15}$$

for all $\boldsymbol{\theta}_0$. By Lemma 1, a Taylor expansion for $\|\widehat{\mathbf{C}}^{1/2}\widehat{\mathbf{u}}\|^2$ around $\boldsymbol{\theta}_0$ eventually gives

$$\left\|\widehat{\mathbf{C}}^{1/2}\widehat{\mathbf{u}}\right\|^2 = \left\|\mathbf{C}^{1/2}\mathbf{u}\right\|^2 + O_P\left\{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t\frac{\partial}{\partial\boldsymbol{\theta}}\left\|\mathbf{C}^{1/2}\mathbf{u}\right\|^2\right\} = \left\|\mathbf{C}^{1/2}\mathbf{u}\right\|^2\{1 + O_P(n^{-1/2})\},$$

which holds for all $\boldsymbol{\theta}_0$. Eventually, let $X_n = \|\mathbf{C}^{1/2}\mathbf{u}\|^2$, $Y_n = O_P(n^{-1/2}X_n - c) = O_P(b_nX_n)$ for all $\boldsymbol{\theta}_0$ with $b_n = n^{-1/2}$ and $a_n = k$. Then,

$$\inf_{\boldsymbol{\theta}_0} P\left(\left\|\widehat{\mathbf{C}}^{1/2}\widehat{\mathbf{u}}\right\|^2 \leq \widehat{k}\right) = \inf_{\boldsymbol{\theta}_0} P\left(\frac{X_n}{1+\xi} + \frac{Y_n}{1+\xi} \leq \frac{k}{1+\xi}\right) = P\left(X_n + Y_n \leq k\right),$$

where the second equality holds as all quantities inside the probability are independent of $\boldsymbol{\theta}_0$. Finally, Lemma 4 gives the claim. □

Lemma 6. *Let model (2) and (A)-(C) hold. Then, for* $k > 0$,

$$\operatorname*{argmin}_{\mathbf{d}\in\{-1,1\}^p} P\Big\{\mathbf{u} \in E\big(\mathbf{C}, k\big)\Big\} = \operatorname*{argmax}_{\mathbf{d}\in\{-1,1\}^p} \left\|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\right\|^2.$$

This result is given in [Ewald and Schneider, 2018, Prop. 4].

Proof (of Theorem 1). Consider $\kappa = \max_{\mathbf{d}} \chi^2_{p,1-\alpha}(\xi)$ with $\xi = \|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2$. Since for $X \sim \chi^2_p(\xi)$ it holds $X = O_P(1+\xi)$ for all $\boldsymbol{\theta}_0$ and by the definition of the quantile $\mathrm{P}\big(X \leq \kappa\big) = 1 - \alpha$ it follows that $\kappa = O(1+\xi)$ for all $\boldsymbol{\theta}_0$ as well.

Now we proceed similar to the proof of Lemma 5. A Taylor expansion for $\|\widehat{\mathbf{C}}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2$ around $\boldsymbol{\theta}_0$ gives for $\boldsymbol{\Omega}_i = \mathbf{C}^{-1}\mathbf{X}^t\mathbf{V}^{-1}\partial\mathbf{V}/\partial\theta_i\mathbf{V}^{-1}\mathbf{X}\mathbf{C}^{-1}/n$ that

$$\begin{aligned}\|\widehat{\mathbf{C}}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2 &= \|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2 + O_P\left\{\sum_{i=1}^{p}\left(\widehat{\theta}_i - \theta_{0,i}\right)^t \left\|\boldsymbol{\Omega}_i^{1/2}\boldsymbol{\Lambda}\mathbf{d}\right\|^2\right\} \\ &= \|\mathbf{C}^{-1/2}\boldsymbol{\Lambda}\mathbf{d}\|^2 + O_P\big(n^{-1/2}\xi\big)\end{aligned}$$

for all $\boldsymbol{\theta}_0$ and together with the first argument it follows that $\widehat{\kappa} = \kappa + O_P(n^{-1/2}\xi)$ for all $\boldsymbol{\theta}_0$. By Lemma 5 it is ensured that the coverage is attained uniformly for both $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$,

$$\inf_{\boldsymbol{\beta}_0,\boldsymbol{\theta}_0} \mathrm{P}\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0\right) \in E\left(\widehat{\mathbf{C}}, \widehat{\kappa}\right)\right\} = \min_{\mathbf{d}\in\{-1,1\}^p} \mathrm{P}\left\{\mathbf{u} \in E\left(\mathbf{C}, \kappa\right)\right\} + O(n^{-1/2}).$$

