August Arnstad

Relative variable importance in Bayesian linear mixed models

TMA4500 Project thesis in Industrial Mathematics
Supervisor: Stefanie Muff
December 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU

ABSTRACT

Statistical inference about the parameters of a model is a fundamental part of statistical analysis. In many applications it is desirable to understand how important each covariate is in explaining the observed response variable. By specifying the importance of each variable, researchers gain a better understanding of the complex relationships in the statistical model, thereby improving their knowledge and quality of research. Currently, there are several methods that try to deduce the importance of variables in a statistical model. These methods include testing a null hypothesis by using parameter estimates, confidence intervals or p-values This thesis aims to provide a method that decomposes the model variance in a Bayesian framework, to emphasize statistical inference rather than threshold-based interpretations, for example regarding a covariate as significant if $p < 0.05$.

We introduce a Bayesian measure of variable importance, which we call the *Bayesian Variable Importance* (BVI) method. The BVI method adapts the key aspects of established methods for decomposing the variance of a random intercept model, into the Bayesian framework. To do this, we use the relative weights method as our starting point. The relative weights method, which can be considered an approximation of more computationally exhausting methods, projects correlated covariates into an uncorrelated space. In the respective space the Bayesian model is fit, and the results are back-transformed to the original covariate space. The results include posterior distributions for each model covariate and the calculated variance that each covariate contributes to the full model variance. Furthermore, the package `BayesianImportance` is developed to implement the Bayesian Variable Importance method in the statistical software R.

From a simulation study it is shown that the Bayesian Variable Importance method provides a computationally fast and proper decomposition of the model variance. Analysis on a single, simulated dataset is shown, highlighting the statistical inference one can obtain from the BVI method. The results, such as posterior distributions, are discussed in further detail and further work is outlined. It is discussed that there are many possibilities with the BVI method going forward and that relative importance in a Bayesian framework is an area where much work can be done in the future.

PREFACE

This thesis, written during the fall semester 2023, is the finalization of the course TMA4500 Specialization Project. The project constitutes 15 ECTS and focuses on introducing a Bayesian variable importance measure. During the spring semester 2024, I will further extend the work presented in this thesis through a 30 ECTS master's thesis.

In the process of creating this thesis, some tools that use artificial intelligence have been used. Mainly, I have used OpenAI's large language model ChatGPT to help me solve problems I encountered when programming, get input on the structure and flow of my text and some help with grammar and spelling. Most importantly, I wish to stress that all text and findings in this thesis are my own work, and that I have not used artificial intelligence tools to generate any content.

I want to thank my supervisor Stefanie Muff for her excellent guidance and support throughout the project. I look forward to continuing our work during the spring of 2024.

CONTENTS

Abstract i

Preface ii

List of Figures iv

List of Tables vi

1 Introduction 2

2 Theory 6
2.1 Linear regression models 6
2.1.1 Multiple linear regression 6
2.1.2 Linear mixed models (LMM's) 7
2.2 Relative variable importance in linear regression 7
2.2.1 Correlation among covariates in linear regression 8
2.2.2 Relative importance measures 8
2.2.3 Naive decompositions 10
2.2.4 LMG - A proper decomposition 10
2.2.5 Relative weights for linear regression 12
2.2.6 R^2 for LMM's 13
2.2.7 Extensions of the LMG and relative weights method 13
2.3 The Bayesian framework 14
2.3.1 General idea 14
2.3.2 Bayesian LMMs 15
2.3.3 Appropriate definition of R^2 for the Bayesian framework . . 15
2.3.4 R^2 for Bayesian linear regression 15
2.3.5 R^2 for Bayesian LMM's 16
2.4 The INLA framework 16

3 Methods 19
3.1 Variable importance in a Bayesian framework 19
3.1.1 Relative variable importance calculations 19
3.1.2 Handling the model fit with INLA 20
3.2 Simulation study 21

4 Results **23**
4.1 Insights from the simulation study 23
4.1.1 Relative importance of the fixed effects 23
4.1.2 Relative importance of the random effects 25
4.1.3 Total variance explained - R^2 estimates 26
4.2 The BVI method 28
4.2.1 Posterior relative importance distributions 28
4.2.2 Posterior R^2 distributions 29

5 Discussion & Further work **31**

6 Conclusions **35**

References **36**

Appendices **38**

A GitHub repository **39**

B Bayesian Variable Importance usage **40**

LIST OF FIGURES

1 Violin plots for the relative importance of the fixed effects X_1, X_2 and X_3 for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG, ERW and the Relaimpo methods. The standardized regressor coefficients are $\boldsymbol{\beta} = \left(\sqrt{1/8}, \sqrt{2/8}, \sqrt{3/8}\right)$, and the true total model variance is $\sigma^2_{\mathbf{y}} = 1$. For the BVI method the distributions of posterior means are shown to compare to the distribution of point estimates from the other three methods. The horizontal line displays the theoretically correct importance of each fixed effect in the case of uncorrelated data.
(a) Relative importance of X_1 as calculated from the four methods. 24
1 (b) Relative importance of X_2 as calculated from the four methods. 25
1 (c) Relative importance of X_3 as calculated from the four methods. 25
2 Violin plots for the relative importance of the random effect $\boldsymbol{\alpha}$, that is, $\hat{\sigma}^2_\alpha$ for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG and the ERW method. For the BVI method the distributions of posterior means of the marginal distribution of $\hat{\sigma}^2_\alpha$ are shown to compare to the point estimates of the other two methods. The horizontal line displays the theoretically correct importance $\sigma^2_\alpha = 0.125$ of the random effect in the case of uncorrelated data. 26
3 Violin plots for the total marginal and conditional variance explained for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG, the ERW and the Relaimpo method(only marginal variance explained can be computed). For the BVI method the posterior means of the sampled posterior distributions of $\boldsymbol{\beta}$ and the marginal distribution of $\hat{\sigma}^2_\alpha$ in each simulation are used to compare to the point estimates of the other two methods. The horizontal lines display the theoretical explained variance for each correlation level ρ as in Table 1. 27

4 Posterior distributions for fixed effects and posterior marginal distributions for random effects from the BVI method on four randomly simulated datasets with different correlation values between the fixed effects. The blue and purple densities are the marginal posteriors for $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\varepsilon$ respectively, whereas the red, yellow and green densities are the sampled posteriors for X_1, X_2 and X_3 respectively. The vertical lines in (a) represent the theoretically correct relative importances. 29

5 Posterior R^2 distribution calculated by the BVI method from the posterior distributions of fixed and random effects on four randomly simulated datasets with different correlation values ρ between the fixed effects. The red, yellow, green and blue distributions correspond to the posterior distribution of R^2 for $\rho = 0, \rho = 0.1, \rho = 0.5$ and $\rho = 0.9$ respectively. The theoretical values can be found in Table 1 and are shown as vertical lines. 30

LIST OF TABLES

1 The theoretically correct marginal variance explained (left column) and conditional variance explained (right column) for different correlation levels between the fixed effects. 22

CHAPTER

ONE

INTRODUCTION

Statistical models that aim to model a response from a set of covariates can be very useful in various fields, if interpreted properly. To be able to interpret the models correctly, it is often of interest to decide on what covariates that provide statistical evidence (Muff et al. 2022) for the response. Determining statistical evidence and quantifying the proportion of variance a covariate explains in the response is no trivial task, which has been and still is a wide topic applicable to many sciences. The term *statistical evidence* is suggested as an alternative to the more popular term *statistical significance* based on the pitfalls that the term statistical significance impose.

Arguably the historically most prominent way of deciding whether or not a covariate is statistically significant is the p-value, which was made widespread by Ronald Fisher (Fisher 1925). The p-value is calculated by performing a hypothesis test on the covariates regression coefficient, which tests the null hypothesis that the regression coefficient is zero against the alternate hypothesis that the regression coefficient is different from zero. The result of the hypothesis test is determined by comparing the p-value to a significance level α, frequently set to be $\alpha = 0.05$. If the p-value is lower than α one rejects the null hypothesis and claims that the regression coefficient is not zero and therefore the covariate is statistically significant in explaining the response, and vice versa if the p-value is larger than α. Following Goodman (2008), the p-value is defined as $\mathbb{P}(X \geq x|H_0)$, where H_0 is the null hypothesis, X is some statistic of the random variable, e.g. mean or variance, and is itself a random variable while x is the observed value of this statistic from the data. This definition can be written in words as *The probability of the observed, or a more extreme, result, if the null hypothesis were true* (Goodman 2008).

Although the p-value is such a popular tool for determining hypothesis tests, this method of determining statistical significance is subject to much criticism. Some of the criticism addresses that the p-value is often highly misinterpreted, as in Goodman (2008) where twelve misconceptions regarding the p-value is discussed. Another line of criticism points to the rigid way of deciding statistical significance. Once a significance level α is chosen, statistical significance is determined in a binary way, based on α. If the p-value is calculated to be smaller than α we say that the covariate is statistically significant and vice versa. This way of concluding can

lead to inflation bias, also known as p-hacking (Head et al. 2015), which is the phenomenon of researchers testing "several statistical analyses and/or data eligibility specifications"(Head et al. 2015), that give significant results based of the p-value and report only these results. In this sense, one can p-hack and eventually obtain significance with respect to p-values for almost any covariate, significant or not (Simmons et al. 2011).

Despite the abundant criticism of determining significance based solely on p-values, we do not dismiss the valuable information that a p-value contains. Rather, the focus should be on supplementing the p-value with methods that provide more insight into how well the covariates explain the response. Instead of determining if a covariate provides evidence or no evidence, it can be preferred to determine the proportion of variance that a covariate explains in the response. By not deciding on the binary outcome evidence or no evidence, and instead referring to predictable variance (Johnson 2000) from the covariates, the interpretations become less rigid and less prone to misinterpretations. This proportion of variance is commonly called the *relative importance* of the covariate and some commonly used tools to quantify the relative importances of covariates are

i **Effect sizes:** An intuitive approach is to look at the size of the regression coefficients, more specifically the squared value of the standardized regression coefficients. This is useful to get a comparable measure between covariates if they are independent, but if the covariates are correlated the uncertainty in coefficient estimates are heavily affected. Therefore, the results become hard to trust and correct interpretations are difficult to obtain.

ii **Confidence intervals:** Given that we have estimated the effect size of a covariate, it would be reasonable to create a confidence interval for the estimate. The confidence interval would give us a range of effect sizes that can be evaluated as statistically consistent with the data. The problem with confidence intervals is that the standard procedure for calculating them has its foundation in the p-value, so relying on a confidence interval would effectively be the same as relying on a p-value.

iii **Information criteria:** A popular variable selection tool in regression models are information criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978), that can even be used to compare non-nested models. Both the AIC and BIC provide a natural measure for the unique information contained in one covariate, but that does not account for information shared between correlated covariates, making it hard to interpret the result in this case.

iv **Methods decomposing the R^2:** The R^2 value is a measure of the variance in the response explained by the statistical model. It is a very popular and intuitive measure of model fit, so if one can find a reasonable decomposition of the R^2 in terms how much the covariates contribute to it, this would be easy for the many researches already familiar with R^2 to interpret. The decomposition of the R^2 value could be done in several ways, for example by looking at the R^2 of the model containing only the covariate of interest, or the difference in R^2 when adding the covariate of interest to a model with multiple covariates. Once

more correlated covariates can pose challenges when interpreting the results, so a decomposition of the R^2 must properly account for correlated covariates.

All the above methods for determining the importance of covariates to a response can be misleading if correlation between covariates is present. As a consequence, a measure of relative covariate importance that can accurately address the difficulty of correlated covariates and at the same time being intuitive enough to become widespread, is needed.

The topic of relative variance importance has been subject to a lot of academic work from different perspectives (Grömping 2007, Johnson 2000). One line of thinking is proposed by Lindeman et al. (1980) which focuses on decomposing the R^2 by assessing models containing different subsets of the covariates. The method, commonly named LMG after the authors, considers all possible orderings of how covariates are added to the null model and then finds the mean increase of R^2 for each covariate. Averaging over orderings is common practice in statistics to reduce variance, so the LMG has established itself as a robust method. Multiple extensions of the LMG has been proposed, including into fields as dominance analysis (Budescu 1993) and game theory (Lipovetsky & Conklin 2001) where it is deduced the LMG method and the Shapley value (Shapley 1953) are equivalent. The downside to averaging over orderings is the computational complexity imposed when the orderings are many, since this requires a great number of permutations.

To address the computational complexity of the LMG method, another approach is to project correlated covariates in to an orthogonal space, use the effect size of covariates in this space and then transform them back onto the correlated covariates. The projection of the original covariates involves well known matrix approximation techniques and in the orthogonal space the problem reduces as the correlation is no longer present. This approach has been proposed, and improved, independently by multiple scientists (Johnson 1966, Fabbris 1980, Genizi 1993, Johnson 2000), and will be called the relative weights (RW) method going forward. The relative weights method can be seen as an approximation of the LMG method which dramatically reduces the complexity, and "might be the method of choice"(Grömping 2015) if the LMG method is not computationally feasible.

A feature that is absent in both the LMG and the relative weights method, is the ability to handle random effects and non-normal responses. Very often, a linear regression model is not sufficient to explain all the information researchers have at their disposal. The data gathered by researchers often contain natural groupings, which can be implemented with a linear mixed model (LMM) as random intercepts or random slopes. Further, the response might be better modelled using a non-normal distribution which can be done by using a generalized linear model (GLM) instead of the standard linear regression. So far an effort to extend the LMG and the RW method to handle random intercepts has been done (Matre 2022) and this effort provides a useful extension. Developing methods capable of handling random slopes and non-normal responses are therefore strongly desirable in the future, as this would help researchers interpret a greater scope of statistical models.

