

August Arnstad

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

TMA4900 Master's 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



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*, in order to quantify, or rank, their impact on the statistical model. 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 reproducibility crisis in the social and biomedical sciences. Our contribution to help remedy this, is to suggest a Bayesian relative variable importance method. The method is designed to make researchers more thoroughly interpret the statistical model and its results, rather than blindly following a threshold to draw conclusions.

Our method, denoted as *Bayesian Variable Importance* (BVI), is implemented by transferring 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 responses. 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 to 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 easily available for researchers across fields, an R package called `BayesianVariableImportance` was made.

Based on the author's previous work for linear mixed models (Arnstad 2024), simulation studies, 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 ofte tilegne kovariatene en *relativ viktighet*, for å kvantifisere, eller rangere, deres bidrag til den statistiske modellen. 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 reproducerbarhetskrise i samfunns- og biomedisinsk forskning. Vårt bidrag for å bøte på dette, er å foreslå en Bayesiansk metode for å beregne relativ variabelviktighet. Denne metoden 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.

Vår metode, betegnet som *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 vektning på kovariatene før en Bayesiansk GLBM konstrueres. 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 lett tilgjengelig for forskere på tvers av fagfelt, ble en R pakke kalt `BayesianVariableImportance` lagd.

Basert på forfatterens tidligere verk for lineære blandingsmodeller (Arnstad 2024), simulasjonsstudier, 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, nylig også i det Bayesianske rammeverket, tror vi at BVV metoden kan bli ytterligere forbedret i fremtiden.

PREFACE

This master's thesis concludes the Master of Science degree obtained from the program Applied 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 master's 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. For their generous help and creative discussions, I want to thank Javier Aguilar at the University of Stuttgart. I also want to express my gratitude to my fellow students at the study program Applied Physics and Mathematics, whom 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 friends. My mom Marit and girlfriend Emma who have supported me throughout my studies, Aasta for inspiration and laughs, Samuel and Hanna for opening their home, Eli for always welcoming me back to Arnstad Vestre, and Ada, Jakob, Iwar, Johannes, and Cathrine for all the good times. You are all great people to have in my life, and I am forever grateful.

Kom igjen Troilljan!

Trondheim, June 2024

August Arnstad

CONTENTS

Abstract	i
Sammendrag	ii
Preface	iii
Contents	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
2 Theory	7
2.1 Linear regression	7
2.1.1 Linear regression	7
2.1.2 Qualitative covariates	8
2.1.3 Correlation among covariates in linear regression	8
2.2 Variable importance in linear regression models	9
2.2.1 Relative importance measures	9
2.2.2 Naive decompositions of R^2	10
2.2.3 The Lindeman, Merenda and Gold(LMG) method	11
2.2.4 Relative weights method	12
2.3 Extensions of the linear regression model	13
2.3.1 Generalized linear models (GLMs)	13
2.3.2 Linear mixed models (LMMs)	15
2.3.3 Generalized linear mixed models(GLMMs)	16
2.4 Extending R^2 to GLMMs	17
2.4.1 R^2 for GLMs	17
2.4.2 R^2 for LMMs and random slope models	18
2.4.3 R^2 for GLMMs	21
2.4.4 Extensions of the LMG and relative weights method	22
2.5 The Bayesian framework	22
2.5.1 General idea	23
2.5.2 Prior and posterior distributions	23
2.5.3 Penalising complexity (PC) priors	24

2.5.4	Hierarchical Bayesian modelling	24
2.5.5	R^2 in the Bayesian framework	25
2.5.6	Variable importance measures in the Bayesian framework	26
2.6	The INLA framework	28
2.6.1	Introduction to INLA	28
2.6.2	Approximating the marginals	29
2.6.3	Parameter estimation and sampling procedure	30
2.7	Quantitative genetics and relative variable importance	31
2.7.1	The Animal Model	31
2.7.2	Heritability	32
2.8	The Animal Model as a Gaussian Markov Random Field	33
3	Methods	35
3.1	Variable importance in the Bayesian framework	35
3.2	Extending the R^2 to Bayesian GLMMs	36
3.3	Decomposing the R^2 value	37
3.3.1	Applying the relative weights method in the Bayesian frame- work	37
3.3.2	Random effects	38
3.3.3	Drawing samples	38
3.4	Gaussian simulation study	39
3.5	Modelling phenotypic traits	40
3.5.1	Heritability as relative variable importance	41
3.5.2	House sparrow study	41
3.6	Non-Gaussian studies	42
3.6.1	Binomial and Poisson simulation studies	42
3.6.2	Binomial and Poisson case studies	45
3.7	Simulation Study with Dirichlet and Generalized Decomposition Priors on R^2	46
4	Results	49
4.1	Gaussian simulation study	49
4.1.1	Fixed effects	49
4.1.2	Random effects	52
4.1.3	R^2 estimates	52
4.2	Modelling house sparrow traits	54
4.3	Binomial simulation	58
4.3.1	Fixed effects	59
4.3.2	Random effect	60
4.3.3	R^2 estimates	61
4.4	Poisson simulation	63
4.4.1	Fixed effects	64
4.4.2	Random effect	65
4.4.3	R^2 estimates	66
4.5	Case study with <code>rptR</code> package	67
4.6	Comparing the BVI method with Dirichlet and Generalized Decom- position Priors on R^2	69
5	Discussion & Further work	73

6 Conclusions	81
References	83
Appendices	89
A GitHub repositories	89
B Bayesian Variable Importance usage	90
C Supplementary Material	97
C.1 Supplementary figures for the house sparrow study	97
C.2 Supplementary tables for the non-Gaussian simulation study	99
D Miscellaneous proofs	101

LIST OF FIGURES

2.1	Illustration of a pedigree as a GMRF	34
4.1	Relative importance of the fixed effects in Gaussian LMM	51
4.2	Relative importance of the random effect α in Gaussian LMM	52
4.3	Marginal and conditional R^2 in Gaussian LMM	53
4.4	Estimated posterior importance of all covariates in the body mass model from the BVI method	55
4.5	Estimated posterior importance of all covariates in the wing length model from the BVI method	56
4.6	Estimated posterior importance of all covariates in the tarsus length model from the BVI method	58
4.7	Relative importance of the fixed effects in binomial GLMM	60
4.8	Relative importance of the random effect α in binomial GLMM	61
4.9	Marginal and conditional R^2 in binomial GLMM	63
4.10	Relative importance of the fixed effects in Poisson GLMM	65
4.11	Relative importance of the random effect α in Poisson GLMM	66
4.12	Marginal and conditional R^2 in Poisson GLMM	67
4.13	Estimated repeatability of colour in male beetles	68
4.14	Estimated repeatability of eggs laid by female beetles	69
4.15	Comparison of the relative importance from the BVI method and the shrinkage prior methods	70
4.16	Comparison of the marginal R^2 from the BVI method and the shrinkage prior methods	71
C.1	Estimated heritability of body mass from grid and CCD strategy	97
C.2	Estimated heritability of wing length	98
C.3	Estimated heritability of tarsus length	98

LIST OF TABLES

3.1	Distribution-specific variance σ_d^2 for the Gaussian, binomial and Poisson distributions	37
3.2	Expected R^2 for Gaussian LMM	40
3.3	Expected relative importance of independent covariates for non-Gaussian GLMMs	44
3.4	Expected R^2 for non-Gaussian GLMMs	45
3.5	Expected R^2 for comparison of BVI and shrinkage prior methods . .	46
4.1	Heritability estimates and confidence intervals	54
C.1	Summary statistics for binomial GLMM simulation study	99
C.2	Summary statistics for Poisson GLMM simulation study	100

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 (McCullagh & Nelder 1989, Ryan 1989), 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 arguably 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, which often mistakenly is interpreted as importance, is very prone to misinterpretations, and 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 complement 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 standardised 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 fall short when correlation makes the coefficient estimates unstable.
- **Confidence intervals:** Confidence intervals can be calculated for the effect sizes, and the intervals can be used to determine a range of values that can be seen as statistically consistent with the data. This can be useful, although confidence intervals are frequently used as a tool to perform hypothesis tests, and the conclusion effectively relies on the p -value.
- **Information criteria:** 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. A general problem with information criteria when used for model selection, is that they tend to introduce bias in the selection. Further, the thresholds used for goodness of fit are often closely related to the p -values (Murtaugh 2014).
- **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 interpretable 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). Going forward, this thesis will consider the R^2 and its decomposition to analyse the statistical evidence of covariates in regression models.

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 various approaches. One of the most rigorous methods is the approach by Lindeman, 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 (Grömping 2007).

Another, approximate method denoted as *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 linear regression. However, for many scenarios, a linear regression is not sufficient to model the relationship between the response and the covariates. Both methods were extended in the likelihood-based framework by Matre (2022), so that they could be applicable for models including random intercepts (that is, linear mixed models), but we are not aware of any further extensions. 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 Matre (2022).

Although general relative variable importance methods are hard to come by for complex models, there are some methods that have been developed for specific purposes. For instance, for hierarchical models, the R package `rptR` is developed in Stoffel et al. (2017) and designed to calculate the intra-class correlation (ICC) of observations belonging to the same hierarchical level. This package is introduced by considering the repeatability of phenotypic traits, which is the ICC in field of ecology and evolution (Stoffel et al. 2017). The repeatability of a trait can be seen as a particular instance of relative variable importance, as it is the result of a variance decomposition. Using the `rptR` package allows for repeatability calculations for binomial and Poisson generalized linear mixed models, as well as for linear mixed models. The package uses the R^2 definition from Nakagawa & Schielzeth (2013) and is a valuable step towards a general relative variable importance measure.

The above discussed methods are all derived and implemented in the frequentist (likelihood) framework. However, 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). Unlike the frequentist framework, the Bayesian framework treats parameters as random variables rather than point estimates and

naturally includes moments such as variance in the posterior distribution. 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 preferable framework for assessing the statistical evidence of covariates. Moreover, there are numerous useful applications for relative variable importance, as decomposing the R^2 is desirable in various fields. For instance, in quantitative genetics, researchers often aim 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. Although point estimates of heritability are fairly straightforward to obtain, it is considerably more difficult to evaluate the uncertainty of the estimates (Stoffel et al. 2017). We believe that the Bayesian framework is the natural choice for quantifying this uncertainty, demonstrating its practical benefits. This particular example has been a great motivation behind our attempt to develop a Bayesian relative variable importance measure.

In the Bayesian framework, there is not much literature or research on relative variable importance. 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 logistic normal distributions (Aguilar & Bürkner 2024). At the time being, the R2D2 and GDR2 priors have been applied only to linear regression, with a focus on obtaining trustworthy predictions. Therefore, these methods have not been used explicitly as relative importance measures, and it is not clear how they generalize to more complex models. Nonetheless, they serve as an important contribution to the Bayesian framework for relative variable importance.

To summarize, the field of relative variable importance offers a wide range of methods to determine the statistical evidence of covariates in the simple linear regression model. However, correlation between the covariates is troublesome and can lead to unreliable results. Further, for more complex models, there is little consensus on how computations should be carried out and therefore a lack of robust methods. Many of the existing methods have been developed for specific purposes and do not pose a general method for relative variable importance. In the Bayesian framework, there has been done little work on relative variable importance, despite its advantageous features for statistical inference.

This thesis aims to develop a general relative variable importance method in the Bayesian framework, which can be easily implemented by researchers in different fields. The method should be applicable to a wide range of regression models with complex structures, emphasizing interpretable and reliable results. The proposed Bayesian Variable Importance (BVI) method uses the relative weights

approach as a basis and applies it within the Bayesian framework. To calculate the relative variable importance measures, the method merges the R^2 definition for GLMMs by Nakagawa & Schielzeth (2013) and Nakagawa et al. (2017), with the Bayesian R^2 for the linear regression model defined by Gelman et al. (2017). The results obtained are in the form of approximate posterior distributions, which allow for the quantification of uncertainty in the estimates. Our hope is that these distributional results are easy to interpret and provide researchers with more informative inference, as opposed to threshold-based methods. 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 virtually unlimited. 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). Here, we particularly highlight and illustrate 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 will be crucial to reach goal 14 and 15 of the United Nations sustainability goals, which concerns 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 simulation studies, case studies 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 Appendix C has some supplementary material. A miscellaneous proof is found in Appendix D.

Please note that this thesis continues the work done by the author in Arnstad (2024), and therefore some sections overlap. Following the guidelines of the Institute of Mathematical Sciences, which state that sections need not be rewritten, much of the theoretical background in Chapter 2 is the same or slightly modified from the project thesis. The specific sections in Chapter 2 with overlapping

content are Sections 2.1.1, 2.1.3, 2.2, 2.3.2, 2.4.2, 2.4.4, 2.5.1, 2.5.5 and 2.6.1. In Chapters 3 and 4, the sections Sections 3.3, 3.4 and 4.1 have significant overlap. However, many of these sections have been extensively edited to fit the context of this master's thesis and to reflect further developments in the work. Additionally, some arguments used to discuss the findings in Arnstad (2024) remains valid and has been incorporated where relevant. The thesis template used is provided by Salvesen (2023).

CHAPTER
TWO

THEORY

2.1 Linear regression

All regression models are based on the assumption that the response variable is influenced by one or more covariates (regressors). 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. Throughout this section, we will mostly follow the derivations of Fahrmeir et al. (2013) and McCullagh & Nelder (1989).

2.1.1 Linear regression

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 modelled by the equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i , \quad (2.1)$$

where β_0 is the intercept, β_1 is the slope, and ε_i is the error term. The error term, or residual, 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, including an intercept, the p covariates in the columns and the n observations in the rows. With this definition, the linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \quad (2.2)$$

where now $\mathbf{y} = (y_1, y_2, \dots, y_n)$ is a vector of n responses, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})$ is a vector of coefficients including the intercept β_0 , and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ 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.*

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

In practice, the coefficients β are estimated from the maximum likelihood estimation (MLE) method, given by

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

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 is 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

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

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

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

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 practical applications. 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 (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 β does not exist. Further, the variance of the MLE of β grows as the correlation between covariates grows (Fahrmeir et al. 2013, p. 116). A larger variance in $\hat{\beta}$ also leads to larger standard errors for $\hat{\beta}$, making it hard to assess the model. Both coefficients and covariates affect the total marginal model variance, which can be decomposed as

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

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 $j = 1, \dots, p$ found along the diagonal of \mathbf{V} and ρ_{jk} the inter-regressor correlations between regressor j and k (Grömping 2007). The middle term in (2.7) consists 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 standardised 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, which can also be used for model comparison. Conceptually, the $R^2 \in [0, 1]$ and quantifies the proportion of variance in the response variable that can be attributed to the covariates in the model. For the linear regression model, the R^2 is defined as

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

where $\bar{\mathbf{y}}$ is the mean vector of responses \mathbf{y} and σ_ε^2 denotes the residual variance. 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 (2.7) gives

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

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 criteria 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 yield

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.

Further, Grömping (2007) introduces some useful notation for considering variable importance, for the explained variance of a subset of regressors S as

$$\text{evar}(S) = \text{Var}(\mathbf{y}) - \text{Var}(\mathbf{y}|\mathbf{x}_j, j \in S) , \quad (2.10)$$

and the added explained variance of a subset M to a model that already contains the covariates in S as

$$\text{svar}(M|S) = \text{evar}(M \cup S) - \text{evar}(S) . \quad (2.11)$$

2.2.2 Naive decompositions of R^2

To make it clear that some simple decompositions fail the criteria 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, \dots, X_p as $R^2(\{1, \dots, p\})$ and the relative importance of regressor X_i as $\text{RI}(\{i\})$.