By Lemma 6, this minimum is in fact attained for the $\mathbf{d} \in \{-1,1\}^p$ for which κ ensures nominal coverage, since $\|\mathbf{C}^{1/2}\mathbf{u}\|^2 \sim \chi^2_p(\xi)$. This proves the claim. □

References

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid Post-Selection Inference. *Annals of Statistics*, 41(2):802–837, 2013.

H. Bondell, A. Krishna, and S. Ghosh. Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, 66:1069–1077, 2010.

E. Demidenko. *Mixed Models: Theory and Applications.* Wiley Series in Probability and Statistics, Hoboken, NJ, 2004.

K. Ewald and U. Schneider. Uniformly Valid Confidence Sets Based on the Lasso. *Electronic Journal of Statistics*, 12:1358–1387, 2018.

C. R. Henderson. Estimation of Genetic Parameters. *The Annals of Mathematical Statistics*, 21:309–310, 1950.

J. Ibrahim, H. Zhu, R. Garcia, and R. Guo. Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67(2):1358–1387, 2011.

J. Jiang. REML Estimation: Asymptotic Behavior and Related Topics. *The Annals of Statistics*, 24(1):255–286, 1996.

P. Juming and J. Shang. A simultaneous variable selection methodology for linear mixed models. *Journal of Statistical Computation and Simulation*, 88(17): 3323–3337, 2019.

P. Kramlinger, T. Krivobokova, and S. Sperlich. Marginal and Conditional Multiple Inference in Linear Mixed Models. *Submitted*, 2020.

H. Leeb and B. Pötscher. Testing in the Presence of Nuisance Parameters: Some Comments on Tests Post-Model-Selection and Random Critical Values. In S. Ahmed, editor, *Big and Complex Data Analysis. Contributions to Statistics*, pages 69–82. Springer, Cham, 2017.

K.-C. Li. Honest Confidence Regions for Nonparametric Regression. *Annals of Statistics*, 17(3):1001–1008, 1989.

J. J. Miller. Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance. *Annals of Statistics*, 5(4):746–762, 1977.

P. A. P. Moran. The Uniform Consistency of Maximum-Likelihood Estimators. *POPS*, 70:435–439, 1970.

S. Müller, J. Scealy, and A. Welsh. Model Selection in Linear Mixed Models. *Staistical Science*, 28(2):135–167, 2013.

H. Peng and Y. Lu. Model Selection in Linear Mixed Models. *Journal of Multivariate Analysis*, 109:109–129, 2012.

D. Pfefferman. New Important Developements in Small Area Estimation. *Statistical Science*, 28(1):40–68, 2013.

J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, NY, 2000.

B. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100:2065–2082, 2009.

N. G. N. Prasad and J. N. K. Rao. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

J. N. K. Rao and I. Molina. *Small Area Estimation*. Wiley, Hoboken, NJ, 2nd edition, 2015.

J. Schelldorfer, P. Bühlmann, and S. van de Geer. Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. *Scandinavian Journal of Statistics*, 38:197–214, 2011.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. Wiley, Hoboken, NJ, 1992.

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *JRSS B*, 58:267–288, 1996.

S. van der Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On Asypmtotically Optimal Confidence Regions and Tests for High-Dimensional Models. *Annals of Statistics*, 42(3):1166–1202, 2014.

A. Wald. Note on the Consistency of the Maximum Likelihood Estimate. *Annals of Statistics*, 20:595–601, 1949.

L. Weiss. Asymptotic Properties of Maximum Likelihood Estimators in Some Nonstandard Cases. *Journal of the American Statistical Association*, 66: 345–350, 1971.

L. Weiss. Asymptotic Properties of Maximum Likelihood Estimators in Some Nonstandard Cases, II . *Journal of the American Statistical Association*, 68 (342):428–430, 1973.