Moreover, both the LMG and the relative weights method are methods for calculating relative importance in a frequentist framework. However, in recent years the advancements made in computational techniques, such as INLA, has lead to an increased interest for the unique possibilities within the Bayesian framework as described in McElreath (2020). The Bayesian framework deals with distributions rather than point estimates and therefore presents a fundamental variance, or uncertainty, in these distributions. This allows researchers to interpret findings with respect to a distribution with some variance, rather than a single point estimate. Consequently, Bayesian methods are increasingly popular (Hackenberger 2019) and can provide attractive alternatives for improving clinical trials (Lee & Chu 2012). For these reasons, we argue that a robust and trustworthy relative importance measure in the Bayesian framework is desirable in the same way as its frequentist counterpart. Furthermore, it would be strongly desirable with a Bayesian analogy to the LMG and RW method that is suited to properly address random effects and non-normal responses.

This thesis will consider the resources outlined above as well as extensions that have been proposed (Matre 2022), and with this basis try to put forward an analogous relative importance measure in the Bayesian framework. For the time being this relative importance measure will be compatible with random intercepts and can hopefully be further extended in the future. Mainly focusing on the relative weights method, we will combine this method with the integrated nested Laplace approximation(INLA) technique. This will allow us to interpret the results from a Bayesian perspective and see how it compares to more established methods from the frequentist framework. The implementation of this Bayesian relative importance analogy will be done in R, where an R package will be developed. Our hopes are that this R package, along with instructions on its usage, will be easy for researchers to use and give insightful information into their research. This R package can be found at `https://github.com/AugustArnstad/BayesianImportance`.

To begin, Chapter 2 will present necessary theory regarding the regression models considered and relative importance measures. In Chapter 3 the methodology for how we develop our importance measure and conduct a simulation study is put forth, and the accompanying results are given in Chapter 4. The empirical results and general considerations are discussed in Chapter 5, where also further work is outlined, and the following conclusion is found in Chapter 6. Lastly, Appendix A and Appendix B contain the GitHub repository in which the R package was built and an example of its usage respectively.

CHAPTER

TWO

THEORY

2.1 Linear regression models

In the following section we will follow the derivations and results of Fahrmeir et al. (2013) and McCullagh & Nelder (1989), however the book might have used scalar notation.

2.1.1 Multiple linear regression

Linear regression aims to model the relationship between a response $y_i, i \in \{1, 2, ..., n\}$ and covariates $\mathbf{x}_i$, where $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$, by estimating the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$. The linear regression can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{2.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is the random error term. It is assumed that the response is independent, and the error term is independent and identically distributed(iid) following a normal distribution with mean zero and variance σ^2, *i.e.* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. By inspection one can state that, given the covariates $\mathbf{X}$,

$$\begin{aligned} \mathbb{E}(\mathbf{y}|\mathbf{X}) = \mathbb{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta} \ , \\ \mathrm{Var}(\mathbf{y}|\mathbf{X}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}) + \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \ , \end{aligned} \tag{2.2}$$

where $\mathbf{I} \in \mathbb{R}^{n \times n}$ is the identity matrix. It is then straightforward to see that $\mathbf{y}$ follows the conditional normal distribution

$$\mathbf{y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{2.3}$$

and from this distribution of $\mathbf{y}$ it can be shown that the likelihood function is expressed as

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto \sigma^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{2.4}$$

and that the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, which maximizes the likelihood, is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \ . \tag{2.5}$$

2.1.2 Linear mixed models (LMM's)

Data often comes in clustered form, for example due to repeated measurements of the covariate over time. Clustered data violate with the assumption of independent responses in linear regression and must be properly accounted for. One solution to this is to introduce random effects that are cluster specific, but independent of the fixed effects and the other clusters. Let the population contain m underlying clusters, with n_j , $j = 1, ..., m$ observations in each cluster, so that $\mathbf{y} \in \mathbb{R}^{(N \times 1)}$ where $N = \sum_{j=1}^{m} n_j$. Assume that we investigate q random effects, including a random intercept and $q - 1$ random slopes, such that the random effects vector can be written as

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m)^T \ , \tag{2.6}$$

where each $\boldsymbol{\alpha}_j \in \mathbb{R}^{q \times 1}$ is assumed independent and represents the random effects for cluster j and has length q. For a cluster j the vector $\boldsymbol{\alpha}_j \sim \mathcal{N}_q(\mathbf{0}, \mathbf{Q})$ where $\mathbf{Q}$ is the $q \times q$ unknown covariance for the random effects, assumed to be positive definite. If the random effects for each cluster are independent of each other, the covariance matrix $\mathbf{Q} = \text{diag}(\tau_0^2, ..., \tau_q^2)$. The linear mixed model now takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{2.7}$$

where $\mathbf{X} \in \mathbb{R}^{N \times p}$ is the design matrix for the fixed effects, $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ are the regression coefficients for the fixed effects, $\mathbf{U} = \text{diag}(\mathbf{U_j}) \ , \in \mathbb{R}^{N \times q}$ is the design matrix for the random effects and $\mathbf{U}_j \in \mathbb{R}^{n_j \times q}$ is the design matrix for cluster j. Since $\boldsymbol{\alpha}$ is a random variable, the parameter to estimate is the variance of each random effect $\mathbf{Q}_{kk} = \tau_k^2$ and their covariance $\mathbf{Q}_{k,l} = \tau_{k,l}$, where $k, l = 1, ..., q$. In this model the independence between clusters are conserved for the response as a whole, but it expresses the correlation that observations of the same cluster have through the random effects. As for the simple linear regression it is assumed that $\mathbf{X}\boldsymbol{\beta}$ is fixed, and that $\mathbf{U}$ is given, so they do not contribute to the model's variance. Therefore, the conditional expectation $\mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \mathbf{X}\boldsymbol{\beta}$ is easily obtained, and the conditional variance can be calculated as

$$\text{Var}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) = \mathbf{U}\text{Var}(\boldsymbol{\alpha})\mathbf{U}^T + \sigma^2\mathbf{I} = \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I} \ , \tag{2.8}$$

where $\mathbf{I} \in \mathbb{R}^{N \times N}$ and $\mathbf{G} \in \mathbb{R}^{mq \times mq}$ is the block diagonal covariance matrix of the random effects, with $\mathbf{Q}_j$ along the diagonal for $j = 1, ..., m$. As we assume that the random effects are independent of the fixed effects, and that the random error term is iid for each observation, the conditional distribution of $\mathbf{y}$ follows that of a sum of independent normal distributions, *i.e.*

$$\mathbf{y}|\mathbf{X}, \mathbf{U} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I}) \ . \tag{2.9}$$

2.2 Relative variable importance in linear regression

In a regression setting with multiple regression coefficients, it is often desirable to be able to assign each coefficient with a measure of its relative importance to the model. The relative importance of predictor X_i is defined in other words as

the contribution to explained variance from X_i. Assigning relative importance is no trivial task, as correlation among covariates poses a challenge in assessing the relative importance of each covariate.

2.2.1 Correlation among covariates in linear regression

Correlation among covariates is to be expected, as it is natural in many scenarios. However, if the correlation is very strong, this poses some serious problems when interpreting the linear regression model. The covariates $\mathbf{x}_i$ in a linear regression are assumed to be linearly independent, so that the design matrix $\mathbf{X}$ has full rank. If the design matrix is not of full rank, that is one or more covariates are perfectly correlated, the model (2.1) is said to be *multicollinear* (Poole & O'Farrell 1971). From equation (2.5) one can see that if the matrix $\mathbf{X}$ is not of full rank, the term $(\mathbf{X}^T\mathbf{X})^{-1}$ is not invertible and the MLE of $\boldsymbol{\beta}$ does not exist. Further, the variance of the MLE of $\boldsymbol{\beta}$ grows as the correlation between covariates grows (Fahrmeir et al. 2013, p. 116). A larger variance in $\hat{\boldsymbol{\beta}}$ also leads to larger standard errors and larger p-values for $\hat{\boldsymbol{\beta}}$, making it hard to assess the model. Both coefficients and covariates affect the total marginal model variance, which can be decomposed as

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta}) + \text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta} + \sigma_\varepsilon^2 = \sum_{j=1}^{p} \beta_j^2 v_j + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_j\beta_k\sqrt{v_j v_k}\rho_{jk} + \sigma_\varepsilon^2 \,, \tag{2.10}$$

(Grömping 2007) where $\mathbf{V} = \text{Cov}(\mathbf{X})$ is the $p \times p$ covariance matrix of the covariates which is assumed to be positive definite, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, v_j the regressor variances for $j = 1, ..., p$ found along the diagonal of $\mathbf{V}$ and ρ_{jk} the inter-regressor correlations between regressor j and k. The middle term in 2.10 consist of the covariance between the covariates and this term makes it hard to assess the relative importance of each covariate. To assign each covariate with an importance, we need to consider relative importance measures that can handle the correlation among covariates.

2.2.2 Relative importance measures

In Grömping (2007) two importance measures are advocated, namely LMG and Proportional marginal variance decomposition (PVMD). Before we describe the LMG, and another measure, relative weights, proposed by Johnson (2000), we will first consider general aspects of decomposing the variance of a linear model. All the methods to be discussed analyze how the regressors compete to compose the models R^2 value. The R^2 is a very popular measure of how much of the variance in the response variable is explained by the model, since it is both intuitive and easy to interpret. In a frequentist framework, the R^2 is defined as

$$R^2 = 1 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{(\mathbf{y} - \bar{\mathbf{y}})^T(\mathbf{y} - \bar{\mathbf{y}})} \,, \tag{2.11}$$

where $\bar{\mathbf{y}}$ is the mean vector of $\mathbf{y}$. Instead of referring to the R^2 value alone, going forward this thesis will focus on decomposing the R^2 of the linear regression model.

This decomposition is done in order to assess the relative importance, or variance explained, of each covariate in the model. In the case of uncorrelated covariates,

$$\mathrm{Var}(\mathbf{y}) = \sum_{j=1}^{p} \beta_j^2 v_j + \sigma_\varepsilon^2 \ , \tag{2.12}$$

and the R^2 is therefore consistent with the total variance of the response variable (Grömping 2007). This consistency provides a natural decomposition of the R^2 in terms of contribution from each covariate, as each predictor X_i contributes $\beta_i^2 v_i$ to the total response variance. A naive decomposition of covariate importance is then starting with the empty model with $R^2 = 0$, and adding one covariate at a time. The increase in R^2 is then the importance of the covariate added.

Due to the popularity of R^2, it is desirable to also be able to decompose the R^2 in the case of correlated covariates, such that it is consistent with the variance of the response. In (2.10) the response variance is split into three parts, the first two sums which comes from the regressors and the latter term which is the variance of the error. It is the middle term that poses the problem of assigning importance to each covariate, since it contains the covariance between the covariates. The literature has established some conditions that relative importance measures should fulfill, so that they can be interpreted and compared in a sensible manner (Grömping 2007). As listed in Grömping (2007), the methods should have

1. **Proper decomposition**: The model variance should be decomposed into shares for each regressor that sum up to the total variance, and the method shall allocate the shares to each regressor.

2. **Non-negativity**: Each share of the variance should be non-negative.

3. **Exclusion**: If a regressor is excluded from the model, $\beta_j = 0$, its share of the variance should be zero.

4. **Inclusion**: If a regressor is included in the model, $\beta_j \neq 0$, its share of the variance should be positive.

Before moving on the such methods, some notation in accordance with Grömping (2007) will be introduced, namely

$$\mathrm{evar}(S) = \mathrm{Var}(Y) - \mathrm{Var}(Y \ | X_j, j \in S) \tag{2.13}$$

and

$$\mathrm{svar}(M|S) = \mathrm{evar}(M \cup S) - \mathrm{evar}(S) \ , \tag{2.14}$$

where S is a subset of regressors, $\mathrm{Var}(Y \ | X_j, j \in S)$ denotes the variance of Y conditioned on $X_j, j \in S$ being fixed, $\mathrm{evar}(S)$ is the explained variance of the regressors in S and $\mathrm{svar}(M|S)$ is the gain in variance explained by adding regressors from the subset M to the model that already contains the regressors S.

2.2.3 Naive decompositions

To make it clear that some simple decompositions fail the conditions of relative importance measures, we will consider two naive approaches for decomposing the R^2. We denote the R^2 of a linear regression with regressors $X_1, \ldots, X_p$ as $R^2(\{1, \ldots, p\})$ and the relative importance of regressor X_i as $\mathrm{RI}(\{i\})$

The first naive method is to fit a model with all regressors p, and then fit a model with all regressors excluding regressor i. The relative importance of X_i is then the difference $R^2(\{1, \ldots, p\}) - R^2(\{1, \ldots, p\} \setminus i)$. To show how this fails the conditions of relative importance measures, an example from Matre (2022) is discussed. The example considers the simple case

$$Y = X_1 + X_2 \ , \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1 \ , \mathrm{Cov}(X_1, X_2) = 0.9 \ . \tag{2.15}$$

The R^2 of the model with both covariates is $R^2(\{1, 2\}) = 1$, since the covariates X_1, X_2 explain fully the response Y. Then one would expect that the importance of X_1 and X_2 is 0.5 each, since they both explain half of the response variance. Using the proposed decomposition, one would calculate

$$\mathrm{Ri}(\{2\}) = R^2(\{1, 2\}) - R^2(\{1\}) = 1 - \frac{\mathrm{Cov}(Y, X_1)^2}{\mathrm{Var}(Y)\mathrm{Var}(X_1)} = 1 - \frac{1.9^2}{3.8} \approx 0.05 \ , \tag{2.16}$$

where it is used that for the simple linear regression, the R^2 is given by the squared correlation coefficient between the response and the regressor. By symmetry $\mathrm{Ri}(\{1\}) = \mathrm{Ri}(\{2\})$, so the sum of the relative importances is 0.1. However, the total explained variance of the model is 1, so this decomposition violates the proper decomposition condition. This decomposition only assign importances to the regressor based on the information that the regressor does not share with any other regressors. Therefore, it does not take into account the shared information and the importance estimated is too low.

