August Arnstad

Relative variable importance in Bayesian generalized linear mixed models with applications in quantitative genetics

TMA4900 Masters thesis in Industrial Mathematics
Supervisor: Stefanie Muff
June 2024

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

NTNU

ABSTRACT

As one of the most widely used statistical methods, regression models have a fundamental position in statistics. Obtaining inference on the covariates used to model the response is a key part of regression analysis, and often it is desirable to assign the covariates with a *relative importance*. To do so, numerous methods from multiple perspectives exist. Despite this, a consensus has not been reached, and the traditional methods using p-values have created a reproduction crisis in the social and biomedical sciences. Our contribution to remedy this, is to put forth a Bayesian relative variable importance measure. The measure is designed to make researchers more thoroughly interpret the statistical model and its results, rather than blindly following a threshold to draw conclusions.

Our measure, coined *Bayesian Variable Importance* (BVI), is implemented by transfering the logic of more established, frequentist methods, to the Bayesian framework. The BVI method is applicable to generalized linear mixed models (GLMMs) with continuous, Binomial and Poisson distributed respones. The core of the method, is to utilize the relative weights method on the covariates of before fitting a Bayesian GLMM and performing calculations with respect the Bayesian framework. This produces posterior distributions of the relative importance of all covariates present in the model, as well as the estimated distributions of the marginal and conditional R^2. To make the methodology and measure easily available for researchers across fields, an R package called `BayesianVariableImportance` was made.

Based on the authors previous work for linear mixed models (Arnstad 2024), a simulation study, case studies and a real world application, we have shown that the BVI method is a viable analogue to the existing frequentist methods. The method is able to produce plausible results for GLMMs with a complex covariance structure, while being simultaneously being computationally efficient. Hopefully, the BVI method can be used across various field and help researchers in their work. With relative variable importance being a topic of much interest and active research, recently also in the Bayesian framework, we believe that the BVI method can be further improved in the future.

SAMMENDRAG

Som en av de mest brukte statistiske metodene, har regresjonsmodeller en fundamental posisjon i statistikk. En nøkkeldel av regresjonsanalysen er å skaffe inferens om kovariatene som brukes til å modellere responsvariabelen, og tilegne kovariatene en *relativ viktighet*. For å gjøre dette, eksisterer flere metoder fra ulike perspektiver. Til tross for mange forskjellige metoder, har det ikke blitt oppnådd en konsensus, og den tradisjonelle fremgangsmåten med p-verdier har skapt en reproduserbarhetskrise i samfunns- og biomedisinsk forskning. Vårt bidrag for å bøte på dette, er å legge frem et Bayesiansk mål for relativ variabelviktighet. Dette målet er designet for at forskere skal tolke den statistiske modellen og dens resultater grundigere, i stedet for å slå seg til ro med konklusjoner basert på en forhåndsbestemt terskel.

Vårt mål, døpt *Bayesiansk Variabel Viktighet* (BVV), er implementert ved å overføre logikken fra mer etablerte, frekventistiske metoder, til det Bayesianske rammeverket. BVV er anvendbart på generaliserte lineære blandingsmodeller (GLBM) som har kontinuerlige, binomiske og Poisson fordelte responser. Kjernen i metoden er å benytte relativ vekting på kovariatene før en Bayesiansk GLBM blir konstruert. Dette produserer posteriore fordelinger av den relative viktigheten til alle kovariatene i modellen, samt de estimerte fordelingene til den marginale og betingede R^2. For å gjøre metodikken og målet lett tilgjengelig for forskere på tvers av fagfelt, ble en R pakke kalt `BayesianVariableImportance` laget.

Basert på forfatterens tidligere verk Arnstad (2024) for lineære blandingsmodeller, en simulasjonsstudie, case studier og en anvendelse på reelle data, har vi vist at BVV metoden er en levedyktig analog til eksisterende frekventistiske metoder. Metoden er i stand til å produsere plausible resultater for GLBM med komplekse kovariansstrukturer, samtidig som den er beregningsmessig effektiv. Forhåpentligvis kan BVV metoden bli brukt på tvers av ulike fagfelt og hjelpe forskere i deres arbeid. Med tanke på at relativ variabelviktighet er et område av stor interesse og aktiv forskning, også i det Bayesiansks rammeverket, tror vi at BVV metoden kan bli ytterligere forbedret i fremtiden.

PREFACE

This masters thesis concludes the Master of Science degree obtained from the program Physics and Mathematics, with a specialization in Industrial Mathematics, at the Norwegian University of Science and Technology (NTNU). In combination with the project thesis (Arnstad 2024), the masters thesis constitutes 45ECTS, and has been developed during the spring of 2024.

First and foremost I want to thank my supervisor Stefanie Muff, who has been critical in developing the thesis and has provided excellent guidance. I also want to express my gratitude to my fellow students at the study program Physics and Mathematics, who I have become close friends with and have learned a lot from. My time at NTNU has been fantastic, and something I will cherish for the rest of my life. Lastly, I want to thank my family and Emma, who have supported me throughout my studies. You have all been great people to have around, and I am forever grateful.

Kom igjen Troilljan!

CONTENTS

[1]This subsection is slightly modified from the project thesis (Arnstad 2024).
[2]This subsection is the same as in the project thesis (Arnstad 2024).

[3]This subsection is slightly modified from the project thesis (Arnstad 2024).

[4]This subsection is slightly modified from the project thesis (Arnstad 2024).

[5]A method for calculating the R^2 for Bayesian LMMs was proposed in Arnstad (2024, Chapter 2), however we see it fitting to include this extension in the methods chapter as it has been developed by the author for this thesis.

LIST OF FIGURES

LIST OF TABLES

CHAPTER

ONE

INTRODUCTION

Statistics as a mathematical field has a long history as a tool for characterizing social, economic and scientific phenomena. One of the most used statistical methods is regression analysis (Grömping 2015), which is used to model the relationship between a response variable and one or more covariates. To understand this relationship, researchers often want to determine whether a covariate is associated with the response, and to what extent. The exploration of this fundamental question has lead to a number of statistical methods trying to answer it. An agreement on a single method has not been reached, with the topic still being debated and actively researched.

A pioneer in statistics, Ronald Fisher, introduced the concept of the p-value almost one hundred years ago (Fisher 1925). To this day, the p-value is arguable the most widespread and used method, to determine if a covariate is *statistically significant* with respect to the response. Popularly, to determine statistical significance, a hypothesis test is performed and the resulting p-value is compared to a threshold level α. Typically, if the p-value is smaller than α, the covariate is considered statistically significant. However, this way of determining statistical significance is very prone to misinterpretations, and is subject to great criticism (Benjamin et al. 2018).

Recently, the social and biomedical sciences have been subject to a reproducibility crisis, in which published results cannot be reproduced (Blakeley B. McShane & Gelman 2019). One possible solution is suggested by 72 authors in Benjamin et al. (2018), which is to lower the typical significance level from $\alpha = 0.05$ to $\alpha = 0.005$. This could be a solution, however Blakeley B. McShane & Gelman (2019) sees this as a quick fix, which will not solve the underlying problem. Instead, it is proposed to simply abandon the term statistical significance, and not force results to be based on a threshold value which gives rigid and binary answers. The remedy, according to Blakeley B. McShane & Gelman (2019), is to rather interpret the p-value as a continuous measure of evidence, among many others. Going forward, the thesis will not consider statistical significance but rather statistical evidence to avoid the hazards by using such rigid interpretations.

There exists many other measures to compliment the p-value when assessing a

statistical model. As in Arnstad (2024), we list some of the most common measures

- **Effect sizes:** By looking at the squared value of the standardized regression coefficients, one can determine the effect size of the covariates. For independent covariates, the effect size coincides with the proportion of variance explained by the covariate. The effect sizes are a good measure for uncorrelated covariates, but falls short when correlation makes the coefficient estimates unstable.

- **Confidence intervals:** Confidence can be calculated for the effect sizes, and the interval can be used to determine a range of values that can be seen as statistically consistent with the data. The downside to using confidence intervals is that one effectively relies on the p-value to determine it, and therefore it does not provide any additional information than using the p-value.

- **Information critera:** The AIC (Akaike 1974) and the BIC (Schwarz 1978) are information criteria that use the likelihood function to compute a goodness of fit statistic. These can be used to assess the information contained in the model, and can be used to assess the unique information that one covariate contributes to the model. Also here, the critera fall short when correlation is present, as it cannot take shared information between covariates into account.

- **Bayes factor:** As an alternative to the p-value, Bayes factor can be used to assess the evidence from the data to either support a hypothesis or not. Bayes factor is therefore less rigid than the p-value, and is rather a continuous measure of evidence.

- **Decomposing the R^2:** The R^2 measures the variance explained in the response by the covariates. As such, it is a goodness of fit measure and can be decomposed into a share from each covariate in the model to determine the variance explained by each covariate. The R^2 is widely used and intuitive, but a proper decomposition of the value with correlated covariates is not straightforward.

The methods listed all have in common that they become less reliable for correlated covariates. As correlation is a common feature of many datasets, especially from the real world, this is a general problem that regression models are not well suited to handle (Grömping 2015).

By decomposing the R^2 value and assigning each variable with a share of explained variance in the response, we have a measure of the relative importance of each covariate (Grömping 2007). The problem of decomposing the R^2 in a sensible manner has lead to many different approaches. One of the most rigorous methods is the approach by Lindemann, Merenda and Gold (LMG) (Lindeman et al. 1980), which decomposes the R^2 value by considering all possible orderings of the covariates. As covariates are added to the null model, the average increase in R^2 for each permutation is calculated and each covariate is assigned a share of relative

variable importance. As the LMG method is very popular, it has been applied in dominance analysis (Budescu 1993) and coincides with the Shapley value in game theory ((Shapley 1953), Lipovetsky & Conklin (2001)). The LMG method has proven to be consistent for the linear regression, but is computationally expensive and therefore in some cases not feasible.

The relative weights method (Johnson (1966), Fabbris (1980), Genizi (1993)) can be seen as an approximation of the LMG method, to remedy the computational burden. By projecting the covariates into an orthogonal space and then conducting the analysis on the projected covariates, before transforming them back to the original covariate space, the relative weights method efficiently decomposes the R^2 value. At the cost of approximating rather than being more rigorous, the relative weights method is able to handle larger models and is therefore preferred if the LMG is not feasible (Grömping 2007).

Both the LMG method and the relative weights method are originally designed to decompose the R^2 for the linear regression. However, for many scenarios, a linear regression is not sufficient to model the relationship between the response and the covariates. Both models were extended in Matre (2022), so that they could be applicable for the random intercept models. Another problem is that there is little consensus on how to properly define the R^2 for more complex methods. A simple and intuitive definition of the R^2 for more general regression models was suggested by Nakagawa & Schielzeth (2013) and serves as the basis for the extensions of the LMG and relative weights method.

In the Bayesian framework, the field of relative variable importance is small. The LMG and relative weights method are both based on the frequentist framework, and the Bayesian framework has not been explored to the same extent. One possible line of action are the Generalized decomposition priors on R^2 (GDR2) (Aguilar & Bürkner 2024), which are based on the R^2-induced Dirichlet decomposition (R2D2) priors (Zhang et al. 2020). By placing a prior on the R^2 value, the R2D2 priors proceed with a Dirichlet decomposition of the R^2 value to be able to assign each covariate with a share of relative variable importance. The GDR2 priors are a generalization of the R2D2 priors which performs the decomposition using logsitic normal distributions (Aguilar & Bürkner 2024). At the time being, the GDR2 priors have been applied to the linear regression, with a focus on obtaining trustworthy predictions. Therefore, it is not clear how the GDR2 priors can be used for relative variable importance and if they can be extended to calculate relative variable importance for more complex regression models. Nonetheless, they serve as an important contribution to the Bayesian framework for relative variable importance.

The Bayesian framework has significant advantages when compared to the frequentist framework (Robert 2007) and has had a surge in popularity due to the recent years advancements in computational capabilities (Hackenberger 2019). The treatment of parameters as random variables in the Bayesian framework, rather than point estimates, naturally includes moments such as variance in the estimation procedure. As the Bayesian framework has become more available, researchers

can obtain more inference and thereby make more informed decisions. Therefore, we believe that the Bayesian framework provides a better platform for assessing the statistical evidence of covariates. A specific example of when Bayesian relative variable importance can be applied, is found in quantitative genetics. Decomposing the variance of regression models is done, and has been for a long time, to determine the heritability of phenotypic traits. Heritability is a key measure used to explain how the mean value of a trait changes, and can thereby help us better understand evolution. This particular example has been a great motivation behind our attempt to develop a relative variable importance measure in the Bayesian framework.

In this thesis, the field of relative variable importance will be explored, and we propose a Bayesian method for calculating the relative variable importance. The author proposed a Bayesian analogue to the extensions of the LMG and relative weights method in Arnstad (2024), which focused on linear mixed models. To address some of its limitations, we have now considered the larger class of generalized linear mixed models and for this class defined what we believe to be a sensible definition of the R^2. The method has applied the relative weights method in the Bayesian framework to fit a GLMM, returning results in the form of distributions. Our hope is that these distributional results will be easy to interpret and allow researchers to interpret the uncertainty, rather than using a threshold based method. For the method to be easily used, it has been implemented as a package in R. The package is called `BayesianVariableImportance` and is available, along with installation and usage examples, on the authors Github `https://github.com/AugustArnstad/BayesianVariableImportance`.

As regression models are perhaps the most used statistical modelling tool, the span of applications for the BVI method is wide. Given that the regression model is a suitable choice, the BVI method can be useful in fields such as biology, economics, social sciences, medicine and more. Research in many of these fields is crucial if we are to reach the 17 goals for sustainability set by the United Nations (United Nations 2023). We highlight the application of the BVI method for use in quantitative genetics. The overarching question for such analysis is to better understand evolution and how species develop when subject to different and changing environments. Answers to such questions is very useful to reach goal 14 and 15 of the United Nations sustainability goals, which is about life in the ocean and on land. The aim of goal 14 is to *Conserve and sustainably use the oceans, seas and marine resources for sustainable development* and goal 15 is about *Protecting, restoring and promoting sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss* (United Nations 2023). Further, we believe that the BVI method can give insight that can be useful for numerous of these goals, such as goal 3 about *Good health and well-being*, goal 7 about *Affordable and clean energy* and goal 13 about *Climate action*.

The structure of the thesis is as follows. In Chapter 2 we look at some background theory and put forth the theoretical results that will be used in the method. To describe our calculations, Chapter 3 presents the methodology and logic of our

contribution. Evaluating the method is done in Chapter 4, where we look at a simulation study, a case study and apply the method to a real dataset. We discuss the findings in Chapter 5 and conclude the thesis in Chapter 6. In Appendix A we give the link to the authors Github repository for the R package and the thesis, Appendix B contains a usage example of the R package and some miscalculations in the Appendix C.

CHAPTER

TWO

THEORY

Some sections in this chapter overlap with the authors project thesis (Arnstad 2024), which lead up to the masters thesis. Following the guidelines of the Institute of Mathematical Sciences, stating that sections need not be rewritten, some sections are the same (or slightly modified) as in the project thesis. To avoid problems relating to self plagirazation and clarify what is new in this thesis, the sections that are the same as in the project thesis have been assigned a footnote with the reference to the project thesis and a brief comment.

2.1 Linear regression

All regression models are based on the assumption that the response variable is influenced by one or more covariates. The relationship between the response and the covariate is assumed not to be deterministic, so we expect our modelling of the response to be influenced by some random error (Fahrmeir et al. 2013). This means that the response is treated as a random variable, and it is desirable to decompose the response into systematic components and random components.

2.1.1 Linear regression[1]

Assuming that an observed response y_i has a linear relationship with a covariate x_i is the basis for the simple linear regression. This can be modeled by the equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \tag{2.1}$$

where β_0 is the intercept, β_1 is the slope, and ε_i is the error term. The error term, or residuals, is assumed to be normally distributed with mean zero and variance σ^2, i.e. $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Generalizing to multiple covariates is straightforward by defining the $n \times p$ matrix $\mathbf{X}$ as a design matrix with the, including an intercept, p covariates in the columns and the n observations in the rows. With this definition, the linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \ , \tag{2.2}$$

[1]This subsection is slightly modified from the project thesis (Arnstad 2024).

where now $\mathbf{y} = (y_1, y_2, ..., y_n)$ is a vector of n responses, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ is a vector of coefficients including the intercept β_0, and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is a vector of error terms. The error terms are assumed to be independent and identically distributed (i.i.d.) with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of size $n \times n$. Consequently, the response $\mathbf{y}$ is conditionally independent given the covariates $\mathbf{X}$, i.e.

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

In practice, the coefficients $\boldsymbol{\beta}$ are estimated from the maximum likelihood estimation (MLE) method, given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \ . \tag{2.4}$$

2.1.2 Qualitative covariates

In many cases the covariates are qualitative, meaning they are categorical variables that can be grouped into different levels or factors. Qualitative covariates, unlike quantitative, cannot be measured numerically, and we must adjust our modelling to account for this. A common approach to model qualitative data is to include dummy variables, which are assigned a value 1 if the observation is in the respective category(factor) and 0 otherwise. Given N factors, it is standard practice to model $N - 1$ dummy variables and let one factor be captured by the intercept to uniquely determine the model. Dummy encoding in this way retains the properties of the linear regression, and are limited by the same assumptions. The model for the response y_i, assuming no quantitative covariates, from group j with dummy encoding is then given by

$$y_i = \beta_0 + \sum_{j=1}^{N-1} \beta_j x_{i,j} + \varepsilon_i \ , \tag{2.5}$$

where β_j denotes the factor coefficient of observation i and the dummy variable

$$x_{i,j} = \begin{cases} 1 & \text{if observation } i \text{ is in group } j \\ 0 & \text{otherwise} \end{cases} \ . \tag{2.6}$$

This way of modelling qualitative covariates is intuitive and easy to interpret, but it also assumes that factor specific effects are uniform and fixed across all levels and becomes cumbersome with many categorical covariates.

2.1.3 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 (2.2) is said to be *multicollinear* (Poole & O'Farrell 1971). From equation (2.4) 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

$$\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{X}\boldsymbol{\beta}) + \text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\beta} + \sigma_\varepsilon^2 = \sum_{j=1}^{p} \beta_j^2 v_j + \sum_{j=1}^{p-1}\sum_{k=j+1}^{p} \beta_j\beta_k\sqrt{v_j v_k}\rho_{jk} + \sigma_\varepsilon^2 \,, \tag{2.7}$$

(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 2.7 consist of the covariance between the covariates and the variance contribution from a single covariate is not immediately clear.

2.2 Variable importance in linear regression models

In a regression setting with multiple regression coefficients, it is often desirable to be able to assign each covariate with a measure of its relative importance with respect to the model. The relative importance of covariate $\mathbf{x}_i$ is defined as the standardized contribution to explained variance in the response $\mathbf{y}$ from $\mathbf{x}_i$ (Grömping 2007). 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 Relative importance measures

The coefficient of determination, R^2, is a widely used and intuitive summary statistic of goodness-of-fit and can also be used in model comparison. Conceptually, the R^2 quantifies how much variance in the response variable can be attributed to the covariates in the model. For the linear regression model, the R^2 is defined as

$$R^2 = 1 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{(\mathbf{y} - \overline{\mathbf{y}})^T(\mathbf{y} - \overline{\mathbf{y}})} = \frac{\text{Var}(\mathbf{y}) - \sigma_\varepsilon^2}{\text{Var}(\mathbf{y})} \,, \tag{2.8}$$

where $\overline{\mathbf{y}}$ is the mean vector of responses $\mathbf{y}$. Instead of referring to the R^2 value alone, going forward this thesis will focus on decomposing of the R^2 value and allocate a proportion of R^2 to the model covariates. This decomposition is done in order to assess the relative importance, or variance explained, of each covariate in the model. The special case of uncorrelated covariates in $\mathbf{X}$ gives

$$\text{Var}(\mathbf{y}) = \sum_{j=1}^{p} \beta_j^2 v_j + \sigma_\varepsilon^2 \,. \tag{2.9}$$

and provides a natural decomposition of the R^2 in terms of contribution from each covariate, as each predictor $\mathbf{x}_i$ contributes $\beta_i^2 v_i$ to the total response variance

(Grömping 2007). In (2.7) however, 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. As mentioned, it is the middle term that poses the problem of assigning importance to each covariate, since it is not immediately clear how to distribute the total response variance to each covariate, as some variance contributions in the response variance are shared among 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.

2.2.2 Naive decompositions of R^2

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

$$Y = X_1 + X_2 \ , \mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1 \ , \mathrm{Cov}(X_1, X_2) = 0.9 \ . \tag{2.10}$$

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

$$\mathrm{Ri}(\{2\}) = R^2(\{1, 2\}) - R^2(\{1\}) = 1 - \frac{\mathrm{Cov}(Y, X_1)^2}{\mathrm{Var}(Y)\mathrm{Var}(X_1)} = 1 - \frac{1.9^2}{3.8} \approx 0.05 \ , \tag{2.11}$$

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

$$\mathrm{Ri}(\{1\}) = R^2(\{1\}) - R^2(\{\emptyset\}) = \frac{\mathrm{Cov}(Y, X_1)^2}{\mathrm{Var}(Y)\mathrm{Var}(X_1)} = \frac{1.9^2}{3.8} \approx 0.95 \ . \tag{2.12}$$

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.

2.2.3 The Lindemann, Merenda and Gold(LMG) method

A method that handles correlation among covariates, and is frequently reinvented (Grömping 2007) from different approaches, is the Linemann, Merenda and Gold (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 $\mathrm{svar}(\{1\}|\emptyset)$. X_2 is added and allocated a share of $\mathrm{svar}(\{2\}|\{1\})$, and lastly X_3 is added and allocated a share of $\mathrm{svar}(\{3\}|\{1, 2\})$.

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 $\mathrm{svar}(\{1\}|\emptyset)$. X_3 is added and allocated

a share of svar($\{3\}|\{1\}$), and lastly X_2 is added and allocated a share of 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),

$$\text{LMG}(1) = \frac{1}{p!} \sum_{S \subseteq \{2,\ldots,p\}} n(S)!(p - n(S) - 1)!\text{svar}(\{1\}|S) \ , \tag{2.13}$$

where $n(S)$ is the number of regressors in S. Equation (2.13) averages the increase in R^2, 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 the great computational expense that the permutations require when p is large. The complexity is 2^{p-1} summations (Grömping 2007), and therefore, the LMG is not suitable for high dimensional models.

2.2.4 Relative weights method

A method that takes advantage of the straightforward decomposition of the variance when the fixed 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 design matrix $\mathbf{X}$ of the fixed effects 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 in equation (2.9).

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

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T} \ . \tag{2.14}$$

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

$$\|\mathbf{X} - \mathbf{Z}\|_F^2 = tr\left((\mathbf{X} - \mathbf{Z})^T(\mathbf{X} - \mathbf{Z})\right) \ , \tag{2.15}$$

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 $\boldsymbol{\beta_Z}$ as

$$\begin{aligned} \boldsymbol{\beta_Z} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}\mathbf{y} \\ &= \left((n-1)\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{V}^T\right)^{-1}\sqrt{n-1}\mathbf{V}\mathbf{U}^T\mathbf{y} \\ &= \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{U}^T\mathbf{y} \ . \end{aligned} \tag{2.16}$$

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 $\boldsymbol{\beta_Z}^{[2]}$. 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

$$\boldsymbol{\Lambda} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X} = (\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T \ , \tag{2.17}$$

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)

$$\mathrm{RI}(\mathbf{X}) = \boldsymbol{\Lambda}^{[2]}\boldsymbol{\beta_Z}^{[2]} \ , \tag{2.18}$$

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 $\boldsymbol{\Lambda}$.

2.3 Extenstions of the linear regression model

The linear regression model is a popular tool in many sciences, but it has limitations when one wants to model more complex structures between the response and covariates. We now generalize the concept of linear regression to be able to model more complex data structures.

2.3.1 Generalized linear models (GLMs)

The first step in expanding the linear regression model, is to allow the responses to be non-Gaussian. Instead of considering only the normal distribution as the distribution of the response, one can consider general responses belonging to the

exponential family. Assume that each we have N observations of the response y_i, where $i = 1, ..., N$, that are conditionally independent given the fixed effects. Then, y_i belongs to the univariate exponential family if

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y, \phi)\right) , \tag{2.19}$$

for some functions $a(\cdot), b(\cdot)$ and $c(\cdot)$, where θ_i is the parameter of the distribution, ϕ is a dispersion parameter and θ_i is a canonical parameter if ϕ is known (McCullagh & Nelder 1989). It is required that the function $b(\cdot)$ is twice differentiable, that the density function $f(y_i|\theta_i, \phi)$ is normalizable and that the support of $f(y_i|\theta_i, \phi)$ is not dependent on θ. Two key properties, expectation and variance, of the exponential family are given by

$$\begin{aligned} \mathbb{E}(Y|\theta) &= b'(\theta) \\ \mathrm{Var}(Y|\theta) &= a(\phi)b''(\theta) , \end{aligned} \tag{2.20}$$

where $b''(\theta)$ may also be refered to as the variance function (Fahrmeir et al. 2013) we have left out indexing, and a proof can be found in Appendix C. In the canonical form, the parameter θ_i coincides with the linear predictor η_i defined as

$$\theta_i = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} . \tag{2.21}$$

To connect the linear predictor η_i to the response, we define a monotonic, differentiable link function $g(\cdot)$ such that

$$\eta_i = g(\mu_i) = g(\mathbb{E}(Y_i)) . \tag{2.22}$$

For normally distributed responses, one typically uses the identity function as the link function, which yields the linear regression model. If one considers a binary response, the perspective changes. In a binary regression, one wishes to analyze how the covariates influence the probability

$$\pi_i = \mathbb{P}(y_i = 1|\mathbf{x}_i) = \mathbb{E}[y_i] . \tag{2.23}$$

This requires that $\mathbb{E}[y_i]$ lies in the interval $[0, 1]$ as it represents a probability measure. Therefore, the inverse of the link function must transform the linear predictor in such a way that the expectation fulfills this criteria (Fahrmeir et al. 2013). A popular choice of inverse link function is the logistic response function

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} , \tag{2.24}$$

yielding the logit link function

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i , \tag{2.25}$$

which will be further investigated later on. An intuitive interpretation of the coefficients can be made by noticing that odds

$$\frac{\pi_i}{1 - \pi_i} = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{1,i})...\exp(\beta_p x_{1,p}) , \tag{2.26}$$

is affected by the covariates in an exponential-multiplicative form (Fahrmeir et al. 2013). Another common regression type is regressing on count data. The most common way of modelling count data is by using the Poisson distribution, which assumes that the events occuring in a time interval or spatial region follow a Poisson process (McCullagh & Nelder 1989). The count of how many events y_i that happen in this time interval or region is said to follow a Poisson distribution with some rate $\lambda_i = \mathbb{E}[y_i]$. As the number of events occuring cannot be negative, the rate is also restricted to positive values. The common choice of inverse link function is therefore

$$\lambda_i = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{1,i})...\exp(\beta_p x_{1,p}) \ , \tag{2.27}$$

which means that the link function is then the logarithm of the rate (Fahrmeir et al. 2013), i.e.

$$\ln(\lambda_i) = \eta_i \ . \tag{2.28}$$

2.3.2 Linear mixed models (LMMs) [2]

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

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m)^T \ , \tag{2.29}$$

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 $\boldsymbol{\alpha}_j \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}) = \mathcal{N}_q(\mathbf{0}, \mathbf{Q}^{-1})$ where $\boldsymbol{\Sigma}$ is the $q \times q$ unknown covariance for the random effects assumed to be positive definite and $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ the corresponding precision matrix. If the random effects for each cluster are independent of each other, the covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_0^2, ..., \sigma_q^2)$. The linear mixed model now takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{2.30}$$

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 $\boldsymbol{\Sigma}_{kk} = \sigma_k^2$ and their covariance $\boldsymbol{\Sigma}_{k,l} = \sigma_{k,l}$, where $k, l = 1, ..., q$. In practice it is often easier to estimate the precision rather than the variance, so calculations often involve the precision matrix $\mathbf{Q}$ rather than the covariance matrix $\boldsymbol{\Sigma}$. 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

[2] This subsection is the same as in the project thesis (Arnstad 2024).