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 $R^2(\{1, \dots, p\}) - R^2(\{1, \dots, p\} \setminus i)$. To show how this fails the criteria of relative importance measures, an example from Matre (2022) is discussed. The example considers the simple case

$$Y = X_1 + X_2, \quad \text{Var}(X_1) = \text{Var}(X_2) = 1, \quad \text{Cov}(X_1, X_2) = 0.9 . \quad (2.12)$$

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

$$\text{RI}(\{2\}) = R^2(\{1, 2\}) - R^2(\{1\}) = 1 - \frac{\text{Cov}(Y, X_1)^2}{\text{Var}(Y)\text{Var}(X_1)} = 1 - \frac{1.9^2}{3.8} \approx 0.05 , \quad (2.13)$$

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 $\text{RI}(\{1\}) = \text{RI}(\{2\})$, 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

$$\text{RI}(\{1\}) = R^2(\{1\}) - R^2(\{\emptyset\}) = \frac{\text{Cov}(Y, X_1)^2}{\text{Var}(Y)\text{Var}(X_1)} = \frac{1.9^2}{3.8} \approx 0.95 . \quad (2.14)$$

Once more by symmetry we have $\text{RI}(\{2\}) = \text{RI}(\{1\})$, so the sum of the relative importances is 1.9, violating the proper decomposition condition. In contrast 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 Lindeman, Merenda and Gold(LMG) method

A method that handles correlation among covariates, and is frequently reinvented from different approaches, is the Lindeman, Merenda and Gold (LMG) method (Grömping 2007). We shall therefore 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 $\{1, \dots, p\}$ 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 two correlated regressors, whose importance share varies depending on which is added first, would receive an averaged importance. Averaging over orders is a statistical tradition (Kruskal 1987) and gives a robust assessment of each regressor's importance by considering different orderings of how they are added to the model. The iterative process for the regressors $\{X_0, X_1, X_2, X_3\}$, where X_0 is the intercept, would be

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

2. Considering $\{X_1, X_3, X_2\}$, X_1 is added to the model, and the share of explained variance allocated to X_1 is $svar(\{1\}|\emptyset)$. 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_i by the LMG method with p regressors,

$$\text{LMG}(i) = \frac{1}{p!} \sum_{S \subseteq \{1, \dots, p\} \setminus i} n(S)! (p - n(S) - 1)! svar(\{i\}|S), \quad (2.15)$$

where $n(S)$ is the number of regressors in S (Grömping 2007). Equation (2.15) 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

$$\mathbf{X} = \mathbf{UDV}^T. \quad (2.16)$$

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 $\mathbf{Z} = \mathbf{UV}^T$ of rank $k \leq r$ such that the difference under the squared Frobenius norm

$$\|\mathbf{X} - \mathbf{Z}\|_F^2 = \text{tr}((\mathbf{X} - \mathbf{Z})^T(\mathbf{X} - \mathbf{Z})) , \quad (2.17)$$

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

$$\begin{aligned} \boldsymbol{\beta}_{\mathbf{Z}} &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z} \mathbf{y} \\ &= ((n-1) \mathbf{V} \mathbf{U}^T \mathbf{U} \mathbf{V}^T)^{-1} \sqrt{n-1} \mathbf{V} \mathbf{U}^T \mathbf{y} \\ &= \frac{1}{\sqrt{n-1}} \mathbf{V} \mathbf{U}^T \mathbf{y} . \end{aligned} \quad (2.18)$$

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 $\beta_{Z,i}^2$. 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 from the matrix

$$\boldsymbol{\Lambda} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{X} = (\mathbf{V} \mathbf{U}^T \mathbf{U} \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{U}^T \mathbf{U} \mathbf{D} \mathbf{V}^T = \mathbf{V} \mathbf{D} \mathbf{V}^T , \quad (2.19)$$

where \mathbf{Z} is orthogonal. The contribution from a column of \mathbf{z}_i with respect to a column \mathbf{x}_j is given by the squared entry Λ_{ij}^2 , with the squared columns of $\boldsymbol{\Lambda}$ summing to one and acting as weights. The contribution from a column \mathbf{x}_j with respect to the response \mathbf{y} , *i.e.* the relative importance, is then estimated as the matrix product (Johnson 2000)

$$\text{RI}(\mathbf{X}) = \boldsymbol{\Lambda}^{[2]} \boldsymbol{\beta}_{\mathbf{Z}}^{[2]} , \quad (2.20)$$

with RI as a column vector where each entry j contains the estimate of the relative importance corresponding to column j of \mathbf{X} . 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 $\boldsymbol{\xi}$. In Matre (2022, Section 2.5.3) it is shown that the relative weights method fulfills the 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 Extensions 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 we have N observations of the response y_i , where $i = 1, \dots, N$, that are conditionally independent given the fixed effects. Then, y_i belongs to the univariate exponential family if

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y, \phi)\right), \quad (2.21)$$

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

$$\begin{aligned} \mathbb{E}(Y|\theta) &= b'(\theta) \\ \text{Var}(Y|\theta) &= a(\phi)b''(\theta), \end{aligned} \quad (2.22)$$

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

$$\theta_i = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (2.23)$$

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

$$\eta_i = g(\mu_i) = g(\mathbb{E}(y_i)). \quad (2.24)$$

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 analyse how the covariates influence the probability

$$\pi_i = \mathbb{P}(y_i = 1|\mathbf{x}_i) = \mathbb{E}[y_i]. \quad (2.25)$$

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, $h(\eta_i)$ must transform the linear predictor in such a way that the expectation fulfills this criterion (Fahrmeir et al. 2013). A popular choice of inverse link function is the logistic response function

$$\pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad (2.26)$$

yielding the logit-link function

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i, \quad (2.27)$$

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

$$\frac{\pi_i}{1 - \pi_i} = \exp(\eta_i) = \exp(\beta_0) \exp(\beta_1 x_{1,i}) \dots \exp(\beta_p x_{1,p}), \quad (2.28)$$

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 occurring 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 occurring cannot be negative, the rate is also restricted to positive values. The common choice of inverse link function is therefore

$$\lambda_i = \exp(\eta_i) = \exp(\beta_0) \exp(\beta_1 x_{1,i}) \dots \exp(\beta_p x_{1,p}) , \quad (2.29)$$

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

$$\ln(\lambda_i) = \eta_i . \quad (2.30)$$

2.3.2 Linear mixed models (LMMs)

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 $n_j, j = 1, \dots, m$ 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

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m)^T , \quad (2.31)$$

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 $\boldsymbol{\Sigma} = \text{diag}(\sigma_0^2, \dots, \sigma_q^2)$. The linear mixed model now takes the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} , \quad (2.32)$$

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 with zero expectation, the parameters to estimate are the variance of each random effect $\boldsymbol{\Sigma}_{kk} = \sigma_k^2$ and their covariance $\boldsymbol{\Sigma}_{k,l} = \sigma_{k,l}$, where $k, l = 1, \dots, q$. 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 is assumed that $\mathbf{X}\boldsymbol{\beta}$ is fixed, and that \mathbf{U} is given, so they do not contribute to the model's variance. Therefore, the conditional

expectation $\mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \mathbf{X}\boldsymbol{\beta}$ is easily obtained, and the conditional variance can be calculated as

$$\text{Var}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) = \mathbf{U}\text{Var}(\boldsymbol{\alpha})\mathbf{U}^T + \sigma^2\mathbf{I} = \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I}, \quad (2.33)$$

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 Σ_j along the diagonal for $j = 1, \dots, m$. As we assume that the random effects are independent of the fixed effects, and that the random error term is i.i.d. for each observation, the conditional distribution of \mathbf{y} follows that of a sum of independent normal distributions, *i.e.*

$$\mathbf{y}|\mathbf{X}, \mathbf{U} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I}). \quad (2.34)$$

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

$$\theta_{i,j} = \eta_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta} + \mathbf{u}_{i,j}^T \boldsymbol{\alpha}_j, \quad (2.35)$$

where $j = 1, \dots, m$ denotes the cluster and $i = 1, \dots, n_j$ denotes observation i 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

$$\ln\left(\frac{\pi_{i,j}}{1 - \pi_{i,j}}\right) = \ln\left(\frac{\mathbb{P}(y_{i,j} = 1|\mathbf{x}_{i,j}, \mathbf{u}_{i,j}, \boldsymbol{\alpha}_j)}{\mathbb{P}(y_{i,j} = 0|\mathbf{x}_{i,j}, \mathbf{u}_{i,j}, \boldsymbol{\alpha}_j)}\right) = \eta_{i,j}, \quad (2.36)$$

and for regression on count responses we have $y_{i,j}$ following a Poisson distribution conditional on $\mathbf{x}_{i,j}, \mathbf{u}_{i,j}$ and $\boldsymbol{\alpha}_j$ with the link function

$$\ln(\lambda_{i,j}) = \mathbf{x}_{i,j}^T \boldsymbol{\beta} + \mathbf{u}_{i,j}^T \boldsymbol{\alpha}_j, = \eta_{i,j}. \quad (2.37)$$

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

$$\lambda_j = \exp(\mathbf{x}_j^T \boldsymbol{\beta} + \alpha_{0,j}), \quad (2.38)$$

where $\alpha_{0,j} \sim \mathcal{N}(0, \tau_0^2)$, 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 Markov chain Monte Carlo (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 (2.8), we now generalize this to the GLMs. This topic has been subject to significant research, (Maddala 1983, Cameron & Windmeijer 1997, Menard 2000, Nakagawa & Schielzeth 2013) and has evolved through various methodologies. The methods first suggested was based on the likelihood function of the model to be analysed. 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 (scaled) 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

$$\begin{aligned}\mathcal{D}(\mathbf{y}|\hat{\theta}) &= -2 \left(\ln(\mathcal{L}(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})) - \ln(\hat{\mathcal{L}}(\hat{\boldsymbol{\beta}}, \sigma^2 | \mathbf{y})) \right) = -2 \left(l(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \sigma^2 | \mathbf{y}) \right) \\ &= -2 \left(-\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{n}{2} \ln(2\pi\sigma^2) \right) \\ &= \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}\tag{2.39}$$

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**, which sets $\mathbf{X}\boldsymbol{\beta} = \bar{\mathbf{y}}$, one obtains for the linear regression

$$\frac{\mathcal{D}(\mathbf{y}|\hat{\theta})}{\mathcal{D}(\mathbf{y}|\theta_0)} = \frac{\frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\frac{1}{\sigma^2} (\mathbf{y} - \bar{\mathbf{y}})^T (\mathbf{y} - \bar{\mathbf{y}})} = 1 - R^2\tag{2.40}$$

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. In Nakagawa & Schielzeth (2013), two different, likelihood based, R^2 measures are given as

$$R_G^2 = \left[1 - \left(\frac{\mathcal{L}_0}{\mathcal{L}_M} \right)^{2/n} \right] \frac{1}{1 - (\mathcal{L}_0)^{2/n}} \quad (2.41)$$

and

$$R_D^2 = 1 - \frac{-2 \ln(\mathcal{L}_M)}{-2 \ln(\mathcal{L}_0)} \quad (2.42)$$

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 . Consider the simple random intercept model in scalar form;

$$y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \alpha_j + \varepsilon_{i,j}, \quad (2.43)$$

where $y_{i,j}$ denotes the i th observation in cluster j , β_0 is the fixed intercept, $x_{i,j}$ is the covariate for the i th observation in cluster j , β is the global coefficient of the fixed effect, α_j is the random intercept for cluster j and $\varepsilon_{i,j}$ is the residual error for the i th observation in cluster j . The two R^2 components can then be expressed in two ways, with the first being

$$R_1^2 = 1 - \frac{\text{Var}(y_{i,j} - \hat{y}_{i,j})}{\text{Var}(y_{i,j})} = 1 - \frac{\sigma_\varepsilon^2 + \sigma_\alpha^2}{\sigma_{\varepsilon 0}^2 + \sigma_{\alpha 0}^2} \quad (2.44)$$

$$\hat{y}_{i,j} = \beta_0 + \mathbf{x}_{i,j}^T \boldsymbol{\beta},$$

where $\sigma_{\varepsilon 0}^2$ and $\sigma_{\alpha 0}^2$ denote the residual and random effect variances of the null model respectively and $\hat{y}_{i,j}$ denotes the fitted value of observation i in the j th

cluster (Nakagawa & Schielzeth 2013). Similarly, the second component is defined as

$$\begin{aligned} R_2^2 &= 1 - \frac{\text{Var}(y_j - \hat{y}_j)}{\text{Var}(\bar{y}_j)} = 1 - \frac{\sigma_\varepsilon^2 + \sigma_\alpha^2/k}{\sigma_{\varepsilon 0}^2 + \sigma_{\alpha 0}^2/k} \\ k &= \frac{M}{\sum_{j=1}^M \frac{1}{m_j}} , \end{aligned} \quad (2.45)$$

where \bar{y}_j is the mean for each observed value of the j th cluster, \hat{y}_j is the mean of the fitted values for the j th cluster, k is the harmonic mean of the number of observations per cluster, m_j is the number of observations for the j th 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 use 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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} , \quad (2.46)$$

with the parameters of interest being $\boldsymbol{\beta}$ and the variance components σ_ε^2 and σ_r^2 for each of the $r = 1, \dots, q$ random intercepts. Then define the variance of the fixed effects as

$$\sigma_f^2 = \text{Var}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^T \text{Var}(\mathbf{X}) \boldsymbol{\beta} , \quad (2.47)$$

and further define the R^2 for the LMM as

$$R_{\text{LMM(m)}}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \sigma_r^2 + \sigma_\varepsilon^2} . \quad (2.48)$$

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

$$R_{\text{LMM(c)}}^2 = \frac{\sigma_f^2 + \sum_{r=1}^q \sigma_r^2}{\sigma_f^2 + \sum_{r=1}^q \sigma_r^2 + \sigma_\varepsilon^2} . \quad (2.49)$$

By inspection, it is clear that this definition will never lead to negative values of the R_{LMM}^2 . It may occur that the R_{LMM}^2 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,

$$y_{i,j} = \beta_0 + \beta_1 x_{i,j} + \alpha_{0,j} + \alpha_{1,j} x_{i,j} + \varepsilon_{i,j}, \quad (2.50)$$

where the same notation is used as in (2.43) 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

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{\alpha_0}^2 & \sigma_{\alpha_0, \alpha_1} \\ \sigma_{\alpha_0, \alpha_1} & \sigma_{\alpha_1}^2 \end{pmatrix} \right), \quad (2.51)$$

where $\sigma_{\alpha_0}^2$ and $\sigma_{\alpha_1}^2$ 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

$$\begin{aligned} \text{Var}(\alpha_{0,j} + \alpha_{1,j} x_{i,j}) &= \text{Var}(\alpha_{0,j}) + 2x_{i,j} \text{Cov}(\alpha_{0,j}, \alpha_{1,j}) + x_{i,j}^2 \text{Var}(\alpha_{1,j}) \\ &= \sigma_{\alpha_0}^2 + 2x_{i,j} \sigma_{\alpha_0, \alpha_1} + x_{i,j}^2 \sigma_{\alpha_1}^2 =: \sigma_{r,i,j}^2, \end{aligned} \quad (2.52)$$

where we define $\sigma_{r,i,j}^2$ as the variance of the random effect $\boldsymbol{\alpha}$ for observation i in the j th 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 $\mathbf{x}_{i,j}$ 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.*

$$\bar{\sigma}_r^2 = \frac{1}{N} \sum_{j=1}^q \sum_{i=1}^{N_j} (\sigma_{r,i,j}^2). \quad (2.53)$$

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

$$\bar{\sigma}_r^2 = \text{Tr}(\mathbf{U}_j \mathbf{Q} \mathbf{U}_j^T), \quad r = 1, \dots, q. \quad (2.54)$$

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

$$R_{\text{LMM(m)}}^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2 + \sigma_\varepsilon^2}, \quad (2.55)$$

and

$$R_{\text{LMM(c)}}^2 = \frac{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2}{\sigma_f^2 + \sum_{i=1}^q \bar{\sigma}_r^2 + \sigma_\varepsilon^2}, \quad (2.56)$$

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

$$\overline{\sigma_r^2} = \frac{1}{N} \sum_{j=1} \sum_{i=1} (\sigma_{r,i,j}^2) = \sigma_{r,i,j}^2 = \sigma_r^2 . \quad (2.57)$$

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_ε^2 into three components, namely distribution specific variance, multiplicative dispersion and additive dispersion (Nakagawa & Schielzeth 2013). The distribution specific variance, σ_d^2 , is inherited from the link function used, and formulas for multiple distributions can be found in Nakagawa et al. (2017). 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

$$\sigma_\varepsilon^2 = \omega \sigma_d^2 \quad (2.58)$$

or

$$\sigma_\varepsilon^2 = \sigma_d^2 + \sigma_e^2 , \quad (2.59)$$

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

$$R^2_{\text{GLMM(m, m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \overline{\sigma_r^2} + \omega \sigma_d^2} , \quad (2.60)$$

and

$$R^2_{\text{GLMM(m, a)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^q \overline{\sigma_r^2} + \sigma_d^2 + \sigma_e^2} , \quad (2.61)$$

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

$$R^2_{\text{GLMM}(\text{c, m})} = \frac{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2}{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2 + \omega \sigma_d^2}, \quad (2.62)$$

and

$$R^2_{\text{GLMM}(\text{c, a})} = \frac{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2}{\sigma_f^2 + \sum_{r=1}^q \bar{\sigma}_r^2 + \sigma_d^2 + \sigma_e^2}, \quad (2.63)$$

completing the generalization.

2.4.4 Extensions of the LMG and relative weights method

Extensions for both the LMG and the relative weights method have been proposed so that they can decompose the R^2 proposed by Nakagawa & Schielzeth (2013) also for random intercept models (Matre 2022). The extended LMG, denoted as the ELMG, method uses the same permutations as described for the regular LMG, and now decomposes the R^2 value for the random intercept model. For this decomposition, the only extension needed for the LMG formula is to include also the random intercepts as model components, which for covariate i gives

$$\text{LMG}(i) = \frac{1}{(p+q)!} \sum_{\substack{S \subseteq \{1, \dots, p \\ p+1, \dots, p+q\} \setminus i}} n(S)!((p+q) - n(S) - 1)! \text{svar}(\{i\}|S), \quad (2.64)$$

where p denotes fixed effects and q denotes random effects (Matre 2022). It is equivalent to the original LMG method (2.15) except that here the random intercepts are treated as categorical fixed effects, where we do not consider the columns but rather the whole predictor, either completely in the model or not.

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

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

2.5 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 this thesis will be developed in.