Another naive decomposition would be to compare the relative importance of a model with one regressor i to the empty model, *i.e.* the model with no covariates. The empty model has an $R^2 = 0$ and therefore for X_1 in the above example we would have

$$\mathrm{Ri}(\{1\}) = R^2(\{1\}) - R^2(\{\emptyset\}) = \frac{\mathrm{Cov}(Y, X_1)^2}{\mathrm{Var}(Y)\mathrm{Var}(X_1)} = \frac{1.9^2}{3.8} \approx 0.95 \ . \tag{2.17}$$

Once more by symmetry we have $\mathrm{Ri}(\{2\}) = \mathrm{Ri}(1)$, so the sum of the relative importances is 1.9, violating the proper decomposition condition. Conversely to the first naive approach, this decomposition assigns importances based on the full information contained in the regressor. Therefore it overestimates the importance of each variable, since the shared information is accounted for twice.

As we have seen from these naive approaches, the task of decomposing the R^2 value is far from trivial, and calls for more sophisticated methods.

2.2.4 LMG - A proper decomposition

A method that handles correlation among covariates, and is frequently reinvented(Grömping 2007) from different approaches, is the LMG method. Therefore we shall discuss

it, as it serves an important role as a leading method for assigning relative variable importance. The LMG method takes use of averaging over orders, meaning that it permutes the index set $\{1, ..., p\}$ of the regressors $(p-1)!$ times, excluding the intercept, and sequentially adds the regressors to the model for each permuted index set. By adding regressors sequentially for each permutation, one can investigate how the importance of the regressors vary depending on what other regressors are included, which is useful when they are correlated. This is justified by the assumption that there is no relevant ordering of the regressors in the index set (Kruskal 1987). For each regressor added, starting with none, it allocates a share of explained variance, or importance, and then adds a new regressor. The final allocated share to the regressor is the average of the allocated shares to that regressor for all permutations of the set of regressors indices. This would mean that for two correlated regressors whose importance share varies depending on which is added first, would receive an averaged importance. Averaging over orders is a statistical tradition (Kruskal 1987) and gives a robust assessment of each regressor's importance by considering different orderings of how they are added to the model. The iterative process for the regressors $\{X_0, X_1, X_2, X_3\}$, where X_0 is the intercept, would be

1. Considering $\{X_1, X_2, X_3\}$, X_1 is added to the model, and the share of explained variance allocated to X_1 is $\text{svar}(\{1\}|\emptyset)$. X_2 is added and allocated a share of $\text{svar}(\{2\}|\{1\})$, and lastly X_3 is added and allocated a share of $\text{svar}(\{3\}|\{1, 2\})$.

2. Considering $\{X_1, X_3, X_2\}$, X_1 is added to the model, and the share of explained variance allocated to X_1 is $\text{svar}(\{1\}|\emptyset)$. X_3 is added and allocated a share of $\text{svar}(\{3\}|\{1\})$, and lastly X_2 is added and allocated a share of $\text{svar}(\{2\}|\{1, 3\})$.

The above iteration is repeated for all 6 possible permutations of orderings among regressors to obtain the final result. This iterative process gives rise to the general formula for share of explained variance allocated to X_1 by the LMG method with p regressors (Grömping 2007),

$$\text{LMG}(1) = \frac{1}{p!} \sum_{S \subseteq \{2,...,p\}} n(S)!(p - n(S) - 1)!\text{svar}(\{1\}|S) \ , \tag{2.18}$$

where $n(S)$ is the number of regressors in S. Equation (2.18) averages the increase in R^2, $\text{svar}(\{X_i\})$, when adding the covariate of interest, X_i, over all possible orderings of covariates. This mean increase over orderings is assigned as the proportion of R^2 explained by X_i. The LMG method fulfills all but the exclusion criteria described previously (Grömping 2007), but Grömping (2007) argues that this "must be seen as a natural result of model uncertainty" and therefore that this criterion is not indispensable. Therefore, we find it also suitable for our purposes to focus on the three other criteria. The setback of the LMG method is naturally the great computational expense that the permutations require, namely 2^{p-1} summations (Grömping 2007).

2.2.5 Relative weights for linear regression

A method that takes advantage of the straightforward decomposition of the variance when the covariates are uncorrelated is the relative weights method (Johnson 2000), which will now be discussed.

The relative weights method proposes an alternative to the LMG, which is significantly less computationally expensive. Intuitively, the relative weights method projects the matrix $\mathbf{X}$ into an orthogonal column space, resulting in a matrix $\mathbf{Z}$ with orthogonal columns. The matrix $\mathbf{Z}$ is then an approximation of $\mathbf{X}$ and will be used as the design matrix in the regression. Since the columns of the design matrix $\mathbf{Z}$ are orthogonal, each covariate is uncorrelated. This allows us to decompose the variance in the straightforward manner as mentioned earlier.

In relative weights one uses the singular value decomposition (Nimon & Oswald 2013), to project the real-valued design matrix $\mathbf{X}$ into an orthonormal matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$ containing the eigenvectors of $\mathbf{X}\mathbf{X}^T$, an $n \times p$ diagonal matrix $\mathbf{D}$ containing the singular values of $\mathbf{X}$ and another orthonormal matrix $\mathbf{V} \in \mathbb{R}^{p\times p}$ containing the eigenvectors of $\mathbf{X}^T\mathbf{X}$ such that

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}} \ . \tag{2.19}$$

From the Eckhart-Young-Mirsky theorem (Mirsky 1960) and following the derivations of Johnson (1966), one can state that the matrix $\mathbf{X}$, of rank r, can be approximated by a matrix $\mathbf{Z} = \mathbf{U}\mathbf{V}^T$ of rank $k \leq r$ such that the difference under the squared Frobenius norm

$$\|\mathbf{X} - \mathbf{Z}\|_F^2 = tr\left((\mathbf{X} - \mathbf{Z})^T(\mathbf{X} - \mathbf{Z})\right) \ , \tag{2.20}$$

is minimized. The relative weights approximation now utilizes the matrix (Johnson 2000) $\frac{1}{\sqrt{n-1}}\mathbf{Z}$, where the factor $\frac{1}{\sqrt{n-1}}$ is the standardization factor for $\mathbf{Z}$ (Matre 2022), and regresses on $\mathbf{Z}$ to find the MLE $\boldsymbol{\beta}_{\boldsymbol{Z}}$ as

$$\begin{aligned} \boldsymbol{\beta}_{\boldsymbol{Z}} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}\mathbf{y} \\ &= \left((n-1)\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{V}^T\right)^{-1}\sqrt{n-1}\mathbf{V}\mathbf{U}^T\mathbf{y} \\ &= \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{U}^T\mathbf{y} \ . \end{aligned} \tag{2.21}$$

As $\mathbf{Z}$ is orthogonal, the relative importance for each column $\mathbf{z_i}$ with respect to the response $\mathbf{y}$ can be found as the square of $\beta_{Z,i}^2$, denoted as $\boldsymbol{\beta}_{\boldsymbol{Z}}{}^{[2]}$. The notation $\boldsymbol{\xi}^{[2]}$ for some $\boldsymbol{\xi}$ represents the Schur product of $\boldsymbol{\xi}$ with itself, *i.e.* element wise squaring of each element in $\boldsymbol{\xi}$. Once these importances are obtained, Johnson (2000) argues that we should regress $\mathbf{X}$ on $\mathbf{Z}$ to obtain the weights that relate the importance of each column of $\mathbf{Z}$ to each column of $\mathbf{X}$. These weights can be calculated as the matrix

$$\boldsymbol{\Lambda} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X} = (\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T \ , \tag{2.22}$$

and since $\mathbf{Z}$ is orthogonal, the contribution from a column of $\mathbf{z_i}$ with respect to a column $\mathbf{x}_j$ is the squared entry $\boldsymbol{\Lambda}_{ij}^2$. The contribution from a column $\mathbf{x}_j$ with

respect to the response $\mathbf{y}$, *i.e.* the relative importance, is then estimated as the matrix product (Johnson 2000)

$$\text{RI}(\mathbf{X}) = \boldsymbol{\Lambda}^{[2]}\boldsymbol{\beta}_{\boldsymbol{Z}}{}^{[2]} \ , \tag{2.23}$$

with RI as a column vector where each entry j contains the estimate of the relative importance corresponding to column j of $\mathbf{X}$. In Matre (2022, section 2.5.3) it is shown that the relative weights method fulfills the criteria same three criteria as the LMG method, because $\mathbf{Z}$ and $\mathbf{X}$ are linear combinations of each other and due to the properties of $\boldsymbol{\Lambda}$. The relative weights method will be considered later on, when we use the transformation of $\mathbf{X}$ to $\mathbf{Z}$ in a Bayesian setting.

2.2.6 R^2 for LMM's

The R^2 value introduced earlier refer to a linear regression setting, and so it is desirable to obtain an analogous value for random effect models. When extending a model to include random effects, one must decide whether one also wants to account for the variance explained by these random effects in the calculations of the R^2 value. A simple and intuitive way is presented in Nakagawa & Schielzeth (2013). To ease notation we write, $\sigma_f^2 = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}}\boldsymbol{\beta}$ for the variance captured by the fixed effects. Similarly we write σ_α^2 for the total variance of random effects and σ_ε^2 for the variance of the random error. The proposed definition of the marginal R^2 for LMM's is then

$$R^2_{\text{marg}} = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2} \ . \tag{2.24}$$

and similarly for the conditional R^2 we have

$$R^2_{\text{cond}} = \frac{\sigma_f^2 + \sigma_\alpha^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2} \ , \tag{2.25}$$

2.2.7 Extensions of the LMG and relative weights method

Extensions for both the LMG and the relative weights method have been proposed so that they can decompose the R^2 also for random intercept models (Matre 2022). The extended LMG, denoted as the ELMG, method uses the same permutations as described for the regular LMG, and now decomposes the R^2 value for the random intercept model instead. This R^2 is described in Section 2.2.6 and effectively divides the variance of the response into the variance of the fixed effects, the random effects and the random error. From this decomposition, the only extension needed for the LMG formula is to include also the random intercepts as model components, which gives (Matre 2022)

$$\text{LMG}(1) = \frac{1}{(p+q)!}\sum_{S\subseteq\ \{2,\ldots,p\}} n(S)!((p+q)-n(S)-1)!\text{svar}(\{1\}|S) \ , \tag{2.26}$$

where p denote fixed effects and q denotes random effects. It is equivalent to the original LMG method (2.18) except that here the random intercepts are treated as categorical fixed effects, where we do not consider the columns but rather the whole predictor, either completely in the model or not.

To create the extended relative weights (ERW) method, Matre (2022) uses the same transformation of data as for the relative weights method to project the covariates into an orthogonal space. Then the fixed effects are treated as one separate block, either in the model or not, and then uses the LMG approach to distribute a share of R^2 to each random intercept. The fixed effects will receive a joint share, which is distributed by using the relative weights method. Since the LMG approach is used for the q random effects and the block of fixed effects, the complexity of the ERW will follow that of the LMG method for $q+1$ covariates.

Both extensions described comes with new considerations (Matre 2022, for full details), for example that the inclusion criteria now should read "If a regressor $\beta_j \neq 0$, or a random intercept α with $\sigma^2(\alpha) > 0$, is included in the model then its share of the variance should be positive".

2.3 The Bayesian framework

The above discussion is rooted in the so called frequentist framework, implying that the parameters are treated as fixed and the uncertainty is quantified by the sampling distribution of the data. The Bayesian framework, on the other hand, treats the parameters as random variables and the uncertainty is quantified by the posterior distribution of the parameters.

2.3.1 General idea

The Bayesian framework is based on a generalization of Bayes theorem (Bayes & Price 1763, see Proposition 3) to functions, which states that

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \ , \tag{2.27}$$

where $\pi(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution of the parameters $\boldsymbol{\theta}$ given the data $\mathbf{y}$, $\pi(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\boldsymbol{\theta})$ is the prior distribution of the parameters and $\pi(\mathbf{y})$ is the marginal distribution of the data. These distributions give rise to many new perspectives and interpretations. Often one only considers the posterior in terms of being proportional to the product of the likelihood and prior, namely

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ . \tag{2.28}$$

The proportionality is useful because the marginal distribution of the data is often intractable, and thus the posterior is only known up to a constant. From well established sampling methods, such as Markov Chain Monte Carlo, this is enough to eventually be able to sample effectively from the posterior distribution. If a prior is proposed, and one can express the likelihood, the posterior can be computed and also updated as more data becomes available.

The fundamental perspective of distributions instead of point estimates are what that separates the Bayesian framework from a frequentist setting, and allows for different interpretations. In natural sciences, measurements are often performed

by professionals over a time period, and it is therefore useful to have a model that can adjust as more data becomes available. This is what the likelihood function allows for as it models the parameters as a function of the current data. Further, the prior allows for inclusion of some prior information about the parameters, which is often the case in natural sciences. These can be specified by experts in the field or by previous studies. Lastly, the fundamental uncertainty of the Bayesian framework is very useful. Since everything is modelled as a distribution, a corresponding variance is calculated. This variance can be a useful quantity for making statistical inference about the parameters one wishes to estimate. It also allows for capturing the fundamental uncertainty of measuring physical quantities.

2.3.2 Bayesian LMMs

When one wants to apply the idea of linear mixed models in the Bayesian framework some key aspects change, and we will follow the logic of Gelman et al. (2015) to explain this theory in our setting. Considering a model as in (2.7) we have four parameters, namely $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ_α^2 and σ_ε^2, where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are model parameters dependent on σ_α^2 and σ_ε^2 which are called hyperparameters. In a Bayesian framework these parameters are treated as random variables instead of values with a true, but unknown value, meaning that we must specify a distribution for the parameters. The posterior distribution of the model parameters will depend on the hyperparameters and the latent structure we assume the model to have. To define the prior distributions $\pi(\sigma_\alpha^2)$ and $\pi(\sigma_\varepsilon^2)$ of the hyperparameters, one assumes they are independent and chooses a distribution based on the prior information available. In this thesis we will use the Penalised Complexity priors, or PC priors (Simpson et al. 2017). If one assumes independence of the random effects and the fixed effects these priors will allow us, through methods discussed later in section 2.4, to derive marginal posterior distributions for the model parameters and sample from these. From these distributions we can obtain statistics such as posterior means and modes, posterior variances and credible intervals.