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

$$\mathrm{Var}(\mathbf{y}|\mathbf{X},\mathbf{U}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta}+\mathbf{U}\boldsymbol{\alpha}+\boldsymbol{\varepsilon}) = \mathbf{U}\mathrm{Var}(\boldsymbol{\alpha})\mathbf{U}^T+\sigma^2\mathbf{I} = \mathbf{U}\mathbf{G}\mathbf{U}^T+\sigma^2\mathbf{I}\ , \quad (2.31)$$

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 $\boldsymbol{\Sigma}_j$ 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.*

$$\mathbf{y}|\mathbf{X},\mathbf{U} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{U}\mathbf{G}\mathbf{U}^T+\sigma^2\mathbf{I})\ . \quad (2.32)$$

2.3.3 Generalized linear mixed models(GLMMs)

Now that we have expanded the linear regression in two different ways, the final step to complete the regression framework is to combine the LMM and GLM to obtain the GLMM. This is done by adding random effects to the linear predictor, such that

$$\theta_{i,j} = \eta_{i,j} = \mathbf{x}_{i,j}^T\boldsymbol{\beta} + \mathbf{u}_{i,j}^T\boldsymbol{\alpha}_j\ , \quad (2.33)$$

where $j = 1, ..., m$ denotes the cluster and $i = 1, ..., n_j$ denotes the observations in cluster j, $\mathbf{x}_{i,j}$ and $\mathbf{u}_{i,j}$ are the i-th columns of the submatrices $\mathbf{X}_j$ and $\mathbf{U}_j$ of the larger design matrices $\mathbf{X}$ and $\mathbf{U}$ respectively, for cluster j. The assumption of conditional independent observations $y_{i,j}$ is now conditional on the random effect as well as the covariates, and the conditional distribution of $y_{i,j}$ is still assumed to belong to the exponential family. The conditioning on the random effects is also present when choosing the appropriate link function, since one must now, in general, relate $\mathbb{E}[y_{i,j}|\mathbf{x}_{i,j}, \mathbf{u}_{i,j}, \boldsymbol{\alpha}_j]$ to the linear predictor $\eta_{i,j}$ (Fahrmeir et al. 2013). For the binary regression, this now means that the link function takes the form

$$\ln\left(\frac{\pi_{i,j}}{1-\pi_{i,j}}\right) = \ln\left(\frac{\mathbb{P}(y_{i,j}=1|\mathbf{x}_{i,j},\mathbf{u}_{i,j},\boldsymbol{\alpha}_j)}{\mathbb{P}(y_{i,j}=0|\mathbf{x}_{i,j},\mathbf{u}_{i,j},\boldsymbol{\alpha}_j)}\right) = \eta_{i,j}\ . \quad (2.34)$$

For the Poisson random intercept model with log link however, it is possible to define the model without conditioning on the random effects (Fahrmeir et al. 2013). This is done by noting that

$$\lambda_j = \exp(\mathbf{x}_j\boldsymbol{\beta} + \alpha_{0,j})\ , \quad (2.35)$$

where $\alpha_j \sim \mathcal{N}(0, \tau_0^2)$, has a log-normal distribution. This is a special case, in which the marginal model can be determined analytically. In general however, the marginal model is not analytically tractable and so obtaining statistical inference on the GLMMs become increasingly complex when compared to the LMM. Parameter estimation therefore calls for numerical methods such as iterated reweighted least squares in the likelihood framework, or MCMC methods in the Bayesian framework, to obtain inference.

2.4 Extending R^2 to GLMMs

As we generalized the linear regression to LMMs, GLMs and GLMMs, we have to find a generalization of the concept of R^2 in order to generalize the concept of variable importance. This is fundamental to be able to propose a method for decomposing the R^2 and thereby assigning relative importance to covariates. However, the task of determining the R^2, and decomposing it, is not a trivial task in the linear regression case and becomes even more complex in the case of GLMMs. Many extensions have been proposed, but due to a variety of theoretical problems and/or computational difficulties, no consensus has been reached on a framework for calculating the R^2 for GLMMs (Nakagawa & Schielzeth 2013). To get an overview of the status quo for R^2, we will follow the paper by Nakagawa & Schielzeth (2013) and go through the different components added to the linear regression to compose the GLMMs.

2.4.1 R^2 for GLMs

Recalling the definition of the R^2 from Equation (2.8), we now generalize this to the GLMs. This topic has been subject to significant research, (see for example Maddala (1983), Cameron & Windmeijer (1997), Menard (2000), Nakagawa & Schielzeth (2013)). The methods first suggested was based on the likelihood function of the model to be analyzed. We will not implement such methods, as they are not suitable for the full generalization to be made later on, however they are important in building a framework for the R^2 value and are therefore included. To illustrate the likelihood based generalization of the R^2 value to GLMs, consider the deviance $\mathcal{D}(\mathbf{y}|\theta)$ function which is defined as twice the difference between the log likelihood of the **saturated model** and the log-likelihood of the model of interest (McCullagh & Nelder 1989). The saturated model denotes the model of the maximum achievable log likelihood, and therefore fits the data perfectly. For a linear regression, with $\theta = (\boldsymbol{\beta}, \sigma^2)$, we would therefore obtain

$$\begin{aligned}
\mathcal{D}(\mathbf{y}|\hat{\theta}) &= -2\left(\ln(\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\mathbf{y})) - \ln(\hat{\mathcal{L}}(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}|\mathbf{y}))\right) = -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}|\mathbf{y})\right) \\
&= -2\left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{n}{2}\ln(2\pi\sigma^2)\right) \\
&= \frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= 1 - R^2 \ ,
\end{aligned} \tag{2.36}$$

where $\hat{\mathcal{L}}$ denotes the saturated model. Optimally, it is desirable to have as small deviance as possible while at the same time having a model that is not too complex. The best practice of the deviance is not as model fit, but rather model comparison, where one compares models through the reduction in deviance (McCullagh & Nelder 1989). Since the model of interest is nested within the saturated model, the deviance coincides with the likelihood ratio test. By comparing the model of interest to the **null model**, the simplest fit possible, one obtains for the linear

regression

$$\begin{aligned}\mathcal{D}(\mathbf{y}|\hat{\theta}) - \mathcal{D}(\mathbf{y}|\theta_0) &= -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}|\mathbf{y})\right) + 2\left(l(\boldsymbol{\beta}_0, \sigma_0^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \hat{\sigma^2})\right) \\ &= -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\boldsymbol{\beta}_0, \sigma_0^2|\mathbf{y})\right) \\ &= -\frac{2}{2\sigma^2}\left(-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{y} - \overline{\mathbf{y}})^T(\mathbf{y} - \overline{\mathbf{y}})\right) \\ &= 1 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\ &= R^2 \ .\end{aligned} \tag{2.37}$$

This is the foundation for the definitions of the generalization of R^2 to GLMs (Nakagawa & Schielzeth 2013), which primarily rely on a ratio of the maximum likelihood of the model of interest and null model. However, in Nakagawa & Schielzeth (2013), two different R^2 measures are proposed as

$$R_G^2 = \left[1 - \left(\frac{\mathcal{L}_0}{\mathcal{L}_M}\right)^{2/n}\right] \frac{1}{1 - (\mathcal{L}_0)^{2/n}} \tag{2.38}$$

and

$$R_D^2 = 1 - \frac{-2\ln(\mathcal{L}_M)}{-2\ln(\mathcal{L}_0)} \tag{2.39}$$

where n denotes the total sample size, $\mathcal{L}_0$ is the likelihood of the null model and $\mathcal{L}_M$ is the likelihood of the model of interest. The reason why we will not apply likelihood based R^2 measures is that when generalizing to the larger class of GLMMs, it is often desirable to do parameter estimation using the restricted maximum likelihood (REML) instead of the maximum likelihood (ML) (Fahrmeir et al. 2013). The REML estimator transforms the data, meaning that models cannot be compared when fitted, and therefore the proposed measure of R^2 is not applicable to the REML framework (Nakagawa & Schielzeth 2013). However, the extension of the R^2 measure to the larger class GLMMs will also cover the special case of GLMs, and is discussed further below in Section 2.4.3.

2.4.2 R^2 for LMMs and random slope models

In the LMMs, as opposed to the linear regression, one wishes to estimate two or more variance components instead of only the residual error variance. This increases complexity and makes the task of assigning relative importance to the covariates even more challenging. Initially, a definition was proposed for the R^2 in the LMMs that included fixed effects separately and then estimated the reduction in each variance component (Nakagawa & Schielzeth 2013, refering to Raudenbush & Bryk 1986, 1992). This violated a key condition, as adding a covariate could decrease σ_ε^2 while at the same time increasing σ_α^2, which can lead to a negative R^2. To handle this problem, Snijders & Bosker (1994) (Nakagawa & Schielzeth 2013) proposed a new definition of the R^2, dividing it into two components R_1^2 and R_2^2. Considering the simple random intercept model in scalar form;

$$y_{i,j} = \beta_0 + \mathbf{x}_{i,j}^T\boldsymbol{\beta} + \alpha_j + \varepsilon_{i,j} \ , \tag{2.40}$$

where $y_{i,j}$ denotes the ith observation in cluster j, β_0 is the fixed intercept, $\mathbf{x}_{i,j}$ is the column vector containing the covariates for the ith observation in cluster j, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effects, α_j is the random intercept for cluster j and $\varepsilon_{i,j}$ is the residual error for the ith observation in cluster j, the two R^2 components can be expressed in two ways, with the first being

$$\begin{aligned} R_1^2 &= 1 - \frac{\text{Var}(y_{i,j} - \hat{y}_{i,j})}{\text{Var}(y_{i,j})} = 1 - \frac{\sigma_\varepsilon^2 + \sigma_\alpha^2}{\sigma_{\varepsilon 0}^2 + \sigma_{\alpha 0}^2} \\ \hat{y_{i,j}} &= \beta_0 + \mathbf{x}_{i,j}^T \beta \ , \end{aligned} \tag{2.41}$$

where $\sigma_{\varepsilon 0}^2$ and $\sigma_{\alpha 0}^2$ denote the residual and random effect variances of the null model respectively (Nakagawa & Schielzeth 2013) and $\hat{y}_{i,j}$ denotes the fitted value of observation i in the jth cluster. Similarily, the second component is defined as

$$\begin{aligned} R_2^2 &= 1 - \frac{\text{Var}(y_j - \hat{\overline{y}_j})}{\text{Var}(\overline{y_j})} = 1 - \frac{\sigma_\varepsilon^2 + \sigma_\alpha^2/k}{\sigma_{\varepsilon 0}^2 + \sigma_{\alpha 0}^2/k} \\ k &= \frac{M}{\sum_{j=1}^{M} \frac{1}{m_j}} \ , \end{aligned} \tag{2.42}$$

where $\overline{y_j}$ is the mean for each observed value of the jth cluster, $\hat{\overline{y}_j}$ is the mean of the fitted values for the jth cluster, k is the harmonic mean of the number of observations per cluster, m_j is the number of observations for the jth cluster and M is the total number of clusters (Nakagawa & Schielzeth 2013). Note that we have formulated the above definitions in a notation corresponding to our previous formulation of the LMM, and therefore uses clusters in general, whereas Nakagawa & Schielzeth (2013) refers to a cluster as being individuals with repeated measurements. The reason for dividing the R^2 into two components, is that intuitively the R_1^2 measures the within cluster variance explained and the R_2^2 measures the between cluster variance explained (Nakagawa & Schielzeth 2013). However, three problems arise when using this definition of the R^2 for LMMs. Firstly, the R_1^2 and R_2^2 can decrease in large models, secondly, R_1^2 and R_2^2 have not been generalized to more complex LMMs with more than one random effect and lastly, it is not clear how to generalize the R_1^2 and R_2^2 to GLMMs (Nakagawa & Schielzeth 2013). To overcome these obstacles, Nakagawa & Schielzeth (2013) proposes a new formulation of the R^2 measure. Consider a general random intercept model as defined in Section 2.3.2, with q random intercepts, as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{2.43}$$

with the parameters of interest being $\boldsymbol{\beta}$ and the variance components σ_ε^2 and σ_i^2 for the $i = 1, ..., q$ clusters. Then define the variance of the fixed effects as

$$\sigma_f^2 = \text{Var}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^T \text{Var}(\mathbf{X})\boldsymbol{\beta} \ , \tag{2.44}$$

and further define the R^2 for the LMM as

$$R^2_{\text{LMM(m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{i=1}^{q} \sigma_i^2 + \sigma_\varepsilon^2} \ . \tag{2.45}$$

This definition of the R^2_{LMM} represents the marginal R^2_{LMM}, denoted by (m), as it measures the proportion of the variance explained by the fixed effects alone, whereas the conditional R^2_{LMM} can be defined as

$$R^2_{\text{LMM(c)}} = \frac{\sigma_f^2 + \sum_{i=1}^{q} \sigma_i^2}{\sigma_f^2 + \sum_{i=1}^{q} \sigma_i^2 + \sigma_\varepsilon^2} \ . \tag{2.46}$$

By inspection it is clear that this definition will never lead to negative values of the R^2_{LMM}. It may occur that the R^2_{LMM} value may decrease when adding more covariates to the model, although Nakagawa & Schielzeth (2013) argues that this is unlikely. This definition now covers the random intercept model, but has not taken into account the possibility of having a LMM with a random slope. To further extend the R^2 to the random slope model, Johnson (2014) proposes a method for computing the mean random effect variance. Consider the simple random intercept and slope model,

$$y_{i,j} = \beta_0 + \mathbf{x}_{i,j}^T \boldsymbol{\beta} + \alpha_{0,j} + \alpha_{1,j} x_{i,j} + \varepsilon_{i,j} \ , \tag{2.47}$$

where the same notation is used as in (2.40) with $\boldsymbol{\alpha_j} = (\alpha_{0,j}, \alpha_{1,j})$ being the random effect, $\alpha_{0,j}$ denoting the random intercept and $\alpha_{1,j}$ now denoting the random deviation from the global slope β_1, for cluster j. The general assumption on the random effects are that

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\alpha_0} & \sigma_{\alpha_0,\alpha_1} \\ \sigma_{\alpha_0,\alpha_1} & \sigma^2_{\alpha_1} \end{pmatrix} \right) \ , \tag{2.48}$$

where $\sigma^2_{\alpha_0}$ and $\sigma^2_{\alpha_1}$ are the variances of the random intercept and random slope respectively, and $\sigma_{\alpha_0,\alpha_1}$ is the covariance between the random intercept and random slope. Thus, we have three variance components of interest ($\frac{q(q+1)}{2}$ for q random effects) to estimate. When inspecting the variance of the random part in the model, we see that it has a dependence on the covariates, as illustrated by

$$\begin{aligned} \text{Var}(\alpha_{0,j} + \alpha_{1,j} x_{i,j}) &= \text{Var}(\alpha_{0,j}) + 2x_{i,j}\text{Cov}(\alpha_{0,j}, \alpha_{1,j}) + x^2_{i,j}\text{Var}(\alpha_{1,j}) \\ &= \sigma^2_{\alpha_0} + 2x_{i,j}\sigma_{\alpha_0,\alpha_1} + x^2_{i,j}\sigma^2_{\alpha_1} =: \sigma^2_{r,i,j} \ , \end{aligned} \tag{2.49}$$

where we define $\sigma^2_{r,i,j}$ as the variance of the random effect $\boldsymbol{\alpha}$ for observation i in the jth cluster. The method proposed by Johnson (2014) is to first estimate all the variance components, and then view the specific random effect as a normal mixture distribution of the random intercept and random slope. This mixture distribution is characterized as having a common mean of zero, and, if all values of the associated covariate $x_{i,j}$ are unique, having N different variances with N being the total number of observations. A mixture distribution with constant mean, has a variance which equals the mean of the individual variances in the distribution (Johnson 2014, citing Behboodian 1970). The proposed variance of the random effect $\boldsymbol{\alpha}$, is therefore the mean of the variance components in $\boldsymbol{\alpha}$, *i.e.*

$$\overline{\sigma^2_r} = \frac{1}{N} \sum_{j=1} \sum_{i=1} \left(\sigma^2_{r,i,j} \right) \ . \tag{2.50}$$

This formulation can be generalized in the case of q random effects, where each random effect has an associated design matrix $\mathbf{U_j}$ and covariance matrix $\mathbf{Q}$ as in Section 2.3.2, so that for each random effect r we have

$$\overline{\sigma^2_r} = \text{Tr}(\mathbf{U_j Q U_j}^T), \quad r = 1, ..., q \ . \tag{2.51}$$

To finally obtain the proposed R^2 for the general LMM, Johnson (2014) uses this estimate in the definition given by Nakagawa & Schielzeth (2013), to obtain

$$R^2_{\text{LMM(m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \sigma_\varepsilon^2} \ , \tag{2.52}$$

and

$$R^2_{\text{LMM(c)}} = \frac{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2}}{\sigma_f^2 + \sum_{i=1}^{q} \overline{\sigma_r^2} + \sigma_\varepsilon^2} \ , \tag{2.53}$$

as the marginal and conditional R^2_{LMM} respectively. For the random intercept model with $\sigma^2_{r,i,j} = \sigma^2_r$, this definition corresponds to the definition by Nakagawa & Schielzeth (2013) as

$$\overline{\sigma_r^2} = \frac{1}{N} \sum_{j=1} \sum_{i=1} \left(\sigma^2_{r,i,j} \right) = \sigma^2_{r,i,j} = \sigma^2_r \ . \tag{2.54}$$

The R^2_{LMM} proposed by Johnson now lets us compute the R^2 for general LMMs, however it is argued in Johnson (2014) whether the improved R^2 estimate by taking the random slope into account is worth the added complexity and computational cost.

2.4.3 R^2 for GLMMs

The final step towards a complete generalization for the R^2 value of regression models is to extend it to the GLMMs. When considering non-normal responses, the link function introduces an aspect not yet discussed, which is to define the residual variance. One can divide the residual variance σ^2_ε into three components, namely distribution specific variance, multiplicative dispersion and additive dispersion (Nakagawa & Schielzeth 2013). The distribution specific variance is inherited from the link function used, and is therefore known before analysis is done. However, the multiplicative and additive dispersion is modelled to account for the variance present that exceeds the distribution specific variance, *i.e.* overdispersion (Nakagawa & Schielzeth 2010). Therefore, one must specify upon implementation on what scale the overdispersion is to be modelled. The multiplicative dispersion, denoted by ω, is overdispersion on the response (data) scale and modelled as a distinct parameter of the assumed distribution of the response $\mathbf{y}$ (Nakagawa & Schielzeth 2010). Conversely, the additive dispersion, denoted by e, is overdispersion on the latent scale and introduced to the model as an additional random effect in the linear predictor (Nakagawa & Schielzeth 2010). Defining the residual variance now depends on the choice of dispersion modelling, and is either defined as

$$\sigma^2_\varepsilon = \omega \sigma^2_d \tag{2.55}$$

or

$$\sigma^2_\varepsilon = \sigma^2_d + \sigma^2_e \ , \tag{2.56}$$

for multiplicative and additive dispersion respectively. With the residual variance defined, the generalization to of the R^2 to GLMMs (thereby also the GLMs) follows

the same logic as the LMMs, and $R^2_{\text{GLMM (m)}}$ is defined as

$$R^2_{\text{GLMM(m, m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \omega\sigma_d^2} \ , \tag{2.57}$$

and

$$R^2_{\text{GLMM(m, a)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \sigma_d^2 + \sigma_e^2} \ , \tag{2.58}$$

where the same notation as before is used and the subscripts (m, m) and (m, a) denote the multiplicative and additive dispersion respectively. The conditional R^2_{GLMM} can be defined in a similar manner,

$$R^2_{\text{GLMM(c, m)}} = \frac{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2}}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \omega\sigma_d^2} \ , \tag{2.59}$$

and

$$R^2_{\text{GLMM(c, a)}} = \frac{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2}}{\sigma_f^2 + \sum_{r=1}^{q} \overline{\sigma_r^2} + \sigma_d^2 + \sigma_e^2} \ , \tag{2.60}$$

completing the generalization.

2.5 The Bayesian framework

So far, we have introduced statistical concepts without considering the framework in which they are used. We now expand the theory to consider the Bayesian framework, which is the framework used in this thesis.

2.5.1 General idea

The Bayesian framework stems from the notorious theorem developed by Thomas Bayes, (Bayes & Price 1763), which states that for events A and B, with nonzero probability of occuring, we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \ . \tag{2.61}$$

This can be generalized to also apply to distributions of continuous random variables, namely that

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \ , \tag{2.62}$$

where $\pi(\boldsymbol{\theta}|\mathbf{y})$ is called the posterior distribution of $\boldsymbol{\theta}$, $\pi(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, or sampling, distribution of $\mathbf{y}$, $\pi(\boldsymbol{\theta})$ is the prior distribution of the parameters and $\pi(\mathbf{y}) = \int \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) < \infty$ is the marginal distribution of the data, which is required to be finite in order to have a proper posterior distribution (Gelman et al. 2015). In practice, the marginal distribution is often omitted and one only consider the proportionality of (2.62), i.e.

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ . \tag{2.63}$$

In the context of statistical analysis, with $\boldsymbol{\theta}$ being the parameter vector of the family of models for the random variable Y under investigation, $\pi(\boldsymbol{\theta}|\mathbf{y})$ is interpreted as the distribution of the parameters given the data $\mathbf{y}$. This is the key element that separates the Bayesian framework from the frequentist framework, as the parameter $\boldsymbol{\theta}$ is now treated as random variable instead of being point estimates.

2.5.2 Prior and posterior distributions

Generally, a Bayesian model is built by first introducing some prior knowledge through the prior distribution $\pi(\boldsymbol{\theta})$ and supplementing this with the likelihood function $\pi(\mathbf{y}|\boldsymbol{\theta})$. The prior distribution must be chosen based on the prior knowledge available, and can either be informative, noninformative or weakly informative (Gelman et al. 2015). As a compromise of the information in the prior and the likelihood of the data, the posterior distribution is obtained. The resulting posterior will be different from analysis to analysis, but some general relations between the prior and posterior are discussed in Gelman et al. (2015). In particular, it is stated that *the posterior variance is on average smaller than prior variance by an amount that depends on the variation in posterior means over the distribution of possible data* (Gelman et al. 2015). This further means that if one wishes to reduce the variability in the posterior, the potential for this lies in reducing the variation of possible posterior means. The posterior distribution will therefore, in general, be a compromise between the prior and the likelihood, which with increasing sampling size will be increasingly influenced by the likelihood (Gelman et al. 2015).

2.5.3 Penalising complexity (PC) priors

Prior distributions pose a great feature by allowing for inclusion of prior information, but also a great challenge in that they must be chosen with care. As the theory of this is wast and out of the scope for this thesis, we will be mostly concerned with the penalising complexity priors proposed in Simpson et al. (2017). In this paper, four main principles are desirable to follow when choosing a prior distribution, namely

1. **Occams razor** - If there is no evidence for a complex mode, a base model should be preferred.

2. **Measure of complexity** - The measure of model complexity is defined as $d(f||g) = \sqrt{2\text{KLD}(f||g)}$ where $\text{KLD}(f||g)$ denotes the Kullback-Leibler divergence (Simpson et al. 2017, for more information).

3. **Constant rate penalisation** - The penalisation, i.e. the decay of prior mass, grows as the complexity grows, but it is desirable that this growth is constant.

4. **User defined scaling** - Assuming that the user has an idea of the magnitude of the parameter of interest, the user should be able to scale the prior accordingly.

The PC priors therefore pose interpretable, applicable priors which are consistent with the above principles, and are therefore a practical choice for the Bayesian framework (Simpson et al. 2017). Particularly, for the case of a linear mixed model with a Gaussian random effect $\alpha \sim \mathcal{N}(0, \sigma^2\mathbf{R}) = \mathcal{N}(0, \tau^{-1}\mathbf{Q}^{-1})$, the base model of the PC priors corresponds to the case where the precision $\tau = 0$ and the prior for τ takes the form

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right), \quad \tau, \lambda > 0 \ . \tag{2.64}$$

To specify λ, the user is required to supply the values (U, a) such that $\mathbb{P}(1/\sqrt{\tau} > U) = a$. This defines the scaling parameter of principle 4 and leads to $\lambda = -\ln(a)/U$ (Simpson et al. 2017). When fitting additive models, thereby modelling additive overdispersion, using PC priors is a natural choice (Gómez-Rubio 2020).