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 occurring, we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} . \quad (2.65)$$

This was generalized by Pierre Simon Laplace (Laplace 1774), to also apply to distributions of continuous random variables, namely

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} , \quad (2.66)$$

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 (2.66), *i.e.*

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) . \quad (2.67)$$

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 parameters in $\boldsymbol{\theta}$ are now treated as random variables 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 caution. As the theory of priors is vast 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. **Occam's 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 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

$$\pi(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp(-\lambda \tau^{-1/2}), \quad \tau, \lambda > 0. \quad (2.68)$$

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 parameters $\boldsymbol{\theta}$ 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

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}), \quad (2.69)$$

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 obtain inference on the parameters of interest in $\boldsymbol{\theta}$ using the population distribution, instead of considering each component 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

$$y_i|\theta_i \sim \text{Po}(\theta_i), i = 1, \dots, n, \quad (2.70)$$

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

$$\theta_i|\boldsymbol{\phi} \sim \text{Gamma}(\alpha, \beta), \quad (2.71)$$

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

$$\alpha \sim \text{Exp}(a) \text{ and } \beta \sim \text{Gamma}(b, c), \quad (2.72)$$

The full posterior density now reads

$$\pi(\boldsymbol{\theta}, \alpha, \beta | \mathbf{y}) \propto \underbrace{\prod_{i=1}^n \theta_i^{y_i} e^{-\theta_i}}_{\text{Po}(\theta_i)} \underbrace{\prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\beta)} \theta_i^{\alpha-1} e^{-\beta\theta_i}}_{\text{Gamma}(\alpha, \beta)} \underbrace{\alpha^{a-1} e^{-\alpha}}_{\text{Exp}(a)} \underbrace{\beta^{b-1} e^{-c\beta}}_{\beta \sim \text{Gamma}(b, c)}, \quad (2.73)$$

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} belonging to the exponential family with density function as in (2.21), dispersion parameter ϕ and associated linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha}, \quad (2.74)$$

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, the posterior can be written out as

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho} | \mathbf{y}) &\propto \left(\prod_{j=1}^m \pi(\mathbf{y}_j | \boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho}) \right) \pi(\boldsymbol{\alpha} | \boldsymbol{\rho}) \pi(\boldsymbol{\beta}) \pi(\phi) \pi(\boldsymbol{\rho}), \\ &\propto \exp \left(-\frac{1}{2} \boldsymbol{\alpha}^T \mathbf{Q}(\boldsymbol{\rho}) \boldsymbol{\alpha} + \sum_{j=1}^m \ln \pi(\mathbf{y}_j | \boldsymbol{\beta}, \boldsymbol{\alpha}, \phi) \right) |\mathbf{Q}(\boldsymbol{\rho})|^{1/2} \pi(\boldsymbol{\beta}) \pi(\phi) \pi(\boldsymbol{\rho}), \end{aligned} \quad (2.75)$$

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

2.5.5 R^2 in the Bayesian framework

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 be a function of random variables, and therefore a random variable itself. An intuitive suggestion to the R^2 could be to use the posterior mode of the parameters β 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 β from the posterior distribution. Then, the proposed definition is

$$R_s^2 = \frac{\beta_s^T \Sigma_{\mathbf{X}^T \mathbf{X}} \beta_s}{\beta_s^T \Sigma_{\mathbf{X}^T \mathbf{X}} \beta_s + \sigma_s^2}, \quad (2.76)$$

where $\Sigma_{\mathbf{X}^T \mathbf{X}}$ 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 posterior distributions. Drawing enough samples from (2.76) one would eventually obtain also an approximation of the distribution for the R^2 value (Gelman et al. 2017).

2.5.6 Variable importance measures in the Bayesian framework

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 predictions 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 β such that $\mathbb{E}[\beta] = 0$ and $\text{cov}(\beta) = \sigma^2 \Lambda$ where Λ is a diagonal matrix with the diagonal elements $\lambda_1, \dots, \lambda_p$. From the calculations in Zhang et al. (2020) and following the definition of the R^2 in Gelman et al. (2017), one can write

$$\text{Var}(\mathbf{x}^T \beta) = \sigma^2 \sum_{j=1}^p \lambda_j \quad (2.77)$$

and

$$R^2 = \frac{\text{Var}(\mathbf{x}^T \beta)}{\text{Var}(\mathbf{x}^T \beta) + \sigma^2} = \frac{\sum_{j=1}^p \lambda_j}{\sum_{j=1}^p \lambda_j + 1} := \frac{W}{W + 1} \quad (2.78)$$

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

$R^2 \sim Beta(a, b)$ follows a Beta distribution (Zhang et al. 2020). Further, Zhang et al. (2020) expresses each $\lambda_j = \phi_j \omega$ with $\sum_{j=1}^p \phi_j = 1$ 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 $j = 1, \dots, p$ (Zhang et al. 2020). Now, to decompose the model R^2 based on these priors, it is proposed to place the prior $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p) \sim Dir(a_\pi, \dots, a_\pi)$ where Dir denotes the Dirichlet distribution, and a_π is a concentration parameter (Zhang et al. 2020). This means that ϕ_j is estimated using 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 β 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 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 make it 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 Dirichlet distribution is not very flexible, and therefore struggles to model correlation structures between covariates (Aguilar & Bürkner 2024, and references therein). Moreover, 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 Logistic 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 prior mean $\boldsymbol{\mu}$ and the prior covariance matrix $\boldsymbol{\Sigma}$ (Aguilar & Bürkner 2024). The authors suggest automating 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-Leibler divergence (Kullback & Leibler 1951) between f and g can be minimized. In Aguilar & Bürkner (2024) a closed form for the minimizers is obtained as

$$\begin{aligned}\boldsymbol{\mu}_j^* &= \delta(\alpha_j) - \delta(\alpha_K) \\ \sigma_{jj}^* &= \varepsilon(\alpha_j) + \varepsilon(\alpha_K) \\ \sigma_{jk}^* &= \varepsilon(\alpha_j) \quad j \neq k,\end{aligned}\tag{2.79}$$

for $j = 1, \dots, p$ where α_K is a reference unit, $\delta(\alpha_j)$ denotes the digamma function and $\varepsilon(\alpha_j) = \delta'(\alpha_j)$ 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

As we have seen, the analytical posterior is possible to obtain for some hierarchical structures (*e.g.* (2.73)). 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 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 (Rue et al. 2009).

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. To present INLA, consider the vector of observations $\mathbf{y} = (y_1, \dots, y_n)$, which may also contain missing values. Given an appropriate link function $g(\mu_i) = \eta_i$, we can model the observations as independent given the linear predictor

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i , \quad i = 1, \dots, n , \quad (2.80)$$

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 (Gómez-Rubio 2020). 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

$$\mathbf{x} = (\eta_1, \dots, \eta_n, \alpha, \beta_1, \dots, \beta_n) , \quad (2.81)$$

is a Gaussian Markov Random Field (GMRF) (Gómez-Rubio 2020). This often leads to the model being called a latent Gaussian model (LMG). Further, in addition to assuming that observations are independent given this latent field, it is also assumed that the latent field is distributed with respect to some hyperparameters $\boldsymbol{\theta}$. The structure of the GMRF is given by a precision matrix $\mathbf{Q}(\boldsymbol{\theta})$, 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 joint posterior distribution of the latent field \mathbf{x} is given by

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) , \quad (2.82)$$

where $\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$ is the likelihood, $\pi(\mathbf{x} | \boldsymbol{\theta})$ 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

$$\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i | x_i, \boldsymbol{\theta}) , \quad (2.83)$$

where the index set $\mathcal{I} \subset \{1, 2, 3, \dots, n\}$ only includes actual observed data. By following Gómez-Rubio (2020), the joint posterior can now be written as

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \exp \left(-\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i \in \mathcal{I}} \ln \pi(y_i | x_i, \boldsymbol{\theta}) \right) |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \pi(\boldsymbol{\theta}) , \quad (2.84)$$

which is similar to the expression obtained in (2.75). However, instead of considering the full joint posterior distribution, INLA attempts to estimate the marginals of the latent effects, and the hyperparameters, to construct the approximate joint posterior. These marginals are given by

$$\pi(x_l | \mathbf{y}) = \int \pi(x_l | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} , \quad (2.85)$$

and

$$\pi(\theta_k | \mathbf{y}) = \int \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-k} , \quad (2.86)$$

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 (2.85) and (2.86) 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 (2.85) and (2.86) as

$$\tilde{\pi}(x_l | \mathbf{y}) = \int \tilde{\pi}(x_l | \boldsymbol{\theta}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta} , \quad (2.87)$$

and

$$\tilde{\pi}(\theta_k | \mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-k} , \quad (2.88)$$

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

$$\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y}) \propto \frac{\tilde{\pi}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x} | \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} , \quad (2.89)$$

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 (2.89) 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

$$\tilde{\pi}_G(x_l | \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}(\mu_l(\boldsymbol{\theta}), \sigma_l^2(\boldsymbol{\theta})) , \quad (2.90)$$

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, more costly, approach is to use a Laplace approximation so that

$$\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\tilde{\pi}(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-l}|x_l, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-l}=\mathbf{x}_{-l}^*(x_l, \boldsymbol{\theta})} \quad (2.91)$$

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

$$\tilde{\pi}_{LA}(x_l|\boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}(\mu_l(\boldsymbol{\theta}), \sigma_l^2(\boldsymbol{\theta})) \exp(\text{cubic spline}(x_l)) \quad (2.92)$$

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 a standardised variable \mathbf{z} such that

$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z} , \quad (2.93)$$

to more effectively explore the hyperparameter space (Gómez-Rubio 2020). Then, the hyperparameter space is explored using \mathbf{z} with either a regular grid with some stepsize h or a central composite design (CCD) (Gómez-Rubio 2020, and references therein). The exploration is done to obtain a set $\{\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^K\}$ that captures the 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

$$\pi(x_l|\mathbf{y}) \simeq \sum_{k=1}^K \tilde{\pi}(x_l|\boldsymbol{\theta}^{(k)}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}^{(k)}|\mathbf{y}) \Delta_k . \quad (2.94)$$

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 from INLA on random effects will be given as precision rather than variance. Throughout the thesis, when INLA is applied, we will place PC priors on the model parameters, as recommended by Simpson et al. (2017) and Gómez-Rubio (2020).

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.

2.7.1 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 colour, height or behaviour. 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

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2 , \quad (2.95)$$

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 **dom-**

inance 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

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha}, \quad (2.96)$$

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 $\boldsymbol{\alpha}_A$ contains the so-called breeding values (Wilson et al. 2010), which are defined as the effect of an individual's 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 correlation 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).

2.7.2 Heritability

As mentioned, we are particularly concerned 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,

$$\frac{\sigma_A^2}{\sigma_P^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2}. \quad (2.97)$$

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 of much misinterpretation, 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 (2.60), (2.61), (2.62) and (2.63), one can quickly notice that the definition of heritability is very similar. In fact, generalizing the definition of variable importance from Grömping (2007) to also yield for random effects, 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 $\mathbf{X} = (X_1, \dots, X_n)$ 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 2.1, right), and in the animal model the pedigree structure (Figure 2.1, 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 2.1 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 allows us to use the INLA framework for model fitting, which can be used to effectively sample and obtain parameter estimates from the quantities in the animal model (Steinsland & Jensen 2010). The pedigree can also be used to construct the relatedness matrix \mathbf{A} , previously defined as the expected correlation between relatives, and the gives rise to the sparse precision matrix $\mathbf{Q} := (\sigma_A^2 \mathbf{A})^{-1}$, reflecting the Markov structure (Steinsland & Jensen 2010). Note that only \mathbf{A}^{-1} 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).

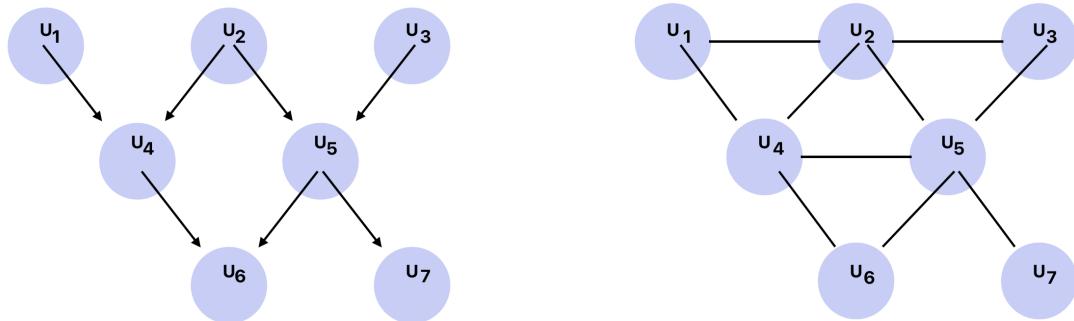


Figure 2.1: Illustration of a pedigree as a GMRF, figure and figure text inspired by Figure 1 in Steinsland & Jensen (2010). On the left, a pedigree structure is depicted as a directed acyclic graph (DAG), where birds U_1 and U_2 are the parents of bird U_4 , birds U_2 and U_3 the parents of bird U_5 , 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 and Poisson responses, 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.

For the complete model formulation of all methods used in this thesis, all files are uploaded to GitHub, with a link in Appendix A. If categorical covariates with more than two levels are contained in the fixed effects, they should be dummy 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. Firstly, the characteristics of the Bayesian framework must be considered. When fitting a GLMM in the frequentist framework, point estimates of the fixed regression coefficients 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. There-

fore, we consider only GLMMs modelled 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. 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

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

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

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

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

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 σ_ε^2 is explicitly modelled as a parameter. The residual variance coincides with the overdispersion in the model, and the distributional variance with the identity link is zero (Nakagawa & Schielzeth 2013). However, for non-Gaussian responses, the residual variance of the model when considering additive overdispersion is defined as

$$\sigma_\varepsilon^2 = \sigma_e^2 + \sigma_d^2, \quad (3.3)$$

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 3.1. Given that we can obtain samples for the variance components, we define for a sample s the marginal and conditional R^2 , on the latent scale, for the Bayesian GLMM as

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

respectively, where $\sigma_{\varepsilon,s}^2 = \sigma_{e,s}^2 + \sigma_{d,s}^2$ is the sampled residual variance. 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 $s = 1, \dots, S$ respectively.

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

Table 3.1: Distribution-specific variance σ_d^2 for the Gaussian, binomial and Poisson distributions with link functions. The full expression $\mathbb{E}[\lambda]$ is given in (3.18). 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, if present, 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 $\boldsymbol{\beta}_{\mathbf{Z},s}, s = 1, \dots, S$ and the matrix $\boldsymbol{\Lambda}$ can be used to approximate a sample of the importance of the columns \mathbf{X} , with the squared entries of $\boldsymbol{\Lambda}$ acting as weights from the projected space to the original covariate space. Using equations (2.19) and (2.20), we calculate this sample as

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

where $\text{IMP}(\mathbf{X})_s$ is a column vector containing the approximated importance of column k of \mathbf{X} on the k -th entry for $k = 1, \dots, p$ and $\boldsymbol{\xi}^{[2]}$ again denotes the Schur product of $\boldsymbol{\xi}$ with itself. To calculate the relative variable importance, note that we estimate $\sigma_{f,s}^2$ in (3.4) by

$$\sigma_{f,s}^2 \simeq \sum_{k=1}^p \text{IMP}(\mathbf{X})_{s,k}. \quad (3.6)$$

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

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

where $\sigma_{\alpha_i,s}^2$ and $\sigma_{\varepsilon,s}^2$ are defined as in Section 3.2. For sufficiently large S , we believe these samples can be used to construct an approximation 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 for sample s is calculated as

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

By sampling the relative importance for the random effects, we construct the approximated posterior distributions of relative importance. 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 α_i in (Nakagawa et al. 2017) as

$$ICC = \frac{\sigma_{\alpha_i}^2}{\sum_{k=1}^q \sigma_{\alpha_k}^2 + \sigma_{\varepsilon}^2}. \quad (3.9)$$

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

$$\text{ICC}_s = \frac{\sigma_{\alpha_i,s}^2}{\sum_{k=1}^q \sigma_{\alpha_k,s}^2 + \sigma_{\varepsilon,s}^2}, \quad (3.10)$$

and obtain an approximate posterior distribution of the ICC. It is also worth noting that the distribution of relative importance of the distributional variance is calculated in the same manner. Although not a quantity of interest, it can give an indication of how influential the added variance from the specific distribution is with respect to the covariates.

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 posterior mode of the precision distribution to obtain the variance. Another way of estimating the variance is to take the variance of the sampled values for the random vector α . 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 in performing the calculations required by the BVI method, is obtaining samples from the joint posterior distribution. To achieve 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 depends on how well the numerical integration is carried out (Gómez-Rubio 2020). INLA provides several integration options, allowing one to choose the desired resolution

at the cost of increased computational complexity. The BVI method is implemented using 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 larger dimension hyperparameter vectors (Martino & Riebler 2019). Further, if the model fit is poor or if the model is misspecified, the samples will also reflect these issues. Recall that INLA assumes a Gaussian latent layer, so this condition is crucial for obtaining a representative set of samples. We apply the simplified Laplace approximation when estimating the posterior marginals of the latent layer conditional on the hyperparameters, as is the default in INLA. However, one can choose among the three options provided in INLA, as described in Section 2.6.2 and Section 2.6.3. Lastly, note that INLA is a tool that is continuously in development, and therefore the method is subject to changes. For instance, the authors state that a skewness correction is in the works (Gómez-Rubio 2020), and more features are likely to be added in the future.

3.4 Gaussian simulation study

To evaluate the performance of our proposed method a simulation study was conducted in Arnstad (2024), which we will reproduce here to provide a comprehensive overview. The study investigates how the BVI compares to the relative importance decomposition (`relaimpo`, see package description in Grömping & Lehrkamp (2023)) as presented in Grömping (2007) and the two methods presented in Matre (2022). The `relaimpo` package uses the LMG decomposition and considers only fixed effects and can therefore only be compared with the BVI in the fixed effects. The two methods in Matre (2022), ELMG and the ERW, are extensions of the LMG and relative weights methods respectively, to include random intercepts. These extensions allow us to compare the results for the random intercept model to our BVI method.

To simulate the data we consider the model as in (3.1), with the link function $g(\cdot)$ being the identity function. We simulate $N = 10^4$ responses, let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ where $\alpha_j \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$ and $\sigma_\alpha^2 = 1$ as a single random intercept for $m = 200$ clusters of $n_j = 50$ observations each, $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^{n \times p}$, where $\boldsymbol{\mu} = (1, 2, 3)$, $\Sigma_{ii} = 1$, $\Sigma_{i,k} = \rho_{i,k}$, $k \neq i$ and $p = 3$ consisting of three fixed effects, \mathbf{U} as a design matrix of appropriate dimension and a random error $\boldsymbol{\varepsilon}_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$. Further, the true vector of regression coefficients is set to be set to be $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$ so the total model, including an intercept column of ones, can be written as

$$\mathbf{y} = \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \quad (3.11)$$

To investigate how different correlations between the fixed effects are handled by the method, we consider four different correlation levels between the fixed covariates in our data. That is achieved by letting $\rho_{1,2} = \rho_{1,3} = \rho_{2,3} = \rho$ take on the values $\rho \in \{0, 0.1, 0.5, 0.9\}$. For each correlation level, we simulate $N_{\text{sim}} = 1000$ datasets and fit each of the four methods BVI, LMG, ELMG and ERW. We place penalizing complexity (PC) priors on all covariates, with parameters $U = 1$ and $a = 0.01$. To get a comparable measure from the Bayesian framework to the

frequentist framework, we use the posterior means of the sampled posterior distribution of $\text{RI}(\mathbf{X})$ by the BVI method when estimating the quantities in (3.7) and (3.8).