2.3.3 Appropriate definition of R^2 for the Bayesian framework

We wish to estimate relative importance in a Bayesian framework and report the distribution of R^2. To do this we must first consider how R^2 can be correctly defined and generalized in the Bayesian framework.

2.3.4 R^2 for Bayesian linear regression

When working in the Bayesian framework, the definition of R^2 is not as straightforward as in the classical framework. The classical definition of R^2 for linear regression is written as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\ , \tag{2.29}$$

where $\hat{y}_i$ is the predicted value of y_i and $\bar{y}$ is the mean of the observed values of y. However, if one was to compare models based on this metric in the Bayesian framework, the denominator would not be fixed. With a variable denominator one

cannot accurately interpret a change in R^2 value when comparing models. Gelman et al. (2017) proposed a definition of the R^2 for the Bayesian linear regression that will be considered in the following. Consider a draw s of the parameters $\boldsymbol{\beta}$ from the posterior distribution. Then, the proposed definition is

$$R_s^2 = \frac{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s + \sigma_s^2} \ , \tag{2.30}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}}$ is the covariance matrix of the design matrix $\mathbf{X}$ and σ_s^2 is the variance of the error term which can be sampled from the posterior distribution. Contrary to the classical definition this definition of R^2 contains only the estimated values from our model and not the observed values. The reasoning behind this is to carry this inherent uncertainty in the Bayesian framework by not using point estimates from the posterior mean, but rather averaging over a posterior distribution. Drawing enough samples from (2.30) one would eventually obtain also a distribution for the R^2 value.

2.3.5 R^2 for Bayesian LMM's

Since the Bayesian framework allows us to sample from the posterior distributions of both random and fixed effects, one can extend the conditional and marginal R^2 proposed by Gelman et al. (2017) to the LMM case. The respective generalization can be found directly as

$$R_{s,\text{marg}}^2 = \frac{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s + \sigma_{\alpha,s}^2 + \sigma_{\varepsilon,s}^2} \ , \tag{2.31}$$

and

$$R_{s,\text{cond}}^2 = \frac{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s + \sigma_{\alpha,s}^2}{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s + \sigma_{\alpha,s}^2 + \sigma_{\varepsilon,s}^2} \ , \tag{2.32}$$

where the subscript s denotes samples from the marginal posteriors of the parameters in question, *i.e.* $\boldsymbol{\beta}, \sigma_\alpha^2$ and σ_ε^2. These definitions of the R^2 highlight exactly the fundamental advantage of the Bayesian framework. Since the R^2 is also treated as a random variable, it has a distribution which can be used for statistical inference. Moreover, one can relate the R^2 more directly to the frequentist framework by using the posterior means or modes from the distributions of $\boldsymbol{\beta}, \sigma_\alpha^2$ and σ_ε^2 in (2.25) and (2.24). This approach can be favorable for comparing methods.

2.4 The INLA framework

As we have seen, the analytical posterior is possible to obtain for the Bayesian linear regression model. However, in the case of GLMMs, the posterior distribution is not in general analytically tractable (Fong et al. 2010). This calls for the use of numerical methods, such as Markov Chain Monte Carlo (MCMC) methods, to be able to sample from the posterior distribution. Such methods are computationally expensive, and require careful analysis to justify convergence and mixing of the Markov chains to the posterior distribution. Therefore it is desirable, under

certain conditions, to look at other methods that are more computationally efficient. In this thesis we will consider the alternative, namely the Integrated Nested Laplace Approximation (INLA) method (Gómez-Rubio 2020).

The INLA method is an alternative to the classical Marko Chain Monte Carlo methods, that has significant advantages at the cost of assuming a certain structure. In order to apply INLA, consider the vector of observations $\mathbf{y} = (y_1, ..., y_n)$, which may also contain missing values. Given an appropriate link function $g(\mu_i) = \eta_i$, we can model the observations as independent given the linear predictor

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i \ , \qquad i = 1, ..., n \ , \tag{2.33}$$

where α is the intercept, β_j are the regression coefficients for the covariates z_{ji}, $f^{(k)}$ are random effects for the vector of covariates $\{\mathbf{u_k}\}_{k=1}^{n_f}$ and ε_i is the error term. This gives rise to the key assumption that the INLA method needs in order to be applicable, namely that the latent field $\mathbf{x}$, denoted as

$$\mathbf{x} = (\eta_1, ..., \eta_n, \alpha, \beta_1, ..., \beta_n) \ , \tag{2.34}$$

is a Gaussian Markov Random Field (GMRF). Further, it is assumed that observations are independent given this latent field and the latent field is distributed according to some hyperparameters $\boldsymbol{\theta}$. The structure of the GMRF is given by a precision matrix $\mathbf{Q}(\theta)$, which is sparse and can be represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. This along with the assumed conditional independence makes computations very fast and is why INLA is effective. Now, the posterior distribution of the latent field $\boldsymbol{x}$ is given by

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ , \tag{2.35}$$

where $\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is the likelihood, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is the posterior of the latent field and $\pi(\boldsymbol{\theta})$ is the prior. Since it is assumed that observations are independent given the latent field, we can further express

$$\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}) \ , \tag{2.36}$$

where the index set $\mathcal{I} \subset \{1, 2, 3, \ldots, n\}$ only includes actual observed data. The INLA method now attempts to estimate the marginals of the latent effects and the hyperparameters. These marginals are given by

$$\pi(x_l|\mathbf{y}) = \int \pi(x_l|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ , \tag{2.37}$$

and

$$\pi(\theta_k|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-k} \ , \tag{2.38}$$

(Gómez-Rubio 2020) respectively, $\boldsymbol{\theta}_{-k}$ is the vector of hyperparameters excluding element θ_k and the latter integral is possible to integrate numerically due to the low dimension of $\boldsymbol{\theta}$ (Rue et al. 2009). The approximations of these integrals are

omitted, see Rue et al. (2009) for the full details. Lastly, the joint posterior distribution can be approximated from the so-called Skew Gaussian Copula class, as specified in Chiuchiolo et al. (2021), and allows for sampling from the joint distribution. The INLA method is implemented in the R-package `R-INLA` (Gómez-Rubio 2020) and is used in this thesis to fit the models and draw from the obtained posteriors. We note that for the random effects INLA outputs the precision for the parameters involved, which is defined as the inverse covariance matrix. For the posterior marginal distribution of variance for the random effects the package has a function for transforming the precision marginal to the variance marginal. The priors used for the models in this thesis follow the recommendations of penalizing priors by Simpson et al. (2017).

CHAPTER

THREE

METHODS

3.1 Variable importance in a Bayesian framework

We now propose a method for calculating relative variable importance in a Bayesian framework, which we call Bayesian Variable Importance (BVI). The BVI method is based on the idea that the relative weights approach, introduced in Section 2.2.5, combined with its extension (Matre 2022) can be used to fit a Bayesian LMM. With this basis we believe that it is possible to create a Bayesian relative importance measure, by taking the new framework into account. When considering the fixed effects there are multiple established methods to compare results with. Here, we compare our results to the previously discussed methods LMG (Grömping 2007), the extended LMG (ELMG) and the extended relative weights (ERW) (Matre 2022). The ELMG and the ERW are extensions of the LMG and relative weights methods respectively, to be compatible with the linear mixed models. This thesis will focus on linear mixed models containing only random intercepts. More general methods with random slopes and GLMM's with a link function not corresponding to a normal distribution are not in the scope of this thesis.

3.1.1 Relative variable importance calculations

First, we must incorporate the matrix transformation from the relative weights method for the Bayesian framework. The orthogonal matrix $\mathbf{Z}$ is generated from only the fixed effects $\mathbf{X}$, which corresponds to the variables for the fixed effects also in the Bayesian framework. Therefore, we can apply this transformation before the Bayesian analysis is performed. This transformation includes standardizing the data to be centered around zero and a standard deviation of one unit variance, as well as orthogonalizing the design matrix.

The variance decomposition of the random intercept model takes the form

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta}_{\mathbf{X}} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}_{\mathbf{X}}^T \text{Var}(\mathbf{X})\boldsymbol{\beta}_{\mathbf{X}} + \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma_\varepsilon^2 \mathbf{I} = 1 \ , \quad (3.1)$$

where $\mathbf{U}$ is now a sparse block matrix where each block $\mathbf{U}_j \in \mathbb{R}^{n_j \times q}$ is a matrix of ones, n_j the number of observations in cluster j, q the number of random intercepts and $\mathbf{G}$ is a blockdiagonal matrix with $\mathbf{Q}$ on the diagonal. In the special case of an independent and identically distributed (iid) random intercept model,

this variance decomposition can be simplified. With several independent random intercepts, each intercept $\boldsymbol{\alpha}_j$ contributes $\sigma^2_{\alpha_j}$ to the total model variance, making $\mathbf{G}$ a diagonal matrix, so the total variance of the model can be decomposed as

$$\text{Var}(\mathbf{y}) = \sum_{i=1}^{p} \beta^2_{i,\mathbf{X}} v_{i,\mathbf{x}_i} + 2\sum_{i=1}^{p}\sum_{k=i+1}^{p} \beta_{i,\mathbf{X}}\beta_{k,\mathbf{X}}\sqrt{v_{i,\mathbf{x}_i} v_{k,\mathbf{x}_k}}\rho_{ik} + \sum_{j=1}^{q} \sigma^2_{\alpha_j} + \sigma^2_\epsilon = 1 \ , \quad (3.2)$$

where $\beta_{i,\mathbf{X}}$ is the vector of regression coefficients when regressing $\mathbf{y}$ on $\mathbf{X}$, $v_{i,\mathbf{x}_i}$ is the variance of the i'th column $\mathbf{x}_i$ and ρ_{ik} is the correlation between the i'th and k'th columns. However, with the relative weights method, we approximate the design matrix $\mathbf{X}$ with the orthogonal matrix $\mathbf{Z}$. Consequently, the variance of each column in $\mathbf{Z}$, $v_{i,\mathbf{z}_i}$, is equal to one and the correlation ρ_{ik} between the columns is zero. Therefore, the total variance in (3.2) can be approximated by

$$\text{Var}(\mathbf{y}) \approx \sum_{i=1}^{p} \beta^2_{i,\mathbf{Z}} + \sum_{j=1}^{q} \sigma^2_{\alpha_j} + \sigma^2_\epsilon \ , \quad (3.3)$$

where $\beta_{i,\mathbf{Z}}$ is the vector of regression coefficients when regressing $\mathbf{y}$ on $\mathbf{Z}$, $\beta^2_{i,\mathbf{Z}}$ represents the variance of column $\mathbf{z}_i$, $\sigma^2_{\alpha_j}$ the variance of $\boldsymbol{\alpha}_j$ and σ^2_ε the variance of $\boldsymbol{\varepsilon}$. To map the importances of each column in $\mathbf{z}_i$ back to the importance of the columns $\mathbf{x}_i$, we regress $\mathbf{Z}$ on $\mathbf{X}$ to obtain the matrix $\boldsymbol{\Lambda}$ as in Section 2.2.5. The final model variance obtained from this setup, can then be estimated as

$$\text{Var}(\mathbf{y}) \approx \sum_{i=1}^{p} \text{RI}(\mathbf{x}_i) + \sum_{j=1}^{q} \sigma^2_{\alpha_j} + \sigma^2_\epsilon \approx \sum_{i=1}^{p} (\Lambda^{[2]}\beta^{[2]}_{\mathbf{Z}})_i + \sum_{j=1}^{q} \sigma^2_{\alpha_j} + \sigma^2_\epsilon \ , \quad (3.4)$$

where $\mathbf{x}_i$ is column i of $\mathbf{X}$. Since the response is standardized the relative importance, or proportion of variance explained, of regressor X_i is given by $\Lambda^{[2]}\beta^{[2]}_{\mathbf{Z}}$ and the relative importance of random intercept $\boldsymbol{\alpha}_j$ is given by $\sigma^2_{\alpha_j}$.

3.1.2 Handling the model fit with INLA

Once $\mathbf{Z}$ has been created from the projection of $\mathbf{X}$ into the orthogonal space, a model of the response, which has the form as explained in Section 2.1.2, is fit using $\mathbf{Z}$ as the design matrix and INLA. INLA is our preferred computational tool to fit the Bayesian LMM since it is very efficient, especially for large data sets and complex models. After the model is fit, INLA provides the approximate marginal posterior distribution of each component in $\boldsymbol{\beta}_{\mathbf{Z}}$, the precision of $\boldsymbol{\alpha}$, the precision of $\boldsymbol{\varepsilon}$ and allows us to sample from the joint posterior distribution $\pi(\boldsymbol{\beta}_{\mathbf{Z}}, \hat{\sigma}^2_\alpha, \hat{\sigma}^2_\varepsilon|\mathbf{y})$. Since each random intercept and the random errors are assumed to be iid, the marginal distributions of the precisions represent the full distributions of the precisions. The respective distributions are then inverted, so that they correspond to the distribution of the variances for $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$. Since the response is standardized, these variance distributions correspond to a distribution of the proportion of variance explained by each random intercept, i.e. a distribution of their relative importance. For the fixed effects, samples of $\boldsymbol{\beta}_{\mathbf{Z}}$ from the joint posterior are drawn, squared and transformed with (2.23), to represent a sampled distribution of the individual relative importance. We must sample $\boldsymbol{\beta}_{\mathbf{Z}}$ from the approximate

joint distribution since the marginal distributions only describe the behavior of individual components of $\beta_{i,\mathbf{Z}}$ in isolation, and therefore does not take correlation between them into account. Using INLA functionality, for a sample $\boldsymbol{\beta}_{s,\mathbf{Z}}$ from the approximate joint distribution we can estimate the relative importance for each column in $\mathbf{X}$ by

$$\mathrm{RI}(\mathbf{X})_s = \boldsymbol{\Lambda}_s^{[2]} \boldsymbol{\beta}_{s,\mathbf{Z}}^{[2]} \ , \tag{3.5}$$

where, $\mathrm{RI}(\mathbf{X})_s$ is a column vector where entry j is the estimate of the relative importance of column j in $\mathbf{X}$ from the sample $\boldsymbol{\beta}_{s,\mathbf{Z}}$. By repeating this process for multiple samples, one can obtain an estimate of the posterior distribution of each element in $\mathrm{RI}(\mathbf{X})$. From the estimated posterior distribution of elements in $\mathrm{RI}(\mathbf{X})$ and the marginal variance distributions of $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ we can obtain the model R^2 distribution in the Bayesian framework as described by equations (2.31) and (2.32). The advantage of having the posterior distributions instead of point estimates is prominent in natural sciences where one might have complex data structures, since it carries a fundamental uncertainty.