2.5.4 Hierarchical Bayesian modelling

When modelling in the Bayesian framework, the posterior distribution of the parameter $\boldsymbol{\theta}$ given the data is what one wants to infer. For many applications, $\boldsymbol{\theta}$ is a high dimensional vector, with naturally connected entries (Gelman et al. 2015). It may therefore be reasonable to assume that the parameters themselves are drawn from a population distribution, which can further be modelled by what is called hyperparameters. The main idea is that the prior $\pi(\boldsymbol{\theta})$ itself contains a hierarchical structure and can be split into levels of conditional prior distributions, i.e. $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})$ for some hyperparameter $\boldsymbol{\phi}$ (Robert 2007). Assuming that the data $\mathbf{y}$ depends only on the parameter $\boldsymbol{\theta}$, and that $\boldsymbol{\theta}$ depends on the hyperparameters $\boldsymbol{\phi}$, we can write the joint posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}, \phi)\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) \ , \tag{2.65}$$

where $\pi(\boldsymbol{\phi})$ is a prior placed on the hyperparameters. This hierarchical structure allows us to first estimate the population distribution using the hyperparameters, and then estimate the parameters of interest using the population distribution, instead of estimating each component of $\boldsymbol{\theta}$ separately (Gelman et al. 2015). It may be practical to view the model in three parts and consider an example with a tractable posterior distribution. Let the observational model be $\pi(\mathbf{y}|\boldsymbol{\theta})$ be defined as

$$y_i|\theta_i \sim \text{Po}(\theta_i) \ , i = 1, ..., n \ , \tag{2.66}$$

for conditionally independent observations y_i given the parameters θ_i. Define then the latent model $\pi(\boldsymbol{\theta}|\boldsymbol{\phi})$ as

$$\theta_i|\phi \sim \text{Gamma}(\alpha, \beta) \ , \tag{2.67}$$

for conditionally independent parameters θ_i given the hyperparameters α, β. Lastly, consider the hyperpriors $\pi(\boldsymbol{\phi})$ as

$$\alpha \sim \text{Exp}(a) \text{ and } \beta \sim \text{Gamma}(b, c) \ , \tag{2.68}$$

The full posterior density now reads

$$\pi(\boldsymbol{\theta}, \alpha, \beta|\mathbf{y}) \propto \underbrace{\prod_{i=1}^{n} \theta_i^{y_i} e^{-\theta_i}}_{\text{Po}(\theta_i)} \underbrace{\prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\beta)} \theta_i^{\alpha-1} e^{-\beta\theta_i}}_{\text{Gamma}(\alpha,\beta)} \underbrace{\alpha^{a-1}e^{-\alpha}}_{\text{Exp}(a)} \underbrace{\beta^{b-1}e^{-c\beta}}_{\beta\sim\text{Gamma}(b,c)} \ , \tag{2.69}$$

which can be used to make inference about the parameters of interest. This hierarchical structure is similar to that of the GLMM and is therefore a natural way of modelling a Bayesian GLMM. To set up a Bayesian GLMM, consider again observations $\mathbf{y}$ having the density function in (2.19) with dispersion parameter ϕ and associated linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} \ , \tag{2.70}$$

where we assume that $\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{Q}^{-1})$ for some precision matrix $\mathbf{Q} = \mathbf{Q}(\boldsymbol{\rho})$ dependent on the hyperparameter $\boldsymbol{\rho}$. Then, to define the model, a prior must be assigned to the likelihood specific parameter ϕ, the fixed effects coefficients $\boldsymbol{\beta}$, and the variance components of the random effects $\boldsymbol{\rho}$. For a general GLMM belonging to the exponential family defined in (2.19), the posterior can be written out as

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho}|\mathbf{y}) &\propto \left(\prod_{j=1}^{m} \pi(\mathbf{y}_j|\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho}) \right) \pi(\boldsymbol{\alpha}|\boldsymbol{\rho})\pi(\boldsymbol{\beta})\pi(\phi)\pi(\boldsymbol{\rho}) \ , \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}(\boldsymbol{\rho})\boldsymbol{\alpha} + \sum_{j=1}^{m} \ln \pi(\mathbf{y}_j|\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) \right) |\mathbf{Q}(\boldsymbol{\rho})|^{1/2}\pi(\boldsymbol{\beta})\pi(\phi)\pi(\boldsymbol{\rho}) \ , \end{aligned} \tag{2.71}$$

where the vector $\mathbf{y}_j$ denotes the jth cluster of observations (Fong et al. 2010).

2.5.5 R^2 in the Bayesian framework[3]

When working in the Bayesian framework, the definition of R^2 for the linear regression is not as straightforward as in the classical framework. As parameters are not treated as fixed, but as random variables, the R^2 value will also be a random variable. A possible remedy to this could be to use the posterior mode of the parameters $\boldsymbol{\beta}$ in (2.8), however Gelman et al. (2017) states two conflicts that this poses. Firstly, the use of point estimates to calculate statistics in the Bayesian framework rejects the fundamental uncertainty of the Bayesian framework. Secondly, when the parameters are estimated in a Bayesian framework, there is no guarantee that the $R^2 \in [0, 1]$, reducing its intuitive interpretability. In Gelman et al. (2017) a definition of the R^2 for the Bayesian linear regression is proposed. Consider a draw s of the parameters $\boldsymbol{\beta}$ from the posterior distribution. Then, the proposed definition is

$$R_s^2 = \frac{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\intercal\mathbf{X}} \boldsymbol{\beta}_s + \sigma_s^2} \ , \tag{2.72}$$

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 (2.72) one would eventually obtain also an approximation of the distribution for the R^2 value (Gelman et al. 2017).

[3]This subsection is slightly modified from the project thesis (Arnstad 2024).

2.5.6 R2D2 and GDR2 priors as variable importance measures

Although the field of Bayesian variable importance is small, there are some methods that can be contextualized so that they can be used as Bayesian variable importance measures. More specifically, the R^2-induced Dirichlet decomposition (R2D2) priors are originally applied as shrinkage priors and used to obtain reliable predicitons in high dimensional linear regression models, but can be interpreted as a variable importance measure (Zhang et al. 2020). Using the same definition of R^2 as that of Gelman et al. (2017), the R2D2 prior directly places a prior on the marginal or conditional R^2 value. We will consider the marginal R^2 prior, which assumes the marginal R^2 value to follow a Beta distribution (Zhang et al. 2020). The scenario introduced in Zhang et al. (2020) is a linear regression model where we consider a prior for $\boldsymbol{\beta}$ such that $\mathbb{E}[\boldsymbol{\beta}] = 0$ and $\text{cov}(\boldsymbol{\beta}) = \sigma^2\Lambda$ where Λ is a diagonal matrix with the diagonal elements $\lambda_1, ..., \lambda_p$. From the calculations in Zhang et al. (2020) and following the definition of the R^2 in Gelman et al. (2017), one can write

$$\text{Var}(\mathbf{x}^T\boldsymbol{\beta}) = \sigma^2 \sum_{j=1}^{p} \lambda_j \tag{2.73}$$

and

$$R^2 = \frac{\text{Var}(\mathbf{x}^T\boldsymbol{\beta})}{\text{Var}(\mathbf{x}^T\boldsymbol{\beta}) + \sigma^2} = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{p} \lambda_j + 1} := \frac{W}{W+1} \tag{2.74}$$

where W is the sum of the diagonal elements of Λ. Assuming that $W \sim BP(a, b)$ where BP denotes the Beta prime distribution, is equivalent to assuming that $R^2 \sim Beta(a, b)$ value follows a Beta distribution (Zhang et al. 2020). Further, Zhang et al. (2020) expresses each $\lambda_j = \phi_j\omega$ with $\sum_{j=1}^{p} \lambda_j = 1$ such that $W = \sum_{j=1}^{p} \phi_j\omega = \omega$. This is the key element that makes the R2D2 prior analogous to a variable importance measure, because ω represents the total prior variability and ϕ_j represents the proportion of the total variability allocated to covariate j for $j = 1, ..., p$ (Zhang et al. 2020). Now, to decompose the model R^2 based on these priors, it is proposed to let $\boldsymbol{\phi} = (\phi_1, ..., \phi_p) \sim Dir(a_\pi, ..., a_\pi)$ where Dir denotes the Dirichlet distribution, and a_π is a concentration parameter (Zhang et al. 2020). This means that ϕ_j is estimated via a Dirichlet decomposition which is analogous to assigning each covariate with a share of relative variable importance. The concentration parameters can be seen as an a priori importance for the covariates (Aguilar & Bürkner 2024) and larger a_π leads to smaller variance of ϕ_j and produces a more uniform $\boldsymbol{\phi}$. Conversely, a smaller a_π leads $\boldsymbol{\phi}$ to have some components with a larger ϕ_j (Zhang et al. 2020). The authors then show the prior on $\boldsymbol{\beta}$ that is induced by the prior on R^2 and further develop the theory and list properties of the R2D2 priors (Zhang et al. 2020). An obvious advantage with the R2D2 priors for the marginal R^2 is that they allow for more extensive asymptotic analysis of both the prior and posterior distributions (Zhang et al. 2020). However, the Dirichlet distribution has some limitations that makes in unsuitable for many applications. Firstly, when multiple covariates compete for the shares of importance, the Dirichlet distribution has a tendency to gravitate this competition towards a negative dependency structure, which is completely determined by the mean of each component (Aguilar & Bürkner 2024). Further, the Dirich-

let distribution is not very flexible, and therefore struggles to model correlation structures between covariates (Aguilar & Bürkner 2024, and references therein). The Dirichlet distribution enforces a high number of constraints on the covariance structure, making some covariance structures impossible to model (Aguilar & Bürkner 2024). Therefore, Aguilar & Bürkner (2024) propose what they call Generalized Decomposition Priors on R^2 (GDR2). The GDR2 priors address some limitations in the R2D2 priors, and suggests to rather place a Logisitic Normal (LN) prior on the parameters $\boldsymbol{\phi}$. This means that $\boldsymbol{\phi} \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and instead of specifying the parameter a_π, we must now specify the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ (Aguilar & Bürkner 2024). The authors suggest to automate the process of choosing prior values for the mean and covariance of $\boldsymbol{\phi}$ by what they call *prior matching*. By letting $f \sim Dir(\alpha)$ for a fixed $\alpha = a_\pi$ and $g \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the Kullback Liebler divergence (Kullback & Leibler 1951) between f and g can be minimized. In Aguilar & Bürkner (2024) a closed form for the minimizers is obtained as

$$\begin{aligned} \boldsymbol{\mu}_k^* &= \delta(\alpha_k) - \delta(\alpha_K) \\ \sigma_{kk}^* &= \varepsilon(\alpha_k) + \varepsilon(\alpha_K) \\ \sigma_{kj}^* &= \varepsilon(\alpha_k) \quad k \neq j \ , \end{aligned} \tag{2.75}$$

for $k = 1, ..., K$ where α_K is a reference unit, $\delta(\alpha_k)$ denotes the digamma function and $\varepsilon(\alpha_k) = \delta'(\alpha_k)$ denotes the first polygamma function (Abramowitz & Stegun 1972, pages 258-260). We will not go further into detail on R2D2 and GDR2 priors, but rather focus on the interpretation of the R2D2 and GDR2 prior as variable importance measures by assessing the values of $\boldsymbol{\phi}$.

2.6 The INLA framework[4]

As we have seen, the analytical posterior is possible to obtain for some hierarchical structures (e.g. (2.69)). 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 proper mixing of the Markov chains to make sure we sample from the steady state 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 Integrated Nested Laplace Approximation (INLA) method (Gómez-Rubio 2020).

2.6.1 Introduction to INLA

The INLA method is an alternative to the classical Marko Chain Monte Carlo methods, that has significant advantages at the cost of some structural assumptions. 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 $g(\mu_i) =$

[4]This subsection is slightly modified from the project thesis (Arnstad 2024).

η_i, we can model the observations as independent given the linear predictor

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i \ , \qquad i = 1, ..., n \ , \tag{2.76}$$

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

$$\mathbf{x} = (\eta_1, ..., \eta_n, \alpha, \beta_1, ..., \beta_n) \ , \tag{2.77}$$

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 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (see Section 2.8 for more details). 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

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ , \tag{2.78}$$

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

$$\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}) \ , \tag{2.79}$$

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

$$\pi(x_l|\mathbf{y}) = \int \pi(x_l|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ , \tag{2.80}$$

and

$$\pi(\theta_k|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-k} \ , \tag{2.81}$$

respectively, where $\boldsymbol{\theta}_{-k}$ is the vector of hyperparameters excluding element θ_k (Gómez-Rubio 2020).

2.6.2 Approximating the marginals

As previously mentioned the marginals in Equation (2.80) and Equation (2.81) are generally not tractable, but INLA uses this form of the marginals, to construct nested approximations (Rue et al. 2009). Consider, as in (Rue et al. 2009), the approximation of the marginals in Equation (2.80) and Equation (2.81) as

$$\tilde{\pi}(x_l|\mathbf{y}) = \int \tilde{\pi}(x_l|\boldsymbol{\theta}, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ , \tag{2.82}$$

and

$$\tilde{\pi}(\theta_k|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-k} \ , \tag{2.83}$$

where $\tilde{\pi}(\cdot,\cdot)$ is an approximation of the density $\pi(\cdot,\cdot)$. To be able to compute the above approximations, we need to first specify the approximations under the integral sign. The first one to consider is $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, which Rue et al. (2009) approximates by

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \left. \frac{\tilde{\pi}(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})} \right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \ , \tag{2.84}$$

where $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ is the Gaussian approximation of the full conditional of $\mathbf{x}$ evaluated at the mode $\mathbf{x}^*(\boldsymbol{\theta})$ of the full conditional for given $\boldsymbol{\theta}$ (Rue et al. 2009). From Equation (2.84) the posterior marginals of hyperparameter k, $\tilde{\pi}(\theta_k|\mathbf{y})$, can be approximated by integrating out $\boldsymbol{\theta}_{-k}$ using numerical integration. However, an approximation for $\tilde{\pi}(x_l|\boldsymbol{\theta},\mathbf{y})$ must be chosen to obtain the posterior marginals of the latent effects. To approximate $\tilde{\pi}(x_l|\boldsymbol{\theta},\mathbf{y})$, (Rue et al. 2009) describe three strategies of varying computational complexity. The cheapest approximation (Gómez-Rubio 2020) is to derive the Gaussian marginals of $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ as

$$\tilde{\pi}_G(x_l|\boldsymbol{\theta},\mathbf{y}) = \mathcal{N}(\mu_l(\boldsymbol{\theta}), \sigma_l^2(\boldsymbol{\theta})) \ , \tag{2.85}$$

where $\mu_l(\boldsymbol{\theta})$ is the mean vector and $\sigma_l^2(\boldsymbol{\theta})$ the corresponding vector with marginal variances of the Gaussian approximation (Rue et al. 2009). The second, and a more costly, approach is to use a Laplace approximation so that

$$\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta},\mathbf{y}) \propto \left. \frac{\tilde{\pi}(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-l}|x_l,\boldsymbol{\theta},\mathbf{y})} \right|_{\mathbf{x}_{-l}=\mathbf{x}^*_{-l}(x_l,\boldsymbol{\theta})} \tag{2.86}$$

where $\tilde{\pi}_{GG}(\mathbf{x}_{-l}|x_l,\boldsymbol{\theta},\mathbf{y})$ is the Gaussian approximation to the density of $\mathbf{x}_{-l}|x_l,\boldsymbol{\theta},\mathbf{y}$ evaluated at the mode $\mathbf{x}^*_{-l}(x_l,\boldsymbol{\theta})$ (Gómez-Rubio 2020). This approximation requires computations for each value x_l, and so a simplified modification

$$\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta},\mathbf{y}) \propto \mathcal{N}(\mu_l(\boldsymbol{\theta}), \sigma_l^2(\boldsymbol{\theta})) \exp(\text{cubic sline}(x_l)) \tag{2.87}$$

with a cubic spline fitted to the difference of $\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta},\mathbf{y})$ and $\tilde{\pi}_G(x_l|\boldsymbol{\theta},\mathbf{y})$ can be used (Rue et al. 2009). The third method, which is implemented as the default strategy in the INLA framework, is named the *simplified* Laplace approximation (Rue et al. 2009). This method uses a series expansion of $\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta},\mathbf{y})$ about the mean $x_l = \mu_l(\boldsymbol{\theta})$ to obtain the approximated density $\tilde{\pi}_{SLA}(x_l|\boldsymbol{\theta},\mathbf{y})$ (Gómez-Rubio 2020). With this expansion, one can correct for skewness and location in the Gaussian approximation, while at the same time maintaining the computational advantages (Gómez-Rubio 2020). For the full derivations of the series expansion and the simplified Laplace approximation, see Rue et al. (2009, chapter 3.2.3).

2.6.3 Parameter estimation and sampling procedure

The parameter estimation procedure in INLA is composed of a number of steps. The mode of the log-likelihood $\ln(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}))$ of the hyperparameters are obtained by maximizing with a quasi-Newton method. Then, to obtain the negative Hessian, $\mathbf{H}$, at the modal configuration $\boldsymbol{\theta}^*$, finite differences are applied (Gómez-Rubio

2020). The negative Hessian is then decomposed by its eigenvalues by $\mathbf{H}^{-1} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$ and the hyperparameters are rescaled using $\mathbf{z}$ such that

$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z} \ , \tag{2.88}$$

to more effectively explore the hyperparameter space (Gómez-Rubio 2020). Then, the hyperparameter space is explored using either a regular grid with some step-size h or a central composite design (CCD) (Gómez-Rubio 2020, and references therein). The exploration is done to obtain a set $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^K\}$ that captures principal portion of the mass in the probability distribution (Martino & Riebler 2019). Once a set of hyperparameters is obtained, $\tilde{\pi}(x_l|\boldsymbol{\theta}, \mathbf{y})$ is approximated by $\tilde{\pi}_G(x_l|\boldsymbol{\theta}, \mathbf{y})$, $\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta}, \mathbf{y})$ or $\tilde{\pi}_{SLA}(x_l|\boldsymbol{\theta}, \mathbf{y})$ and finally one can compute the desired marginal $\pi(x_l|\mathbf{y})$ using a numerical integration scheme on the form

$$\pi(x_l|\mathbf{y}) \simeq \sum_{k=1}^{K} \tilde{\pi}(x_l|\boldsymbol{\theta}^{(k)}, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}^{(k)}|\mathbf{y})\Delta_k \ . \tag{2.89}$$

A similar scheme for numerical integration can be used to obtain the marginals $\pi(\theta_k|\mathbf{y})$. 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 it is common to work with the precision matrix, which is defined as the inverse covariance matrix, rather than the covariance matrix directly. Therefore, all estimates on random effects will be given as precision rather than variance. Throughout the thesis, the

2.7 Quantitative genetics and relative variable importance

An important application of GLMMs, which we will later analyse, is in the context of evolutionary biology and quantitative genetics. More specifically, one wishes to estimate the variance of the random effect which contributes to direct heritage of traits between relatives. Further, with this estimate, one uses its proportion of total model variance to evaluate the interaction between inheritance and environmental factors in developing distinct traits. We will now describe how this can be seen as a special case of wanting to estimate relative variable importance of random effects.

The Animal Model

To introduce the animal model and biological terminology, the section will rely heavily on the work of Kruuk (2004) and Conner & Hartl (2004). The animal model is a mathematical model, used as a tool for quantitative genetic analysis

in evolutionary biology where the aim is to explain the phenotypic variation in a population. A phenotype is defined as *the outward appearance of an organism for a given characteristic* (Conner & Hartl 2004), such as eye color, height or behavior. In an organism, the observed phenotypic trait in an individual is a result of the complex combination of environmental effects and genotype. The genotype of a trait can be defined as *the diploid pair of alleles present at a given locus*, and is the outcome of genetic inheritance (Conner & Hartl 2004). As evolutionary biology seeks to explain diversity among individuals in a population (Kruuk 2004), a decomposition of the phenotypic variance is of great interest. The simplest partition is to define the phenotypic variance as the sum of the genetic variance and environmental variance (Conner & Hartl 2004). However, for species that mate with other individuals in the population rather than self-fertilize, it is common to further decompose the genetic variance into three parts. The **total phenotypic variance** can therefore be partitioned as

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2 \ , \tag{2.90}$$

where σ_P^2 is the total phenotypic variance, σ_G^2 is the **genetic variance**, σ_E^2 is the **environmental variance**, σ_A^2 is the **additive genetic variance**, σ_D^2 is the **dominance genetic variance** and σ_I^2 is the **interaction genetic variance** (Conner & Hartl 2004). The parameter σ_A^2, the variance of the additive genetic effect, is of particular interest, as the additive genetic effects are the only effects directly transferred to the offspring from its parents (Conner & Hartl 2004). Thus, the animal model aims to estimate σ_A^2 to gain inference on how changes in phenotypic values across generations occur, which is defined as phenotypic evolution (Conner & Hartl 2004). The animal model can be stated as a generalized linear mixed model, by letting

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} \ , \tag{2.91}$$

where $\boldsymbol{\mu}$ is the mean of the observations $\mathbf{y}$ of the phenotypic trait(s), $\boldsymbol{\eta}$ is the linear predictor, $\mathbf{X}$ the design matrix of the fixed effects, $\boldsymbol{\beta}$ the population coefficients, $\mathbf{U}$ the design matrix of the random effects and $\boldsymbol{\alpha}$ the vector of random effects. One of the random effects in the animal model, $\boldsymbol{\alpha}_A \sim \mathcal{N}(0, \mathbf{G})$, accounts for the additive genetic effect. The values of the vector $\boldsymbol{\alpha}_A$ containes the so-called breeding values Wilson et al. (2010), which are defined as the effect of an individuals genes on the value of the phenotypic trait in its offspring (Conner & Hartl 2004) As in Section 2.3.2, we let $\mathbf{G}$ denote the covariance matrix of the random effect $\boldsymbol{\alpha}_A$, which in the animal model can be derived from the expected covariance between relatives (Kruuk 2004). This derivation can be done by considering the coefficient of coancestry, $\Theta_{i,j}$, defined as *the probability that an allele drawn at random from an individual i will be identical by descent to an allele drawn at random from individual j* (Kruuk 2004). We use the coefficient of coancestry to define the expected covariance between relatives, or additive relationship matrix, as $\mathbf{A}_{i,j} = 2\Theta_{i,j}$ and consequently $\mathbf{G} = \sigma_A^2\mathbf{A}$ (Kruuk 2004).

Heritability

As mentioned, we are particularly conserned in the additive genetic variance σ_A^2 and functions of it, such as the **heritability**. Heritability in the narrow sense, is

defined as (Wilson 2008) the proportion of the total phenotypic variance that is present due to the additive genetic variance,

$$\frac{\sigma_A^2}{\sigma_P^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2} \ . \tag{2.92}$$

The narrow sense heritability is what one considers for outbreeding species, and therefore, when we refer to heritability, we refer to the narrow sense heritability. In quantitative genetics, heritability is perhaps the most frequently estimated and discussed measure (Conner & Hartl 2004). Heritability has this role, as it can be used to partly explain how quickly the mean phenotypic values evolve, when populations are subject to artificial or natural selection (Conner & Hartl 2004). This is directly linked to the aim of quantitative genetics, which is to explain diversity and the cause of diversity (Kruuk 2004). As a subject to much misinterpretations, it is important to note that the definition of heritability is based purely on variance, and consequently heritability refers only to variation within a population. Further, as heritability is calculated for a specific population, environment and over time, it is not to be viewed a fixed value (Conner & Hartl 2004). Nonetheless, heritability is a widely used quantity to compare populations, species and traits, and is an important tool for understanding the evolutionary forces that drive genetic diversity and thereby evolution (Conner & Hartl 2004). The estimation process is often carried out using the animal model, and as heritability is the result of a variance decomposition of the model fit, we can connect it to variable importance. Recalling our preferred definition of the R^2 in Equation (2.57), (2.58), (2.59) and (2.60), one can quickly notice that the definition of heritability is very similar. In fact, generalizing the definition of variable importance fromGrömping (2007) to also yield for random intercepts, one can define the heritability as the relative variable importance of the additive genetic effect. Therefore, estimating heritability can be seen as a special case of estimating relative variable importance, and serves as a suitable application for variable importance measures.

2.8 The Animal Model as a Gaussian Markov Random Field

INLA is a powerful tool for fitting latent gaussian models (LGMs) as it provides a computationally efficient alternative to the traditional MCMC methods (Rue et al. 2009). To be applicable it relies heavily on the latent field, which is Gaussian, to possess the Markov property and thereby have the structure of a Gaussian Markov Random Field (GMRF). If a Gaussian random variable $\mathbf{X} = (X_1, ..., X_n)$ possesses the Markov property it means that for some $i \neq j$, X_i is independent of X_j conditioned $X_{-i,j}$, where $X_{-i,j}$ denotes all other elements of $\mathbf{X}$ except X_i and X_j (Rue et al. 2009). This property is readily visualized in a conditional independence graph (Figure 1, right), and in the animal model the pedigree structure (Figure 1, left) derived from the family relation can be used as the conditional independence graph (Wermuth & Lauritzen 1983, as cited in Steinsland & Jensen (2010)). The pedigree of a population is a directed acyclic graph (DAG) where each node represents an individual and the directed edges represent the parent-offspring relationship. This gives rise to the conditional independence graph, which can be

found by inserting edges between parents that share offspring and removing the directions in the pedigree (Wermuth & Lauritzen 1983). An individual(node) in this graph will therefore only have edges, meaning it is conditionally dependent on, its parents, the parent(s) of its offspring, and its offspring. For example, in Figure 1 bird U_4 is conditionally dependent on birds U_1 and U_2 as they are its parents, bird U_6 as it is the offspring of U_4 and on bird U_5 as it is the other parent of the offspring of U_4. U_3 and U_7 therefore does not provide additional information on U_4 (Steinsland & Jensen 2010). Therefore, this GMRF structure from the pedigree can be used to effectively sample and obtain parameter estimates from the animal model (Steinsland & Jensen 2010), and allows us to use the INLA framework for model fitting. The pedigree can also be used to construct the relatedness matrix $\mathbf{A}$, previously defined as the expected covariance between relatives, and the gives rise to the sparse precision matrix $\mathbf{Q} := \mathbf{A}^{-1}$ which is needed for calculations. As we consider each node an individual, the corresponding value of that node is its breeding value $\boldsymbol{\alpha_A}$ (Steinsland & Jensen 2010).