From this setup, the total variance of the response is

$$\text{Var}(\mathbf{y}) = \beta_1^2 + \beta_2^2 + \beta_3^2 + 2 \sum_{i=1}^3 \sum_{k=i+1}^3 \beta_i \beta_k \rho_{ik} + \sigma_\alpha^2 + \sigma_\varepsilon^2 , \quad (3.12)$$

and the theoretically expected relative importances for the case $\rho = 0$ are

$$\text{RI}(\mathbf{X}_1) = \text{RI}(\alpha) = \beta_1^2 = \sigma_\alpha^2 = \frac{1}{8} , \quad \text{RI}(\mathbf{X}_2) = \beta_2^2 = \frac{2}{8} , \quad \text{RI}(\mathbf{X}_3) = \beta_3^2 = \frac{3}{8} , \quad (3.13)$$

by following the logic of Grömping (2007) and Matre (2022). For correlated covariates however, the lack of consensus on an analytically sound decomposition of the R^2 value means we do not have theoretical values to compare with.

Further, the theoretically expected marginal R^2 values can be calculated as the variance of the fixed effects divided by the total variance given in 3.12. Similarly, the expected conditional R^2 are given by the sum of variance of the fixed effects and random intercepts divided by the total variance. The expected R^2 values are listed in Table 3.2.

ρ	$\mathbb{E}[R^2_{\text{marg}}]$	$\mathbb{E}[R^2_{\text{cond}}]$
0	0.750	0.875
0.1	0.781	0.890
0.5	0.852	0.926
0.9	0.889	0.945

Table 3.2: The theoretically expected marginal variance explained (R^2_{marg}) and conditional variance explained (R^2_{cond}) for different correlation levels between the fixed effects.

These values provide an empiric way of checking if our method fulfills the proper decomposition criteria listed in Section 2.2.1, by seeing if the relative importances for each effect sum to the model R^2 .

3.5 Modelling 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 is often the heritability of phenotypic traits, which is defined as the ratio of additive genetic variance to total phenotypic variance (Wilson 2008). We now aim to illustrate how the heritability of phenotypic traits becomes readily accessible using the BVI method, and hence why heritability is a special case of variable importance. This also involves modelling a pedigree covariance structure in random effects, which is a pivotal feature of the BVI method.

3.5.1 Heritability as relative variable importance

By comparing (2.97) 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, fixed effects are commonly not featured in formulae for the variance decomposition when estimating heritability (see Kruuk (2004) and Wilson et al. (2010)). Further, the discrimination between fixed and random effects are not always clear in biology. Often, no variance component of fixed effects is calculated. This means that they do not go into the calculation of the total phenotypic variance. However, there may be effects that are modelled as fixed, but still contribute to the phenotypic variance. To avoid confusion on this topic, we have implemented our method such that any covariate that contributes with variance in the model, is included in the calculation of total phenotypic variance. We see this to be the most clear and general way to handle the problem.

3.5.2 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 standardised inbreeding coefficient denoted *FGRM*, the standardised *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, inner islands, outer 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 called *IDC2*. It is the sampled relative importances 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 the inverse as the precision matrix for the additive genetic variance term. Lastly, to account for individual differences we add an i.i.d. random intercept, denoted by *IDC*, 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 penalising complexity (PC) priors with 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 *IDC* having $U = 1$ and $a = 0.05$. The additive genetic effect, *IDC2*, is assigned $U = \sqrt{2}$ and $a = 0.05$. These priors have been chosen through discussion with the supervisor of the thesis and researchers with domain knowledge in biology. We draw $N_{\text{samp}} = 10^4$ samples from the posterior distribution of the Bayesian GLMM fitted with the BVI method to estimate the posterior relative importances of the covariates.

3.6 Non-Gaussian studies

In this section, we present the methodology used to apply the Bayesian Variable Importance method to non-Gaussian GLMMs. This is a key extension of the method, as it allows the method to handle a wider range of models. We will analyse the binomial and Poisson GLMMs, both via a simulation study and a case study.

To evaluate the performance of the BVI method, we will use analytic results when possible, *i.e.* for uncorrelated covariates and R^2 estimates. For correlated covariates, there is no consensus on an analytically correct method to allocate the shared variance among correlated fixed effects (Grömping 2007). Further, we also compare the results with the `rptR` package (Stoffel et al. 2017), implemented in the frequentist framework. This package was originally designed to calculate the repeatability of phenotypic traits, which is closely related to relative variable importance. The main difference in capabilities between the BVI method and that of the `rptR` package, is that the BVI method estimates the relative importance distribution of each fixed effect, whereas the `rptR` package estimates only the sum of fixed effects importances. Due to this difference, our comparisons involve the estimates of relative importance for the random effects. We also compare the marginal R^2 and conditional R^2 with `rptR`, as these measures correspond to the sum of fixed effects importances and the aggregated importances of fixed and random effects, respectively.

3.6.1 Binomial and Poisson simulation studies

There are three 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 second 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 evaluate the spread of the estimates from the BVI method. 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. Lastly, the fundamental challenge that variable importance measures face, is the task of allocating the part of the variance explained due to correlation between covariates. Therefore, we wish to evaluate how our model performs for different correlation levels. This will give insight into how robust it is, and if the method handles correlated covariates in a sensible manner.

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 for the binomial model are set to be $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$. In the binomial model, the random effect $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)$ comes from $m = 100$ clusters, each with $n_j = 100$ observations for $j = 1, \dots, m$. Further, we draw the random effect from a normal distribution such that $\alpha_j \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$, with $\sigma_\alpha^2 = 1$. This means that the total variance of the linear predictor $\boldsymbol{\eta}$ is $\sigma_\eta^2 = 7$. For the Poisson model, to avoid numerical instabilities, it was necessary to standardise the linear predictor used in the simulation study. Thus, the true regression coefficients were set to be $\boldsymbol{\beta} = \frac{1}{\sqrt{7}}(1, \sqrt{2}, \sqrt{3})^T$ and the random effect $\alpha_j \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$ now with $\sigma_\alpha^2 = 1/7$. To investigate the impact of correlated fixed effects, we fit five 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 with parameters $U = 1$ and $a = 0.01$. We fit $N_{\text{sim}} = 500$ binomial and Poisson models with different datasets for each correlation level. For each fitted model, the posterior modes of the quantities of interest are used to estimate the relative importance of the covariates, as well as marginal and conditional R^2 .

In the simulation study, when parameters are simulated so that we know their true value, we can analytically calculate the relative importance of the parameters when they are not correlated. When uncorrelated, the proportion of variance explained by each fixed effect in the linear predictor is equal to the square of the true coefficient divided by the total model variance on latent scale. The proportion of variance explained by the random effect is simply its variance divided by the total model variance on latent scale. By defining $\sigma_{x_k}^2$ as the variance contribution to the linear predictor for fixed effect k and σ_α^2 as the variance contribution of the random effect, we then have for the binomial model with logit-link

$$\sigma_{x_1}^2 = \sigma_\alpha^2 = 1 \quad \text{and} \quad \sigma_{x_2}^2 = 2 \quad \text{and} \quad \sigma_{x_3}^2 = 3 , \quad (3.14)$$

and for the Poisson model with log-link

$$\sigma_{x_1}^2 = \sigma_\alpha^2 = 1/7 \quad \text{and} \quad \sigma_{x_2}^2 = 2/7 \quad \text{and} \quad \sigma_{x_3}^2 = 3/7 . \quad (3.15)$$

Then, the relative importance of the covariates can be calculated by following Section 3.3 as

$$\begin{aligned} \text{RI}(\mathbf{X}_1) = \text{RI}(\alpha_1) &= \frac{\sigma_{x_1}^2}{\sum_{i=1}^3 \sigma_{x_i}^2 + \sigma_{\alpha_1}^2 + \sigma_d^2}, \\ \text{RI}(\mathbf{X}_2) &= \frac{\sigma_{x_3}^2}{\sum_{i=1}^3 \sigma_{x_i}^2 + \sigma_{\alpha_1}^2 + \sigma_d^2}, \\ \text{RI}(\mathbf{X}_3) &= \frac{\sigma_{x_3}^2}{\sum_{i=1}^3 \sigma_{x_i}^2 + \sigma_{\alpha_1}^2 + \sigma_d^2}. \end{aligned} \quad (3.16)$$

In our simulation study, the binomial model with logit-link is assigned $\sigma_d^2 = \pi^2/3$ as in Table 3.1. The distributional variance of the Poisson model with log-link is given by

$$\sigma_d^2 = \ln(1 + 1/\mathbb{E}[\mathbf{y}]) = \ln(1 + 1/\mathbb{E}[\lambda]), \quad (3.17)$$

with

$$\mathbb{E}[\lambda] = \exp(\beta_0 + 0.5\sigma_\tau^2), \quad (3.18)$$

being the quantity used in Table 3.1 and where σ_τ^2 denotes the total model variance on the latent scale (Nakagawa et al. 2017). So we obtain, using a single random intercept, $\sigma_d^2 = 0.4741$ with $\beta_0 = 0$, $\sigma_\tau^2 = 1$. Therefore, we can summarize the expected relative importance of our three models as in Table 3.3. Recall that for correlated covariates, no consensus exists, and we therefore have no analytical results to compare with.

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	0.0969	0.0969	0.1938	0.2907

Table 3.3: The theoretically expected relative importance of the covariates on the latent scale in the different models when they are 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

$$\mathbb{E}[R_{\text{marg}}^2] = \frac{\sum_{i=1}^3 \beta_i^2 + 2 \sum_{i=1}^2 \sum_{k=i+1}^3 \beta_i \beta_k \rho}{\sum_{i=1}^3 \beta_i^2 + 2 \sum_{i=1}^2 \sum_{k=i+1}^3 \beta_i \beta_k \rho + \sigma_\alpha^2 + \sigma_d^2}, \quad (3.19)$$

and similarly for the expected conditional R^2 as

$$\mathbb{E}[R_{\text{cond}}^2] = \frac{\sum_{i=1}^3 \beta_i^2 + 2 \sum_{i=1}^2 \sum_{k=i+1}^3 \beta_i \beta_k \rho + \sigma_\alpha^2}{\sum_{i=1}^3 \beta_i^2 + 2 \sum_{i=1}^2 \sum_{k=i+1}^3 \beta_i \beta_k \rho + \sigma_\alpha^2 + \sigma_d^2}. \quad (3.20)$$

Model Type	Correlation (ρ)	$\mathbb{E}[R_{\text{marg}}^2]$	$\mathbb{E}[R_{\text{cond}}^2]$
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	0.225	0.373
Poisson Log	-0.1	0.518	0.625
Poisson Log	0	0.581	0.678
Poisson Log	0.1	0.634	0.723
Poisson Log	0.4	0.747	0.820

Table 3.4: Theoretically expected marginal and conditional R^2 values on the latent scale for the binomial regression with logit-link (top) and Poisson regression with log-link (bottom) for different correlation levels ρ .

3.6.2 Binomial and Poisson case studies

To further 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). As mentioned, this package was developed, in the frequentist framework, to estimate the repeatability of phenotypic traits. An important clarification for this case study, is that there are multiple formulations of repeatability. Two of the most common ways of looking at repeatability are

$$\begin{aligned} R_1 &= \frac{\text{Additive genetic variance}}{\text{Total variance of covariates}} \\ R_2 &= \frac{\text{Additive genetic variance}}{\text{Total variance of random covariates}} , \end{aligned} \tag{3.21}$$

where the former corresponds to our notion of heritability (Stoffel et al. 2017) and the latter to the ICC (Fahrmeir et al. 2013). 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` (Stoffel et al. 2017). The dataset used in the `rptR` package vignette, introduced for a different purpose, is simulated to replicate a study on twelve different beetle larvae populations (Stoffel et al. 2017). It contains $N = 960$ observations of the covariates *population*, the discrete *habitat* of the larvae, the dietary *treatment* of the larvae, the *sex* and *container* of which the larvae were contained in. The phenotypes to be modelled by the binomial and Poisson distributions are the two distinct male colour morphs 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 by the BVI method on the fixed effect *treatment* and random effects *population* and *container* are PC priors with parameters $U = 1$ and $a = 0.01$. Furthermore, we also here draw $N_{\text{samp}} = 10^4$ samples from the posterior distribution of the Bayesian GLMM fitted with the BVI method to estimate the posterior distributions of the repeatability.

3.7 Simulation Study with Dirichlet and Generalized Decomposition Priors on R^2

As mentioned, the literature on Bayesian variable importance is not very comprehensive. However, as discussed in Section 2.5.6, the R2D2 and GDR2 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 and GDR2 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 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$. By noting that $\text{Var}(\mathbf{y}) = 7$ in the uncorrelated case, it is clear that the expected relative importance of the independent covariates can be calculated as

$$\text{RI}(\mathbf{X})_1 = \frac{1}{7} \quad \text{RI}(\mathbf{X})_2 = \frac{2}{7} \quad \text{RI}(\mathbf{X})_3 = \frac{3}{7}. \quad (3.22)$$

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

$$R^2 = \frac{\text{Var}(\mathbf{y}) - \sigma^2}{\text{Var}(\mathbf{y})}, \quad (3.23)$$

which gives expected values that are summarized in Table 3.5.

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

Table 3.5: Theoretically expected 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 uncorrelated covariates. 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 employ the Gibbs sampler for the marginal R2D2 prior as described in (Zhang et al. 2020, section 5.3) and run the MCMC iteration $N_{\text{samp}} = 10^4$ times, with a burn in of 9000 samples which are discarded. This gives 1000 samples from the posterior distribution of the marginal R^2 as well as 1000 samples of each ϕ_j for $j = 1, 2, 3$. The BVI method draws the same number of samples for the covariates from the posterior distribution of the fitted Bayesian GLMM, using PC priors with $U = 1$ and $a = 0.01$ for all covariates. 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 $N_{\text{samp}} = 10^4$ 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 the corresponding distributions estimated from 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 `relaimpo` (Grömping & Lehrkamp 2023) in R as described in Grömping (2007). As the LMG 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 importance estimates.

RESULTS

4.1 Gaussian simulation study

This section, like Section 3.4, is reproduced from Arnstad (2024) and is included here for completeness. We present the results of the Gaussian simulation study, where the BVI method is compared to the ELMG, ERW, and LMG methods. The ELMG and ERW methods were implemented by Matre (2022), and we use the R package `relaimpo` to apply the LMG method (Grömping 2007). Note that the `relaimpo` package only considers models without random effects.

4.1.1 Fixed effects

Here, we review the relative importance allocated to each fixed effect from the simulations. There are four different distributions for each method, which corresponds to the four different correlation levels. The horizontal dashed line displays the theoretically expected relative importance from (3.13) when the covariates are pairwise independent. We used violin plots to visualize the estimated quantities, as they contain much information in a compact way. The violin plot is analogous to a density plot, but the density is shown along the y -axis and mirrored about the y -axis to form a symmetrical shape.

In general, it can be seen that the distributions in the case of uncorrelated data are unbiased with some variation around the theoretically correct relative importance (Figure 4.1). For a correlation of $\rho = 0.1$ the distributions of the estimates are shifted marginally compared to the uncorrelated case for all methods. The importance attributed to X_1 and X_2 , in Figure 4.1, is larger when compared to the uncorrelated case, whereas the importance attributed to X_3 is smaller. All methods seem to shift the relative importance estimate for the covariate with the same amount in the same direction. This shift is both expected and desirable, when considering the values found in Table 3.2 for the theoretically expected variance explained. Therefore, we should expect our method to assign different shares when we have various levels of covariate correlation, which it does. This trend continues for the correlation level $\rho = 0.5$, where the distributions are shifted further in the same directions as for $\rho = 0.1$. Lastly, for $\rho = 0.9$ we see the largest reallocation of the distributions, which follows the same trend as for the other correlation levels.