3.2 Simulation study

To evaluate the performance of our proposed method, BVI, a simulation study was conducted. The study investigates how the BVI compares to the relative importance decomposition(Relaimpo) presented in Grömping (2007) and the two methods presented in Matre (2022). The Relaimpo method uses the LMG decomposition and considers only fixed effects and can therefore only be compared with the BVI in the fixed effects. The two methods in Matre (2022), ELMG and the ERW, are extensions of the LMG and relative weights methods respectively, to include random intercepts. These extensions allow us to compare the results for the random intercept model to our BVI method. The decompositions of the variance in (3.2) can be used to compare the theoretical variance explained in a model against the results from Bayesian Variable Importance method. Since the response $\mathbf{y}$ is standardized, each relative importance assigned to a distinct effect is analogous to a proportion or percentage of the total variance of $\mathbf{y}$.

To simulate the data we consider the model as in (2.7), with a sample size $n = 10^4$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$ where $\alpha_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$ as a single random intercept for $m = 200$ clusters of $n_j = 50$ observations each, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^{n \times p}$, where $\boldsymbol{\mu} = (1, 2, 3)$, $\Sigma_{ii} = 1, \Sigma_{i,k} = \rho_{i,k}, k \neq i$ and $p = 3$ consisting of three fixed effects, $\mathbf{U}$ as a desgin matrix of appropriate dimension and a random error $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 = 1)$. Further, the true vector of regression coefficients is set to be set to be $\boldsymbol{\beta}_{\mathbf{X}} = (1, \sqrt{2}, \sqrt{3})^T$ so the total model, including an intercept column of ones, can be written as

$$\mathbf{y} = \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_{\mathbf{X}} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{3.6}$$

The data is standardized, meaning that $\sigma_{\mathbf{y}}^2 = 1$, $\boldsymbol{\beta}_{\mathbf{X}} = (\sqrt{1/8}, \sqrt{2/8}, \sqrt{3/8})^T$ and $\sigma_\alpha^2 = \sigma_\varepsilon^2 = 1/8$. This standardization allows us to easily compare and interpret results as proportions of the total variance of $\mathbf{y}$.

From this setup, the theoretical variance of the response is

$$\text{Var}(\mathbf{y}) = \beta_{1,\mathbf{X}}^2 + \beta_{2,\mathbf{X}}^2 + \beta_{3,\mathbf{X}}^2 + 2\sum_{j=1}^{3}\sum_{k=j+1}^{3} \beta_{j,\mathbf{X}}\beta_{k,\mathbf{X}}\rho_{jk} + \sigma_\alpha^2 + \sigma_\varepsilon^2 \ . \tag{3.7}$$

as in (3.2) and the theoretically correct relative importances for uncorrelated data are

$$\text{RI}(\mathbf{x}_1) = \beta_{1,\mathbf{X}}^2 = \text{RI}(\alpha) = \sigma_\alpha^2 = \frac{1}{8}\ , \ \text{RI}(\mathbf{x}_2) = \beta_{2,\mathbf{X}}^2 = \frac{2}{8}\ , \ \text{RI}(\mathbf{x}_3) = \beta_{3,\mathbf{X}}^2 = \frac{3}{8}\ . \tag{3.8}$$

Further, the theoretically expected marginal and conditional R^2 values can be calculated from 3.7 as the variance of the fixed effects divided by the total variance and the variance of the fixed effects and random intercepts divided by the total variance respectively. The R^2 values are listed in Table 1. These values provide

ρ	R^2_{marg}	R^2_{cond}
0	0.750	0.875
0.1	0.781	0.890
0.5	0.852	0.926
0.9	0.889	0.945

Table 1: The theoretically correct marginal variance explained (left column) and conditional variance explained (right column) for different correlation levels between the fixed effects.

an empiric way of checking if our method fulfills the proper decomposition criteria listed in Section 2.2.2, by seeing if the relative importances for each effect sum to the model R^2.

To investigate how different correlations between the fixed effects are handled by the method, we consider four different correlation levels between the fixed covariates in our data. That is achieved by letting $\rho_{1,2} = \rho_{1,3} = \rho_{2,3}$ take on the values $\{0, 0.1, 0.5, 0.9\}$. For each correlation level, we simulate $N = 1000$ datasets and fit each of the four methods BVI, Relaimpo, ELMG and ERW. To get a comparable measure from the Bayesian framework to the frequentist framework, we use the posterior means of the sampled posterior distribution of $\text{RI}(\mathbf{X})$ when estimating (3.4). It can here be noted that in the BVI method the approximated posterior marginals for each predictor, as well as the sampled posterior distribution of $\boldsymbol{\beta}_\mathbf{Z}$ and $\text{RI}(\mathbf{X})$, are available for each dataset.

CHAPTER

FOUR

RESULTS

4.1 Insights from the simulation study

To present the results of the simulation study we consider each effect separately in different plots, that is we show results for the importance of variables X_1, X_2, X_3 and the random effect $\boldsymbol{\alpha}$, in distinct plots. We used violin plots to visualize the estimated quantities, as they contain much information in a compact way. The violin plot is analogous to a density plot, but the density is shown along the y-axis and mirrored about the y-axis to form a symmetrical shape. Each violin therefore displays the distribution of our simulated estimates. Lastly, we consider how our BVI model estimates the conditional and marginal R^2 values from a Bayesian perspective, comparing these results to the R^2 value obtained from the decompositions by the Relaimpo, ELMG, and ERW methods, where the first method only considers models without random effects.

4.1.1 Relative importance of the fixed effects

In Figure 1 the distribution of the relative importance allocated to each fixed effect from the simulations are shown. There are four different distributions for each method, which corresponds to the four different correlation levels. The horizontal dashed line displays the theoretically correct relative importance when the covariates are pairwise independent.

In general, it can be seen that the distributions in the case of uncorrelated data are unbiased with some variation around the theoretically correct relative importance. For a correlation of $\rho = 0.1$ the distributions of the estimates are shifted marginally compared to the uncorrelated case for all methods. The importance attributed to X_1 and X_2, in Figure 1a and Figure 1b respectively, is larger when compared to the uncorrelated case, whereas the importance attributed to X_3 in Figure 1c is smaller. All methods seem to shift the relative importance estimate for the covariate with the same amount in the same direction. This shift is both expected and desirable, when considering the values found in Table 1 for the theoretically correct variance explained. Therefore, we should expect our method to assign different shares when we have various levels of covariate correlation, which it does. This trend continues for the correlation level $\rho = 0.5$, where the distribu-

tions are shifted further in the same directions as for $\rho = 0.1$. Lastly, for $\rho = 0.9$ we see the largest reallocation of the distributions, which follows the same trend as for the other correlation levels.

The rise in importance for X_1 and X_2 for increasing correlation can be understood by the relation $\mathbf{Z\Lambda} = \mathbf{X}$ in the relative weights method. When the matrix $\mathbf{X}$ is not correlated, $\mathbf{\Lambda}$ is close to the identity matrix, but with an increase in correlation the diagonal elements grows smaller and off diagonal elements grow larger. An increase in off diagonal values would for X_1 and X_2 imply that a larger value is multiplied with β_3^2, which is larger than β_1^2 and β_2^2. Therefore, it is expected to see a rise in importance as correlation increases for X_1 and X_2, and the opposite for X_3. In all figures, the BVI method is in agreement with the other methods when allocating importance for different correlation levels. The width of the distributions seem to become lower as the correlation increases, most notably for $\rho = 0.9$, where the distributions exhibit significantly smaller dispersion than for $\rho = 0$. Generally all methods seem to follow the same trends and produce similar results for all three fixed effects. As correlation increases the trend is that the relative importance assigned to X_1 and X_2 increases, in contrast to the decrease in relative importance assigned to X_3.

In general, the BVI method is in agreement with the theoretical results for uncorrelated data derived in Chapter 3 and is consistent with the other three methods for correlated fixed effects.

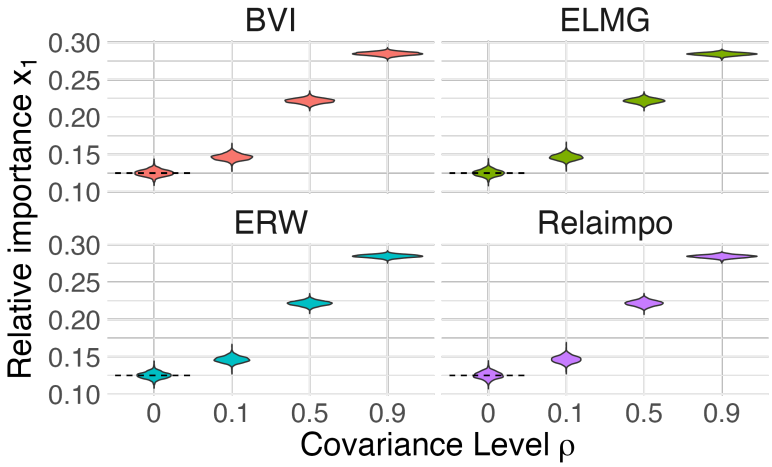


Figure 1: Violin plots for the relative importance of the fixed effects X_1, X_2 and X_3 for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG, ERW and the Relaimpo methods. The standardized regressor coefficients are $\boldsymbol{\beta} = \left(\sqrt{1/8}, \sqrt{2/8}, \sqrt{3/8}\right)$, and the true total model variance is $\sigma_y^2 = 1$. For the BVI method the distributions of posterior means are shown to compare to the distribution of point estimates from the other three methods. The horizontal line displays the theoretically correct importance of each fixed effect in the case of uncorrelated data. (a) Relative importance of X_1 as calculated from the four methods.

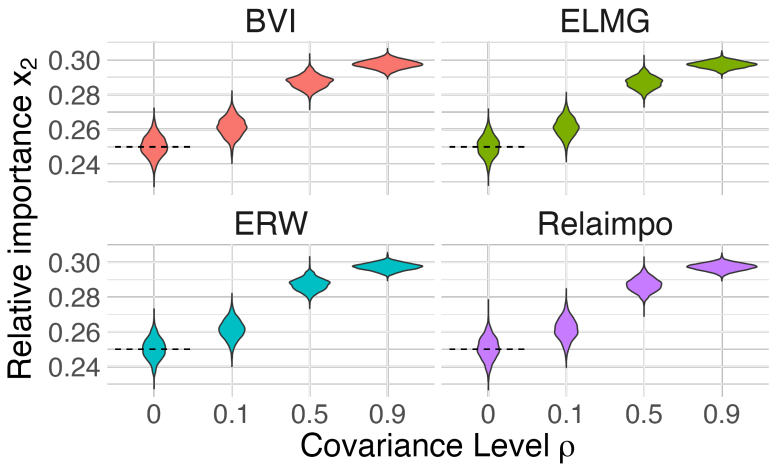


Figure 1: (b) Relative importance of X_2 as calculated from the four methods.

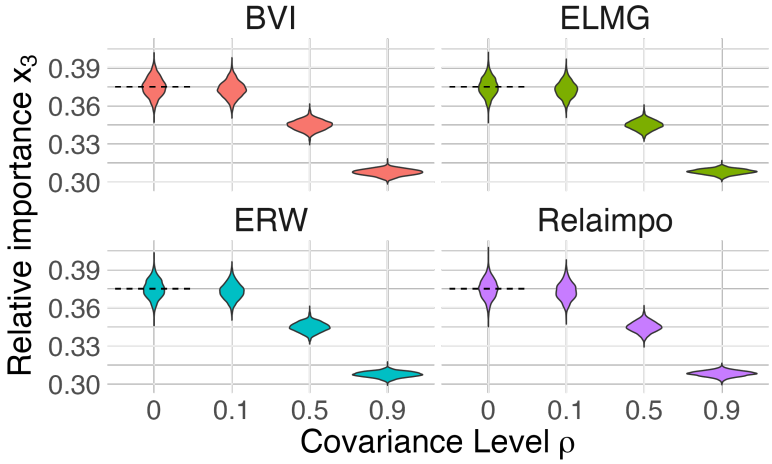


Figure 1: (c) Relative importance of X_3 as calculated from the four methods.

4.1.2 Relative importance of the random effects

Considering a model with one random intercept, we can no longer compare our model with the Relaimpo method, which is only implemented for the linear regression in the relaimpo R package (Grömping 2007). Therefore, we now compare the BVI method only with the ELMG and ERW methods, which have been extended from the Relaimpo method (Matre 2022). Figure 2 shows the distribution of the relative importance, or variance, assigned to the random intercept $\boldsymbol{\alpha}$ for different correlation levels. The random intercept $\boldsymbol{\alpha}$ follows a univariate normal distribution with variance equal to $1/8$ and the standard deviation of the response is $\sigma_{\mathrm{y}}^2 = 1$. As before the horizontal line shows the theoretical relative importance that $\boldsymbol{\alpha}$ has in the model when the fixed effects are uncorrelated.