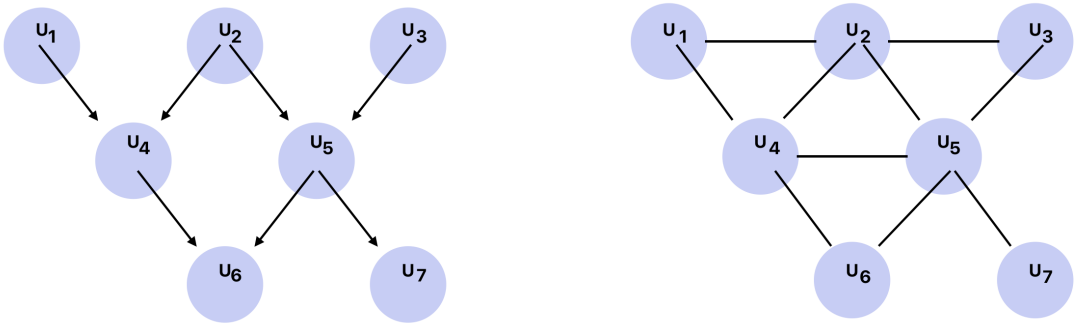

Figure 1: Illustration of a pedigree as a GMRF, figure and figure text inspired by Figure 1 in Steinsland & Jensen (2010). On the left, a pedigree structure is depicted as a directed acyclic graph (DAG), where birds U_1 and U_2 are the parents of bird U_4, birds U_2 and U_3 the parents of bird U_5, and birds U_4 and U_5 the parents of bird U_6. Bird U_7 has one known parent in U_5, and one unknown. On the right, the conditional independence graph of the pedigree structure is given, where the parents sharing offspring is assigned an edge and the direction is removed.

CHAPTER

THREE

METHODS

Based on the presented background theory, we now present our novel method for combining this into a relative variable importance tool for Bayesian GLMMs called Bayesian Variable importance (BVI). The proposed method is an extension of the method presented in Arnstad (2024) so that it now applies to GLMMs modelled with Binomial, Poisson in addition to Gaussian responses. The BVI method assumes the distinct random effects to be independent and does not include variable importance for random slopes. If categorical covariates with more than two levels are contained in the fixed effects, they should be encoded using distinct names in order to make sure the method can handle them correctly.

3.1 Variable importance in the Bayesian framework

There are a few considerations necessary in order to calculate variable importance on GLMMs in a Bayesian framework. First of all, the characteristics of the Bayesian framework must be considered. When fitting a GLMM in the frequentist framework, point estimates of the coefficients of the fixed effects as well as point estimates of the variance of the random effects are obtained. These estimates are then used to calculate relative variable importance measures. In contrast, a Bayesian GLMM tries to estimate the joint posterior distribution of parameters. From the posterior distribution, one can obtain samples of all parameters, that can be used to approximate a posterior distribution for each parameter. It is these samples that we will use for further calculations.

Secondly, we argue that the most intuitive way to calculate variable importance is on the link (or latent) scale. The reasoning behind this is the definition of residual variance for models with additive overdispersion in Nakagawa & Schielzeth (2013). This definition makes variable importance calculations on GLMMs analogous to that of LMMs, thus supporting a unified approach to both types of models. Therefore, we consider only GLMMs modeled with additive overdispersion, although we believe the method could be extended to handle multiplicative overdispersion as well. These considerations are the basis of our proposed method for calculating relative variable importance in Bayesian GLMMs. The presented method can han-

dle categorical variables with more than two categories as long as they are dummy encoded. Random slopes are excluded from our method due to the added computational complexity and the debatable improvement of GLMMs and R^2 values with random slopes as mentioned in Johnson (2014). We now go in to detail on how the different components of the GLMM model are handled in our method, to finally develop a relative importance measure for GLMMs.

3.2 Extending the R^2 to Bayesian GLMMs[1]

The core of our Bayesian variable importance measures is a decomposition of the R^2 value so that each covariate is assigned a share of relative variable importance. We now combine the definition of the R^2 for GLMMs presented Section 2.4 and the R^2 for the Bayesian linear regression from Section 2.5.5 to yield our proposed distribution of the R^2 for Bayesian GLMMs. Consider the linear predictor

$$g(\mathbb{E}[\mathbf{y}]) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} \ , \tag{3.1}$$

for some response $\mathbf{y}$ and monotonic and differentiable link function $g(\cdot)$. The variance components of the linear predictor can be decomposed into variance from the fixed effects and the random effects. Define the variance of the fixed effects as

$$\sigma_f^2 = \text{Var}(\mathbf{X}\boldsymbol{\beta}) \ , \tag{3.2}$$

and let $\sigma_{\alpha_i}^2$ denote the variance of the i-th random effect, with random effects assumed independent. For Gaussian responses corresponding to an LMM, the residual variance is defined as σ_ε^2 and is explicitly modelled. However, for non-Gaussian responses, the residual variance of the model when considering additive overdispersion is defined as

$$\sigma_\varepsilon^2 = \sigma_e^2 + \sigma_d^2 \ , \tag{3.3}$$

where σ_e^2 is the additive dispersion and σ_d^2 is the distributional variance (Nakagawa & Schielzeth 2013). A table containing the distributional variances for the link functions used in this thesis can be found in Table 1. Given that we can obtain samples for the variance components, we define for a sample s the marginal and conditional R^2 for the Bayesian GLMM as

$$R_{s,m}^2 = \frac{\sigma_{f,s}^2}{\sigma_{f,s}^2 + \sum_{i=1}^q \sigma_{\alpha_i,s}^2 + \sigma_{\varepsilon,s}^2} \quad \text{and} \quad R_{s,c}^2 = \frac{\sigma_{f,s}^2 + \sum_{i=1}^q \sigma_{\alpha_i,s}^2}{\sigma_{f,s}^2 + \sum_{i=1}^q \sigma_{\alpha_i,s}^2 + \sigma_{\varepsilon,s}^2} \ , \tag{3.4}$$

respectively, where $\sigma_{\varepsilon,s}^2 = \sigma_{e,s}^2 + \sigma_d^2$ is the sampled residual variance and σ_d^2 is distribution specific and the same for all samples. The posterior distribution of the marginal and conditional R^2 will then be approximated by the distribution of the samples of $R_{s,m}^2$ and $R_{s,c}^2$ for $s = 1, ..., S$ respectively.

[1] A method for calculating the R^2 for Bayesian LMMs was proposed in Arnstad (2024, Chapter 2), however we see it fitting to include this extension in the methods chapter as it has been developed by the author for this thesis.

Distribution	Link Function	Parameter	σ_d^2
Binomial	Logit	$0 < p < 1$	$\pi^2/3$
Poisson	Log	$\lambda > 0$	$\ln(1 + 1/\mathbb{E}[\lambda])$
Poisson	Square Root	$\lambda > 0$	0.25

Table 1: Distribution-specific variance σ_d^2 for the Binomial and Poisson distributions for some common link functions. The full expression $\mathbb{E}[\lambda]$ is given in (3.15). Distributional variances correspond to the variances in Nakagawa & Schielzeth (2013) and the calculation for the log-link Poisson follow the recommendations of Nakagawa et al. (2017).

3.3 Decomposing the R^2 value

We now seek to decompose the proposed R^2 value and assign each covariate with a proportion of the variance explained, i.e. assign each covariate with a *relative variable importance*. Recall that the fixed and random effects are assumed to be independent, so that one can consider the variances of the fixed and random effects separately. Further, the residual variance is also considered as independent of both fixed and random effects.

3.3.1 Applying the relative weights method in the Bayesian framework

To remedy the problems of calculating importance of correlated covariates, we will apply the relative weights method to the fixed effects before fitting the model. Following Section 2.2.4, we project the design matrix $\mathbf{X}$ of the fixed effects to obtain the matrix $\mathbf{Z}$. The model is fit using $\mathbf{Z}$ as an approximated design matrix of fixed effects, and from the joint posterior distribution samples of the coefficients $\boldsymbol{\beta}_{\mathbf{Z}}$ can be drawn. Each sample $\boldsymbol{\beta}_{\mathbf{Z},s}, s = 1, ..., S$ and the matrix Λ can be used to approximate a sample of the importance of the columns $\mathbf{X}$, with the matrix Λ acting as weights from the projected space to the original covariate space. Using equations (3.15) and (2.18), we calculate this sample as

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

where $\mathrm{IMP}(\mathbf{X})_s$ is a column vector containing the approximated importance of column k of $\mathbf{X}$ on the k-th entry for $k = 1, ..., p$. To calculate the relative variable importance, note that we estimate $\sigma^2_{f,s}$ in (3.4) by

$$\sigma^2_{f,s} \simeq \sum_{k=1}^{p} \mathrm{IMP}(\mathbf{X})_{s,k} \ . \tag{3.6}$$

Therefore, we define the relative importance of column k of $\mathbf{X}$ in our method as

$$\mathrm{RI}(\mathbf{X})_{s,k} = \frac{\mathrm{IMP}(\mathbf{X})_{s,k}}{\sum_{j=1}^{p} \mathrm{IMP}(\mathbf{X})_{s,j} + \sum_{i=1}^{q} \sigma^2_{\alpha_i,s} + \sigma^2_{\varepsilon,s}} \ , \tag{3.7}$$

where $\sigma^2_{\alpha_i,s}$ and $\sigma^2_{\varepsilon,s}$ are defined as in Section 3.2. For sufficiently large S, we believe these samples can be used to construct an approxmation of the posterior distribution of the relative importance for each fixed effect.

3.3.2 Random effects

The presented background theory on relative variable importance has mostly been developed for linear regression models. As long as the random effects are assumed not to be correlated, introducing random effects does not change the general idea. For each random effect, an approximation of the posterior distribution is constructed from the samples of the joint posterior distribution. Then, the proportion of variance explained by random effect i is calculated as

$$\mathrm{RI}(\alpha_i)_s = \frac{\sigma^2_{\alpha_i,s}}{\sum_{k=1}^{p} \mathrm{IMP}(\mathbf{X})_{s,k} + \sum_{k=1}^{q} \sigma^2_{\alpha_k,s} + \sigma^2_{\varepsilon,s}} \ . \tag{3.8}$$

In addition to the relative importance of the random effects, a quantity of interest is the intraclass correlation, often also called the within cluster correlation or repeatability (Fahrmeir et al. 2013). The ICC represents the correlation between observations within the same cluster, and is defined for a random effect $\boldsymbol{\alpha}_i$ in (Nakagawa et al. 2017) as

$$ICC = \frac{\sigma^2_{\alpha_i}}{\sum_{k=1}^{q} \sigma^2_{\alpha_k} + \sigma^2_{\varepsilon}} \ . \tag{3.9}$$

Thus, following the same logic as before we can sample the ICC as

$$\mathrm{ICC}_s = \frac{\sigma^2_{\alpha_i,s}}{\sum_{k=1}^{q} \sigma^2_{\alpha_k,s} + \sigma^2_{\varepsilon,s}} \ , \tag{3.10}$$

and obtain an approximate posterior distribution of the ICC.

As previously mentioned, it is common to report the precision of random effects rather than the variance. Since the random effects are assumed to be independent, one can invert the precision estimate to obtain the variance. Another way of estimating the variance is to take the variance of the sampled values for the random vector $\boldsymbol{\alpha}$. Both methods seem to give very similar results as long as the sample size is large enough, and we therefore see both methods as fit for estimating the variance of random effects.

3.3.3 Drawing samples

A critical part of performing the calculations the BVI method requires, is to obtain samples from the joint posterior distribution. To do this, we utilize the built-in function from the INLA framework called `inla.posterior.sample()`. This function uses the approximation of the posterior distribution fitted with INLA by numerical integration, and therefore the accuracy of the samples is dependent on how well the numerical integration is carried out (Gómez-Rubio 2020). INLA provides several integration options, so one can choose the resolution one desires, but this comes at the cost of computational complexity. In this thesis, we use the default integration strategy in INLA, which is either the grid strategy for a hyperparameter vector of dimension less than or equal to two or the central composite design (CCD) for a larger dimension hyperparameter vector (Martino & Riebler 2019). Further, if the model fit is poor or if the model is misspecified,

the samples will suffer from this as well. Recall that INLA assumes a Gaussian latent layer, so this condition is crucial to obtain a representative set of samples. Lastly, INLA is a tool that is continuously in development, and the authors state that a skewness correction is in the works (Gómez-Rubio 2020).

3.4 Heritability of phenotypic traits

As we have seen in Section 2.7, the concept of variance decomposition in GLMMs is not new and has been used in quantitative genetics with the animal model for many years (e.g. Kruuk (2004)). The main quantity of interest in such studies has been the heritability of phenotypic traits, which is defined as as the ratio of additive genetic variance to total phenotypic variance (Wilson 2008). We now aim to illustrate how we calculate the heritability of phenotypic traits using the BVI method, and hence illustrating why heritability is a special case of variable importance.

3.4.1 Heritability as relative variable importance

By comparing (2.92) with (3.8), it is clear that the way we have defined relative variable importance of a random effect coincides with the definition of heritability, if the random effect is the additive genetic effect and one assumes the total phenotypic variance σ_P^2 to be captured by the other fixed and random effects present. Therefore, when applying the BVI method to model a phenotypic trait, the relative variable importance of the random effect accounting for additive genetic variance can be interpreted as the heritability of the phenotypic trait. This is a highly relevant and useful application of our method and has been a major motivation for the development of the BVI method. It should be mentioned here that in the frequentist framework, fixed effects are assumed to not have an associated variance. Therefore, they are commonly not featured in formulae for the variance decomposition when estimating heritability (see Kruuk (2004) and Wilson et al. (2010)). FINISH WITH WHY WE DO IT THE WAY WE DO IT AND WHY THAT WORKS, IN CONTRAST TO STENSLAND. POSSIBLY ALSO MENTIONED THAT THE FIXED EFFECTS USED ARE LEVEL/FACTOR EFFECTS, THEREFORE THE VARIANCE STRUCTURE IS A BIT FUNKY.

3.4.2 House sparrow study

We now apply the BVI method to a dataset gathered on house sparrows (*Passer domesticus*) from a study on the coast of Helgeland, Norway (Steinsland & Jensen 2010). The entire bird population on five islands have been surveyed since 1993 and several morphological traits have been measured. Blood samples were drawn to determine the relatedness between birds and we therefore have a pedigree structure for the birds (Steinsland & Jensen 2010, citing Jensen et al., 2003, 2004, 2008). In the dataset we use we have $N = 3116$ birds with one or more observations on the traits and covariates. For a more thorough description of the house sparrow study, see Steinsland & Jensen (2010, and references therein). We model three phenotypic traits using a Gaussian LMM, namely the body mass, wing length and tarsus length. The fixed effects in the model consist of observations of *sex*,

a standardized inbreeding coefficient denoted *FGRM*, the standardized *month* of the year (measurements were made during May-August), the *age* of each bird, and dummy variables encoding the location of the *native island* group of the bird (three levels, outer islands, inner islands or other islands). In addition, we model the *hatchyear* as an independent and identically distributed (i.i.d.) random intercept. To account for the correlation between relatives, we include a random effect for the additive genetic variance. It is the sampled variances of the additive genetic random effect that will determine the heritability of each trait. We derive the relatedness matrix of the birds from our pedigree, and specify this as the covariance matrix for the additive genetic variance term. Lastly, to account for individual differences we add an i.i.d. random intercept for the individual bird. We prefer to use the INLA framework, described in Section 2.6, to fit our LMM as it is computationally efficient and easy to use. Each prior is internally parametrized in INLA by $\theta = \ln(\tau)$ with τ being the precision of the prior. This means when placing priors, they are always placed on the scale of the internal parameter θ, and if we want to place a prior on the external scale we must take this into account. For the fixed effects, we place penalizing complexity (PC) priors with the initial value being $\ln(0.5)$ on the external scale and parameters $U = \sqrt{2}$ and $a = 0.05$ as the input parameters discussed in section Section 2.5.3. Similarly, we place PC priors on each random effect, with the effects *hatchyear* and *individual differences* having $U = 1$ and $a = 0.05$. The initial value of the priors for *hatchyear* and *individual differences* is set to be $\tau_0 = \ln(1)$ to correspond to 1 on the external scale. The additive genetic effect is assigned $U = \sqrt{2}$ and $a = 0.05$, with $\tau_0 = \ln(0.5)$. These priors have been chosen through discussion with the supervisor of the thesis and researchers with domain knowledge in biology. The approximation of the posterior marginals $\tilde{\pi}(x_l|\mathbf{y}\boldsymbol{\theta})$ will be made using the simplified Laplace approximation for all components, as described in Section 2.6.2 and Section 2.6.3.

3.5 Non-Gaussian case studies

For the complete model formulation of all methods used in the case studies, all files will be uploaded to the Github, with a link in Appendix A. A simulation study on Gaussian responses (LMMs) can be found in Arnstad (2024), where the same methodology is used and results are reported and discussed.

3.5.1 Binomial and Poisson case studies

To investigate how well the BVI method generalizes to non-Gaussian responses, we perform a case study using the setup described in the vignette of the R-package `rptR`, found at `https://cran.r-project.org/web/packages/rptR/vignettes/rptR.html` (Stoffel et al. 2017). An important clarification for this case study, is that there are multiple formulations of repeatability. Two most common ways of looking at repeatability are

$$\begin{aligned} \mathrm{R}_1 &= \frac{\text{Additive genetic variance}}{\text{Total variance of covariates}} \\ \mathrm{R}_2 &= \frac{\text{Additive genetic variance}}{\text{Total variance of random covariates}}, \end{aligned} \tag{3.11}$$

where the former corresponds to our notion of heritability (Stoffel et al. 2017). We choose to look at the notion corresponding to heritability, and to obtain the result from `rptR` so that they match this, each model must be fit with the argument `adjusted=FALSE`. The dataset, introduced for a different purpose, is simulated to replicate a study on twelve different beetle larvae populations (Stoffel et al. 2017). It contains the covariates *population*, the discrete *habitat* of the larvae, the dietary *treatment* of the larvae, the *sex* and *container* of which the larvae was contained in. The phenotypes to be modeled by the Binomial and Poisson distributions are the two distinct male colour morph and the number of eggs laid by female larvae respectively. Both models use *treatment* as the only fixed effect and place i.i.d. random intercepts on the *population* and *container* covariates. Note that a more complex covariance structure could be modelled by the BVI method, but the `rptR` package does not allow for this, so for comparing the methods we see it as suitable with i.i.d. random intercepts. As before, our modelling is carried out using INLA, whereas the models in the vignette are calculated from functions in the `rptR` package. The priors placed on the fixed effect *treatment* and random effects *population* and *container* are PC priors with initial values $\tau_0 = \ln(1)$ on external scale and parameters $U = 1$ and $a = 0.01$ for all effects. As before, our preferred approximation of the posterior marginals of the latent field conditioned on the observations and set of hyperparameters is the simplified Laplace approximation.

3.5.2 Binomial and Poisson simulation studies

There are two primary reasons why we wish to conduct a simulation study with our method. The first being the ability to evaluate how well our method assigns relative variable importance to all covariates in the model. The real life case studies available mostly have the heritability, or some other function of the additive genetic variance, as the objective of analysis (Steinsland & Jensen 2010). We aim to provide the heritability, but at the same time provide information on the relative variable importance of all covariates present in the model. The latter motivation is that the Bayesian framework is stochastic, and so is our method. We wish to assess the variability of this stochasticity by simulating different datasets with the same underlying structure, and see the spread of the estimates. We hope that this can provide signs that any fitted model can be seen as a random sample of a distribution centered around the true value.

We simulate $N = 10^4$ responses from a Binomial distribution with a logit-link and from a Poisson distribution with log-link. The linear predictor contains three fixed effects and one random intercept. The fixed effects are, for simplicity but without loss of generalization, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0, 0, 0)^T$, $\Sigma_{i,i} = 1$ and $\Sigma_{i,k} = \rho, k \neq i$. The true regression coefficient are set to be $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$. The random effect comes from $m = 100$ clusters, each with $n_j = 100$ observations for $j = 1, ..., m$. Further, we draw the random effect from a normal distribution with mean zero and variance $\sigma_\alpha^2 = 1$. To investigate the impact of correlated fixed effects, we fit four different models letting ρ vary for each model by taking on the values $\rho \in \{-0.4, -0.1, 0, 0.1, 0.4\}$. The INLA framework is used to fit the GLMMs and the methodology described used to calculate the relative importance. All fixed and random effects receive the same PC prior as used in the comparison

with the `rptR` package, that is with initial values of $\tau_0 = \ln(1)$ and parameters $U = 1$ and $a = 0.01$. As has been done throughout the thesis, the simplified Laplace approximation is used to approximate the posterior marginals of the latent field conditioned on the observations and hyperparameters. We fit $N_{\text{sim}} = 500$ Binomial and Poisson models with different datasets for each correlation level.

In the simulation study, when parameters are simulated so that we know their true value, we can empirically calculate the relative importance of the parameters when they are not correlated. When uncorrelated, the proportion of variance explained by each covariate in the linear predictor is equal to the square of the true coefficient. By defining $\sigma^2_{x_k}$ as the variance contribution to the linear predictor (the latent scale) for fixed effect k and σ^2_α as the variance contribution of the random effect, we then have

$$\sigma^2_{x_1} = \sigma^2_\alpha = 1 \quad \text{and} \quad \sigma^2_{x_2} = 2 \quad \text{and} \quad \sigma^2_{x_3} = 3 \ . \tag{3.12}$$

Then, the relative importance of the covariates can be calculated as

$$\begin{aligned} \text{RI}(\mathbf{X}_1) = \text{RI}(\alpha_1) &= \frac{\sigma^2_{x_1}}{\sum_{i=1}^3 \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d}, \\ \text{RI}(\mathbf{X}_2) &= \frac{\sigma^2_{x_3}}{\sum_{i=1}^3 \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d}, \\ \text{RI}(\mathbf{X}_3) &= \frac{\sigma^2_{x_3}}{\sum_{i=1}^3 \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d} \ . \end{aligned} \tag{3.13}$$

In our simulation study, the binomial model with logit-link is assigned $\sigma^2_d = \pi^2/3$. The distributional variance of the Poisson model with log-link is given by

$$\sigma^2_d = \ln(1 + 1/\mathbb{E}[\mathbf{y}]) = \ln(1 + 1/\mathbb{E}[\lambda]) \ , \tag{3.14}$$

where

$$\mathbb{E}[\lambda] = \exp\left(\beta_0 + 0.5(\sum_{k=1}^{q} \sigma^2_{\alpha_k} + \sigma^2_e)\right) \ , \tag{3.15}$$

is the quantity used in Table 1 (Nakagawa et al. 2017). So we obtain, using a single random intercept, $\sigma^2_d = 0.4741$ with $\beta_0 = 0$, $\sigma^2_\alpha = 1$ and $\sigma^2_e = 0$. Therefore, we can summarize the expected relative importance of our three models as in Table 2.

Model	$\mathbb{E}[\text{RI}(\boldsymbol{\alpha})]$	$\mathbb{E}[\text{RI}(\mathbf{X}_1)]$	$\mathbb{E}[\text{RI}(\mathbf{X}_2)]$	$\mathbb{E}[\text{RI}(\mathbf{X}_3)]$
Binomial, logit	0.0972	0.0972	0.1944	0.2915
Poisson, log	0.1338	0.1338	0.2676	0.4014

Table 2: The expected relative importance of the covariates in the different models when uncorrelated.

In practice, the distributional variance of the Poisson model should be calculated using the estimated values, and the distributional variance will therefore be dependent on the fitted model (Nakagawa et al. 2017).

In addition to the expected importance of covariates in the uncorrelated case, we can calculate the expected marginal and conditional R^2 values for all correlation levels on the latent scale. Recalling that each of the $p = 3$ columns of $\mathbf{X}$ is initialized to have variance equal to 1, the expected marginal R^2 can be calculated as

$$\mathbb{E}[R^2_{\text{marg}}] = \frac{\sum_{i=1}^{3} \beta_i^2 + 2\sum_{i=1}^{2}\sum_{k=i+1}^{3} \beta_i\beta_k\rho}{\sum_{i=1}^{3} \beta_i^2 + 2\sum_{i=1}^{2}\sum_{k=i+1}^{3} \beta_i\beta_k\rho + \sigma_\alpha^2 + \sigma_d^2} \ , \tag{3.16}$$

and similarly for the expected conditional R^2 as

$$\mathbb{E}[R^2_{\text{cond}}] = \frac{\sum_{i=1}^{3} \beta_i^2 + 2\sum_{i=1}^{2}\sum_{k=i+1}^{3} \beta_i\beta_k\rho + \sigma_\alpha^2}{\sum_{i=1}^{3} \beta_i^2 + 2\sum_{i=1}^{2}\sum_{k=i+1}^{3} \beta_i\beta_k\rho + \sigma_\alpha^2 + \sigma_d^2} \ . \tag{3.17}$$

Model Type	**Correlation (ρ)**	$\mathbb{E}[R^2_{\text{marg}}]$	$\mathbb{E}[R^2_{\text{cond}}]$
Binomial Logit	-0.4	0.262	0.434
Binomial Logit	-0.1	0.532	0.641
Binomial Logit	0	0.583	0.680
Binomial Logit	0.1	0.624	0.712
Binomial Logit	0.4	0.709	0.777
Poisson Log	-0.4	0.508	0.842
Poisson Log	-0.1	0.768	0.925
Poisson Log	0	0.803	0.937
Poisson Log	0.1	0.828	0.945
Poisson Log	0.4	0.877	0.960

Table 3: Expected marginal and conditional R^2 values for the binomial regression with logit-link (top) and Poisson regression with log-link (bottom) for different correlation levels ρ.