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

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

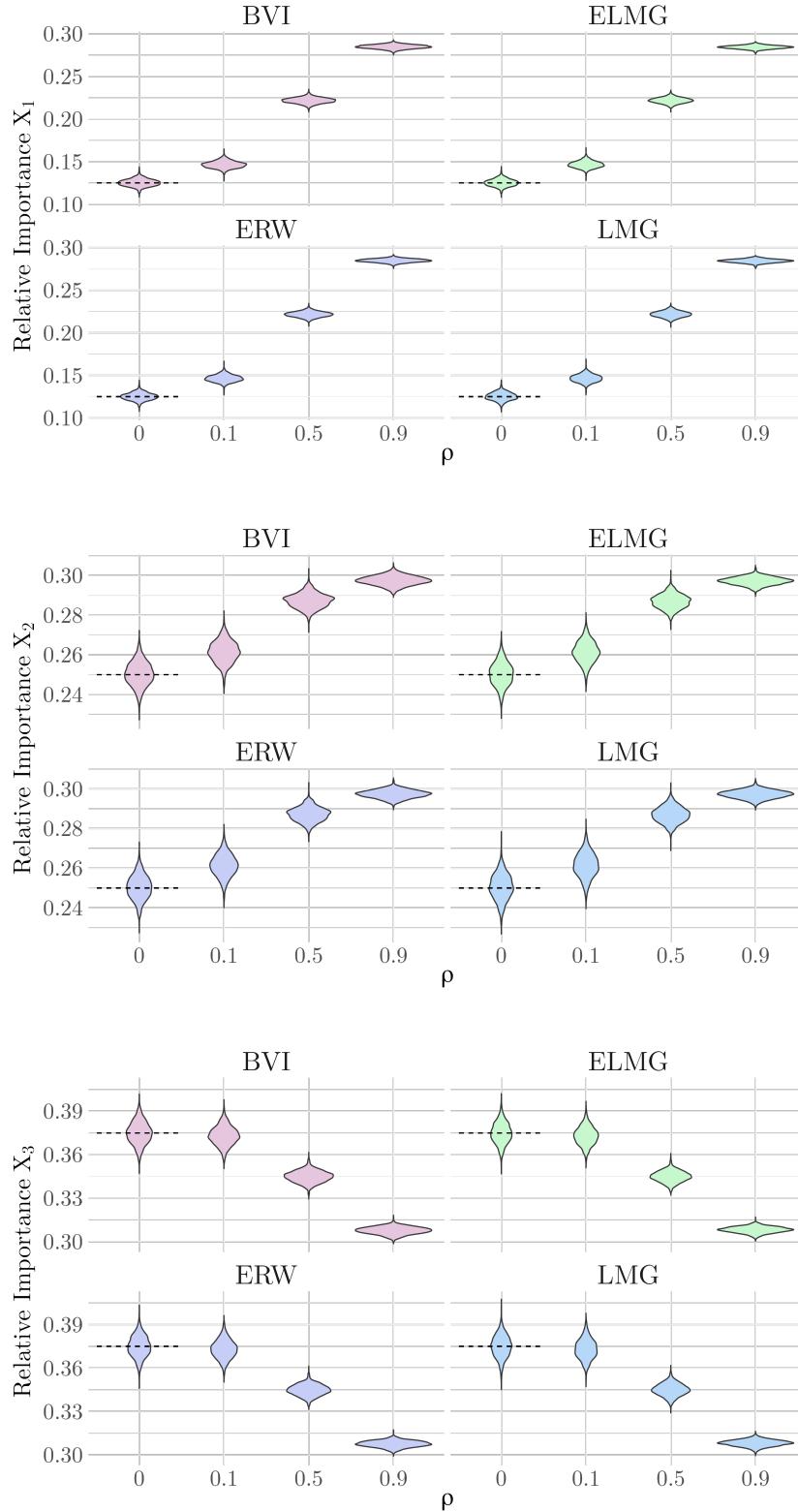


Figure 4.1: Violin plots for the relative importance of the fixed effects X_1 , X_2 and X_3 for different correlation levels displayed along the x-axis, calculated from the ensemble of $N_{\text{sim}} = 10^4$ simulated datasets by the BVI, ELMG, ERW, and the LMG methods. The horizontal line displays the theoretically expected importance of each fixed effect in the case of uncorrelated data. For the BVI method, the distributions of posterior means are shown to compare to the distribution of point estimates from the other three methods.

4.1.2 Random effects

Considering a model with one random intercept, we can no longer compare our model with the LMG method, which is only implemented for the linear regression in the `relaimpo` R package (Grömping 2007). Therefore, we now compare the BVI method only with the ELMG and ERW methods, which are extensions of the LMG method (Matre 2022). We display (Figure 4.2) the distribution of the relative importance, or variance, assigned to the random intercept α for different correlation levels. The random intercept α follows a univariate normal distribution with mean zero and variance equal to 1. As before the horizontal line shows the theoretical relative importance from (3.13) that α has in the model when the fixed effects are uncorrelated.

It is apparent that both the location and width of the relative importance distribution of all methods are largely indistinguishable (Figure 4.2). The distributions take on a moderately smaller value when $\rho = 0.1$ and the location of the estimates is further decreased for $\rho = 0.5$ and $\rho = 0.9$. For the latter correlation level, the distributions are located around a value that is less than half of the value of the centering when the fixed effects are uncorrelated. To re-emphasize, this is both expected and desirable since the increase in response variance comes solely from the correlation of fixed effects, so the random effects now contribute to explain a smaller proportion of the variance, *i.e.* the importance is lower.

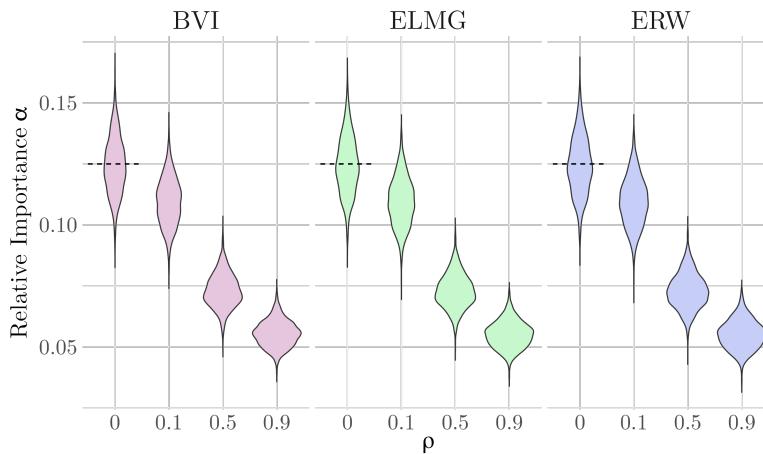


Figure 4.2: Violin plots for the relative importance of the random effect α , that is, $\hat{\sigma}_\alpha^2$ for different correlation levels displayed along the x-axis, calculated from the ensemble of $N_{\text{sim}} = 10^4$ simulated datasets by the BVI, ELMG and the ERW method. For the BVI method the distributions of posterior means of the marginal distribution of $\hat{\sigma}_\alpha^2$ are shown to compare to the point estimates of the other two methods. The horizontal line displays the theoretically expected importance σ_α^2 of the random effect in the case of uncorrelated data.

4.1.3 R^2 estimates

As a useful by-product from the previous results we can get the total variances explained by our model (Figure 4.3). The marginal variance explained, R_{marg}^2 , is

the variance explained by the fixed effects, and we get results for all four methods, including the LMG method. Below, the marginal variance explained, and the total conditional variance explained, R^2_{cond} , is displayed. This is the variance given all the fixed effects and the random effect. To complement the conditional and marginal variances explained, a horizontal line is drawn for each correlation level corresponding to the theoretically expected variance explained, found in Table 3.2, for the correlation level.

We see that the four methods produce very similar results of R^2_{marg} for the fixed effects across all correlation values, albeit a slightly larger width for the BVI method can be seen. When considering the conditional variance R^2_{cond} , the dispersion of the BVI method is strikingly larger compared to the other methods. It is not immediately clear why, but it could be a result of the relatively large dispersion of the estimated posterior marginal variance of α . Both the marginal and the conditional variance are centered around the theoretically correct value with some variability, particularly visible for conditional variance of the BVI method. The centering of the distributions for both the marginal and conditional variances resemble each other for all methods, regardless of correlation level.

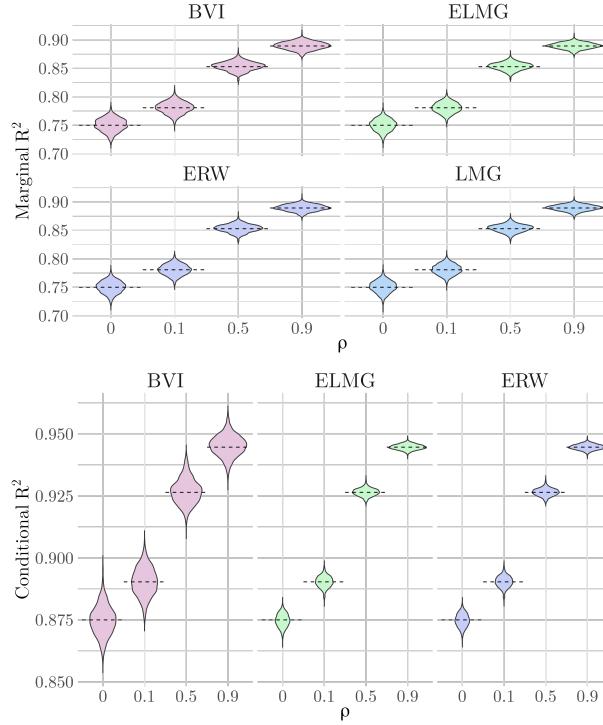


Figure 4.3: Violin plots for the total marginal (top) and conditional (bottom) variance explained (R^2) for different correlation levels displayed along the x-axis, calculated from the ensemble of $N_{\text{sim}} = 10^4$ simulated datasets by the BVI, ELMG, the ERW and the LMG method(only marginal variance explained can be computed). For the BVI method the posterior means of the sampled posterior distributions of β and the marginal distribution of $\hat{\sigma}_\alpha^2$ in each simulation are used to compare to the point estimates of the other two methods. The horizontal lines display the theoretical explained variance for each correlation level ρ as in Table 3.2.

4.2 Modelling house sparrow traits

We now investigate the results of applying our method to the house sparrow dataset. The presented results contain the sampled relative importance distributions of the covariates used to model body mass, wing length and tarsus length. By including the full analysis, we want to demonstrate that our method enables a comprehensive analysis by providing the relative variable importance of all covariates. A particular analysis is done on the heritability, as this is a special case of relative importance (Section 2.7.2) and is the focus of much research. The samples of heritability are sampled from the variance component that captures additive genetic variance (*IDC2*), and we use the results from Silva et al. (2017) and Muff et al. (2019) for comparison. More detailed plots of the estimated posterior heritability distributions are available in Appendix C.1. For the house sparrow analysis, 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 4.1 the mean of sampled heritability along with quantile values is presented, as well as the corresponding measures from the comparable studies.

	h_{mass}^2		h_{wing}^2		h_{tarsus}^2	
	Est.	CI	Est.	CI	Est.	CI
Silva et al. (2017)	0.300	[0.231, 0.369]	0.388	[0.353, 0.461]	0.415	[0.333, 0.497]
Muff et al. (2019)	0.288	[0.219, 0.371]	0.344	[0.294, 0.409]	-	-
BVI	0.284	[0.255, 0.315]	0.356	[0.334, 0.380]	0.401	[0.363, 0.438]

Table 4.1: Heritability estimates and confidence intervals 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 intervals, and the mean and quantile values of the heritability samples obtained from the BVI method, for the phenotypic traits: body mass, wing length, and tarsus length. Note that in Muff et al. (2019) no estimate of tarsus length heritability was provided.

For the sampled heritability of body mass, we have a mean of 0.284 and the interval [0.255, 0.315] covers the 95th percentile (Table 4.1). In general, the posterior distribution of heritability has a Gaussian shape, centered around the mean value (Figure 4.4). The posterior distribution of relative importance of the fixed effects predominantly shows very small values. Recall that importance calculations of fixed effects include squaring the coefficients, and so no negative importance can be allocated to a covariate. Notably, the relative importance of covariates *FGRM* and *other* exhibits a negative exponential decay pattern, while the covariates *sex*, *age* and *outer* seem to have a skewed normal distribution. The posterior distribution of relative importance allocated to *month* looks normally distributed. Considering the random effects, it is evident that they have been allocated a much larger share of importance compared to the fixed effects. The importance of the Gaussian observations have a normal shape and are often allocated over 50% of the total importance. The distribution of importance for *IDC*, which is the covariate accounting for individual differences, does not seem to follow any standard

distribution. Lastly, *hatchyear* is allocated a quite small importance, and the distribution looks like a skewed normal distribution. To fit the model and obtain the samples, the BVI method took about 73 seconds. Overall, the heritability values for body mass are in close agreement with the estimates from Silva et al. (2017) and Muff et al. (2019), and the posterior importance distributions for all covariates seem to be reasonable.

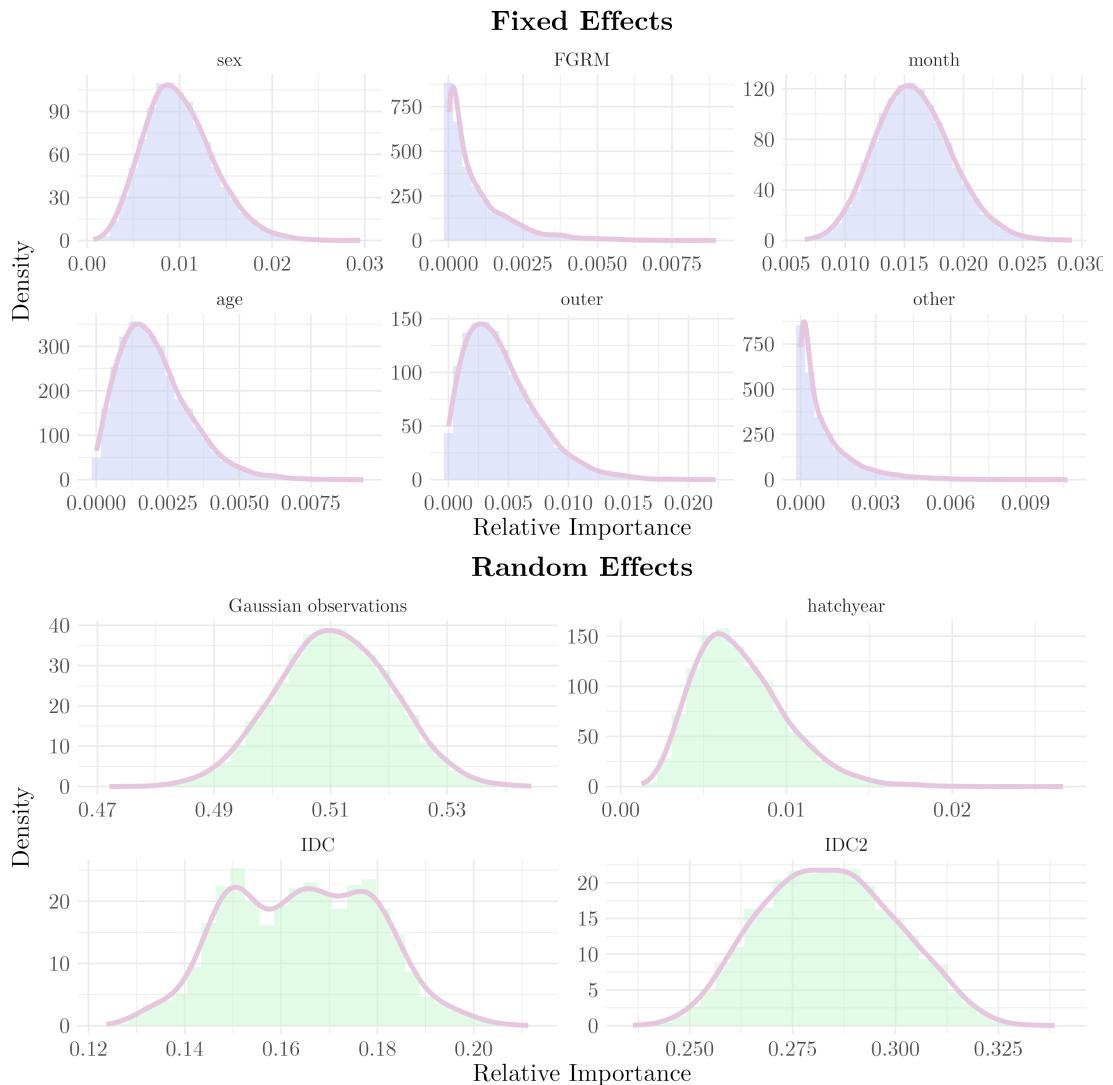


Figure 4.4: Histogram depicting the sampled posterior importance values of all covariates used to model body mass by the BVI method for the house sparrow dataset. The heritability estimates correspond to the random effect *IDC2*, and the estimates from the BVI method, Silva et al. (2017) and Muff et al. (2019) of heritability are found in Table 4.1.

The samples of wing length heritability form a symmetric bell curve, around the mean of 0.356 (Figure 4.5 and Table 4.1). Further, the 95th percentile is roughly the interval [0.334, 0.380], which is more narrow than the quantile for the heritability of body mass. Contrary to what was the case for body mass, when modelling wing length, some fixed effects are allocated a significant share of importance. For instance, the distribution of relative importance for *sex* and *age* are centered

around 0.340 and 0.055 respectively, while the others are allocated an importance of the same order of magnitude as for body mass. Again, some fixed covariates exhibit a decay pattern close to zero (*e.g.* *FGRM* and *age*), and some seem to have a skewed normal shape (*e.g.* *sex* and *outer*). For the random effects, we notice that the Gaussian observations are normally distributed and allocated roughly 17.5% of the total importance. Again the *hatchyear* effect is allocated a small importance, and the posterior importance distribution for *IDC* is not easily interpretable, but contains smaller estimates than for body mass. For wing length, the BVI method took about 76 seconds to fit the model and draw the samples. The heritability estimate from the BVI method lies between the estimates of Silva et al. (2017) and Muff et al. (2019), and again we find the posterior importance distributions for all covariates to be plausible.