From Figure 2 it is apparent that both the location and width of the relative importance distribution of all methods are largely indistinguishable. The distributions take on a moderately smaller value when $\rho = 0.1$ and the location of the estimates is further decreased for $\rho = 0.5$ and $\rho = 0.9$. For the latter correlation

level, the distributions are located around a value that is less than half of the value of the centering when the fixed effects are uncorrelated. To re-emphasize, this is both expected and desirable since the increase in response variance comes solely from the correlation of fixed effects, so the random effects now contribute to explain a smaller proportion of the variance, *i.e.* the importance is lower.

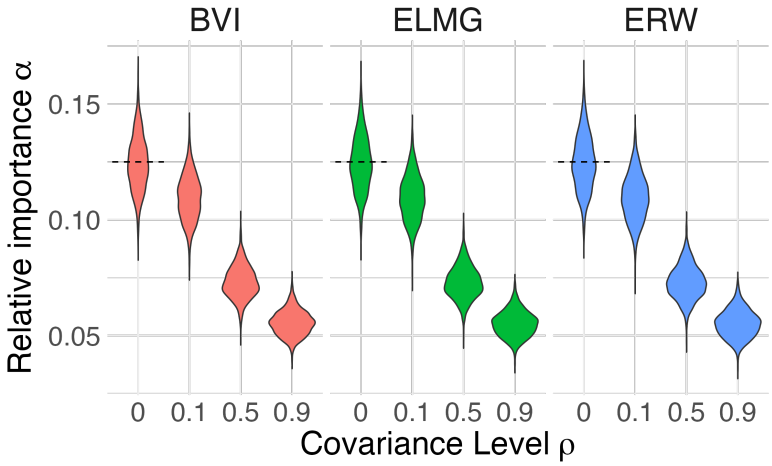


Figure 2: Violin plots for the relative importance of the random effect $\boldsymbol{\alpha}$, that is, $\hat{\sigma}^2_\alpha$ for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG and the ERW method. For the BVI method the distributions of posterior means of the marginal distribution of $\hat{\sigma}^2_\alpha$ are shown to compare to the point estimates of the other two methods. The horizontal line displays the theoretically correct importance $\sigma^2_\alpha = 0.125$ of the random effect in the case of uncorrelated data.

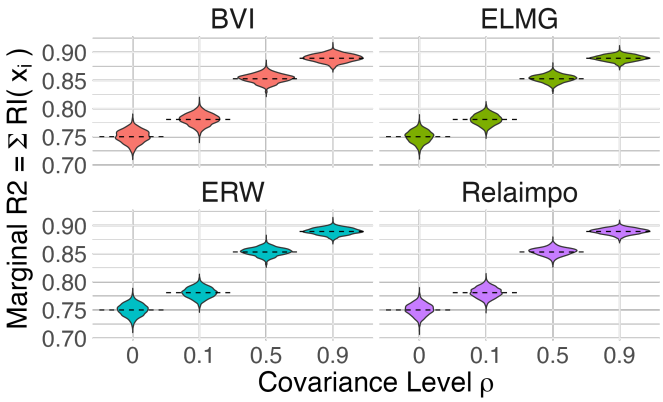
4.1.3 Total variance explained - R^2 estimates

As a useful by-product from the previous results we can get the total variances explained by our model (Figure 3). The marginal variance explained is the variance explained by the fixed effects (Figure 3a), and we get results for all four methods, including Relaimpo. In Figure 3b the total conditional variance explained, R^2_{cond}, is displayed. This is the variance given all the fixed effects and the random effect. To complement the conditional and marginal variances explained, a horizontal line is drawn for each correlation level corresponding to the theoretically correct variance explained, found in Table 1, for the correlation level.

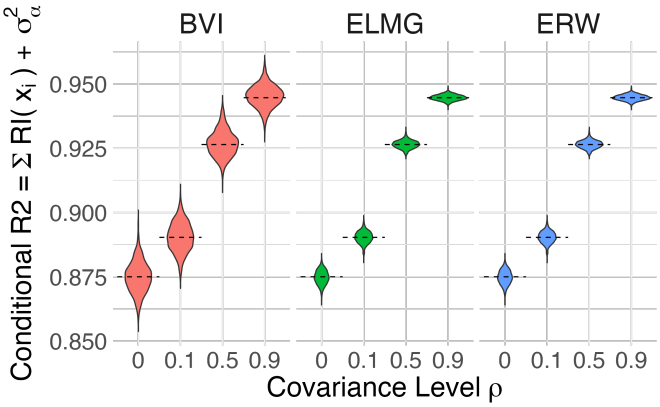
Figure 3a shows that the four methods produce very similar results of R^2_{marg} for the fixed effects across all correlation values, albeit a slightly larger width for the BVI method can be seen. When considering the conditional variance R^2_{cond} in Figure 3b, the dispersion of the BVI method is strikingly larger compared to the other methods.

Both the marginal and the conditional variance are centered around the theoretically correct value with some variability, particularly visible for conditional variance of the BVI method. The centering of the distributions for both the

marginal and conditional variances resemble each other for all methods, regardless of correlation level.



(a) Total marginal variance R^2_{marg}.



(b) Total conditional variance R^2_{cond}.

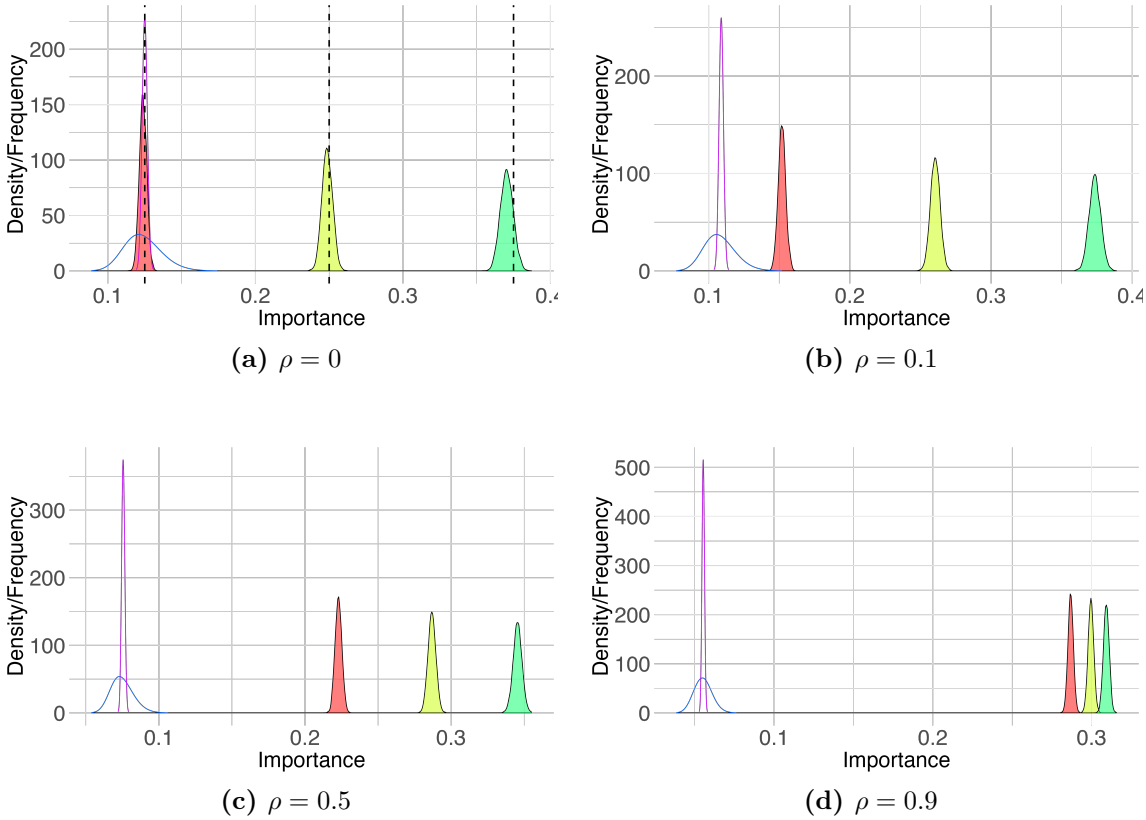
Figure 3: Violin plots for the total marginal and conditional variance explained for different correlation levels calculated from the ensemble of simulated datasets by the BVI, ELMG, the ERW and the Relaimpo method(only marginal variance explained can be computed). For the BVI method the posterior means of the sampled posterior distributions of $\boldsymbol{\beta}$ and the marginal distribution of $\hat{\sigma}^2_\alpha$ in each simulation are used to compare to the point estimates of the other two methods. The horizontal lines display the theoretical explained variance for each correlation level ρ as in Table 1.

4.2 The BVI method

The results from the BVI method for each dataset consists of the posterior marginal distributions of the variances of $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$, as well as the approximated distribution of RI($\mathbf{X}$) from samples $\boldsymbol{\beta}_{s,\mathbf{Z}}$ drawn from the joint posterior distribution $\pi(\boldsymbol{\beta}_{\mathbf{Z}}, \hat{\sigma}^2_\alpha, \hat{\sigma}^2_\varepsilon|\mathbf{y})$. The derivation of these results are described in detail in Section 3.1.2 and provides a more informative basis to make inference on compared to point estimation. As this is the key advantage of the Bayesian framework, we also take a look on what results our method provides for a single model fit for different correlation levels. We calculate these posterior distributions for the same four different correlation levels as we use in the simulation study. Further, we use the samples $\boldsymbol{\beta}_{s,\mathbf{Z}}$ from the joint distribution and the marginal posterior variances of $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ to calculate the distribution of R^2 in accordance with (2.31) and (2.32).

4.2.1 Posterior relative importance distributions

Approximate posterior marginal distributions for the variances of the random effects, and sampled approximate posterior distributions for the fixed effects are available for the BVI method for each dataset. These are featured in Figure 4 for one realization of the four different datasets of different correlations. All posteriors for one correlation level are shown in the same subplot, with one subplot for each correlation level. For the correlation level $\rho = 0$ the theoretically correct relative importances are shown as vertical lines. Note here that it is expected and desired that the distributions from a single dataset are stochastic and therefore deviate somewhat from the theoretical values. When data are uncorrelated (Figure 4a) the posterior distributions are coincidentally centered close to the theoretical value. The posterior distribution of $\hat{\sigma}^2_\alpha$ is slightly skewed to the left and demonstrates a distinctively larger degree of dispersion than the other effects. For $\rho = 0.1$ (Figure 4b), a small shift of all posterior distributions is noticeable, in accordance with the results from Figure 1 and Figure 2. This repositioning is more clear for $\rho = 0.5$ and $\rho = 0.9$ (Figure 4c and Figure 4d), and it is clear to see that the posterior distribution of the variance of $\hat{\sigma}^2_\alpha$ follows the trend of $\hat{\sigma}^2_\varepsilon$ closely. When correlation is increased, it seems that the width of all posterior distributions narrow and become tighter. The trend of increasing relative importance assigned to X_1 and X_2 while decreasing importance assigned to X_3 is highlighted by the posterior distributions of the respective effects as they align closer for each increase in correlation. These are the types of results that we would expect to see, as they deviate from the theoretical values with a plausible amount when considering that they are stochastic.



(a) $\rho = 0$ **(b)** $\rho = 0.1$

(c) $\rho = 0.5$ **(d)** $\rho = 0.9$

Figure 4: Posterior distributions for fixed effects and posterior marginal distributions for random effects from the BVI method on four randomly simulated datasets with different correlation values between the fixed effects. The blue and purple densities are the marginal posteriors for $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\varepsilon^2$ respectively, whereas the red, yellow and green densities are the sampled posteriors for X_1, X_2 and X_3 respectively. The vertical lines in (a) represent the theoretically correct relative importances.

4.2.2 Posterior R^2 distributions

From the above discussed posterior distributions, the posterior distributions of marginal (Figure 5a) and conditional (Figure 5b) variance explained, *i.e.* R^2, can also be obtained. As was the case for the posterior distributions of fixed and random effects, the R^2 distributions are also desired to possess a degree of stochasticity causing them to deviate moderately from the theoretical value. We notice that in the uncorrelated case the R^2 is evenly distributed around the theoretically expected variance explained. For $\rho = 0.1$ and $\rho = 0.9$ the distribution of R^2 seem to be a skewed to the right when comparing it to the theoretically expected value and the distribution for $\rho = 0.5$ skewed to the left. The width of the distributions seem to narrow as correlation levels increase, as a result of the decreasing width in the individual posterior distributions of random and fixed effects.

Moving on to the conditional R^2 (Figure 5b), which are displayed on the same scale as the marginal R^2, it is obvious that these distributions are tighter than the marginal R^2 distributions. The scale is arguably too large to be used to display

the conditional R^2 when one sees how narrow they are, but we chose to do so because of the comparison to the marginal R^2. To the contrary, for $\rho = 0.1$ and $\rho = 0.9$ the distributions are skewed to the right. The marginal and conditional R^2 for the same model will be closely related so one should expect to see the same behaviour in both, as we do.

The posterior distributions of R^2 reflect our expectations as they exhibit the same stochasticity as seen for the posterior distributions of fixed and random effects.