3.6 Simulation study with R2D2 and GDR2 priors

As mentioned, the Bayesian variable importance field is not very large. However, as discussed in Section 2.5.6, the R2D2 priors decompose the R^2 value and can therefore be interpreted as a variable importance measure. We find it sensible to try and compare the resulting variable importance distributions, even though the R2D2 priors have not been developed for this goal specifically. To be able to evaluate the two measures, we simulate a linear regression model with $p = 3$ covariates and $n = 1000$ observations. The covariates are for simplicity simulated as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (0, 0, 0)^T$, $\boldsymbol{\Sigma}_{i,i} = 1$ and $\boldsymbol{\Sigma}_{i,j} = \rho$ for $i \neq j$. As before, we vary the correlation by letting $\rho \in \{-0.4, -0.1, 0, 0.1, 0.4\}$. The true regression coefficients are initialized as $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})$, and we simulate a random error by $\varepsilon \sim \mathcal{N}(0, \sigma^2 = 1)$. By noting that $\text{Var}(\mathbf{y}) = 7$, it is clear that in the uncorrelated case, the relative importance of the covariates can be calculated as

$$\text{RI}(\mathbf{X})_1 = \frac{1}{7} \quad \text{RI}(\mathbf{X})_2 = \frac{2}{7} \quad \text{RI}(\mathbf{X})_3 = \frac{3}{7} \ . \tag{3.18}$$

Further, the R^2 for this model is by the definition in (2.8)

$$R^2 = \frac{\text{Var}(\mathbf{y}) - \sigma^2}{\text{Var}(\mathbf{y})} \ , \tag{3.19}$$

which gives values that are summarized in Table 4.

ρ	R^2
-0.4	0.604
-0.1	0.830
0	0.857
0.1	0.877
0.4	0.913

Table 4: R^2 values for the correlation levels ρ used in the linear regression for analyzing the BVI method in comparison to the R2D2 priors.

To fit the linear regression using R2D2 priors for the marginal R^2 we use functions from the Github repository Zhang (2024) by the author of Zhang et al. (2020). For the GDR2 priors, we utilize the Stan code available on Romero & Bürkner (2024) by the authors of Aguilar & Bürkner (2024). The hyperparameters for the marginal $R^2 \sim \text{Beta}(a, b)$ are set so that $\mathbb{E}[R^2] \simeq 0.857$ which is approximately the theoretical R^2 of 6/7. This is done for the R2D2 priors by using the default value $b = 0.5$ from Zhang (2024) and noting that the expected value of the $\text{Beta}(a, b)$ distribution is $a/(a+b)$ (Tjelmeland et al. 2000). We follow the Gibbs sampler for the marginal R2D2 prior as described in (Zhang et al. 2020, section 5.3) and run the MCMC iteration $N = 10^4$ times, with a burn in of 9000 samples. This gives 1000 samples from the posterior distribution of the marginal R^2 as well as 1000 samples of each ϕ_j for $j = 1, 2, 3$. For the GDR2 priors, the implementation requires a prior on the expected value and the precision of the R^2 value directly. These are calculated by letting $a_\pi = 0.7$, the reference unit α_K be zero, the expected R^2 equal 6/7 and then solving for the precision τ according to the properties of the Beta distribution given in Aguilar & Bürkner (2024). The choice of a_π corresponds to a scenario in which one assumes the covariates to be approximately equally important and the difference between the GDR2 prior and R2D2 prior is substantial (Aguilar & Bürkner 2024). Similarly, as for the R2D2 case, we run the MCMC iteration $N = 10^4$ times, with a burn in of 9000 samples, but we also fit four Markov chains this time to ensure proper mixing. This means we obtain 4000 samples for the GDR2 priors. The samples of ϕ_j from the linear regression using R2D2 and GDR2 priors are then seen as the posterior distribution of relative variable importance of $\mathbf{x}_j$, and compared to that of the BVI method. We will refer to the results by using R2D2 and GDR2 priors as the R2D2 method and the GDR2 method respectively. To evaluate all methods, we fit a frequentist linear regression model and evaluate the importance metrics according to the LMG method by using the package Grömping & Lehrkamp (2023) in R as described in Grömping (2007). As the LMG method is feasible in this context, it poses perhaps the most robust and reliable benchmark available. We draw 1000 bootstrap samples of the LMG metrics and use this to create confidence intervals for the LMG method.

CHAPTER

FOUR

RESULTS

4.1 Simulation study

In this section, we lay forth the results of our simulation study on a binomial and a Poisson regression. We note that it has been difficult to find suitable methods to compare the non-Gaussian models with. In parallel to fitting our model as described in Section 3.5.2, we fit a model using the `rptR` package with 100 bootstrap samples. This allows us to directly compare the importance of the random effect and the marginal and conditional R^2 values. However, it does not compute the importance of each isolated fixed effect. For a simulation study on Gaussian responses with our model, we refer to the result and discussion chapters of Arnstad (2024).

4.1.1 Binomial simulation

We begin by presenting the results obtained from the simulation study. The first model to be analyzed is the Binomial regression on binary response, modelled with the logit-link function. As mentioned, we fit the model for five different correlations $\rho = (-0.4, -0.1, 0, 0.1, 0.4)$. For each correlation level, we fit $N_{\text{sim}} = 500$ models, and use the Bayesian Variable Importance method to estimate the relative importance of all covariates in each model. The simulation study was somewhat halted for the positive correlation levels, as INLA was not always able to fit all the models. Of the 500 simulations, we had 2 model fitting failures for $\rho = 0.1$ and 84 failures for $\rho = 0.4$, whereas the other correlation levels had no failures. A summary of the 500 estimated importances are shown in Table 1, which contains the mean and values for the lower and upper 95% quantile.

Measure		$\rho = \mathbf{0}$	$\rho = \mathbf{0.1}$	$\rho = \mathbf{-0.1}$	$\rho = \mathbf{0.4}$	$\rho = \mathbf{-0.4}$
Random Importance	Average	0.0971	0.0861	0.1067	0.0663	0.1688
	2.5%	0.0732	0.0621	0.0796	0.0472	0.1266
	97.5%	0.1262	0.1148	0.1381	0.0885	0.2087
Fixed Importance X1	Average	0.0972	0.1170	0.0775	0.1728	0.0199
	2.5%	0.0824	0.1005	0.0643	0.1578	0.0168
	97.5%	0.1124	0.1326	0.0916	0.1873	0.0239
Fixed Importance X2	Average	0.1945	0.2103	0.1761	0.2390	0.0767
	2.5%	0.1731	0.1891	0.1541	0.2227	0.0632
	97.5%	0.2155	0.2323	0.1972	0.2584	0.0898
Fixed Importance X3	Average	0.2919	0.2983	0.2805	0.2991	0.1662
	2.5%	0.2653	0.2724	0.2571	0.2792	0.1456
	97.5%	0.3134	0.3229	0.3047	0.3183	0.1874
R_m^2	Average	0.5836	0.6256	0.5342	0.7108	0.2628
	2.5%	0.5533	0.5951	0.5020	0.6850	0.2337
	97.5%	0.6119	0.6532	0.5648	0.7348	0.2918
R_c^2	Average	0.6807	0.7117	0.6409	0.7772	0.4316
	2.5%	0.6512	0.6889	0.6132	0.7562	0.3944
	97.5%	0.7072	0.7354	0.6696	0.7971	0.4695

Table 1: Summary of simulation study results for the quantiles of relative importance estimates of the Logit model across different correlation levels.

4.1.1.1 Fixed effects

The sampled posterior distribution of relative importance allocated to the three fixed effects X_1, X_2 and X_3 are shown for each correlation level (Figure 1). We see that the distributions generally form a normal shape around the mean, with somewhat varying spread. As correlation levels go from negative to positive, meaning that the variance contribution from the fixed effects increase, the importances of the fixed effects also increase. This is expected, as the shared covariance increases and is spread across the correlated fixed effects. The difference is quite substantial, with the average relative importance allocated to X_1 for $\rho = -0.4$ being 0.0199 compared to 0.1728 for $\rho = 0.4$. The same pattern is seen for X_2 and X_3, with the average relative importance increasing from 0.0767 to 0.2390 for X_2 and from 0.1662 to 0.2991 for X_3 when going from $\rho = -0.4$ to $\rho = 0.4$. For $\rho = 0$ (middle plot of Figure 1), it is clear that the average estimate for relative importance of all fixed effects is very similar to the expected importance (Table 2) shown as a dashed green line.

We notice that the covariates X_1 and X_2 are allocated a significantly larger share when correlation goes from $\rho = 0$ to $\rho = 0.4$, whereas X_3 is almost unchanged for the same correlation levels. This was also experienced in the simulation study on LMMs from Arnstad (2024), and explained by the fact that off diagonal elements of $\mathbf{\Lambda}$ increase positively when the fixed effects are positively correlated, while the diagonal elements decrease. In the uncorrelated case, $\mathbf{\Lambda}$ should be equal to the

identity matrix. The columns of $\mathbf{\Lambda}$ therefore act as weights and due to this, when $\rho = 0.4$, X_1 will receive an importance estimate where β_2^2 and β_3^2 will have positive weights contrary to $\rho = 0$ where the only weight is put on β_1^2. Since β_1^2 is smaller than β_2^2 and β_3^2, the higher positive correlation level yields a higher importance estimate for X_1. The same pattern is seen for X_2, where β_1^2 is smaller and β_3^2 is larger than β_2^2. This means the importance of X_2 is estimated to be higher for $\rho = 0.4$ than for $\rho = 0$, but the increase is smaller than the increase for X_1 (Arnstad 2024). In contrast, the importance of X_3 is then estimated with more weight on β_1^2 and β_2^2, which are both smaller than β_3^2, and thus the importance is not notably increased. If one had introduced a larger positive correlation level than $\rho = 0.4$, we would therefore expect the importance of X_3 to even decrease, as was seen in Arnstad (2024). It is hard to say, based on these results, whether the inverse pattern can be seen for negative correlation levels, but it could be noted that the decrease in importance is less for X_3 compared to X_1 and X_2 when ρ changes from 0 to -0.1.

Generally, it seems that the method is able to capture the expected effects of varying correlation levels, and is in close agreement with the expected theoretical values when the fixed effects are uncorrelated.

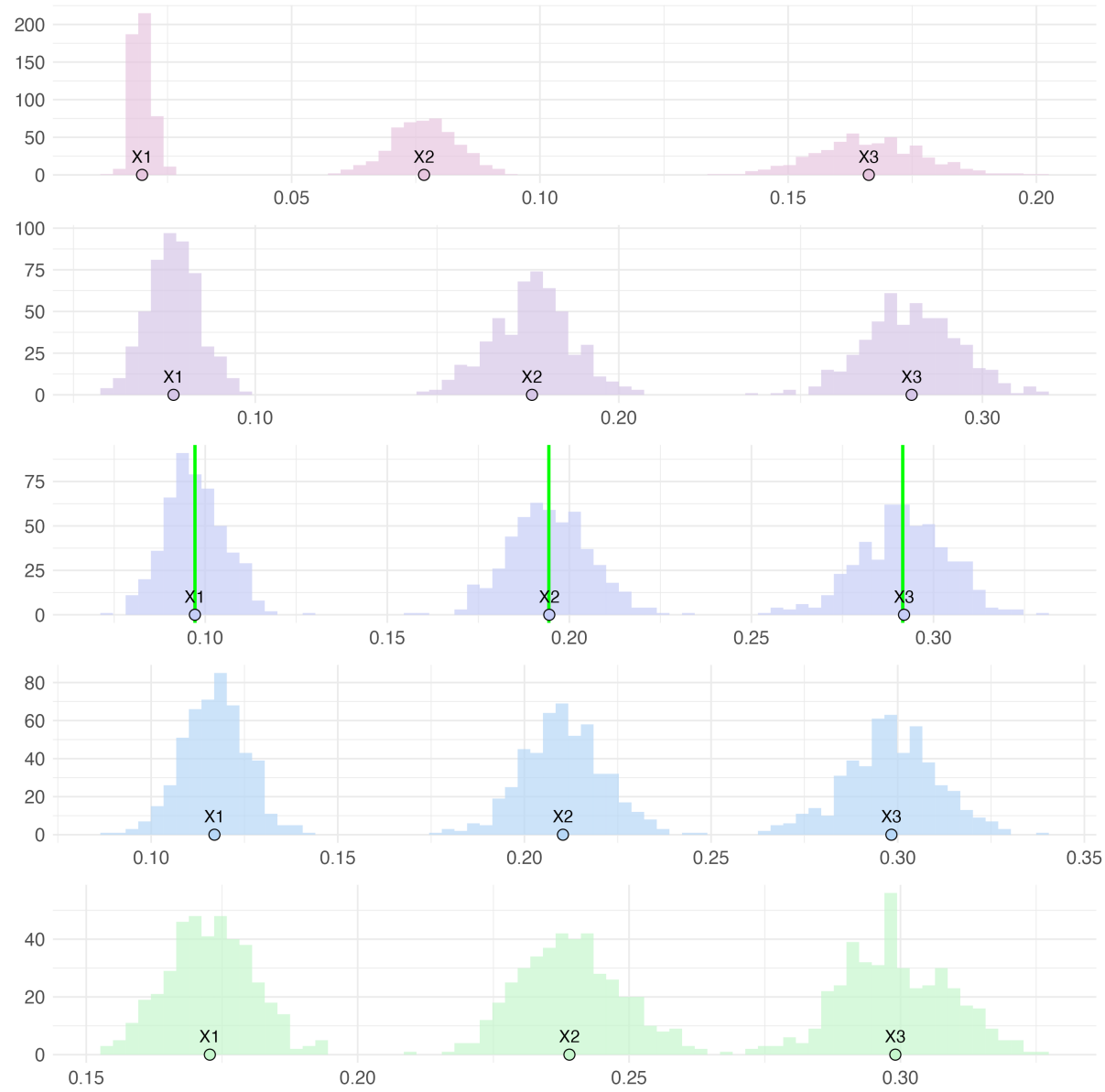

Figure 1: Histogram with relative importance of the fixed effects present in the binomial regression for the different correlation levels $\rho = -0.4$ (top), $\rho = -0.1$ (second from the top), $\rho = 0$ (middle), $\rho = 0.1$ (second from bottom) and $\rho = 0.4$ (bottom). The values are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ simulations in the simulation study. The true regression coefficients are $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$ and the vertical green line for $\rho = 0$ displays the expected relative importance in the case of uncorrelated data. The mean of the relative importance for all simulations is denoted at the bottom of each histogram as a circle.

4.1.1.2 Random effect

The sampled posterior importances for the random effect in the logit model (Figure 2) all seem to be roughly normally distributed around the mean. It is clear that when the correlation in fixed effects go from negative to positive, the estimated importance of the random effect shrinks. Specifically, when $\rho = -0.4$ the

average estimate of relative importance for the random effect is 0.1688 compared to only 0.0663 when $\rho = 0.4$. This naturally occurs as the variance contribution from the random effect should be held constant for the correlation levels, and the variance contribution from the fixed effects rise as the correlation increases. Therefore the proportion of variance explained, which is our definition of relative variable importance, will decrease for the random effect. For $\rho = 0$ we see that the average relative importance estimate lies very near the expected value of 0.0972 as shown in Table 2. The orange dot at the bottom of the histograms in Figure 2 displays the estimated relative importance of the fixed effect from the `rptR` package, and we see that the estimates are quite close to the mean of the BVI method. The largest difference from the BVI and the `rptR` package is 0.0351 and are found when $\rho = -0.4$. It can be noted that the authors of the `rptR` package (Stoffel et al. 2017) write in the vignette that any model fit with the package will automatically be fit with a term to account for additive overdispersion. This may explain the difference in the relative importance estimates, as the simulated data is not overdispersed.

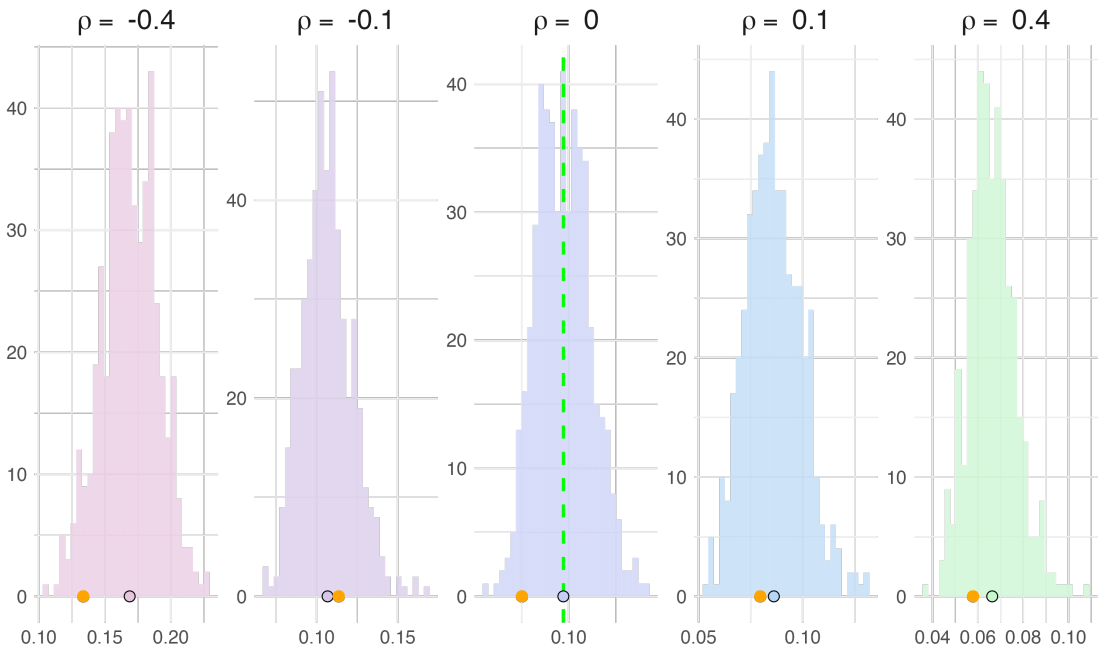

Figure 2: Histogram with values from the BVI method for each of the $N_{\text{sim}} = 500$ simulations, estimating relative importance of the random effect $\boldsymbol{\alpha}$ across the different correlation levels ρ. The mean of the estimated relative importance from all simulations is displayed at the bottom as a circle and the orange dot at the bottom displays the estimate fron the `rptR` package. The vertical green line for $\rho = 0$ is the expected relative importance as in Table 2.

4.1.1.3 R^2 estimates

An important measure in this simulation study is the models sampled posterior distribution of marginal and conditional R^2 (Figure 3). The expected values for the marginal and conditional R^2 are shown in Table 3, and are displayed as vertical

green lines in each plot. It is clear that, regardless of correlation level, the BVI method is able to estimate the marginal and conditional R^2 close to what we expect. The distributions of R^2 values seem to have the shape of a bell curve and are symmetric about the mean value. For the R^2 values, the largest difference between the mean of the simulations from the BVI method and the expected values is 0.002 for the marginal R^2 when $\rho = -0.1$ and 0.003 for the conditional when $\rho = -0.4$. When comparing to the results from the `rptR` package, it seems that the BVI method consistently estimates larger R^2 values both marginally and conditionally.

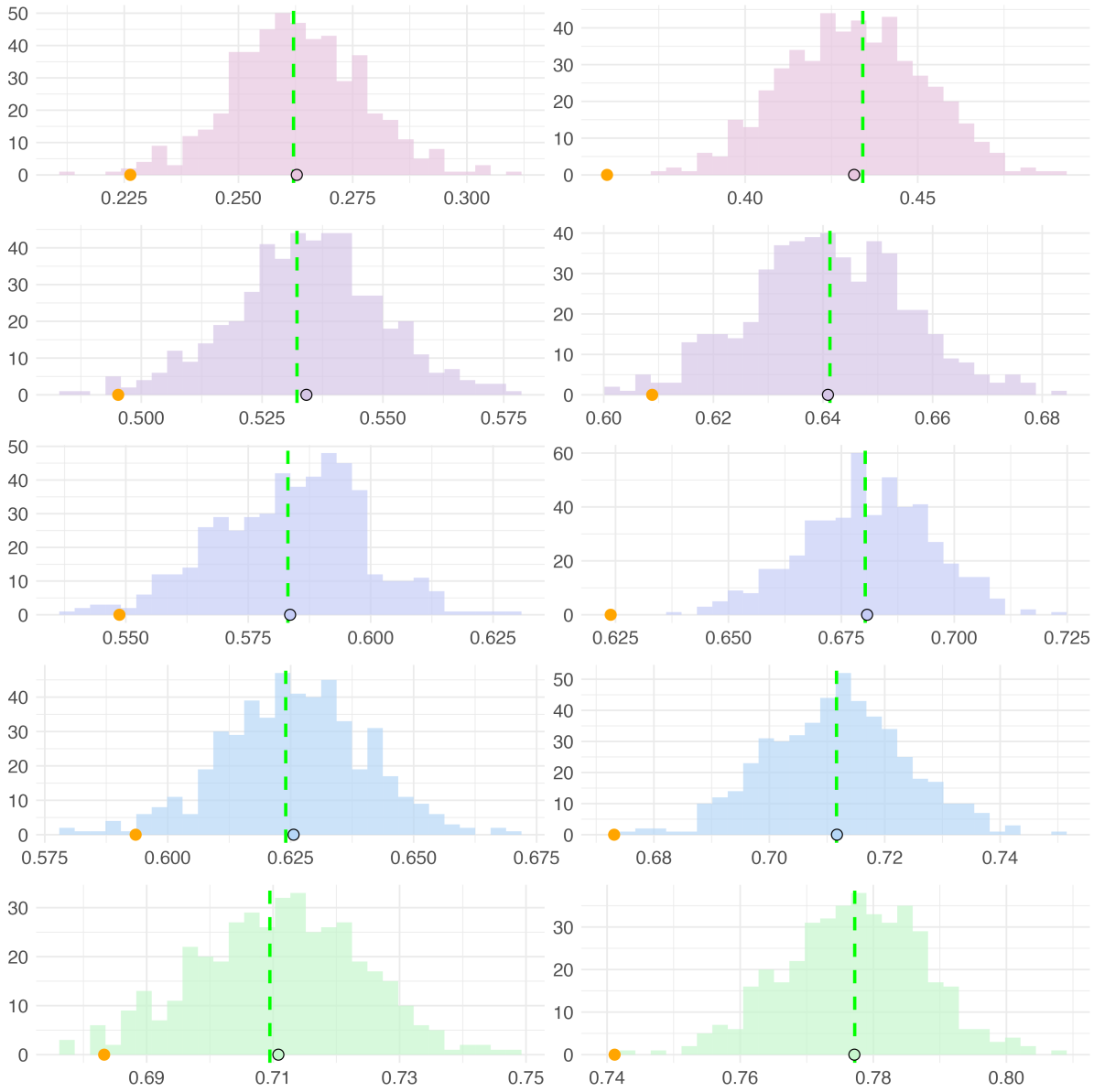

Figure 3: Histograms with the estimated marginal R^2 (left) and conditional R^2 (right) from the BVI method for the binomial regression for the different correlation levels $\rho = -0.4$ (top), $\rho = -0.1$ (second from the top), $\rho = 0$ (middle), $\rho = 0.1$ (second from bottom) and $\rho = 0.4$ (bottom). The values are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ simulations in the simulation study. The expected values are displayed as vertical green lines, and can be found in Table 3, while the orange dot denotes the estimate from the `rptR` package. The mean value of the R^2 values for all simulations is marked with a circle at the bottom of each histogram.

4.1.2 Poisson simulation

The second model fit in the simulation study is a Poisson regression with log-link. The Poisson model was fit with the same correlation levels as the binomial simulation. Further, as was the case for the binomial simulation, we experienced that INLA was not able to fit all $N_{\text{sim}} = 500$ simulations. All models were fit for

$\rho = 0, -0.1$ and -0.4, two models crashed for $\rho = 0.1$ and we had 84 unsuccessful model fits for $\rho = 0.4$. In Table 2, a summary of the mean and 95% quantile values for different correlation levels is displayed.

Measure		$\rho = \mathbf{0}$	$\rho = \mathbf{0.1}$	$\rho = \mathbf{-0.1}$	$\rho = \mathbf{0.4}$	$\rho = \mathbf{-0.4}$
Random Importance	Average	0.1348	0.1159	0.1560	0.0868	0.3299
	2.5%	0.1018	0.0914	0.1212	0.0600	0.2713
	97.5%	0.1689	0.1464	0.1931	0.1072	0.3903
Fixed Importance X1	Average	0.1337	0.1552	0.1112	0.2058	0.0385
	2.5%	0.1251	0.1478	0.1042	0.0002	0.0347
	97.5%	0.1421	0.1625	0.1195	0.2191	0.0425
Fixed Importance X2	Average	0.2672	0.2778	0.2544	0.2853	0.1489
	2.5%	0.2545	0.2647	0.2407	0.0002	0.1342
	97.5%	0.2803	0.2881	0.2679	0.3029	0.1648
Fixed Importance X3	Average	0.4008	0.3957	0.4039	0.3565	0.3237
	2.5%	0.3851	0.3799	0.3838	0.0003	0.2939
	97.5%	0.4185	0.4095	0.4222	0.3784	0.3576
R_m^2	Average	0.8017	0.8286	0.7695	0.8476	0.5111
	2.5%	0.7722	0.7981	0.7363	0.0007	0.4639
	97.5%	0.8304	0.8530	0.8029	0.8960	0.5582
R_c^2	Average	0.9365	0.9445	0.9255	0.9344	0.8410
	2.5%	0.9258	0.9354	0.9136	0.2112	0.8149
	97.5%	0.9467	0.9532	0.9358	0.9660	0.8658

Table 2: Summary of simulation study results for the quantiles of relative importance estimates the Poisson model across different correlation levels.

4.1.2.1 Fixed effects

As for the binomial, we first look at the fixed effects. The estimates of posterior relative importance for the fixed effects (Figure 4) are very similar in shape as the binomial model. The first thing to note is that for the case $\rho = 0.4$, we see that the distribution of relative importance of all covariates contains values very close to zero. After some investigation of the fitted models, it turns out that INLA sometimes returns a model that estimates all fixed coefficients to be 0, but estimates the random effect almost as expected. One might expect that such a model is not sensible and that it should have crashed, but it seems that the model does not crash and has therefore been included in results of the simulation study. With the coefficients estimated to be close to zero, the model will consequently estimate the relative importance of the fixed effects close to zero. It is therefore plausible to believe that the method correctly estimates relative importance, and that this model fit might be caused by some problem with the simulated dataset or INLA.