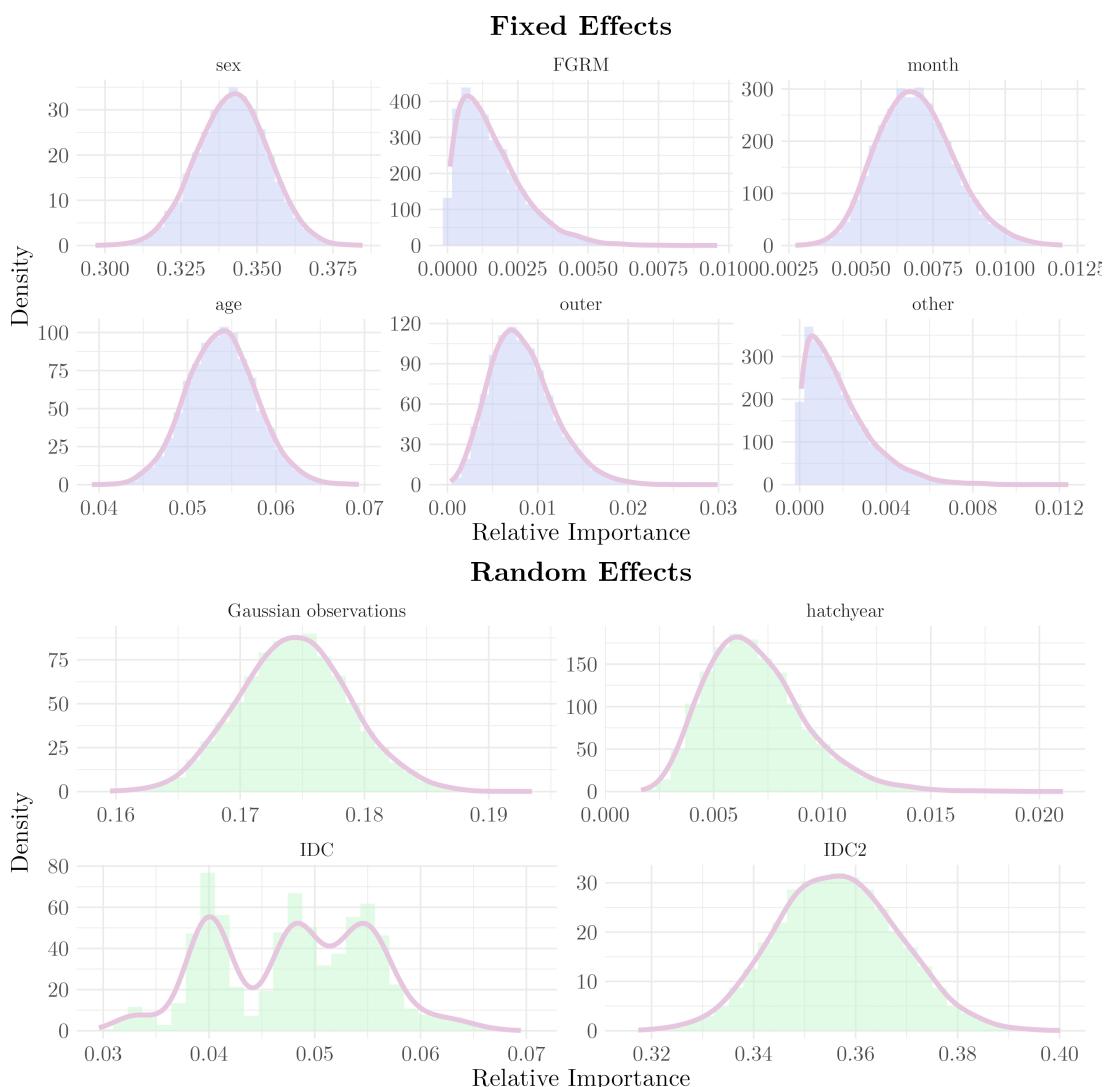


Figure 4.5: Histogram of the sampled posterior importance values for the covariates modelling wing length from the house sparrows, estimated by the BVI method. The heritability estimates correspond to the random effect *IDC2*, and the estimates from the BVI method, Silva et al. (2017) and Muff et al. (2019) of heritability are found in Table 4.1.

The heritability samples of tarsus length has a mean of 0.401 and the estimated 95th quantile is [0.363, 0.438] (Table 4.1). This is the largest quantile for all traits, and the posterior distribution of tarsus length heritability seems normal with a broader and flatter peak than for body mass and wing length (Figure 4.6). As for the body mass model, the posterior distribution of the relative importance allocated to the fixed covariates used to model tarsus length mainly consist of very low values. Here, a strong decay pattern is observable for *age*, *outer* and *other*, and a moderate decay for *sex* and *FGRM*. The only fixed covariate that looks to have a normal shape is *month*. Conspicuously, the Gaussian observations are given a small share of importance, with a bell curve centered between 0.031 and 0.032. The importance of the random effect *hatchyear* is also for tarsus length very small, and the distribution of importance allocated to *IDC* is now a bell curve with a relatively large mean of around 0.538. To model tarsus length and draw samples, the BVI reported a runtime of 74 seconds. No heritability estimate from Muff et al. (2019) was available, however the estimate from the BVI method is close to that of Silva et al. (2017). The posterior importance distributions for all covariates seem reasonable, but the small importances allocated to the random error (Gaussian observations) should be met with caution.

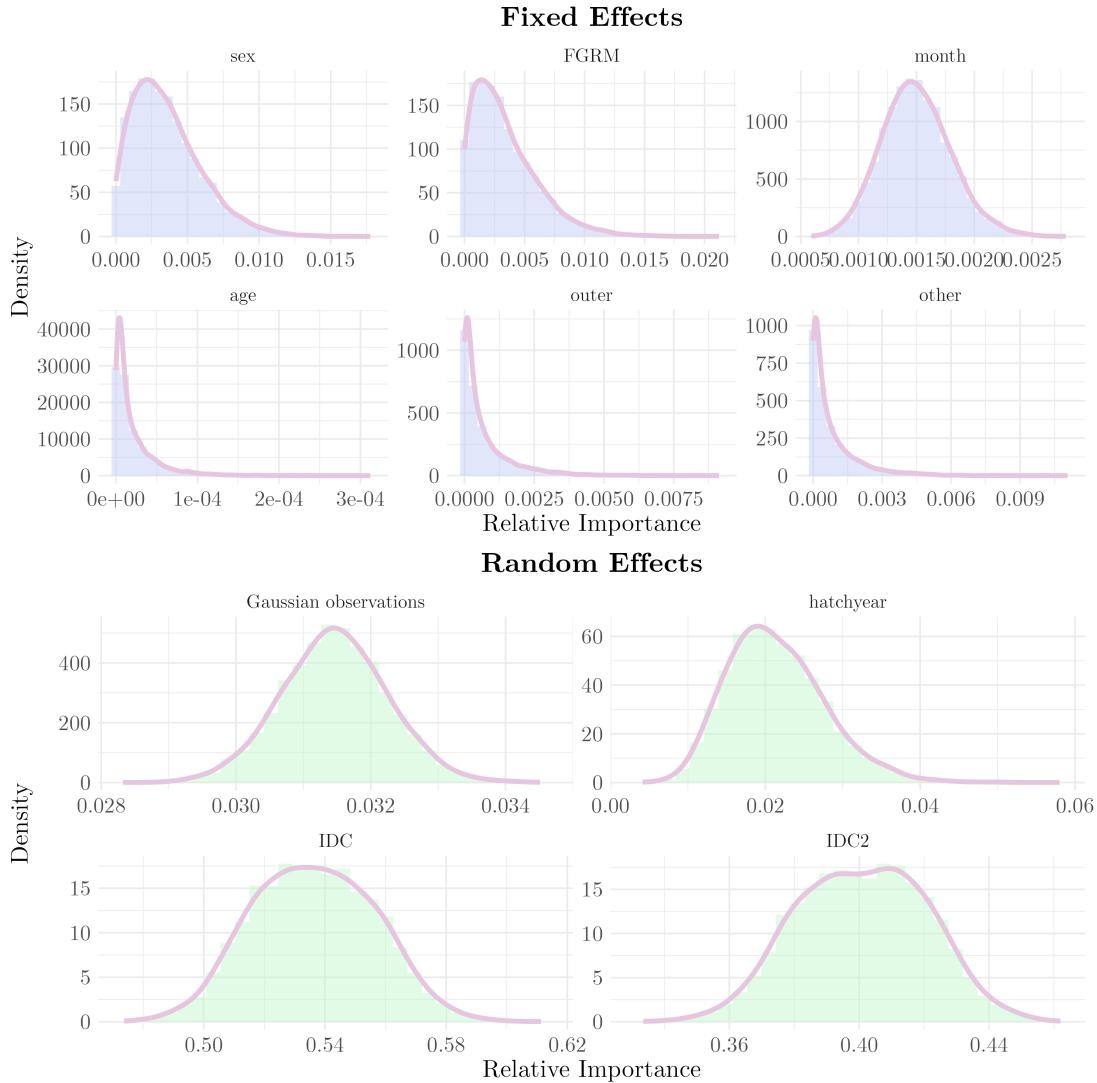


Figure 4.6: Histogram of the sampled posterior importance values for covariates used in the tarsus length model of the house sparrows, estimated by the BVI method. The heritability estimates correspond to the random effect *IDC2*, and the estimates from the BVI method and Silva et al. (2017) of heritability are found in Table 4.1. No estimate from Muff et al. (2019) was available.

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

4.3 Binomial simulation

In this section, we display the results of our simulation study of a binomial regression on binary responses. We note that it has been difficult to find suitable methods to compare the non-Gaussian models with, as we are not aware of any method that calculates relative variable importance for all covariates in the same manner as the BVI method. In parallel to fitting our model as described in Sec-

tion 3.6.1, 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 fixed effect. The model to be analysed is the binomial regression on binary response, modelled with the logit-link function. As mentioned, we fit $N_{\text{sim}} = 500$ models for each correlation level using the Bayesian Variable Importance method. Then, the BVI method extracts the posterior modes and calculates the derived measures as described in Chapter 3 to estimate the mode of relative importance for all covariates in each model. A summary of the 500 estimated modes is shown in the supplementary material (Appendix C, specifically Table C.1), which contains the mean value and values for the lower and upper 95th quantile.

4.3.1 Fixed effects

The sampled distribution for the posterior modes of relative importance allocated to the three fixed effects X_1 , X_2 and X_3 are shown for each correlation level (Figure 4.7). 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 marginal R^2 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 0.020 compared to 0.173 for $\rho = 0.4$. The same pattern is seen for X_2 and X_3 , with the average relative importance increasing from 0.077 to 0.239 for X_2 and from 0.166 to 0.299 for X_3 when going from $\rho = -0.4$ to $\rho = 0.4$. For $\rho = 0$ (middle plot of Figure 4.7), it is clear that the average estimate for relative importance of all fixed effects is very similar to the expected importance (Table 3.3) 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 in Section 4.1 from Arnstad (2024). It is explained by the fact that off-diagonal elements of Λ increase positively when the fixed effects are positively correlated, while the diagonal elements decrease. In the uncorrelated case, Λ should be equal to the identity matrix. The squared columns of Λ sum to one and act as weights. Due to this weighting, 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.

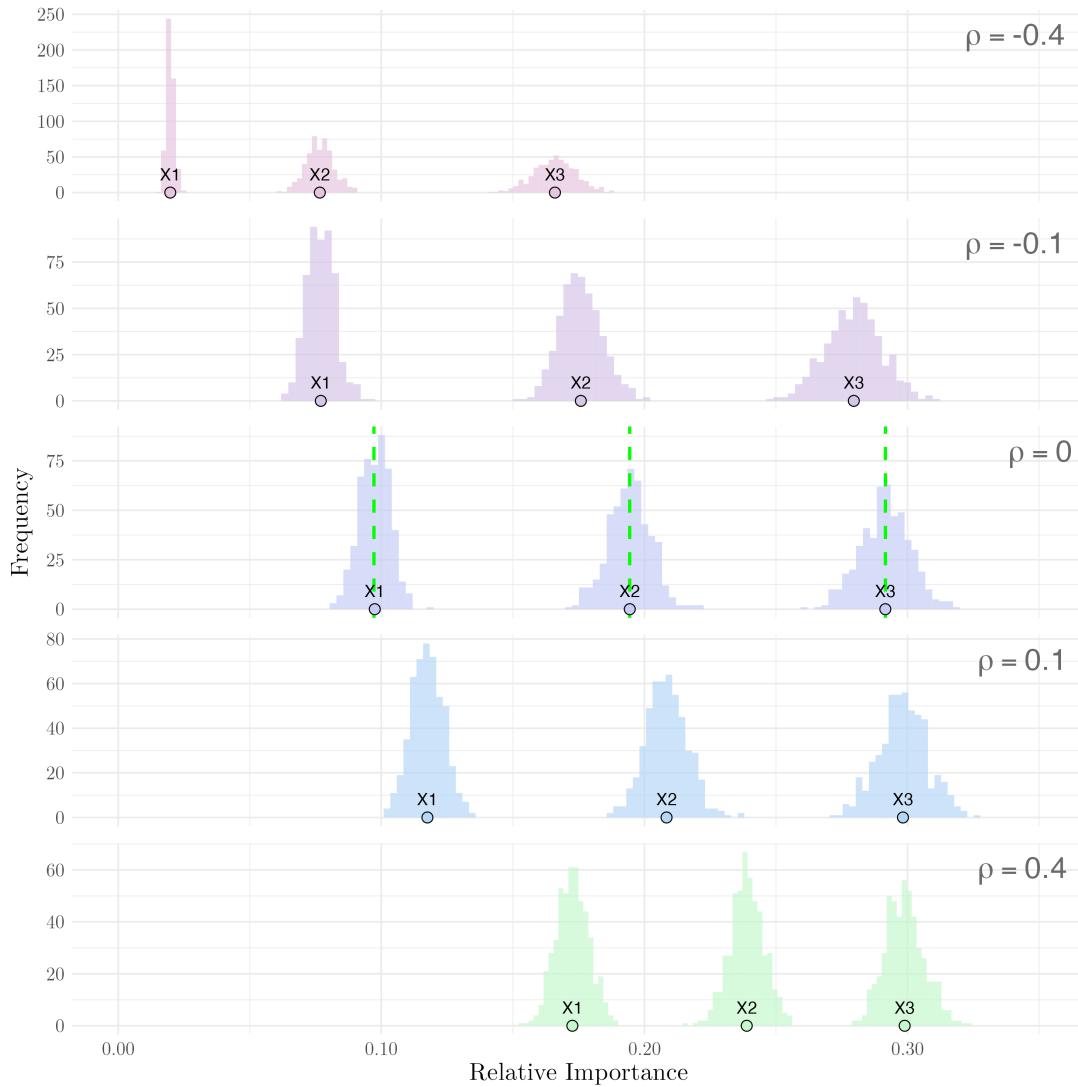


Figure 4.7: Histogram with the distribution of posterior modes of relative importance for the fixed effects present in the binomial regression for the different correlation levels ρ . The modes of relative importance are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ models fit in the simulation study. The vertical green line for $\rho = 0$ displays the expected relative importance in the case of uncorrelated data. The mean of the mode values for all simulations is denoted at the bottom of each histogram as a circle.

4.3.2 Random effect

The sampled posterior modes of importances for the random effect in the logit model all seem to be roughly normally distributed around the mean (Figure 4.8).

The spread of the random effects seem to be larger than the spread seen in the fixed effects for negative correlations, and more similar for independent and positively correlated covariates. One can see 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 0.171 compared to only 0.067 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 0.097 (Table 3.3). The orange dot at the bottom of the histograms in Figure 4.8 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 0.021 and are found when $\rho = 0$. This difference is 22% of the average estimated relative importance from the BVI method, and is therefore not negligible. However, the methods seem to be in agreement for the overall trends and the methods are closer in accordance with each other for the other correlation levels.

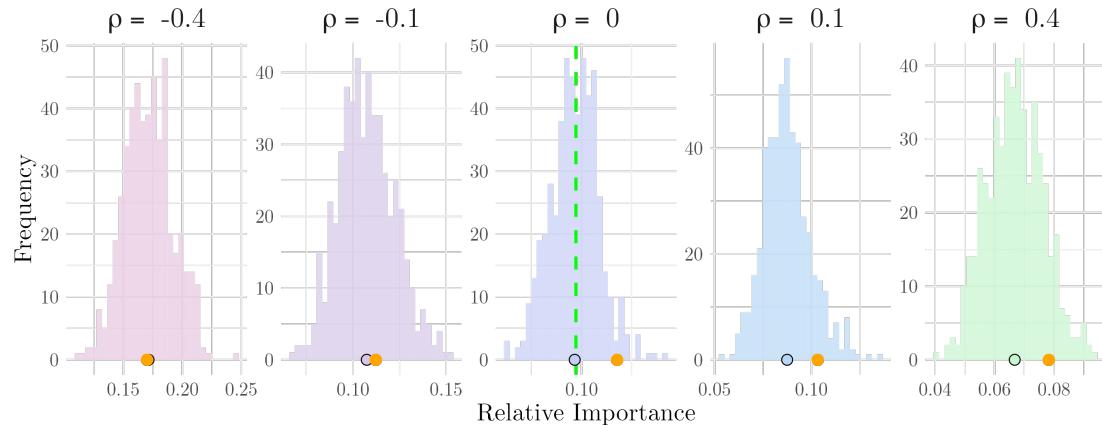


Figure 4.8: Histogram with the posterior modes of the binomial regression from the BVI method for each of the $N_{\text{sim}} = 500$ models fit, estimating relative importance of the random effect α across the different correlation levels ρ . The mean of the mode values from all simulations is displayed at the bottom as a circle and the orange dot at the bottom displays the estimate from the `rptR` package. The vertical green line for $\rho = 0$ is the expected relative importance as in Table 3.3.

4.3.3 R^2 estimates

An important measure in this simulation study is the modes for the sampled posterior distribution of marginal and conditional R^2 (Figure 4.9). The expected values for the marginal and conditional R^2 are shown in Table 3.4, 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 around the mean value. The spread is naturally larger than for the individual fixed effects and random effects, as the R^2 is constructed from these importances. For both the marginal and conditional R^2 estimates, there is only a negligible difference between the results from the BVI method and the expected values. When comparing to the results from the `rptR` package, there are some larger differences. The marginal R^2 deviates the most when $\rho = 0$ with a difference of 0.014, which is 3% of the average estimated value. For the conditional R^2 the largest difference is 0.018 which also makes up 3% of the average estimated value, found for $\rho = 0.1$. Generally, the BVI method seems to be in line with our expectation for R^2 values, deviating slightly more from the `rptR` package estimates.