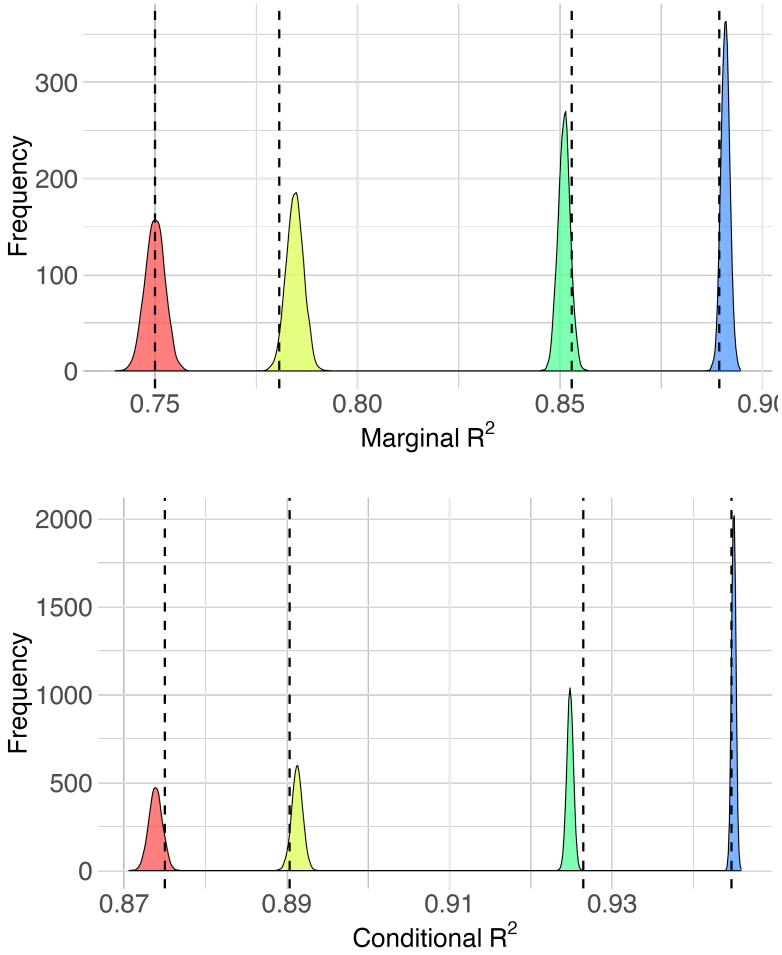


Figure 5: Posterior R^2 distribution calculated by the BVI method from the posterior distributions of fixed and random effects on four randomly simulated datasets with different correlation values ρ between the fixed effects. The red, yellow, green and blue distributions correspond to the posterior distribution of R^2 for $\rho = 0, \rho = 0.1, \rho = 0.5$ and $\rho = 0.9$ respectively. The theoretical values can be found in Table 1 and are shown as vertical lines.

CHAPTER

FIVE

DISCUSSION & FURTHER WORK

In this thesis we have presented a novel method for Bayesian variable importance, the BVI method, a new method for calculating relative importance of covariates in random intercept models. The BVI method broadens the concept of the extended relative weights (ERW) method to the Bayesian framework, which provides a more computationally feasible method as an alternative to the extended LMG (ELMG) method. This extension makes use of the advantageous properties of Bayesian methods that give posterior distributions of each parameter and combines this with the relative weights method (Johnson 2000). The relative weights method projects the covariates into an orthogonal space and approximates the design matrix with the projected covariates (Johnson 2000, Mirsky 1960) for the linear model and is extended in Matre (2022) to also work for random intercept models. Our main inference is done using results from (Gelman et al. 2017) and Nakagawa & Schielzeth (2013) regarding the R^2 for Bayesian linear mixed models. After conducting a simulation study, the results show that the BVI method is comparable to both the ERW and ELMG, as well as the more established Relaimpo method. Due to the BVI method's Bayesian nature, we have also presented the sampled posterior distributions of the fixed covariates and the marginal posteriors of the random effects. All code used in the implementation of the BVI method and used to produce results is available in GitHub with a link in Appendix A. Further, Appendix B contains a usage example of the method so that the reader can easily implement the BVI method in their own work.

Hopefully, the BVI method can be used as a new tool to help researchers in various fields, which rely upon inference about covariates that try to explain a complex relationship with the response. The data often reflects some of this complexity, making it hard to draw conclusions and obtain trustable inference. In addition to its usefulness within natural sciences, a Bayesian variable importance measure is a useful tool in itself as an analogue to the established methods, e.g. in Grömping (2007), for the frequentist framework. The key aspect that separates the BVI method from the other methods discussed is the inference it delivers from a single dataset, in the form of the posterior distributions of the relative importances. These distributions will presumably allow field experts to make more informed statements about the effect on the response caused by each covariate in a random intercept model, which can further help in their research.

As listed in section Section 2.2.2, relative importance measures are compared by a list of criteria, which we consider when evaluating our results while keeping in mind that we are in a Bayesian framework with our model. It has been shown that the relative weights method satisfies the proper decomposition, inclusion and non-negativity criteria (Matre 2022). In Grömping (2007) it is argued that the exclusion criterion does not appear to always be reasonable, and therefore nor do we consider this criterion to be relevant when assessing our BVI method. The BVI method has shown promising results that in posterior expectation it fulfills the proper decomposition criterion, since the BVI method gives very similar results in the simulation study as the other methods, which have been proven to provide a proper decomposition. It could be mentioned that the BVI method sometimes varied more than the established methods. We see this a consequence of the inherit uncertainty in the Bayesian framework, which is expected and in our case desirable. Further, non-negativity is fulfilled by the BVI method by construction as the relative importances are squared. It is noted in Matre (2022) that the ELMG and ERW can theoretically violate the inclusion criterion, however it is seen to be unlikely to happen in practice. We have not fully explored the inclusion criterion for the BVI method, but also regard it as unlikely that this criterion will be violated in practice. In fact, since we consider distributions in the Bayesian framework rather than point estimates, we think that the results should not be subject to the same strict constraints that the frequentist framework can impose. For a small regression coefficient, it should not uncritically be dismissed as a problem if the resulting distribution contains zero. If the posterior importance distribution contains zero, the importance should perhaps be subject to careful interpretation rather than dismissal at first sight. One could even argue that the inclusion of zero in the posterior distribution is no problem at all, and regard it as interesting information without any need to perform model selection. The BVI method produces plausible results for the total variance explained and the decomposition of total variance into relative importance of covariates for one dataset. Further, it produces very similar results compared to the Relaimpo, ELMG and ERW methods over a simulation study. We believe this to be an indication that it provides a proper decomposition, which we consider to be the most important criteria to fulfill. A proof of proper decomposition for the BVI method is perhaps only available in expectation, but would nonetheless be of great interest, and hopefully this can be achieved in the future.

In Matre (2022, section 6.1) it is highlighted that the ERW method can be sensitive to correlation between the fixed covariates, and this causes it to be less trustworthy than the ELMG method. As the BVI method utilizes the relative weights transformation on the data, it is to be expected that this same sensitivity is present in the BVI method. Even though the difference between the ELMG and ERW was found to be relatively small in Matre (2022), one could investigate whether the uncertainty carried from the relative weights approximations might negatively affect the desired uncertainty that the BVI method delivers. From the results presented in this thesis, any systematic error from the transformation of data are also believed to be relatively small, as the BVI compares very similarly to both ERW and ELMG.

As mentioned the main advantage of relative importance in a Bayesian framework is that the methods provide more information on each predictor thanks to posterior marginals. We believe that even though some accuracy is lost in the transformation of data, the BVI method provides a viable analogue to the established methods in the Bayesian framework. Since the BVI method is a Bayesian relative importance measure it also allows for incorporating prior knowledge and because of this may prove to be more robust for small sample sizes. The Bayesian framework avoids the many misinterpretations regarding p-values mentioned and allows researchers to make more direct probability statements regarding the model. The pitfall of a sharp cutoff that the null hypothesis testing with p-values are not present, since the model provides posterior distributions rather than point estimates. Hopefully, this can lead researchers to make more thoughtful conclusions and not blindly follow a threshold.

Going forward, there are some obvious expansions of the BVI method that should be considered. The first thing that is to get a better understanding of how the method performs when applied to real data. So far, the BVI method has only been tested on simulated data and testing the BVI method with real data might be the next natural step in its development. Testing with real data should be fairly straightforward to execute with a suitable dataset, but it is still necessary to further strengthen the credibility of the BVI method. With real data, the challenge of categorical covariates also needs to be addressed. Categorical covariates have been considered out of the scope for this thesis, but, is nonetheless a very important aspect that needs to be explored. One possibility could perhaps be to use dummy encoding.

Another extension that would be of interest is to investigate the analytic properties of the BVI method. This thesis has mainly been focused on creating a relative importance measure in the Bayesian framework, and to be able to do so within a limited amount of time has required us to not thoroughly dive into analytic results. We do believe, based on the promising results presented in this thesis, that some analytic results can be made. Particularly proofs in expectation, which is common in the Bayesian framework, would further underline the consistency of the method.

Furthermore, the random intercept model is just scratching the surface of the much larger field of generalized linear mixed models. From the linear model there are essentially two paths to take, either one can introduce random effects and interaction terms (LMM) or one can generalize the linear model (GLM) to not be restricted to a normal response. The BVI method has been implemented with a random intercept, but it would be a great improvement if one could extend the method to also work for random slopes. Random intercepts add variance to the model uniformly across all clusters, but a random slope adds variance to the effects of a specific predictor. Therefore, the variance associated from a covariate would be a combination of the fixed effects and the variance across groups coming from the random slopes. This not only makes calculations more complicated, but would also require a more sophisticated interpretation of the variable importance. No longer could one say that a covariates importance is its overall attribution to

the response variable, but also how it varies across different clusters. Moreover, adding both random intercepts and random slopes to clusters introduces a covariance structure between the random intercepts and random slopes that would also need to be assessed. It would be essential to investigate the effect of correlated random effects, and the implications this poses when trying to obtain a proper variance decomposition. These new elements pose obstacles that one would need to overcome in order to create a successful relative importance measure for the LMM's.

Going down the other path, the extension to generalized linear models would be a substantial advancement. A GLM extension would allow us to also investigate relative importance in non-normal responses, such as the Poisson distribution and the binomial distribution, by linking the expectation of the response to a linear predictor via a link function. This also implies that the regression coefficients calculated would be on a different scale, depending on the link function, and further different scales for measuring the model variance. As a consequence one would have to find a way to overcome this transformation imposed by the link function, as well as the constraints on model parameters that non-normal responses require. If a dependable relative importance measure could tackle the challenges of linear mixed models and generalized linear models, this would hopefully pave the way for a combination that can tackle the generalized linear mixed models (GLMM's).

At the time being, the author of this thesis is especially interested in testing how the BVI method tackles the animal model (Kruuk 2004), before further exploring its analytical properties and possible extensions. This can hopefully be a topic for further investigation in a master thesis and possibly further work.

Variable importance as a field within mathematics is a debated topic, with some authors, including Ehrenberg (Grömping 2015) criticizing the concept itself. The challenges often arise from the recognition that a correct way of allocating unique importances for correlated covariates can never be agreed upon (Grömping 2015), if a unique definition of variable importance even exists. With these considerations in mind, relative importance measures should be seen as a (possibly) useful tool for a statistical model that relies upon assumptions. The insights of relative importance are limited by the statistical model and cannot account for poor model design.

CHAPTER

SIX

CONCLUSIONS

In this thesis we aimed to provide a novel Bayesian variable importance measure, the BVI method, for the random intercept model. The BVI method projects the fixed covariates into an uncorrelated space to easy computational efforts. These uncorrelated covariates are used when the model is fit with a Bayesian approach, where we relied upon the INLA framework. The fitted model can be used to draw samples from the posterior distribution of the model parameters, which are back-transformed to the original covariate space which is of interest. It is in this original covariance space inference is made.

From a simulation study we have shown that the BVI method provides results that suggest it provides a proper decomposition and performs well when compared to more established methods in the frequentist framework. Further, the BVI method allows for extensive inference through posterior distributions and is implemented in the package `BayesianImportance` in the statistical software R. We lastly outline some desired improvements and extensions of the BVI method, before a Bayesian relative importance framework for the generalized linear mixed models can be accomplished. Hopefully, the Bayesian Variable Importance method is the first step towards this goal.

BIBLIOGRAPHY

Akaike, H. (1974), 'A New Look at the Statistical Model Identification', *IEEE Transactions on Automatic Control* **19**(6), 716 – 723.

Bayes, T. & Price, R. (1763), '*An Essay towards Solving a Problem in the Doctrine of Chances*', *Philosophical Transactions of the Royal Society of London* **53**, 370–418.

Budescu, D. V. (1993), 'Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression', *Psychological Bulletin* **114**(3), 542–551.
URL: *https://doi.org/10.1037/0033-2909.114.3.542*

Chiuchiolo, C., van Niekerk, J. & Rue, H. (2021), 'Joint posterior inference for latent gaussian models with r-inla', *arXiv preprint arXiv:2112.02861* .
URL: *https://arxiv.org/pdf/2112.02861.pdf*

Fabbris, L. (1980), 'Measures of predictor variable importance in multiple regression: An additional suggestion', *Quality and Quantity* **14**, 787–792.
URL: *https://doi.org/10.1007/BF00145808*

Fahrmeir, L., Lang, S., Kneib, T. & Marx, B. (2013), *Regression - Models, Methods and Applications*, Springer Berlin, Heidelberg.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

Fong, Y., Rue, H. & Wakefield, J. (2010), '*Bayesian inference for generalized linear mixed models*', *Biostatistics* **11**, 397–412.
URL: *https://doi.org/10.1093/biostatistics/kxp053*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2015), *Bayesian Data Analysis*, 3 edn, Chapman and Hall/CRC, New York.

Gelman, A., Goodrich, B., Gabry, J. & Ali, I. (2017), R-squared for bayesian regression models, Technical report, Linköping.
URL: *https://www.ida.liu.se/~732G43/bayes_R2.pdf*

Genizi, A. (1993), '*Decomposition of R2 in multiple regression with correlated regressors*', *Statistica Sinica* **3**, 407–420.

Goodman, S. (2008), 'A dirty dozen: Twelve p-value misconceptions', *Seminars in Hematology* **45**, 135–140. All rights reserved.

Grömping, U. (2007), '*Estimators of Relative Importance in Linear Regression Based on Variance Decomposition*', *The American Statistician* **61**, 139–147.
URL: *https://www.jstor.org/stable/27643865*

Grömping, U. (2015), 'Variable importance in regression models', *WIREs Computational Statistics* **7**(2), 137–152.
URL: *https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1346*

Gómez-Rubio, V. (2020), *Bayesian Inference with INLA*, Chapman & Hall/CRC Press, Boca Raton, FL.