Disregarding the close to zero values for $\rho = 0.4$, the overall estimates are larger for the Poisson model, which is mainly due to the log-link function having a smaller associated distributional variance than the logit-link and therefore the contribu-

tion by the fixed effects becomes larger. It seems the estimates form a normal curve about the mean, and for $\rho = 0$ the average estimated importance is close to the expected value. The same influence of varying correlation levels can be seen as in the binomial model, namely that we obtain larger importance of the fixed effects when the correlation increases. Again the difference is notably large, with the average relative importance of X_1 going from 0.0385 for $\rho = -0.4$ to 0.2058 for $\rho = 0.4$. For X_2 and X_3 the average relative importance increases from 0.1489 to 0.2853 and from 0.3237 to 0.3565 respectively.

Further, we see the same pattern of larger increase in importance to X_1 than for X_2 and larger for X_2 than X_3, as for the binomial model, when correlation increases from $\rho = 0$ to $\rho = 0.4$. For the Poisson model, it can be seen that for $\rho = 0$, X_3 is on average allocated an importance of 0.4008, and when $\rho = 0.4$, we have no importance estimates larger than 0.4. This is in line with what one would expect, when the off diagonal elements of $\mathbf{\Lambda}$ become large enough so that the distributed importance to the covariate with the largest coefficient becomes smaller with increasing correlation. Contrary, X_1 is again allocated the largest increase and X_2 is allocated a larger increase in importance than X_3 but smaller than X_1. For the case $\rho = 0$, we see that the mean of our samples and the expected relative importance (Table 2) are very close, and it seems the model express the expected pattern of relative importance for varying correlation levels.

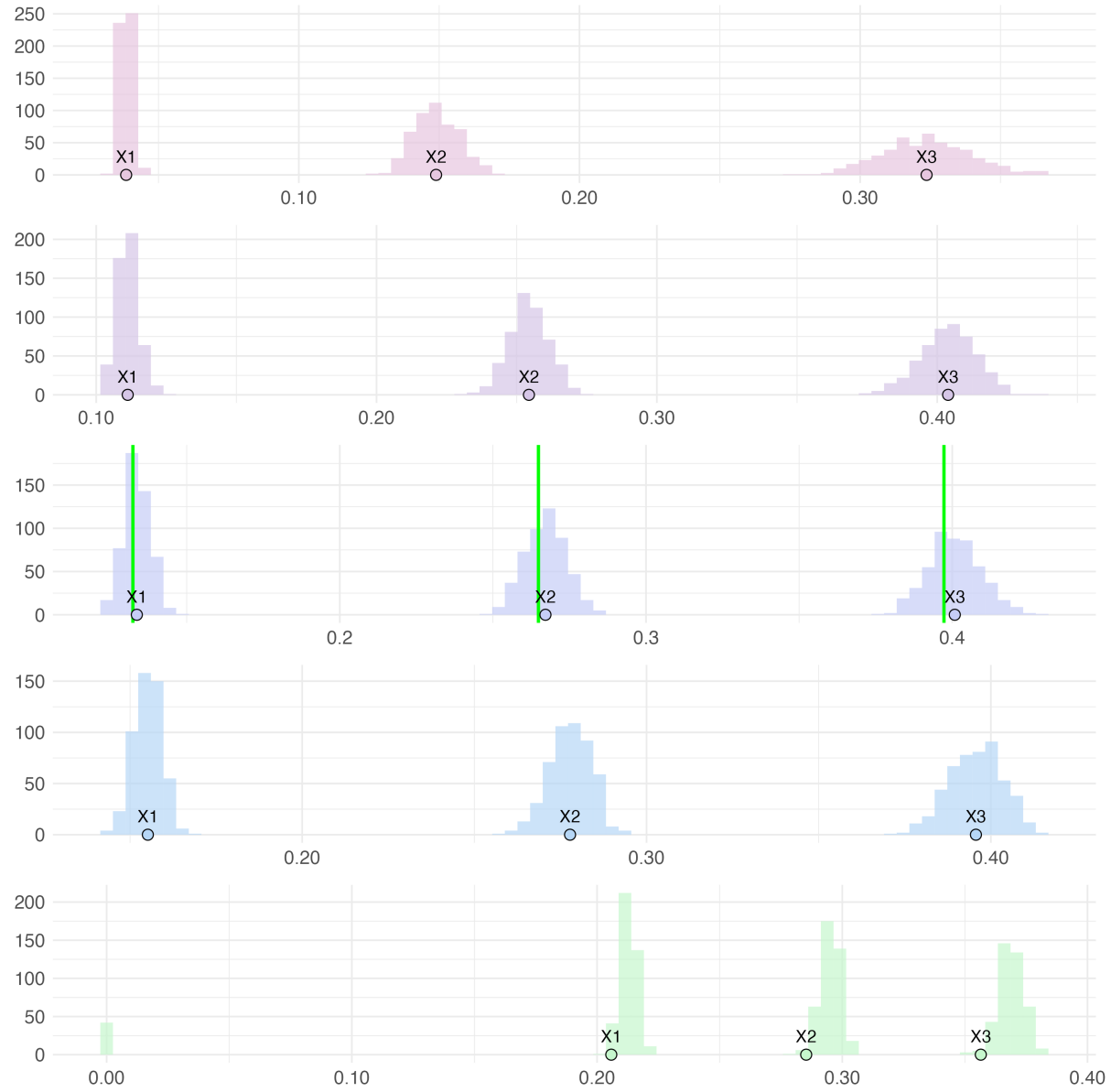

Figure 4: Histogram with relative importance of the fixed effects present in the poisson regression for the different correlation levels $\rho = -0.4$ (top), $\rho = -0.1$ (second from the top), $\rho = 0$ (middle), $\rho = 0.1$ (second from bottom) and $\rho = 0.4$ (bottom). The values are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ simulations in the simulation study. The true regression coefficients are $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$ and the vertical green line for $\rho = 0$ displays the expected relative importance in the case of uncorrelated data. The mean of the relative importance for all simulations is denoted at the bottom of each histogram as a circle.

4.1.2.2 Random effect

When looking at the sampled posterior distribution of relative importance estimates of the random effect in the Poisson mode (Figure 5), we see that they too are in general a bit larger than the same estimates for the binomial case. This is again a consequence of the smaller distributional variance, and so we expect

these results. The shrinkage effect of increasing correlation is also here apparent, with the average relative importance of the random effect going from 0.3299 for $\rho = -0.4$ to 0.0868 for $\rho = 0.4$. The expected value when $\rho = 0$ is 0.1338 as shown in Table 2, and we see that the average estimate is close to this value. The orange dots, denoting the estimates form the `rptR` package are close to the average estimate from the BVI method, with the largest difference being seen for $\rho = -0.1$ with a difference of 0.0260.

As previously mentioned, for $\rho = 0.4$, some models estimated all the fixed effects to be zero. As a consequence of this, the random effect for these models are in turn estimated to have a corresponding larger importance, due to the way we have defined relative importance. Although it is not so clear from Figure 5, it can be seen that we have values of approximately 0.3, 0.9 and 1 when $\rho = 0.4$. Again, albeit unfortunate, we believe that the method correctly estimates the relative importance, based on the given model, and that the fitted models might be results of some problem with the simulated data or INLA.

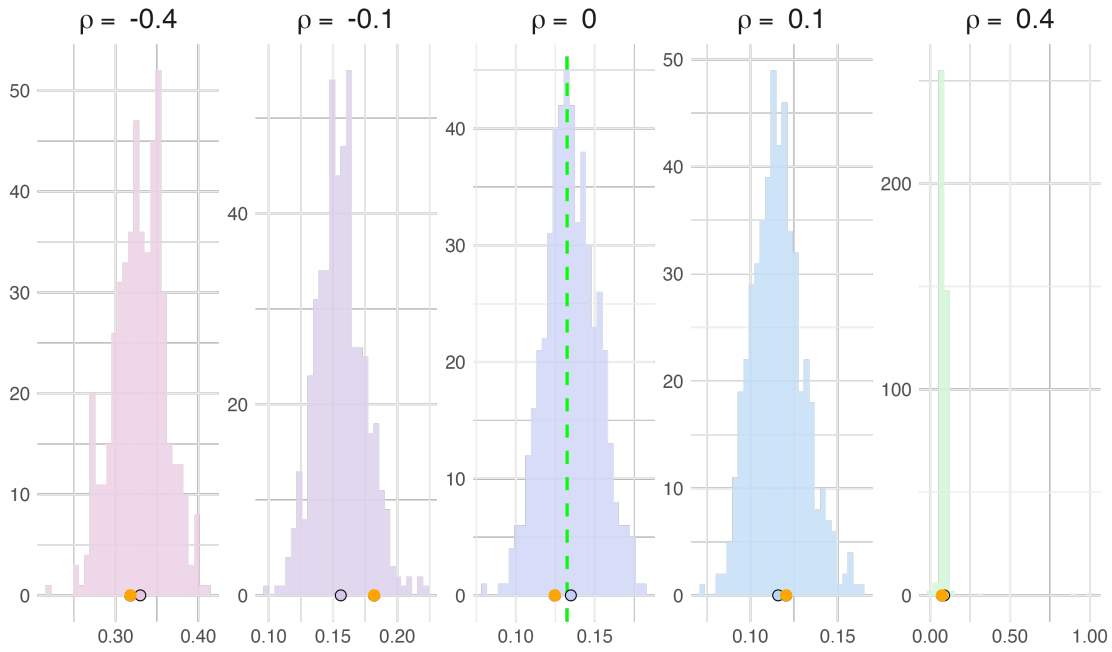

Figure 5: Histogram with relative importance estimates for the random effect $\boldsymbol{\alpha}$ for varying values of ρ calculated by the BVI method. The study conducted $N_{\text{sim}} = 500$ simulations and the mean of the relative importance for all simulations is displayed at the bottom of each histogram as a circle. The vertical green line for $\rho = 0$ is the expected relative importance as in Table 2.

4.1.2.3 R^2 estimates

Moving on to the estimated posterior R^2 distributions for the Poisson model (Figure 6), we see that the expected values from Table 3 are in close agreement with the average marginal and conditional R^2 estimated from the BVI method for all correlation levels. The largest difference in expected values and average values from the BVI method is found for $\rho = 0.4$ and is 0.0286 and 0.0263 for the marginal and

conditional R^2 respectively. As can be seen in Figure 6, we have some values close to zero. This is likely another consequence of the fixed effects sometimes being estimated to be zero, causing our method to estimate the marginal R^2 values close to zero. The conditional R^2 values have some instances where it is much lower than expected, which is likely a consequence of a strange estimate for the importance of $\boldsymbol{\alpha}$ being propagated to the conditional R^2. The estimates from the `rptR` package deviate quite a bit from our method, and are consistently larger for both marginal and conditional R^2. Overall, the distributions seem to form a reasonably symmetric normal curve centered at the mean, close to what we would expect.

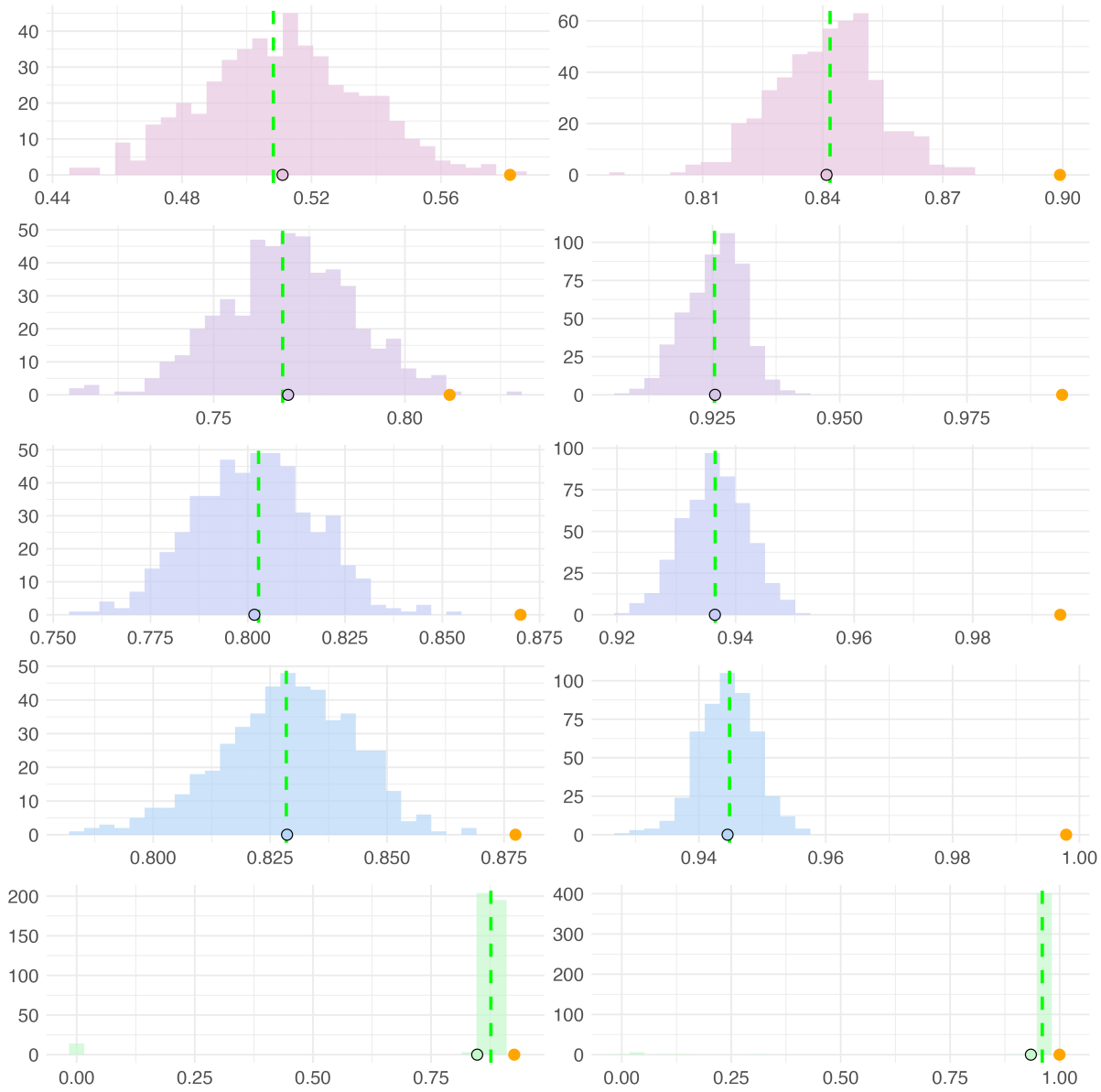

Figure 6: Histograms with the estimated marginal R^2 (left) and conditional R^2 (right) from the BVI method for the binomial regression for the different correlation levels $\rho = -0.4$ (top), $\rho = -0.1$ (second from the top), $\rho = 0$ (middle), $\rho = 0.1$ (second from bottom) and $\rho = 0.4$ (bottom). The values are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ simulations in the simulation study. The expected values are displayed as vertical green lines, and can be found in Table 3, while the orange dot denotes the estimate from the `rptR` package. The mean value of the R^2 values for all simulations is marked with a circle at the bottom of each histogram.

4.2 Comparison with `rptR` package

To further assess our method, a comparison to the vignette for the `rptR` was made. No expected results were available, and so we can only compare our method to the results made by the authors of the vignette. It should however be noted, that

the `rptR` package returns the marginal R^2 as the only measure of importance for the fixed effects, whereas our method directly decomposes this value and assigns a share to each fixed effect. To obtain uncertainty estimates in the likelihood framework, Stoffel, Nakagawa and Schielzeth have built in bootstrap functionality. This is used in our comparison, to evaluate computational complexity and confidence intervals.

The heritability of the color of male beetles is modelled by a binomial GLMM with binary outcome and logit-link. We use the same formulation as in the model `rep11` from the vignette (Stoffel et al. 2017), with the parameter `adjusted=FALSE`. We see that the sampled posterior distribution of heritability Figure 7a from the Bayesian Variable Importance method is centered around a mean of 0.1932, which is very similar to the estimate by Stoffel which is 0.1958. The obtained distribution appears unimodal, with the mode and mean aligning closely. Perhaps a slightly longer tail on the right side can be observed. From 10^3 bootstrap samples, the `rptR` estimates a 95% confidence interval of $[0.051, 0.338]$, which is a bit larger than our estimated 95th percentile of $[0.114, 0.280]$. In terms of computation time, the Bayesian Variable Importance method used 6 seconds to obtain the model fit and 10^4 samples, whereas the `rptR` package used 66 seconds to obtain the model fit and the same number of bootstrap samples.

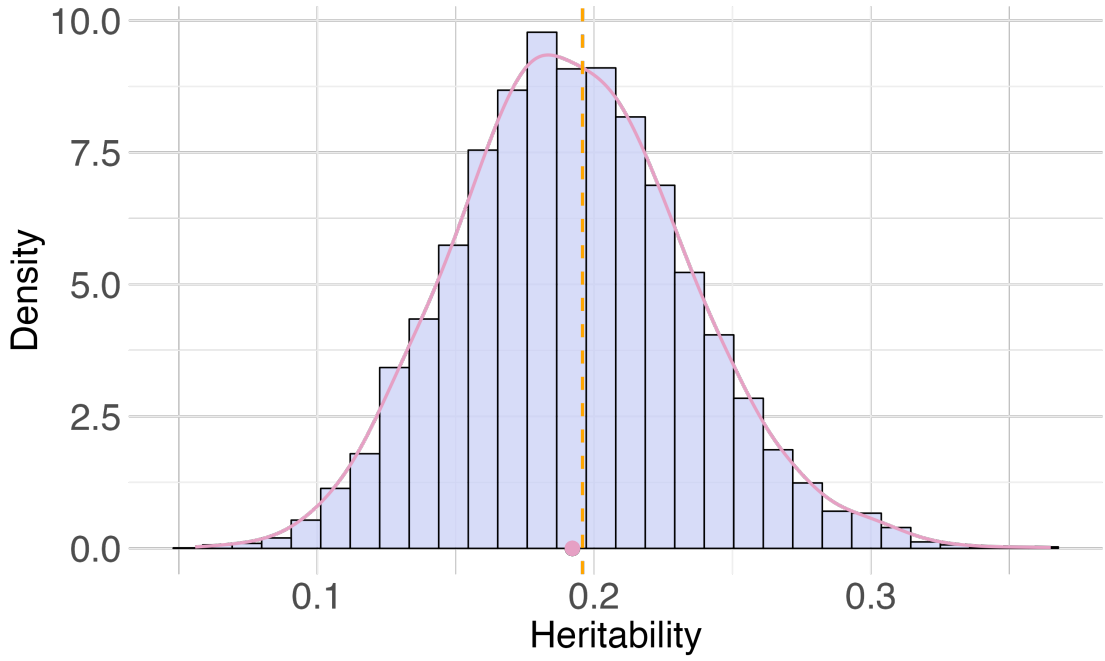

Figure 7: Histogram with heritability values for the color of male beetles from the BVI method, with the estimate from the `rptR` package marked as a dashed line with orange color.

To estimate the heritability of the number of eggs laid by female beetles, we use a Poisson GLMM with log-link. The model used in our method corresponds to `rep9`, but as is described in the vignette after fitting `rep9`, we set the option `expect="latent"` so that the method calculates the distributional variance as in Table 1. This corresponds to the recommendations of Nakagawa et al. (2017) as

previously mentioned. Also this model is estimated from the `rptR` package with `adjusted=FALSE`. From the plotted samples of posterior heritability of eggs laid (Figure 8a), we see a very similar distribution as that of the binomial color model. The distribution is symmetric and centered around a mean of 0.3585. Further, the estimate from the `rptR` model is 0.3795 with a confidence interval of $[0.131, 0.542]$, compared to our 95th percentile of $[0.278, 0.445]$. The 10^3 bootstrap samples and model fit for the `rptR` package took 2 minutes and 13 seconds, whereas the BVI method used 8 seconds to obtain the model fit and 10^4 samples.

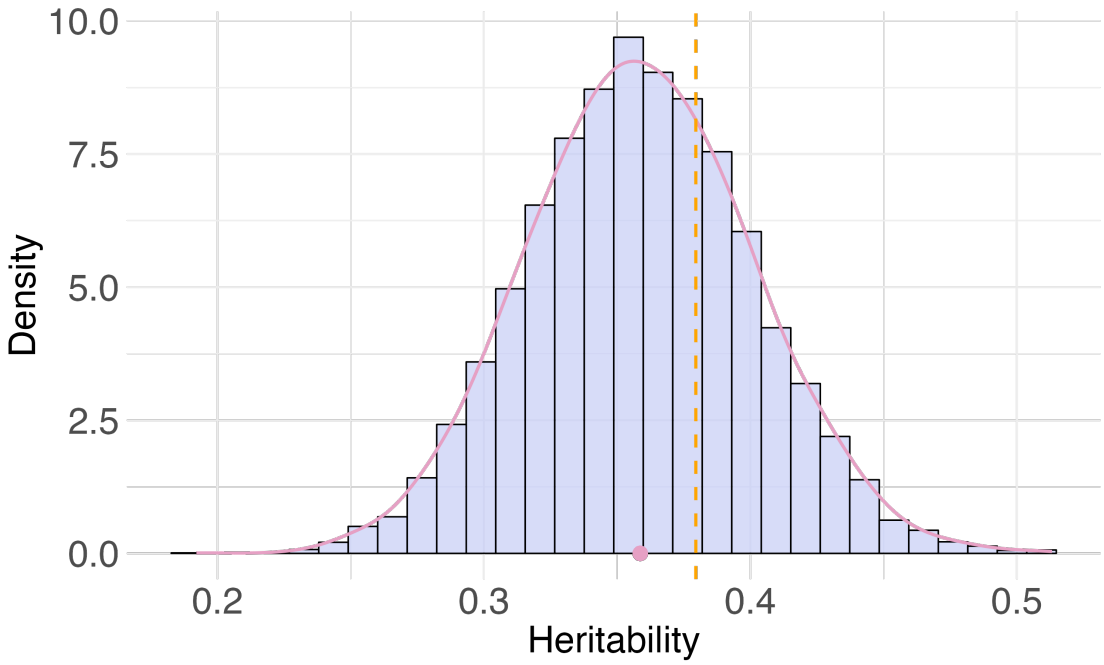

Figure 8: Histogram with heritability values for eggs laid by female beetles from BVI method, with the estimate from the `rptR` package marked as a dashed line with orange color.

It should perhaps be mentioned, that the estimates from the BVI method will vary each time a model is fit, as it is stochastic. Therefore, it could be that another fit from the BVI method might align closer with Stoffels results, but it could also be further off.

4.3 Heritability of house sparrow traits

We now investigate the results of applying our method to the house sparrow dataset. As previously discussed (Section 2.7), estimating the heritability of phenotypic traits can be seen as a special case of relative variable importance and so the findings we present are directly obtained by our method. The samples of relative variable importance presented, are sampled from the variance component that captures additive genetic variance, and we use the results from Silva et al. (2017) and Muff et al. (2019) to compare with. For this case, the covariance structure of the pedigree required us to model more complex random effects than i.i.d. random intercepts, and so the `rptR` package could not be used for comparisons. In Table 3

the mean of sampled heritability along with confidence intervals is presented, as well as the corresponding measures from the comparable studies.

	Silva et al. (2017)		Muff et al. (2019)		BVI	
	Estimate	CI	Estimate	CI	Mean	CI
h^2_{mass}	0.300	[0.231, 0.369]	0.288	[0.219, 0.371]	0.283	[0.232, 0.344]
h^2_{wing}	0.388	[0.353, 0.461]	0.344	[0.294, 0.409]	0.356	[0.322, 0.393]
h^2_{tarsus}	0.415	[0.333, 0.497]	-	-	0.401	[0.330, 0.470]

Table 3: Heritability estimates and confidence interval from Silva et al. (2017), posterior means of additive genetic variance divided by the posterior means of total phenotypic variance in Muff et al. (2019) with corresponding confidence interval and the mean and confidence interval of the heritability samples obtained from the BVI method for the phenotypic traits; body mass, wing length and tarsus length.

For the sampled heritability of body mass (Figure 9), we have a mean of 0.2838 (Table 3) and a distribution that is centered around the mean. The distribution does not seem to be symmetric, and might show signs of being bimodal, with one larger peak at the mean and possibly one smaller peak to the left of the larger peak. Furthermore, the distribution exhibits a longer tail to the right and the 95th percentile is approximately the interval $[0.2324, 0.3430]$. To fit the model and obtain the samples, the method took about 73 seconds.

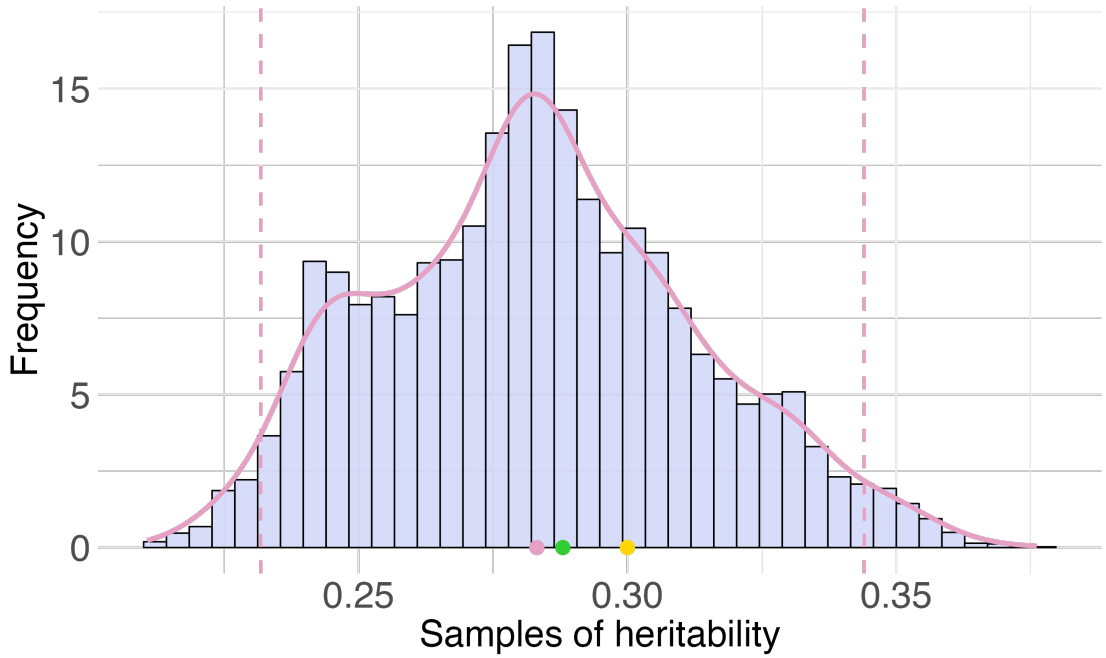

Figure 9: Histogram depicting the estimated heritability values of body mass by the BVI method for the house sparrow dataset. The mean of the samples is marked as a circle at the bottom of the histogram, with the lower and upper value for the 95% percentile marked as dashed lines. The heritability estimate from Silva et al. (2017) and Muff et al. (2019) are marked as gold and green dots respectively at the bottom of the histogram.