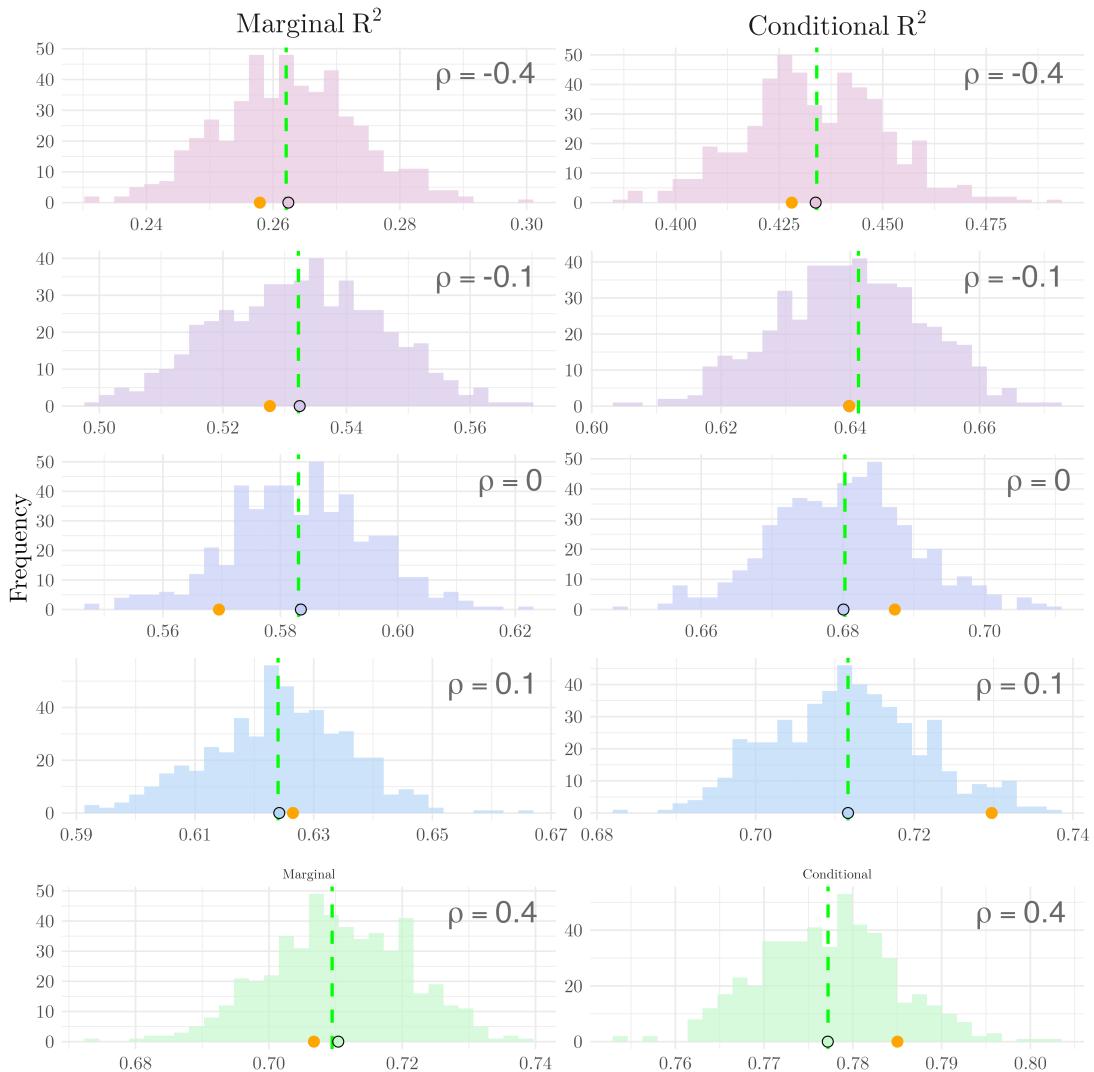


Figure 4.9: Histograms with the posterior modes of the estimated marginal and conditional R^2 from the BVI method for the binomial regression for the different correlation levels ρ . The posterior modes are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ models fit in the simulation study. The expected values are displayed as vertical green lines, and can be found in Table 3.4, while the orange dot denotes the estimate from the `rptR` package. The mean value of the R^2 modes for all simulations is marked with a circle at the bottom of each histogram.

4.4 Poisson simulation

In this section, we present the results of our simulation study of a Poisson regression. Similar to the binomial regression, we faced challenges in finding suitable methods for comparison. To address this, the `rptR` package was used with 100 bootstrap samples, which provided comparable measures in some instances. This approach allowed us to compare the importance of the random effect and the marginal and conditional R^2 values; however, `rptR` does not provide the importance of each fixed effect. We modelled the Poisson regression with count responses

using the log-link function. The Poisson models were fit with the same correlation levels as the binomial simulation and $N_{\text{sim}} = 500$ models for each level were fitted. For each fitted model, the Bayesian Variable Importance method was employed to extract the posterior mode and then calculate the relative variable importance measures as described in Chapter 3. In the supplementary material (Appendix C, specifically Table C.2), a summary of the mean and 95th quantile values of the 500 importance estimates for different correlation levels is given.

4.4.1 Fixed effects

As for the binomial model, we first look at the fixed effects. The estimates of posterior modes for relative importance for the fixed effects (Figure 4.10) are very similar in shape as the binomial model. Again, the spread is somewhat varying for the different correlation levels. Overall, the estimates for the Poisson model are marginally smaller than the binomial for negative correlations, and marginally larger for positive correlations. This is mainly due to the log-link having a distributional variance which is dependent on the correlation of covariates, and the logit-link having a constant distributional variance. When covariates are negatively correlated, the distributional variance of the log-link increases and the importance allocated to covariates decreases. Similarly, the opposite happens when covariates are positively correlated. The quantiles of the Poisson model is marginally more narrow than that of the binomial model. 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 0.017 for $\rho = -0.4$ to 0.181 for $\rho = 0.4$. For X_2 and X_3 the average relative importance for the same correlation levels increases from 0.066 to 0.251 and from 0.143 to 0.314 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 $\rho = 0.4$. Specifically, X_1 increases with 0.084, X_2 with 0.058 and X_3 with 0.024 when we go from uncorrelated covariates to the highest correlation level. This is in line with our expectations, as the diagonal elements of Λ decrease with increasing correlation levels and the off-diagonal elements increase. The inverse pattern is once again not so clear, as the decrease of X_3 is in fact larger than the decrease of X_1 and X_2 . However, if one looks at the relative decrease in importance, it is clear that the decrease is larger for X_1 and X_2 than for X_3 when going from $\rho = 0$ to $\rho = -0.4$. The inverse pattern may therefore be hard to detect, due to the relatively large difference in magnitude of the coefficients. For the case $\rho = 0$, we see that the mean of our samples coincides with the expected relative importance (Table 3.3), and it seems the model express the expected pattern of relative importance for varying correlation levels.

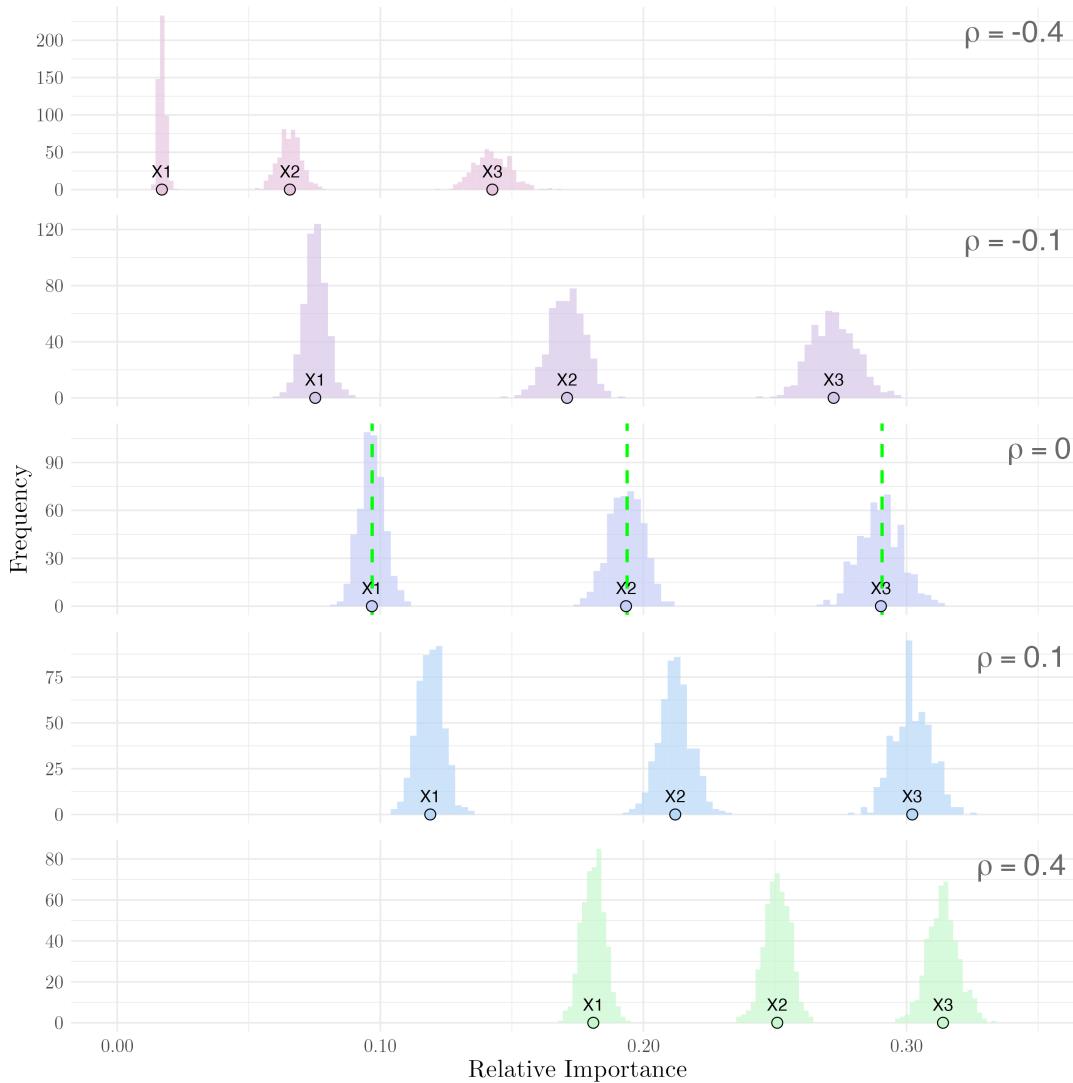


Figure 4.10: Histogram with the posterior modes of relative importance of the fixed effects present in the Poisson regression for the different correlation levels ρ . The modes are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ models fit in the simulation study. The vertical green line for $\rho = 0$ displays the expected relative importance in the case of uncorrelated data, and the mean of the modes for all simulations is denoted at the bottom of each histogram as a circle.

4.4.2 Random effect

When looking at the sampled posterior distribution of relative importance estimates of the random effect in the Poisson model (Figure 4.11), we see that they too are generally smaller than the same estimates for the binomial case when covariates are negatively correlated. Conversely, for positive correlations, the estimates tend to be marginally larger. This is consistent with the pattern observed for the fixed effects, and caused by the distributional variance of the log-link being dependent on the correlation of covariates. Furthermore, the distributions of relative importance allocated to the random effect for different correlation levels all appear to be normal and symmetric about the mean. We notice a somewhat

larger spread in the random effect importances than in the fixed effect importances for the Poisson model. The shrinkage effect of increasing correlation is also here apparent, with the average relative importance of the random effect going from 0.149 for $\rho = -0.4$ to 0.072 for $\rho = 0.4$. We again emphasize that this is natural and anticipated, as the random effect variance is constant across correlation levels, while the variance of the fixed effects varies. The expected value when $\rho = 0$ is 0.097 as shown in Table 3.3, and we see that the average estimate is close to this value. The orange dots, denoting the estimates from the `rptR` package, are close to the average estimate from the BVI method, with the largest difference being 0.017 for $\rho = 0.4$. This is also the largest relative difference, being 24% of the average estimated relative importance. The BVI method seems to be in agreement with the expected values for the relative importance of the random effect, and the `rptR` estimates are fairly close to the BVI method.

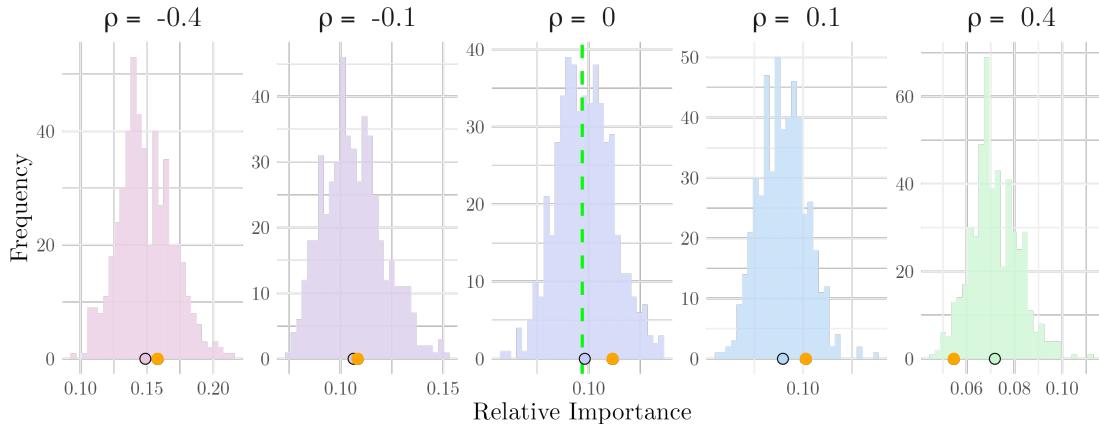


Figure 4.11: Histogram with posterior modes of the relative importance estimates for the random effect α for varying values of ρ in the Poisson regression, calculated by the BVI method. The study conducted $N_{\text{sim}} = 500$ simulations and the mean of the modes of relative importance for all simulations 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 3.3 and the orange dot is the corresponding estimate from the `rptR` package.

4.4.3 R^2 estimates

Moving on to the estimated posterior R^2 distributions for the Poisson model (Figure 4.12), we see that the expected values from Table 3.4 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 $\rho = 0$ and is 0.001 for the marginal and for the conditional R^2 the difference is negligible. The R^2 distributions seem roughly normal and symmetric around the mean value, with a plausible size of the spread. Again, there are some more notable differences between the BVI method and the `rptR` package. For $\rho = 0.1$ we observe the difference to be 0.010 for the marginal R^2 and for the conditional R^2 when $\rho = -0.4$ it is also 0.010. These differences are 2% and 4% of the average estimated values, respectively. In general, we also see here that our method aligns closely with our expectation, and deviates a bit

more from the **rptR** package estimates.

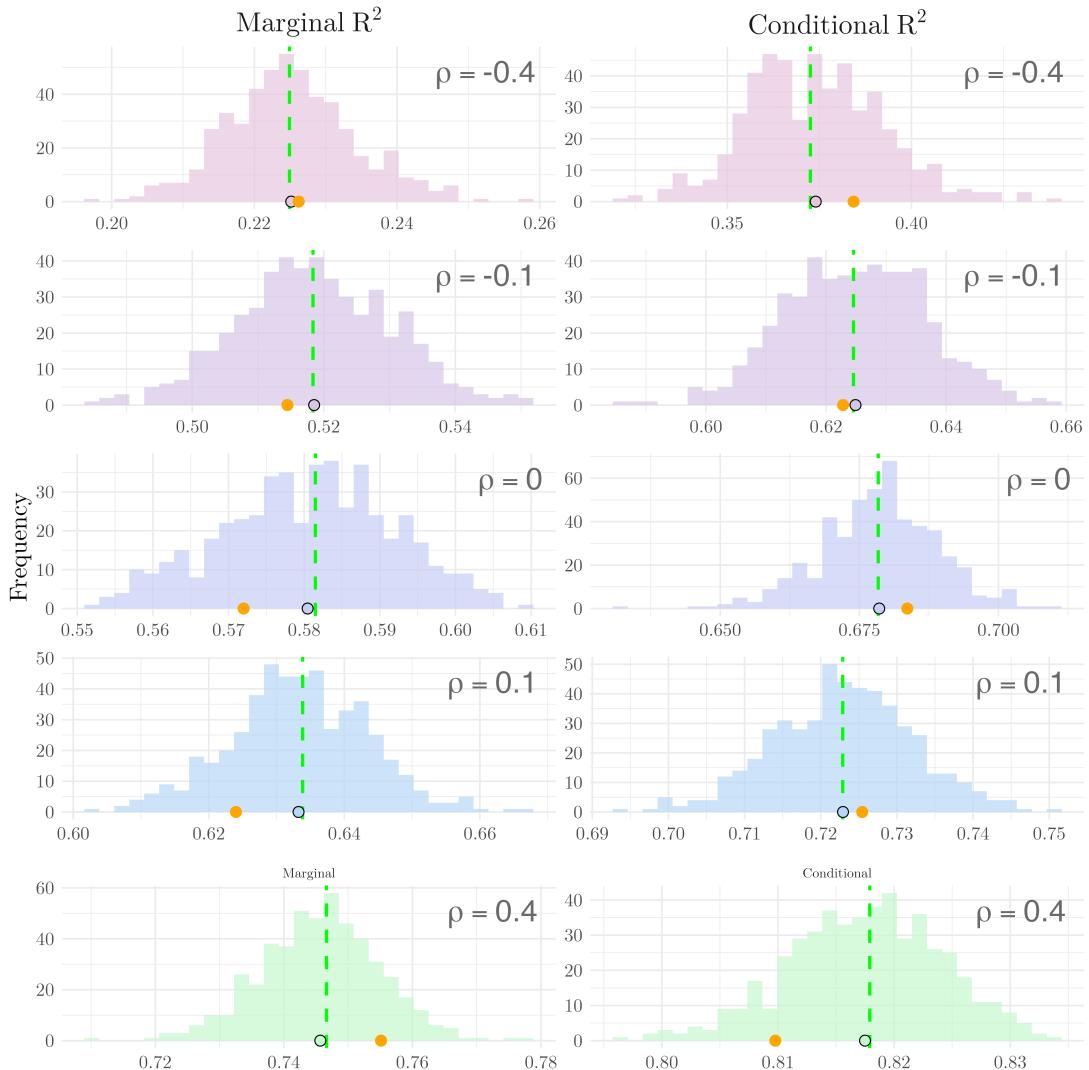


Figure 4.12: Histograms with the posterior modes of the estimated marginal and conditional R^2 from the BVI method for the Poisson regression for the different correlation levels ρ . The modes are calculated by the Bayesian Variable Importance method from the $N_{\text{sim}} = 500$ models fit in the simulation study. The expected values are displayed as vertical green lines, and can be found in Table 3.4, while the orange dot denotes the estimate from the **rptR** package. The mean value of the R^2 modes for all simulations is marked with a circle at the bottom of each histogram.