Hackenberger, B. K. (2019), 'Bayes or not bayes, is this the question?', *Croatian Medical Journal* **60**(1), 50–52.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6406060/*

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. (2015), 'The extent and consequences of p-hacking in science', *PLOS Biology* **13**(3), 1–15.
URL: *https://doi.org/10.1371/journal.pbio.1002106*

Johnson, J. W. (2000), '*A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression*', *Multivariate Behavioral Research* **35**(1), 1–19.
URL: *https://doi.org/10.1207/S15327906MBR3501_1*

Johnson, R. (1966), '*The minimal transformation to orthonormality*', *Psychometrika* **31**, 61–66.
URL: *https://doi.org/10.1007/BF02289457*

Kruskal, W. (1987), 'Relative importance by averaging over orderings', *The American Statistician* **41**(1), 6–10.
URL: *http://www.jstor.org/stable/2684310*

Kruuk, L. E. B. (2004), 'Estimating genetic parameters in natural populations using the 'animal model", *Philosophical Transactions of the Royal Society B: Biological Sciences* **359**(1446), 873–890.
URL: *http://rstb.royalsocietypublishing.org/*

Lee, J. J. & Chu, C. T. (2012), 'Bayesian clinical trials in action', *Statistics in Medicine* **31**(25), 2955–2972.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5404*

Lindeman, R. H., Merenda, P. F. & Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis*, Scott, Foresman and Company, Glenview, IL.

Lipovetsky, S. & Conklin, M. (2001), 'Analysis of regression in game theory approach', *Applied Stochastic Models in Business and Industry* **17**, 319 – 330.

Matre, A. (2022), Relative variable importance approaches for linear models with random intercepts, Master's thesis, NTNU.

McCullagh, P. & Nelder, J. (1989), *Generalized linear models*, 2 edn, Chapman and Hall.

McElreath, R. (2020), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2 edn, Chapman and Hall/CRC.

Mirsky, L. (1960), '*SYMMETRIC GAUGE FUNCTIONS AND UNITARILY INVARIANT NORMS*', *The Quarterly Journal of Mathematics* **11**, 50–59.
URL: *https://doi.org/10.1093/qmath/11.1.50*

Muff, S., Nilsen, E. B., O'Hara, R. B. & Nater, C. R. (2022), 'Rewriting results sections in the language of evidence', *Trends in Ecology & Evolution* **37**(3), 203–210. Published: November 16, 2021.
URL: *https://doi.org/10.1016/j.tree.2021.10.009*

Nakagawa, S. & Schielzeth, H. (2013), 'A general and simple method for obtaining R2 from generalized linear mixed-effects models', *Methods in Ecology and Evolution* **4**, 133–142.
URL: *https://doi.org/10.1111/j.2041-210x.2012.00261.x*

Nimon, K. F. & Oswald, F. L. (2013), '*Understanding the Results of Multiple Linear Regression Beyond Standardized Regression Coefficients*', *Organizational Research Methods* **16**, 650–674.
URL: *https://doi.org/10.1177/1094428113493929*

Poole, M. A. & O'Farrell, P. N. (1971), 'The assumptions of the linear regression model', *Transactions of the Institute of British Geographers* **52**, 145–158.
URL: *http://www.jstor.org/stable/621706*

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 319–392.
URL: *https://doi.org/10.1111/j.1467-9868.2008.00700.x*

Schwarz, G. (1978), 'Estimating the Dimension of a Model', *The Annals of Statistics* **6**(2), 461 – 464.
URL: *https://doi.org/10.1214/aos/1176344136*

Shapley, L. S. (1953), 'Stochastic games*', *Proceedings of the National Academy of Sciences* **39**, 1095 – 1100.
URL: *https://api.semanticscholar.org/CorpusID:263414073*

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011), 'False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science* **22**(11), 1359–1366. PMID: 22006061.
URL: *https://doi.org/10.1177/0956797611417632*

Simpson, D., Rue, H., Riebler, A., Martins, T. G. & Sørbye, S. H. (2017), '*Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors*', *Statistical Science* **32**, 1–28.
URL: *https://doi.org/10.1214/16-STS576*

APPENDIX

A

GITHUB REPOSITORY

All code used to produce results and all latex files in this document are included in the Github repositories linked below. Further explanations are given in the readme-files.

GitHub repository link

- `https://github.com/AugustArnstad/BayesianImportance`
- `https://github.com/AugustArnstad/TMA4500-Specialization-Project`

APPENDIX

B

BAYESIAN VARIABLE IMPORTANCE USAGE

In this section, an example R code snippet is provided for demonstration purposes. This code creates a random intercept model according to the Bayesian Variable Importance method described in the thesis, and includes some relevant plots and shows the main capabilities of the package.

R Code Example

````
## INSTALLING THE PACKAGE
# This section ensures the devtools package is installed, which is required for installing packages from GitHub. We then install the BayesianImportance package directly from GitHub using devtools::install_github(). In the package under the Hello.R file, all functions are defined with corresponding documentation.
```{r}
# If not already installed, install the 'devtools' package
if(!require(devtools)) install.packages("devtools")

# Install BayesianImportance
devtools::install_github("AugustArnstad/BayesianImportance")

```

## SIMULATE DATA
# In this part, we simulate data to demonstrate the functionality of the BayesianImportance package. We generate random variables with different correlation structures, random effects, and an error term. The data is then structured into data frames for further analysis. If you have a suitable dataset you can use this instead.

```{r}
library(BayesianImportance)
````
````

```
library(INLA)
library(mnormt)
library(ggplot2)
library(reshape2)
library(RColorBrewer)
set.seed(1234)

n <- 10000
nclass_gamma <- 200
nclass_eta <- 100

mu <- c(1,2,3)

sigma <- matrix(c(1, 0.3, 0.5, 0.3, 1, 0.4, 0.5, 0.4, 1),
   3, 3)

#Sample a standardized correlated design matrix
X <- rmnorm(n, mu, sigma)

#Add random effects
gamma <- rep(rnorm(nclass_gamma, 0, sqrt(1)), each=n/nclass
   _gamma)
eta <- rep(rnorm(nclass_eta, 0, sqrt(1/2)), each=n/nclass_
   eta)
epsilon = rnorm(n, mean=0, sd=sqrt(1))
beta <- c(1, sqrt(2), sqrt(3))

#Define some formula
Y1<-  1 + beta[1]*X[, 1] + beta[2]*X[, 2] + beta[3]*X[, 3]
   + gamma + epsilon # + eta
Y2<-  1 + beta[1]*X[, 1] + beta[2]*X[, 2] + beta[3]*X[, 3]
   + gamma + eta + epsilon

#Collect as a dataframe
data_bayes1 = data.frame(cbind(Y1, X = X))
data_bayes1 = data.frame(cbind(data_bayes1, gamma=gamma))

data_bayes2 = data.frame(cbind(Y2, X = X))
data_bayes2 = data.frame(cbind(data_bayes2, gamma=gamma))
data_bayes2 = data.frame(cbind(data_bayes2, eta=eta))

names(data_bayes2)
```

```
## USAGE
# Here we demonstrate the usage of the BayesianImportance
   package. We fit two Bayesian models and sample posterior
    distributions for different simulated datasets using
   functions from the package BayesiannImportance. We fit
   one model with a single random intercept and one model
```

```
    with two random intercepts
59  ```{r}
60  set.seed(1234)
61
62  model1 <- run_bayesian_imp(Y1 ~ V2 + V3 + V4 + (1 | gamma),
        data=data_bayes1)
63  posteriors1 <- sample_posteriors(Y1 ~ V2 + V3 + V4 + (1 |
    gamma), data=data_bayes1, 5000, n)
64
65  model2 <- run_bayesian_imp(Y2 ~ V2 + V3 + V4 + (1 | gamma)
    + (1 | eta), data=data_bayes2)
66  posteriors2 <- sample_posteriors(Y2 ~ V2 + V3 + V4 + (1 |
    gamma) + (1 | eta), data=data_bayes2, 5000, n)
67  ```
68
69  ```{r}
70  variance_marginals_list <- lapply(model1$marginals.hyperpar
    , function(x) inla.tmarginal(function(t) 1/t, x))
71
72  res <- variance_marginals_list$`Precision for the Gaussian
    observations`
73
74  inla.mmarginal(res)
75  ```
76
77
78  ## PLOTTING
79  # This code block defines a function plot_posterior that
    visualizes posterior distributions and relative
    importance from the sampled posteriors. The function
    creates density plots and overlays line plots for
    variance marginals. This can be modified for the
    specific problem at hand, and is not included in the
    package since it is a problem specific function.
80  ```{r}
81  plot_posterior <- function(betas, importances, marginals,
    importance=FALSE, theoretical=FALSE, eta=FALSE) {
82
83    if (importance) {
84      mat_long <- melt(as.data.frame(importances))
85    } else {
86      mat_long <- melt(as.data.frame(betas))
87    }
88    names(mat_long) <- c("Variable", "Value")
89
90    df_marginals <- do.call(rbind, marginals)
91    df_marginals$Effect <- as.factor(df_marginals$Effect)
92
93    if (!importance){
94      plot_title = paste("Posterior Distributions")
95    }else{
```

```
96        plot_title = paste("Posterior Relative Importance")
97      }
98
99      if (eta) {
100       random_effect_labels <- c(expression(sigma[eta]^2),
            expression(sigma[alpha]^2), expression(sigma[epsilon
            ]^2))
101       num=6
102     }
103     else{
104       random_effect_labels <- c(expression(sigma[alpha]^2),
            expression(sigma[epsilon]^2))
105       num=5
106     }
107
108     fixed_effect_labels <- if (importance) {
109       c(expression(beta[1]^2), expression(beta[2]^2),
            expression(beta[3]^2))
110     } else {
111       c(expression(beta[1]), expression(beta[2]), expression(
            beta[3]))
112     }
113     colors=rainbow(5)
114
115     p <- ggplot() +
116       geom_density(data = mat_long, aes(x = Value, fill = as.
            factor(Variable)), alpha = 0.5) +
117       geom_line(data = df_marginals, aes(x = x, y = y, color
            = Effect)) +
118       labs(#title = plot_title,
119            x = "Importance", y = "Density/Frequency") +
120       theme_minimal() +
121       theme(legend.position = "right", #USE NONE FOR NO
            LEGEND
122             legend.text = element_text(size = 20),
123             legend.title = element_text(size = 20),
124             plot.title = element_text(size = 20, face = "bold
                "),
125             axis.title.x = element_text(size = 20),
126             axis.title.y = element_text(size = 20),
127             strip.text = element_text(size = 20),
128             axis.text.x = element_text(size = 20),
129             axis.text.y = element_text(size = 20)) +
130       scale_fill_manual(name = "Fixed effects",
131                         values = colors[1:3],
132                         labels = fixed_effect_labels) +
133       scale_color_manual(name = "Random effects",
134                          values = colors[4:num],
135                          labels = random_effect_labels)
136
137     if (theoretical) {
```

```
      if(eta){
        p <- p + geom_vline(xintercept = 1/9, color = "black"
            , linetype = "dashed") +
              geom_vline(xintercept = 1/18, color = "black",
                 linetype = "dashed") +
              geom_vline(xintercept = 2/9, color = "black",
                 linetype = "dashed") +
              geom_vline(xintercept = 3/9, color = "black",
                 linetype = "dashed")
      }
      else{
        p <- p + geom_vline(xintercept = 0.125, color = "
           black", linetype = "dashed") +
              geom_vline(xintercept = 0.25, color = "black",
                 linetype = "dashed") +
              geom_vline(xintercept = 0.375, color = "black",
                  linetype = "dashed")
      }
    }
    return(p)
}

plot1 <- plot_posterior(posteriors1$beta, posteriors1$
   importance, posteriors1$marginals, importance = TRUE)

plot2 <- plot_posterior(posteriors2$beta, posteriors2$
   importance, posteriors2$marginals, importance = TRUE,
   eta=TRUE)

```

```

# These plots are generated for each model's posteriors. We
    create and store the plots in variables plot1, plot2,
   plot3, and plot4 for the respective models.
```{r}
plot1

plot2
```


GELMAN R2
Here, we compute and visualize the Gelman conditional R2
   metrics for each model. This metric gives insight into
   the proportion of variance explained by the fixed
   effects in the models.
```{r}
colors=rainbow(5)
# Create a combined plot with different colors for each
   distribution
marginal_var <- ggplot() +
```
```

```
    geom_density(data = as.data.frame(posteriors1$r2), aes(x = V1), fill = colors[1], alpha = 0.5) +
    #geom_vline(xintercept = variance_explained$R2_marginal[1], color = colors[1], linetype = "dashed") +

    geom_density(data = as.data.frame(posteriors2$r2), aes(x = V1), fill = colors[2], alpha = 0.5) +
    #geom_vline(xintercept = variance_explained$R2_marginal[2], color = colors[2], linetype = "dashed") +

    theme_minimal() +
    theme(legend.position = "none",
          axis.title.x = element_text(size = 20),
          axis.title.y = element_text(size = 20),
          axis.text.x = element_text(size = 20),
          axis.text.y = element_text(size = 20)) +
    labs(x = expression(R^2), y = "Frequency")

  marginal_var
  ```

  ```{r}

  # Create a combined plot with different colors for each distribution
  conditional_var <- ggplot() +
    geom_density(data = as.data.frame(posteriors1$r2_cond), aes(x = V1), fill = colors[1], alpha = 0.5) +
    #geom_vline(xintercept = variance_explained$R2_conditional[1], color = colors[1], linetype = "dashed") +

    geom_density(data = as.data.frame(posteriors2$r2_cond), aes(x = V1), fill = colors[2], alpha = 0.5) +
    #geom_vline(xintercept = variance_explained$R2_conditional[2], color = colors[2], linetype = "dashed") +

    theme_minimal() +
    theme(legend.position = "none",
          axis.title.x = element_text(size = 24),
          axis.title.y = element_text(size = 24),
          axis.text.x = element_text(size = 24),
          axis.text.y = element_text(size = 24)) +
    labs(x = expression(R^2), y = "Frequency")

  conditional_var
  ```
```

**Listing B.1:** Usage of the Bayesian Importance package with plots and examples.
```