The samples of wing length heritability form a more symmetric curve, centered around a mean of 0.3560 (Figure 10 and Table 3). There are fewer signs of another mode for these samples, but one could argue that the right tail is a bit longer also for the heritability of wing length. The 95th percentile is approximated by the interval $[0.3224, 0.3930]$, which is smaller than the same percentile for the body mass by a factor of 0.63. The samples of heritability for wing length therefore exhibits less dispersion than those for heritability of body mass. In this case, the model fit and sampling procedure took 76 seconds.

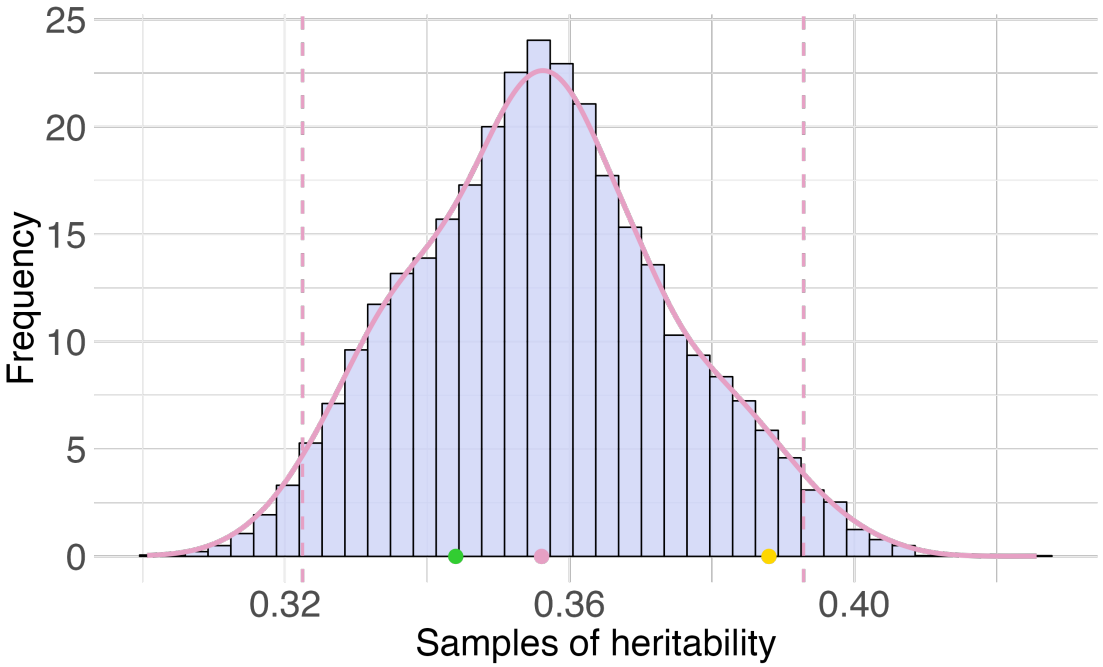

Figure 10: Histogram of heritability values for wing length of the house sparrows estimated by the BVI method. The mean of the samples is marked as a circle at the bottom of the histogram, and the lower and upper value for the 95% percentile are featured as dashed lines. The heritability estimate from Silva et al. (2017) and Muff et al. (2019) are marked as gold and green dots respectively at the bottom of the histogram.

The heritability samples of tarsus length (Figure 11) has a mean of 0.4015 (Table 3) and the distribution is centered around this value. An obvious observation here is that the samples form a trimodal distribution, with three very distinct peaks. The center peak is the highest and centered around the mean and the right and left peak seem to be of equal height and symmetric about the center peak. A possible explanation for this pattern is that for this trait, the grid used for numerical integration might not be fine enough, forcing the sampling to occur most frequently at the three modal values. The 95th percentile is captured by the interval $[0.3316, 0.4669]$ and the time spent to fit the model and draw the samples was reported to be 74 seconds.

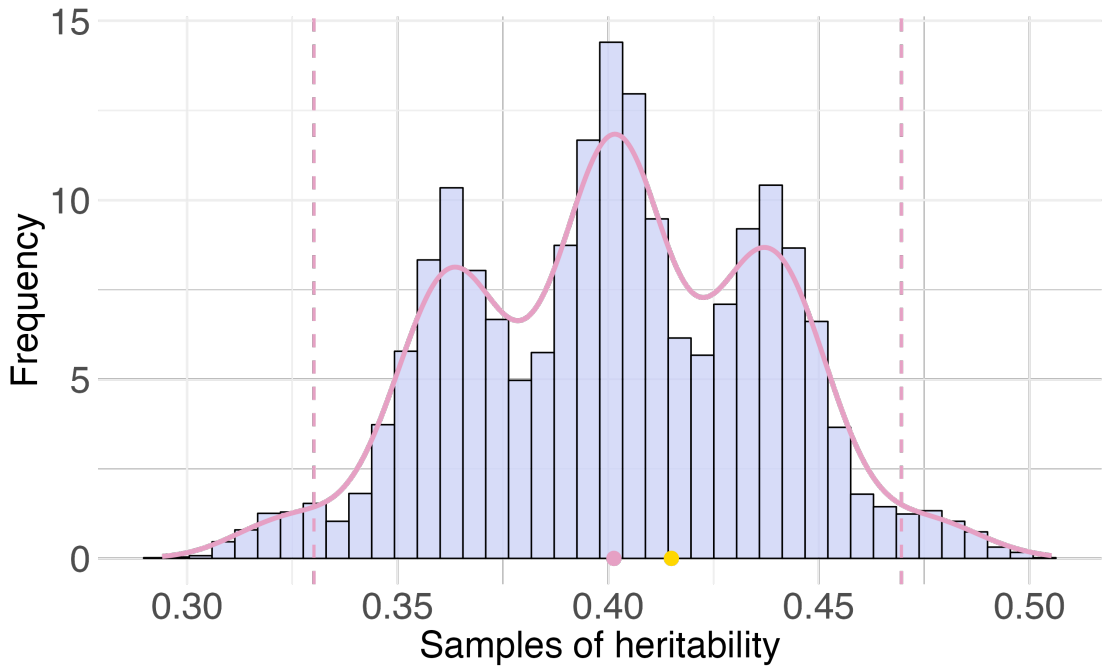

Figure 11: Histogram showing estimated heritability values for tarsus length of the house sparrows from the BVI method. The two dots at the bottom represent the mean of the samples (pink) and the estimate from (Silva et al. 2017) (gold). The dashed lines represent the lower and upper value for the 95% percentile.

We see it as natural that we see different patterns that are hard to fully interpret, as the dataset is from real life and relatively small. Further, the measurements are taken on birds that are quite small, so one should expect measurement error to some degree.

4.4 Comparing the BVI method with R2D2 and GDR2 priors

To explore other possible relative variable importance tools in the Bayesian framework, we have discussed R2D2 and GDR2 priors in Section 2.5.6 and Section 3.6. We now present the results of applying the R2D2 and GDR2 priors to a linear regression, and see how they can be interpreted as relative importance measures. The results are compared to the BVI method and the LMG method serves as a robust benchmark for all methods. Please note once again that the R2D2 and GDR2 priors are originally developed for prediction models, and that the results presented here are based on our interpretation of how the R2D2 and GDR2 priors can be used for relative importance. The theoretical importances for uncorrelated covariates are found in (3.18) and the theoretical R^2 for all correlation levels are found in Table 4.

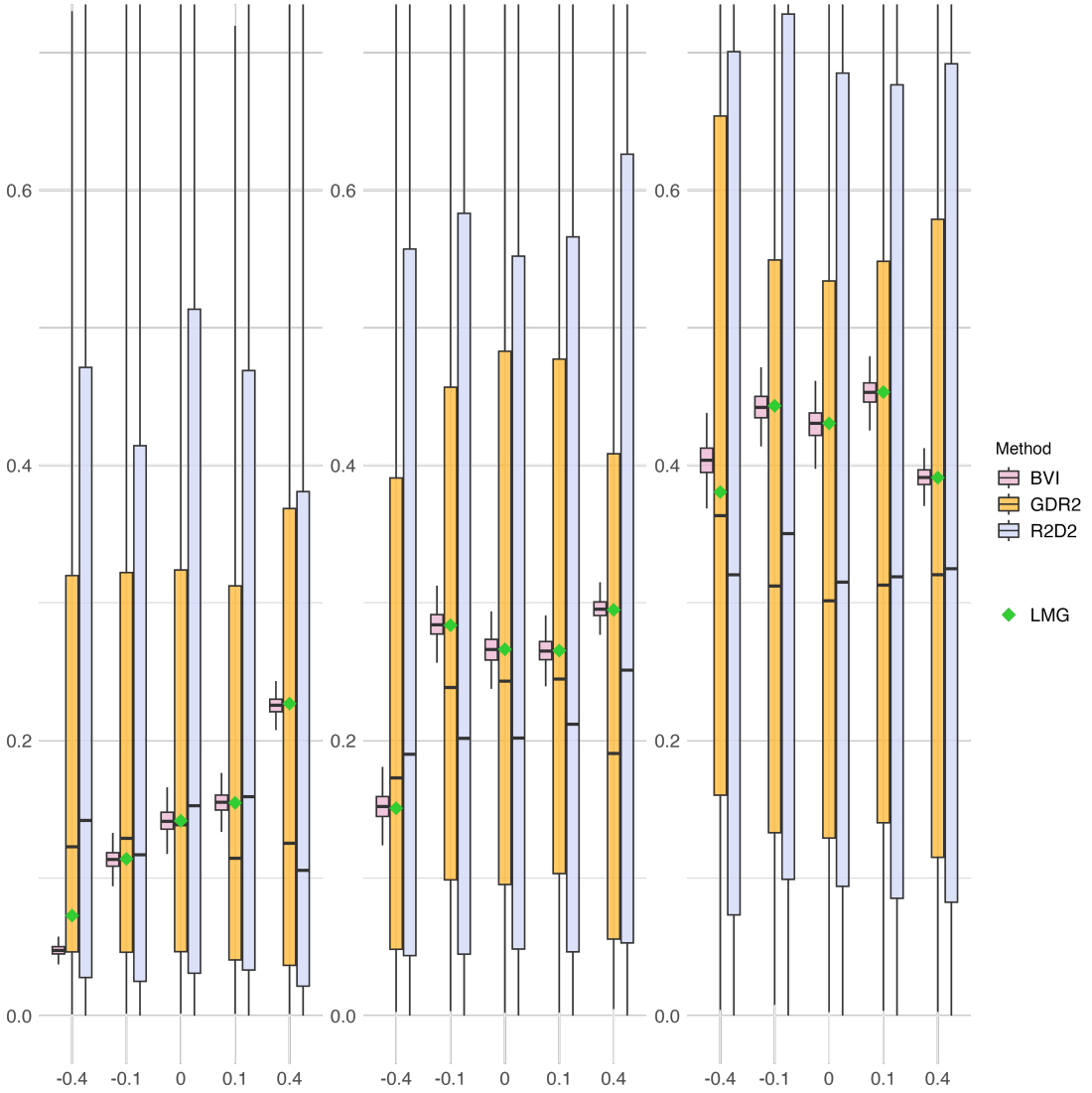

Figure 12: Boxplots of the relative importance distributions for X_1 (left), X_2 (middle) and X_3 (right) for varying correlation levels by the R2D2, GDR2 and BVI methods. The correlation levels ρ are denoted along the x-axis, and the green diamond represents the relative variable importance calculated from the LMG method.

The first thing one notices from the posterior relative importance distributions (Figure 12), is that the spread from the R2D2 method is larger than the spread of the GDR2 method, which again is significantly larger than the spread of the BVI method. In the boxplots, the 25th and 75th quantiles define the interval of each box. The mean values of the relative importance distributions for the R2D2 and GDR2 methods do not seem to follow any pattern at all for varying correlation. They are more similar across correlation levels for all covariates, and do not adjust for correlation as the benchmark LMG method. Although it is hard to find patterns and the R2D2 and GDR methods give very flexible results, it becomes clear from they grasp the larger aspects, in that X_3 is estimated to be the covariate with the largest variance contribution, followed by X_2 and lastly X_1. The BVI method is in close agreement with the LMG method, with deviation for X_1 and X_3 when $\rho = -0.4$. Overall the BVI and LMG methods follow the same

pattern for varying correlation, as we expect and have previously discussed.

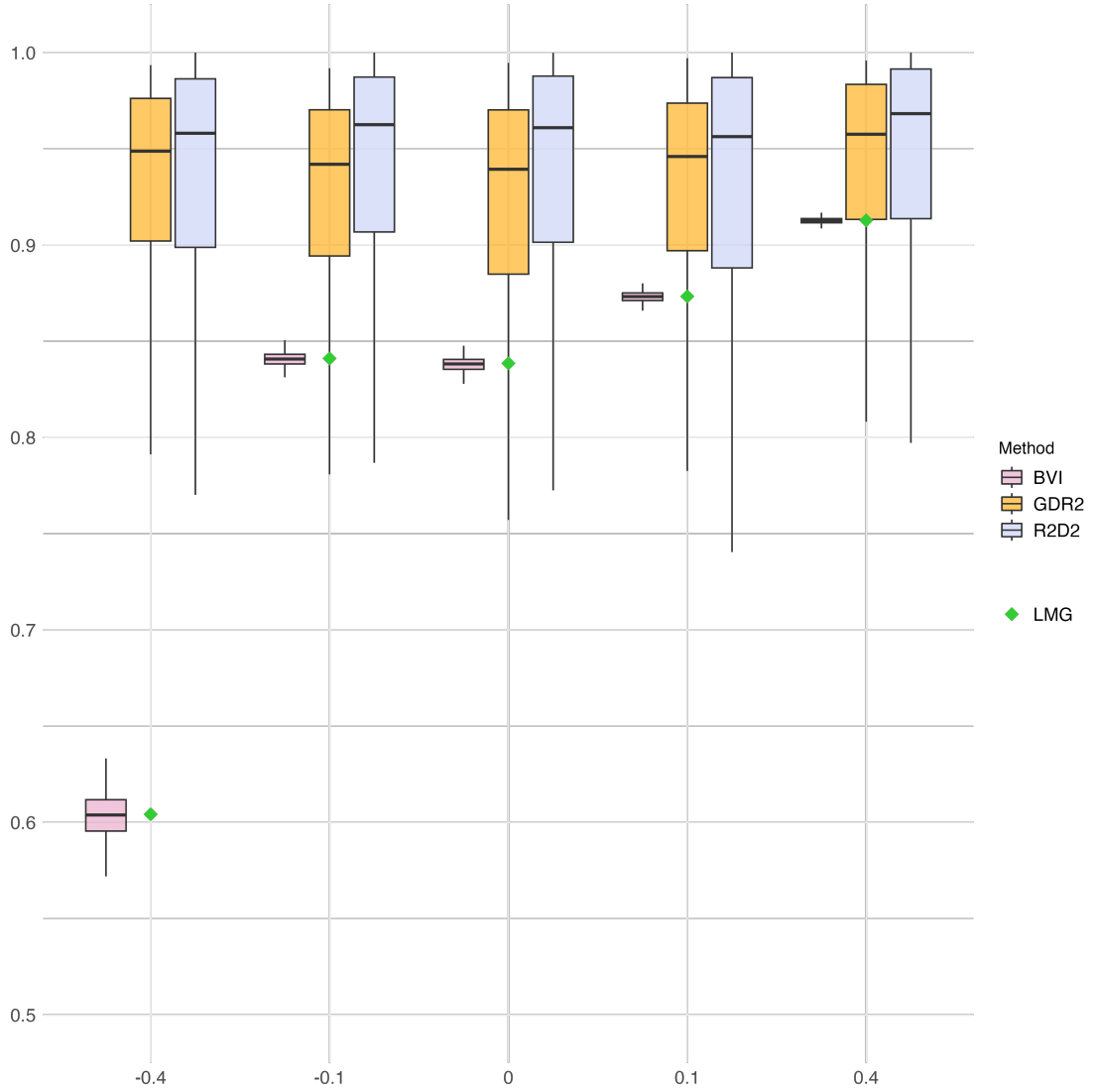

Figure 13: Boxplots of the estimated R^2 distributions for varying correlation levels ρ for the R2D2, GDR2 and BVI methods. The correlation levels ρ are denoted along the x-axis, and the green diamond represents the relative variable importance calculated from the LMG method.

The estimated posterior R^2 distributions for the R2D2 and GDR2 priors (Figure 13) are uniform across the correlation levels, and significantly larger than both the estimates of the BVI and LMG methods. For the R^2, the spread of the R2D2 and GDR2 methods are quite similar, with both being much larger than that of the BVI method. Overall, the R2D2 and GDR2 estimates a much larger R^2 than the BVI and LMG methods, and the means deviate quite much from the expected values. The BVI method is consistent with the LMG method, and they both follow the expected pattern of the R^2. For the expected R^2, the BVI and LMG methods seem to align closely with the expectation, with a small deviation for both methods when $\rho = 0$.

As the R2D2 and GDR2 methods are not specifically relative variable importance

measures, one should also interpret the results with this in mind. The development of the R2D2 and GDR2 priors have been done to produce robust predictions for high dimensional linear regression models, and the results presented here are therefore not an evaluation of the R2D2 and GDR2 priors as relative importance measures. Moreover, the interpretation made to obtain these results was made by the author, and the results should be taken with caution. The main motivation behind this comparison was to explore other possible relative importance measures in the Bayesian framework, as this is a small field with few available methods.

CHAPTER

FIVE

DISCUSSION & FURTHER WORK

The work presented in this thesis partially builds on the authors previous work in Arnstad (2024) and has developed the Bayesian Variable Importance method to put forward a novel framework for estimating relative variable importance in generalized linear mixed models. As a result, the BVI method is now capable of handling complex models, while at the same time proving to be computationally efficient. A Bayesian approach to variable importance has involved utilizing the relative weights method (Johnson 2000) to project the fixed covariates into an orthogonal space. The projection, or approximation, of these covariates are used to fit the model, before a back-transformation is applied to relate the estimated results back to the original covariate space. To obtain inference on the Bayesian GLMMs, we have translated frequentist concepts, such as the R^2 measure, to fit in the Bayesian framework. This translation has been inspired by the work in Gelman et al. (2017), but is also a result of the authors own work. Once the methodology was developed, a simulation study was conducted in which the underlying structure was known. After a simulation study, the method was applied to a case study in which we could compare the model to a similar relative variable importance measure, `rptR`, in the frequentist framework. Lastly the method was applied to a dataset gathered from house sparrows on Helgelandskysten, Norway, to investigate the heritability properties of the sparrow population. The methodology has been implemented in an R package, `BayesianVariableImportance`, which is available in full on the authors Github, with a link to the repository provided in Appendix A. In Appendix B a usage example of the package is supplied, which is also available on the authors Github along with all code used to obtain the results of this thesis.

Being a general method, our aspirations are that the BVI method will be applied by researches across disciplines that are interested in the statistical properties of covariates in GLMMs. The BVI method does not aim to give researchers an exact measure of variable importance, but rather provide posterior distributions of relative importance that should be interpreted by the researcher in the field of application. As the distributions will naturally have an uncertainty, it is advantageous if this uncertainty is assessed and understood as a part of the analysis. Hopefully, this can give broader inference on the importance of the covariates, which will in turn lead to more informed conclusions on the effect of covariates on

a response. In itself, the BVI poses an analogue to the frequentist relative variable importance measure `rptR` for non-Gaussian responses, but with the added benefit of directly estimating the relative importance distributions of fixed effects. Further, for gaussian data, it also poses an analogue to more established methods such as the LMG method (Grömping 2007), the extended LMG method (Matre 2022) and the extended relative weights method (Matre 2022) as discussed in Arnstad (2024). Lastly, the BVI method allows one to specify covariance structures in the random effects, which can be beneficial when modeling complex data structures.

The field of Bayesian variable importance measures for regression models is not very large, but there has been some research on the topic. Specifically, the use of continuous shrinkage priors for linear models of high dimension has attracted attention (Aguilar & Bürkner 2024). Two priors that can be applied as continuous shrinkage priors and that have favorable properties for variable importance are the R^2-induced Dirichlet Decomposition (R2D2) priors (Zhang et al. 2020) and its generalization to Generalized Decomposition R^2 (GDR2) priors. Through a simulation study on a linear regression model, the use of R2D2 and GDR2 priors were compared to the BVI method with the LMG method as a benchmark. The results show that the R2D2 and GDR2 priors are not very rigid, by estimating almost uniform distributions of relative posterior variable importance. One could argue that they prove to be too flexible and therefore not suitable to give accurate results. However, the use of these shrinkage priors are primarily not focused on calculating the specific variable importance, and they could perhaps be developed further, with this as the emphasis, to yield more suitable estimates for posterior relative variable importance. As the BVI method and the shrinkage prior methods have been developed for different purposes, the R2D2 and GDR2 methods differ from the BVI method in some fundamental ways. Firstly, to our knowledge, the shrinkage prior methods have yet to be applied for GLMMs and so direct comparison for the most complex models is not possible. Further, we sample values of coefficients and random effects a posteriori and then estimate the relative importances based on the samples. This means that the estimates from the model are used, which in most cases do not vary greatly for different model fits to the data. On the other hand, the R2D2 and GDR2 priors consider the relative variable importance as a parameter in the model, and therefore places prior values on the relative variable importance directly. When placing the priors directly on the importance, one must keep in mind that the choice of priors are the most criticized topic in Bayesian statistics (Robert 2007). Taking into account that the user is often not informed about the underlying mechanism in the relative variable importance parameter, we assume that they will parameterize the priors in such a way that they reflect the lack of information. Therefore, we see it as sensible that the users initial lack of precision propagates into the final posterior distribution of relative variable importance. This could be a reason why the estimates for the shrinkage priors are more spread out than estimates for the BVI method. Moreover, the priors themselves differ, as the R2D2 and GDR2 priors are continuous shrinkage priors, which are designed to shrink small effects towards zero. We use penalizing complexity priors, which puts the emphasis on the complexity of the model. This means the general idea for the priors used is the same, but the implementation and interpretation of their results are different. Lastly, it should be

mentioned that the results in this thesis are based on the authors interpretation of how the shrinkage priors can be used for relative variable importance. The author gained this knowledge by reading the papers Zhang et al. (2020) and Aguilar & Bürkner (2024) and by discussing the topic with the authors of Aguilar & Bürkner (2024). Therefore, we believe that one could further optimize the use of shrinkage priors for variable importance by further studying the topic. MORE?

For relative variable importance measures, some criteria are found in Section 2.2.1 that it is desirable to fulfill. It was argued that the simulation study in Arnstad (2024) gave promising results of the BVI method fulfilling the proper decomposition criteria. When assessing how the BVI method performs on GLMMs, in which the response variance is not on the same scale as the covariates, this criteria is hard to assess. Instead of aiming to decompose the total model variance, we find it natural to rather aim for a proper decomposition of the models R^2 on the latent scale. From the definition of R^2 for GLMMs in Nakagawa & Schielzeth (2013), the simulation study shows that the posterior distributions of the marginal and conditional R^2 are generally symmetrically distributed around the expected R^2 value. When the correlation between fixed effects is 0.4, we see that the R^2 estimates from the Poisson model are slightly smaller on average than the expected value. The average is of course effected by the model fitting problems causing the estimated R^2 values to be artificially small. Therefore, it is plausible that for a good model fit, the BVI method will give estimate the R^2 values closer to what one might expect. Based on these observations, we argue that the BVI method, in posterior expectation, is capable of providing a satisfactory decomposition of the R^2 in GLMMs. The non-negativity criteria is fulfilled by recalling that the relative importance estimates of fixed effects are squared, and no variance estimate for random effects can be negative. Consequently, the posterior relative importance distributions will not contain negative values. As discussed in Arnstad (2024), the exclusion criteria will not be used in our assessment, as Grömping (2007) argues this is not in general reasonable. Lastly, violating the inclusion criteria is seen as unlikely to occur in practice, although it is mentioned in Matre (2022) that the extensions of the LMG method and the relative weights method can violate this criteria. It has not yet been properly assessed how the inclusion criteria applies to the BVI method. A suggestion that was debated in (Arnstad 2024) is whether one should directly translate the desirable criteria for relative importance measures in the frequentist framework to the Bayesian framework. In the case of the inclusion criteria, we interpret this to mean that if the posterior relative importances of a non-zero regressor contains zero, this is a violation the criteria. The Bayesian framework is designed to provide uncertainty, and therefore subjecting its result against a rigid threshold of containing or not containing zero is not necessarily reasonable. By not considering the inclusion criteria, the inclusion of zero values in the relative importance distributions of a non-zero regressor would require the researcher to carefully assess the covariate. A careful evaluation of the results is in line with what we intend the BVI method to invoke, and therefore the violation of the inclusion criteria might not pose a problem at all. With this in mind, the results of the simulation study show that the BVI method produces results that align well with what we expect, and that the results are plausible. Therefore, we believe that for most practical applications, the general idea behind the criteria

of variable importance measures are fulfilled by the BVI method.

It could be questioned if our investigations of the Bayesian Variable Importance method has been sufficient. For example, in the simulation study we do not allow for more extreme correlation levels than -0.4 and 0.4. This is not a very large value, and therefore some analysis on the method for higher correlation levels could be of interest. The reasoning behind using such moderate correlation levels, is that the model fitting procedure was often compromised for more correlated covariates. As mentioned in the results, we experienced model crashes for a correlation of 0.4, and the frequency of crashes rose when testing with higher correlation. It is therefore not clear how well the model will perform if covariates share much information, but results for an LMM with correlation levels of 0.9 can be found in Arnstad (2024). Further, our initial idea was to also include a binomial model with the probit link function in the simulation thesis. This idea was abandoned due to severe model fitting problems, in which INLA would not converge. This was of course unfortunate, but as the result could not be trusted, we chose to ommit them from the thesis. We do however believe that if the probit model had been a good fit, the method would be able to calculate the relative importances in a similar manner as for the logit model.

Another topic of discussion, regards choice of priors. It would be natural, given more time, to investigate how different priors would perform and also if one could tune the hyperparameters of the priors applied. As priors are a large subject in themselves, a thorough analysis of prior effects on the BVI method was not performed. We chose to follow the recommendations of Simpson et al. (2017) to use penalizing complexity priors, as these had desirable properties and are designed to nicely fit INLA models (Simpson et al. 2017). The parameters of our PC priors follow the default values in the `R-INLA` package, and we have not investigated how these could be tuned to better fit the data. This could be done to further solidify the results of the BVI method, but would also require more time and resources than what was available in the scope of this thesis.

Many of the foundational calculations made in the BVI method relies on approximations. The relative weights method can be viewed an approximation of the Lindemann, Merenda and Gold (LMG) method, and the accuracy of INLA is dependent on how well the marginals are approximated. It is to be expected that the errors made in these approximations are propagated to the outputted results of the BVI method. However, we argue that the results obtained from the BVI method are satisfactory. For the simulation study, the results align well with our expectation in cases where we can give an expectation. Further, the patterns for varying correlation levels seem to be logical, and the results plausible, even though a true value is hard to obtain. A worrying aspect of the simulation study was the results for the highest correlation level in the Poisson model. The models that estimated all fixed coefficients to be zero are not sensible, and their origin is not clear. Investigations into the cause of these problems showed that such models with zero fixed coefficients spontaneously appear when fitting a large number of INLA models. We believe that the trouble here lies with the INLA fitting procedure, and we therefore stress the need to make sure the model is a good fit

before interpreting the posterior relative variable importance distributions. The BVI method seems to provide very similar results to the `rptR` package which is a comparable method. In addition to give similar results, the BVI method proves to be much faster than the frequentist counterpart in `rptR`. The computational gain is mostly due to the use of sampling rather than bootstrapping, which we hope will be a benefit for researchers who wish to apply the method. Perhaps the most positive result, is the performance on real data from the house sparrow study. The method gives strikingly similar estimates as those done by domain experts, further solidifying our belief that the methodology and implementation is sound. However, the posterior distributions of relative variable importance for the house sparrow study are not very smooth and returns a pattern that is hard to interpret. Nonetheless, based on the results we have believe that applying the BVI method is accurate enough to be viewed as drawing samples from a distribution centered around the true value. ADD SOMETHING HERE? NEED TO ASK STEFFI ABOUT WHAT SHE THINKS THIS IS THE RESULT OF.

We are not at the time aware of a similar variable importance tool for Bayesian GLMMs as the BVI method. Therefore, there is still much work to be done in this field, and many opportunities for expanding the BVI method. Currently we have implemented the BVI method to handle Gaussian, Binomial and Poisson distributed responses, but there are a number of other distributions that could be of interest. In Nakagawa et al. (2017), the quasi-Poisson, negative Binomial and Gamma distributions are analysed, so these would be natural extensions. Further, extending the BVI to also handle multiplicative overdispersion would allow the user to specify if the overdispersion should be modelled as additive or multiplicative and would be a valuable addition.

It was also desirable in Arnstad (2024), to go deeper in to the theoretical properties that the BVI method possesses. As the BVI method is first and foremost a tool for researchers, the main focus of this thesis was put on developing a credible variable importance measure and wrap this in an R package so that it could be applied. Due to the complexity of this, the time and resources did not allow for a full theoretical breakdown of the BVI method. Such analysis would be of very high interest, in particular some proofs in expectation for the variable importance estimates would be helpful, to further solidify the credibility of the method.

We did not consider random slopes when developing the BVI method, but this could also be a possibility for further work. As the random slopes are often associated with a fixed effect, the correlation structure one obtains with random slopes is much more complex than that of random intercepts. This could be a difficult challenge to implement, but as discussed in Section 2.4.2, the proposal by Johnson (2014) could be a good starting point. One should however be careful to make sure that the random slope improves the model significantly, and that this improvement outweighs the potential computational burden including random slopes could have.

Conceptually, variable importance is in itself a debated topic. The first question one can ask is what the definition of relative importance is. In Grömping (2007), relative importance is based on variance decomposition and we have cho-

sen to follow this notion. However, this definition has the disadvantage that an agreement of allocation of importances for correlated covariates seems impossible (Grömping 2015). This problematic issue is present in our results when the fixed effects were correlated, making evaluation of the method difficult. For our method, the pattern observed was a consequence of the relative weights method, rather than a general method for distributing the shared variance between covariates. The search for a unified variable importance framework has given us methods such as the LMG (Grömping 2007), Proportional marginal variance decomposition (PMVD) (Grömping 2007), the relative weights method (Johnson 2000) and dominance analysis methods (Budescu 1993). Yet, no one has been able to provide a method that is completely accepted by the field of mathematics. For these reasons, variable importance as a subject, and its methods, have received criticism (Grömping 2007). However, we believe that variable importance methods can give researchers valuable information and spark ideas, and that they therefore should have a place in the statistical toolbox. That being the case, we wish to emphasize that all statistical methods are limited by the assumptions they rely on and the data they are applied to. As Chevan & Sutherland (1991) put it; *Statistical techniques do not build theory - theoreticians do.*

- Summarize the method. Similar to that of the project thesis.
- State that it can be used across diciplines, and that a Bayesian approach is useful when prior information is available.
- Now the discussion really begins. State that the methodology for LMMs proved to imply that we have a proper decomposition of the R^2. Even though this was not a main focus in this thesis, the results of uncorrelated covariates and marginal and conditional R^2 values seem to be in line with the what one would expect.
- State that we have addressed two main points from the project thesis, namely testing the methodology on real data and expanding it to handle GLMMs.
- Further, we allow for a covariance structure in the random effects, which is not possible in the project thesis.
- Emphasize that the results in this thesis are calculated based on theory from a subject that in itself has been subject to criticism. Therefore, the results should be interpreted with caution, especially when we also use the definitions of R^2 from Nakagawa, which may be oversimplified. (See last discussion section of project thesis)

Further work:

- Extend the package to encompass the models with known distributional variance as in Nakagawa & Schielzeth (2013) and Nakagawa et al. (2017).
- No proofs were considered due to time limitation
- Random slopes could be featured, at least if one can say they improve the model.

- Correlated random effects (?)

DET BØR NEVNES AT USIKKERHETEN I RESULTATENE FRA BVI ER GANSKE SMÅ OFTE, SÆRLIG NÅR DATAEN ER SIMULERT, OG AT DET SÅLEDES IKKE KAN SIDESTILLES MED ET KONFIDENSINTERVALL.

HUSK Å NEVNE AT DET ER HELT NATURLIG AT VI FÅR RESULTATER SOM VARIERER LITT FRA GANG TIL GANG. DETTE KAN FORKLARE HVORFOR VI FOR EKSEMPEL IKKE TREFFER STOFFELS ESTIMATER OG BIOLOGENES ESTIMATER, MEN, BASERT PÅ SIMULERINGSSTUDIEN, FØLER VI OSS TRYGGE PÅ AT DENNE SPREDNINGEN IKKE ER FOR STOR, OG AT EN KJØRING AV METODEN KAN SEES PÅ SOM EN TILFELDIG PRØVE FRA EN FORDELING SOM ER SENTRERT RUNDT DEN KORREKTE VERDIEN.

LEGG TIL TO GITHUB LINKER, EN FOR PAKKEN OG EN FOR MASTEREN.

KAN DISKUTERES OM VI BURDE UNDERSØKT BEDRE PRIORS FOR ANIMAL MODEL OG DE ANDRE SIMULERINGENE KAN DISKUTERES OM VI BURDE BRUKT HØYERE KORRELASJON, MEN DA TROR JEG IKKE MODELLENE VILLE BLITT KONVERGENTE JEG TROR METODEN HELT FINT KLARER KATEGORISKE KOVARIATER NÅR DUMMY ENCODING BRUKES

DISKUTER SHRINKAGE PÅ RANDOM EFFECTS MED POSITIVT KORRELERTE FIXED EFFECTS OG INCREASE PÅ NEGATIVT KORRELERTE DISKUTER HVORFOR POISSON MED HØY KORRELASJON GIR SÅ RARE RESULTATER

All code used to produced the presented results can be found on the authors Github, and a link to the repository is provided in **??**. In the Github, the fully developed package is available, with all files found in the repository linked in **??**. To make it easy to apply, the author has provided a usage example of the package in Appendix B. This covers installation, simulates data, formulates and fits a model before drawing samples and obtaining relative importance plots and summary statistics.

CHAPTER

SIX

CONCLUSIONS

BIBLIOGRAPHY

Abramowitz, M. & Stegun, I. A. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th edn, Dover, New York. psi (Digamma) Function.

Aguilar, J. E. & Bürkner, P.-C. (2024), 'Generalized decomposition priors on r^2'. Available at: `https://jear2412.github.io`.

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

Arnstad, A. (2024), Relative variable importance in bayesian linear mixed models, Project report in TMA4500, Department of Mathematical Sciences NTNU – Norwegian University of Science and Technology.

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.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2018), 'Redefine statistical significance', *Nature human behaviour* **2**(1), 6–10.

Blakeley B. McShane, Christian Robert, J. L. T. & Gelman, A. (2019), 'Abandon statistical significance', *The American Statistician* **73**(sup1), 235–245.
URL: *https://doi.org/10.1080/00031305.2018.1527253*

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*

Cameron, A. C. & Windmeijer, F. A. (1997), 'An r-squared measure of goodness of fit for some common nonlinear regression models', *Journal of Econometrics* **77**(2), 329–342.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407696018180*

Chevan, A. & Sutherland, M. (1991), 'Hierarchical partitioning', *The American Statistician* **45**(2), 90–96.
URL: *http://www.jstor.org/stable/2684366*

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*

Conner, J. K. & Hartl, D. L. (2004), *Primer of Ecological Genetics*, Michigan State University Press / Harvard University Press, East Lansing, MI / Cambridge, MA, USA.

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.

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*

Grömping, U. & Lehrkamp, M. (2023), 'relaimpo: Relative importance of regressors in linear models', `https://CRAN.R-project.org/package=relaimpo`. R package version 2.2-7.
URL: *https://CRAN.R-project.org/package=relaimpo*

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/*

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, P. C. (2014), 'Extension of nakagawa & schielzeth's r2glmm to random slopes models', *Methods in Ecology and Evolution* **5**, 944–946.

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/*

Kullback, S. & Leibler, R. A. (1951), 'On Information and Sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79 – 86.
URL: *https://doi.org/10.1214/aoms/1177729694*

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.

Maddala, G. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge.

Martino, S. & Riebler, A. (2019), 'Integrated nested laplace approximations (inla)', *Department of Mathematical Sciences, Norwegian University of Science and Technology* .

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.

Menard, S. (2000), 'Coefficients of determination for multiple logistic regression analysis', *American Statistician* **54**(1), 17–24.

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., Niskanen, A. K., Saatoglu, D., Keller, L. F. & Jensen, H. (2019), 'Animal models with group-specific additive genetic variances: extending genetic group models', *Genetics Selection Evolution* **51**(7).
URL: *https://doi.org/10.1186/s12711-019-0449-7*

Nakagawa, S., Johnson, P. C. & Schielzeth, H. (2017), 'The coefficient of determination r^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded', *J. R. Soc. Interface* **14**, 20170213. Available online at `http://dx.doi.org/10.1098/rsif.2017.0213`.

Nakagawa, S. & Schielzeth, H. (2010), 'Repeatability for gaussian and non-gaussian data: a practical guide for biologists', *Biological reviews* **85**(4), 935–956. Received 08 August 2009; revised 16 April 2010; accepted 24 April 2010.

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*

Robert, C. P. (2007), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer Texts in Statistics, 2 edn, Springer New York, NY.
URL: *https://doi.org/10.1007/0-387-71599-1*

Romero, J. E. A. & Bürkner, P.-C. (2024), 'Generalized decomposition priors on r2', Open Science Framework. Data and code to accompany the paper: Aguilar, J.E, Bürkner, P. C. Generalized Decomposition Priors on R2.
URL: *https://osf.io/ns2cv/*

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*

Silva, C., McFarlane, S., Hagen, I., Rönnegård, L., Billing, A., Kvalnes, T., Kemppainen, P., Rønning, B., Ringsby, T., Sæther, B.-E., Qvarnström, A., Ellegren, H., Jensen, H. & Husby, A. (2017), 'Insights into the genetic architecture of morphological traits in two passerine bird species', *Heredity* **119**, 197–205. Published online 14 June 2017.

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*

Steinsland, I. & Jensen, H. (2010), 'Utilizing gaussian markov random field properties of bayesian animal models', *Biometrics* **66**(3), 763–771.
URL: *http://www.jstor.org/stable/40962447*

Stoffel, M. A., Nakagawa, S. & Schielzeth, H. (2017), 'rptr: repeatability estimation and variance decomposition by generalized linear mixed-effects models', *Methods in Ecology and Evolution* **8**(11), 1639–1644.

Tjelmeland, H., Lang, K. & Terje, J. (2000), *Tabeller og formler i statistikk*, Fagbokforlaget, Kanalveien 51, Bergen.

United Nations (2023), 'Sustainable development goals', `https://sdgs.un.org/goals`. [Accessed: 14-May-2024].

Wermuth, N. & Lauritzen, S. L. (1983), 'Graphical and recursive models for contingency tables', *Biometrika* **70**(3), 537–552. Printed in Great Britain.

Wilson, A. J. (2008), 'Why h2 does not always equal va/vp?', *Journal of Evolutionary Biology* **21**(3), 647–650.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1420-9101.2008.01500.x*

Wilson, A. J., Réale, D., Clements, M. N., Morrissey, M. M., Postma, E., Walling, C. A., Kruuk, L. E. B. & Nussey, D. H. (2010), 'An ecologist's guide to the animal model', *Journal of Animal Ecology* **79**(1), 13–26.
URL: *https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2656.2009.01639.x*

Zhang, Y. (2024), 'R2d2: Bayesian linear regression using the r2-d2 shrinkage prior'. Accessed: 2024-05-15.
URL: *https://github.com/yandorazhang/R2D2*

Zhang, Y. D., Naughton, B. P., Bondell, H. D. & Reich, B. J. (2020), 'Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior', *arXiv preprint arXiv:2007.04558* . Available at: `https://arxiv.org/abs/2007.04558`.

APPENDIX

A

GITHUB REPOSITORY

APPENDIX

B

BAYESIAN VARIABLE IMPORTANCE USAGE

````
## GENERAL SETUP
First, we set up the necessary libraries and configure the
   environment for our analysis. This includes loading
   essential packages and setting options for chunk output
   and plot dimensions."

```{r setup, input=FALSE, echo=FALSE}
library(formatR)
showsol <- FALSE
library(knitr)
library(devtools)
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 68),
                      tidy = TRUE,
                      warning = FALSE,
                      error = FALSE,
                      message = FALSE,
                      echo = TRUE,
                      fig.width=7,
                      fig.height=5,
                      fig.align="center")
```

## INSTALLING THE PACKAGE
This section ensures the devtools package is installed, which
    is required for installing packages from GitHub. We then
   install the BayesianImpGLMM 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")
devtools::install_github("AugustArnstad/BayesianImpGLMM")
library(BayesianImpGLMM)
```
````

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

```{r}
library(remotes)
library(INLA)
library(mnormt)
library(ggplot2)
library(reshape2)
library(RColorBrewer)
library(tidyr)
library(dplyr)

set.seed(1)

simulate_data <- function(n = 10000, n_groups = 100,
   covariance_level=0) {
  # Simulate fixed effects

  sigma <- matrix(c(1, covariance_level, covariance_level,
                    covariance_level, 1, covariance_level,
                    covariance_level, covariance_level, 1),
                       3, 3)

  X <- MASS::mvrnorm(n = n, mu = c(0, 0, 0), Sigma = sigma)
  X1 <- X[, 1]
  X2 <- X[, 2]
  X3 <- X[, 3]

  # Simulate random effects groups
  Z1 <- sample(1:n_groups, n, replace = TRUE)
  random_effect_contributions_z1 <- rnorm(n_groups, mean = 0,
      sd = 1)[Z1]

  # Coefficients for fixed effects
  beta1 <- 1
  beta2 <- sqrt(2)
  beta3 <- sqrt(3)

  # Linear predictor
  eta <- beta1*X1 + beta2*X2 + beta3*X3 + random_effect_
     contributions_z1

  # Binomial with logit link
  p_logit <- exp(eta) / (1 + exp(eta))
  y_logit_bin <- rbinom(n, size = 1, prob = p_logit)
```
```

```
  data_logit <- data.frame(y_logit_bin, X1, X2, X3, Z1)

  # Binomial with probit link
  p_probit <- pnorm(eta)
  y_probit_bin <- rbinom(n, size = 1, prob = p_probit)
  data_probit <- data.frame(y_probit_bin, X1, X2, X3, Z1)

  # Poisson with log link
  lambda <- exp(eta)
  y_pois <- rpois(n, lambda = lambda)
  data_poisson <- data.frame(y_pois, X1, X2, X3, Z1)

  epsilon = rnorm(n, mean=0, sd=sqrt(1))
  y_normal <-  beta1*X[, 1] + beta2*X[, 2] + beta3*X[, 3] +
     random_effect_contributions_z1 + epsilon
  data_normal <- data.frame(y_normal, X1, X2, X3, Z1)


  list(binomial_logit = data_logit,
       binomial_probit = data_probit,
       poisson = data_poisson,
       normal = data_normal)
}

datasets <- simulate_data()



```

USAGE

Here we demonstrate the usage of the BayesianImpGLMM package. We fit Bayesian binomial, Poisson and gaussian models and sample posterior distributions for different simulated datasets using functions from the package.

```{r}
set.seed(1234)

glmm_logit <- y_logit_bin ~ X1 + X2 + X3 + f(Z1, model="iid",
    hyper=list(prec = list(
        prior = "pc.prec",
        param = c(1, 0.01),
        initial = log(1)
      ))
)

glmm_pois <- y_pois ~ X1 + X2 + X3 + f(Z1, model="iid", hyper
   =list(prec = list(
        prior = "pc.prec",
        param = c(1, 0.01),
```

```
        initial = log(1)
      ))
)

lmm <- y_normal ~ X1 + X2 + X3 + f(Z1, model="iid", hyper=
   list(prec = list(
        prior = "pc.prec",
        param = c(1, 0.01),
        initial = log(1)
      ))
)

model_logit <- BayesianImpGLMM::perform_inla_analysis(
   datasets$binomial_logit, glmm_logit, family = "binomial",
   link_func = "logit")
model_pois <- BayesianImpGLMM::perform_inla_analysis(datasets
   $poisson, glmm_pois, family = "poisson", link_func = "log"
   )
model_normal <- BayesianImpGLMM::perform_inla_analysis(
   datasets$normal, lmm, family = "gaussian", link_func = "
   identity")

samples_logit <- BayesianImpGLMM::sample_posterior_count(
   model_logit, glmm_logit, datasets$binomial_logit, n_samp
   =5000, additive_param = "Z1")
samples_pois <- BayesianImpGLMM::sample_posterior_count(model
   _pois, glmm_pois, datasets$poisson, n_samp=5000, additive_
   param = "Z1")
samples_lmm <- BayesianImpGLMM::sample_posterior_gaussian(
   model_normal, lmm, datasets$normal, n_samp=5000, additive_
   param = "Z1")

plots_logit <- BayesianImpGLMM::plot_samples(samples_logit)
plots_pois <- BayesianImpGLMM::plot_samples(samples_pois)
plots_lmm <- BayesianImpGLMM::plot_samples(samples_lmm)
```

PLOTS

These are the default plots that are implemented in the package, displaying the importance of all effects and R^2 metrics.

```{r}
plots_logit$fixed_effects
plots_logit$random_effects
plots_logit$heritability
plots_logit$R2

plots_pois$fixed_effects
plots_pois$random_effects
plots_pois$heritability
plots_pois$R2
```

````
152
153 plots_lmm$fixed_effects
154 plots_lmm$random_effects
155 plots_lmm$heritability
156 plots_lmm$R2
157
158 ```
159
160
161 ## CUSTOM PLOT
162 Cutsomizing plots is often very nice to display information
    in the way you want it. Therefore, we show how one can
    customize the plots using ggplot2 based on the samples
    drawn.
163 ```{r}
164 random <- "Z1"
165
166 random_plot <- ggplot(samples_pois$scaled_random_samples, aes
    (x = !!sym(random))) +
167   geom_histogram(aes(y = ..density..), fill = "#C6CDF7",
      alpha = 0.7, bins = 40, color = "black") +
168   geom_density(color = "#E6A0C4", adjust = 1.5, linewidth
      =1.5) +
169   geom_point(aes(x = mean(samples_mass$scaled_random_samples$
      Z1), y = 0), color = "#E6A0C4", size = 4) +
170   labs(#title = paste("Heritability of mass"),
171        x = "Samples of relative importance of random effect",
172        y = "Frequency") +
173   theme_minimal() +
174   theme(legend.position = "none",
175     axis.title.x = element_text(size = 24),
176     axis.title.y = element_text(size = 24),
177     axis.text.x = element_text(size = 24),
178     axis.text.y = element_text(size = 24)
179     )
180
181 random_plot
182
183 str(samples_pois)
184
185 # Assuming 'samples_pois$scaled_importance_samples' is your
    dataframe
186 data_long <- samples_pois$scaled_importance_samples %>%
187   pivot_longer(cols = c(X1, X2, X3), names_to = "Variable",
      values_to = "Value")
188
189 # Updated plot code
190 fixed_plot <- ggplot(data_long, aes(x = Value)) +
191   geom_histogram(aes(y = ..density..), fill = "#C6CDF7",
      alpha = 0.7, bins = 40, color = "black") +
````

```
  geom_density(color = "#E6A0C4", adjust = 1.5, linewidth
     =1.5) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(x = "Samples of relative importance of random effect",
       y = "Frequency") +
  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))

# Print the plot
fixed_plot

r2_data <- data.frame(
  Marginal_R2 = samples_pois$R2_marginal$`Marginal R2`,
  Conditional_R2 = samples_pois$R2_conditional$`Conditional
     R2`
)

# Reshape the data from wide to long format
r2_long <- pivot_longer(r2_data, cols = c(Marginal_R2,
   Conditional_R2),
                        names_to = "R2_Type", values_to = "
                           Value")

# Create the plot
r2_plot <- ggplot(r2_long, aes(x = Value, fill = R2_Type)) +
  geom_histogram(aes(y = ..density..), alpha = 0.7, bins =
     40, color = "black") +
  geom_density(adjust = 1.5, color = "black", alpha = 0.7) +
  labs(x = "R2 Values", y = "Density") +
  scale_fill_manual(values = c("Marginal_R2" = "#C6CDF7", "
     Conditional_R2" = "#E6A0C4")) +
  theme_minimal() +
  theme(legend.title = element_blank(),
        legend.position = "top",
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12))

# Print the plot
r2_plot

```
```

Listing B.1: Usage of the BayesianImpGLMM package with plots and examples.
```

APPENDIX

C

MISCELLANEOUS PROOFS

We present a joint proof of the expectation and variance of a random variable belonging to the univariate exponential family. For a random variable Y with a normalized probability density function $f(y|\theta,\phi)$ on the form

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \ , \tag{C.1}$$

where θ is the natural parameter and ϕ is the dispersion parameter, the expectation and variance of Y can be expressed as

$$\begin{aligned} \mathbb{E}(Y|\theta) &= b'(\theta) \\ \mathrm{Var}(Y|\theta) &= b''(\theta) \end{aligned} \tag{C.2}$$

This can be shown by considering the following:

$$\frac{df(y)}{d\theta} = \frac{1}{a(\phi)} f(y|\theta\phi)(y - b'(\theta)) \ , \tag{C.3}$$

and

$$\frac{d^2 f(y)}{d\theta^2} = \frac{1}{a(\phi)} f(y|\theta,\phi)\left(\frac{1}{a(\phi)}(y - b'(\theta))^2 - b''(\theta)\right) \ . \tag{C.4}$$

Now, as $\int_{\mathbb{R}} f(y|\theta)dy = 1$, we have

$$\frac{d}{d\theta}\int_{\mathbb{R}} f(y)dy = \int_{\mathbb{R}} \frac{df}{d\theta}dy = 0 \ , \tag{C.5}$$

and

$$\frac{d^2}{d\theta^2}\int_{\mathbb{R}} f(y)dy = \int_{\mathbb{R}} \frac{d^2 f}{d\theta^2}dy = 0 \ . \tag{C.6}$$

Equations (C.5) and (C.6) can be used to derive the relation

$$\begin{aligned} 0 = \int_{\mathbb{R}} \frac{df(y)}{d\theta}dy &= \frac{1}{a(\phi)}\int_{\mathbb{R}} f(y)(y - b'(\theta))dy \\ &= \frac{1}{a(\phi)}\left(\mathbb{E}(Y|\theta) - b'(\theta)\int_{\mathbb{R}} f(y)dy\right) \\ &= \frac{1}{a(\phi)}\left(\mathbb{E}(Y|\theta) - b'(\theta)\right) \\ &\implies \mathbb{E}(Y|\theta) = b'(\theta) \ , \end{aligned} \tag{C.7}$$

and

$$
\begin{aligned}
0 = \int_{\mathbb{R}} \frac{d^2 f(y)}{d\theta^2} dy &= \frac{1}{a(\phi)} \int_{\mathbb{R}} f(y) \left(\frac{1}{a(\phi)} (y - b'(\theta))^2 - b''(\theta) \right) dy \\
&= \frac{1}{a(\phi)} \int_{\mathbb{R}} f(y) \left(\frac{1}{a(\phi)} (y - \mathbb{E}(Y))^2 - b''(\theta) \right) dy \\
&= \frac{1}{a(\phi)} \left(\mathbb{E}[(y - \mathbb{E}(Y))^2] - b''(\theta) \int_{\mathbb{R}} f(y) dy \right) \\
&= \frac{1}{a(\phi)} \mathrm{Var}(Y) - b''(\theta) \\
&\implies \mathrm{Var}(Y|\theta) = a(\phi) b''(\theta) \ \square
\end{aligned}
\tag{C.8}
$$