4.5 Case study with **rptR** package

To further assess our method, a comparison to the vignette for the **rptR** was made. The package described in this vignette estimates the repeatability of phenotypic traits, which for some definitions coincide with heritability and therefore can be seen as a special case of variable importance. Thus, we were able to apply the Bayesian Variable Importance method and compare the results. No expected re-

sults 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 repeatability of the colour 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 repeatability (Figure 4.13) from the Bayesian Variable Importance method is centered around a mean of 0.193, which is very similar to the estimate by Stoffel which is 0.196. 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.

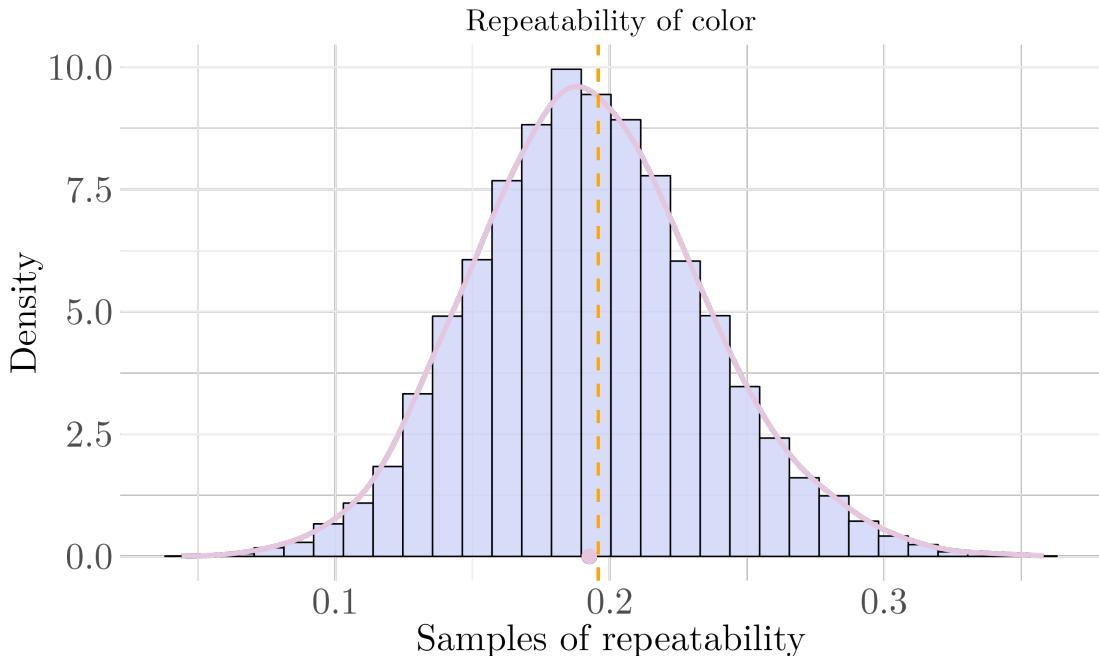


Figure 4.13: Histogram with values sampled from the posterior distribution of repeatability for the colour of male beetles from the BVI method, with the mean of the samples denoted by a pink circle. The estimate from the `rptR` package marked as a dashed line with orange colour.

To estimate the repeatability 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 3.1. This corresponds to the recommendations of Nakagawa et al. (2017) as previously mentioned. Likewise, this model is estimated from the `rptR` package with `adjusted=FALSE`. From the plotted samples of posterior repeatability of eggs laid (Figure 4.14), we see a very similar distribution as that of the binomial colour model. The distribution is symmetric and centered around a mean of 0.369. Further, the estimate from the `rptR` model is 0.380 with a confidence interval of [0.131, 0.543], compared to our 95th percentile of [0.288, 0.452]. 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.

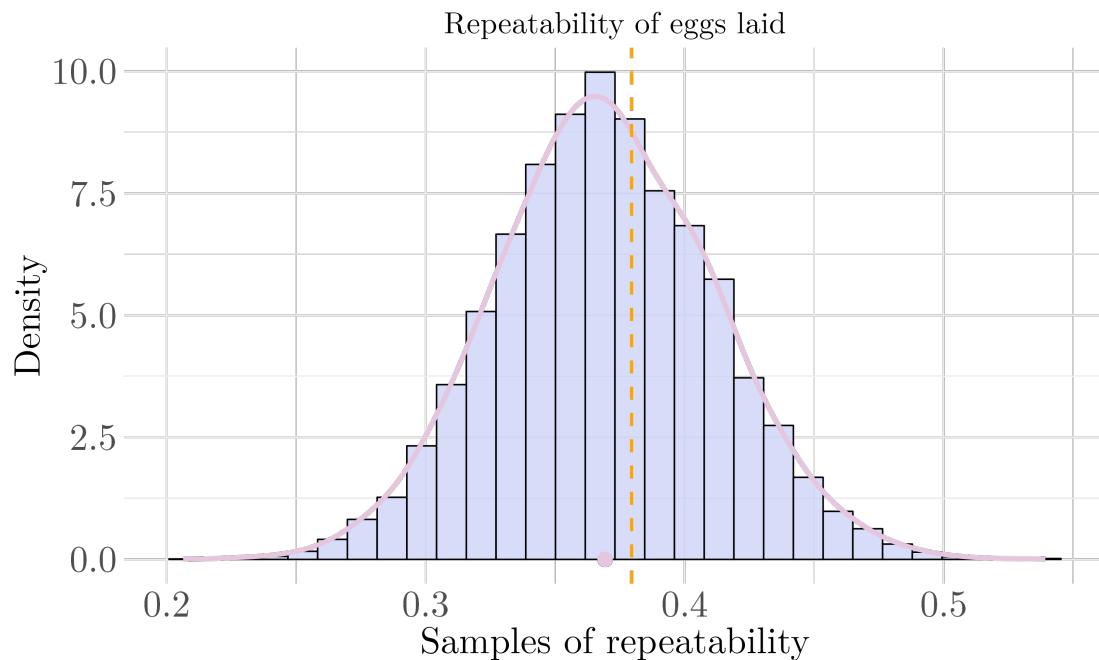


Figure 4.14: Histogram with values sampled from the posterior distribution of repeatability for eggs laid by female beetles from BVI method, with the mean of the samples denoted by a pink circle. The estimate from the `rptR` package marked as a dashed line with orange colour.

Importantly, note that the estimates from the BVI method will vary each time a model is fit, as it is stochastic. In this comparison, we only fit a single Bayesian GLMM with the BVI method. Therefore, it could be that another fit from the BVI method might align closer with the results of Stoffel and Nakagawa, but it could also be further off.

4.6 Comparing the BVI method with Dirichlet and Generalized Decomposition Priors on R^2

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 3.7. 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 sampled posterior relative importances computed by 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 theoretically expected importances for uncorrelated covariates are found in (3.22) and the theoretical R^2 for all correlation levels are found in Table 3.5.

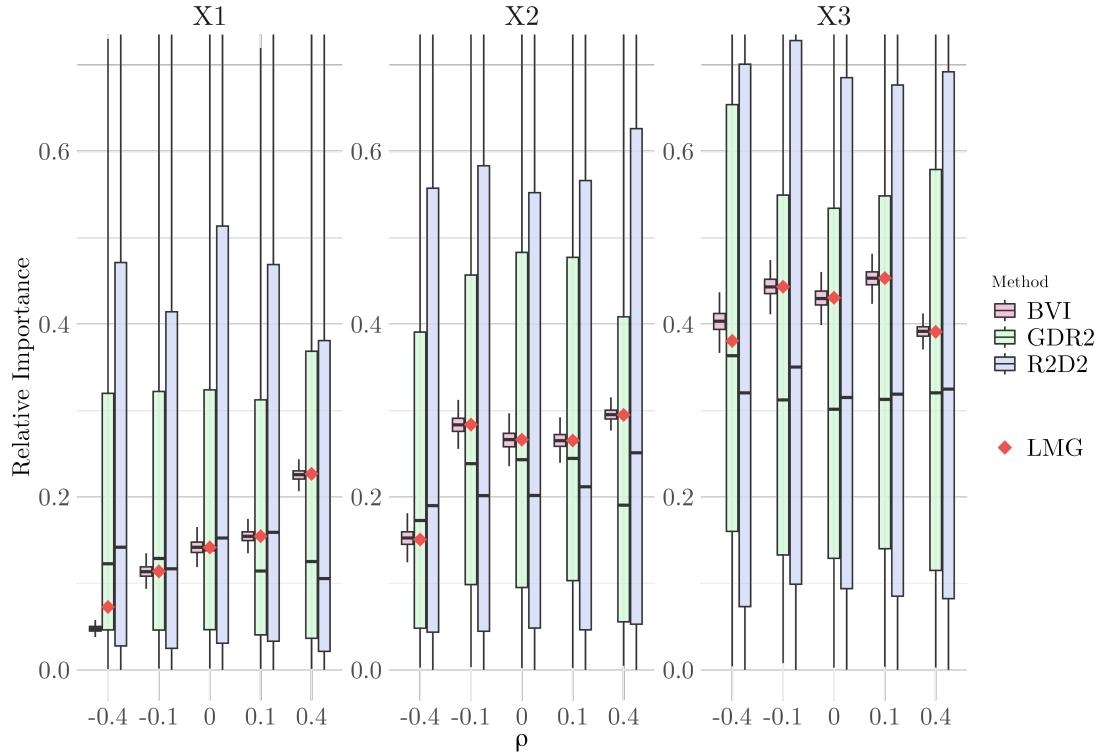


Figure 4.15: Box plots of the relative importance distributions for X_1 , X_2 and X_3 for varying correlation levels by the R2D2, GDR2 and BVI methods. The correlation levels ρ are denoted along the x-axis, and the red diamond represents the relative variable importance calculated from the LMG method.

The first thing one notices from the posterior relative importance distributions (Figure 4.15), 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 box plots, 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 and the BVI method does. Although it is hard to find patterns, and the R2D2 and GDR methods give very flexible results, they clearly grasp the larger aspects, with X_3 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, except for some small deviations 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.

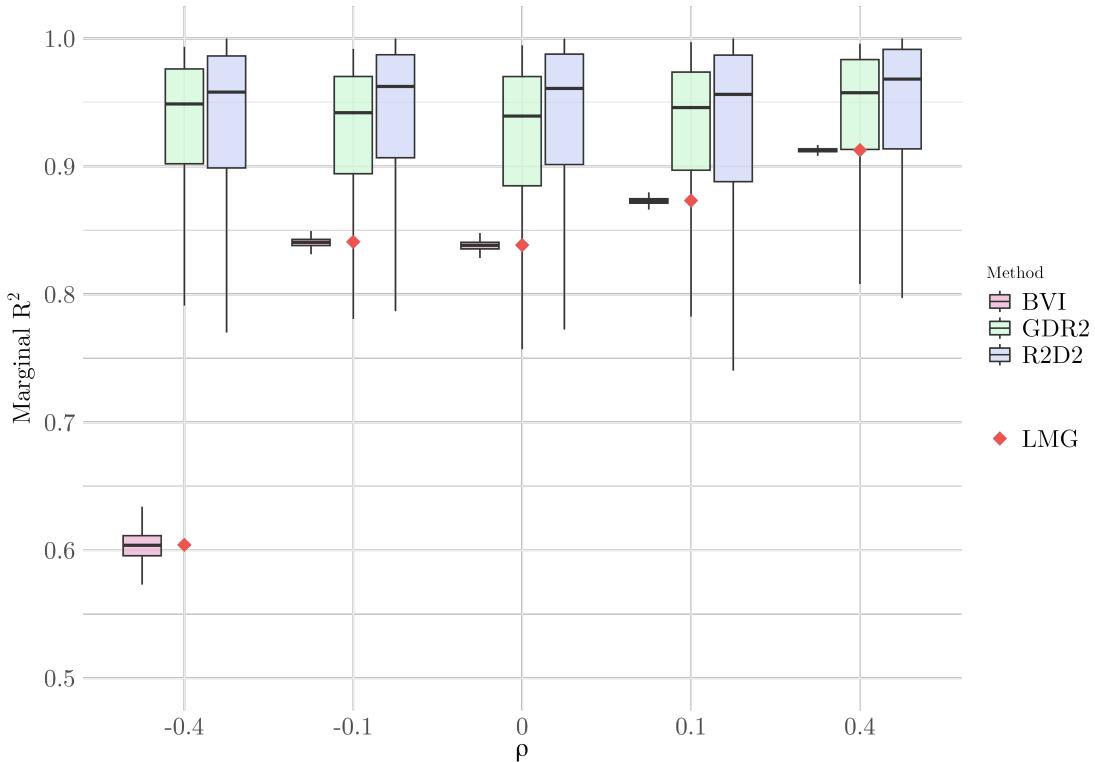


Figure 4.16: Box plots of the estimated marginal R^2 distributions for varying correlation levels ρ for the R2D2, GDR2 and BVI methods. The correlation levels ρ are denoted along the x-axis, and the red diamond represents the marginal R^2 calculated from the LMG method.

The estimated posterior marginal R^2 distributions with the R2D2 and GDR2 priors (Figure 4.16) 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 considerably larger than that of the BVI method. Overall, the R2D2 and GDR2 estimates a substantially larger R^2 than the BVI and LMG methods, and the means deviate significantly from the expected values. The BVI method is consistent with the LMG method, and they both follow the expected pattern of the R^2 when the correlation levels vary. 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 main objectives of this thesis were to develop a general variable importance method applicable to a wide range of GLMMs, allowing complex covariance structures in the random effects, and to provide interpretable and trustworthy results. In addition, the method should be easily accessible for researchers across disciplines, and be computationally feasible in most applications. Our attempt to reach these objectives has culminated in the Bayesian Variable Importance (BVI) method, which is a novel framework for estimating relative variable importance in generalized linear mixed models. The work presented in this thesis is motivated by the increased inference possible in the Bayesian framework and partially builds on the author's previous work in Arnstad (2024) and that of Matre (2022).

Summary of contributions

The development of the BVI method involved utilizing the relative weights method (Johnson 2000) to project the fixed covariates into an orthogonal space. The projection, or approximation, of these covariates is 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 translated frequentist concepts, such as the R^2 measure, to fit in the Bayesian framework. This translation has been inspired by the work of Gelman et al. (2017) and Nakagawa et al. (2017), but is the result of the author's own work. Once the methodology was developed, we revisited a Gaussian simulation study from Arnstad (2024). This simulation study indicated that the BVI method was sound for the LMMs, and so it was applied to a real world dataset. The dataset, gathered from house sparrows on the Helgeland coast, Norway, was used to investigate the heritability properties of the sparrow population. We found that the BVI method was in close agreement with other heritability studies, which was reassuring. In addition to calculating heritability, we demonstrated how the BVI method allows for a more in-depth analysis by sampling the posterior relative importance distributions of all covariates used in the model. Moving from the LMMs to the GLMMs, we conducted a simulation study for the binomial and Poisson GLMMs in which the underlying structure was known. The results from the simulation study were

promising and showed that the BVI method aligned closely with our expectations. In some cases, we compared the results to those obtained using the `rptR` package, a similar relative importance measure in the frequentist framework. This study showed that while the BVI method and the `rptR` method produced similar results, the BVI method allows for a more thorough assessment of the covariates by assigning relative importances to each of the fixed effects separately. Following this, a separate case study using `rptR` was conducted. Again, the estimates were quite similar, with the major difference being that the BVI method was significantly more efficient computationally. Hopefully, the added inference on all covariates possible with the BVI method will be preferable for researchers, as it allows for a more detailed analysis of the covariates. Lastly, we explored how similar the BVI method was to related Bayesian methods that use shrinkage priors. The results showed that the R2D2 and GDR2 methods contain very large uncertainty, and thus we argue that the BVI method is a more suitable choice for estimating relative variable importance. This is not surprising, as the shrinkage prior methods were not specifically developed for variable importance.

The full methodology has been implemented in the statistical software R, and is available as the R package `BayesianVariableImportance` 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 researchers 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 to 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 modelling complex data structures.

Assessment and validation

For relative variable importance measures, some criteria are found in Section 2.2.1 that are desirable to fulfill. The simulation study on Gaussian LMMs (Section 3.4)

shows that the BVI method compares very nicely to the robust LMG method, as well as its extension and the extension of the relative weights method. We argue, as in Arnstad (2024), that this is a promising result. Although no theoretical results were derived, one could argue that the simulation study implies that the BVI method gives a proper decomposition, at least in expectation. This is perhaps the most fundamental criterion to fulfill, as decomposing the R^2 is the main objective of the BVI method. When assessing how the BVI method performs on GLMMs, in which the response variance is not on the same scale as the covariates, this criterion 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. As the R^2 values in our thesis are constructed from the relative importance assigned to covariates, this indicates that the allocation of relative importance is sensible. 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. Further, the results from the simulation studies for the isolated covariates and the R^2 strengthen our belief that the BVI method correctly captures the expected patterns for different correlation levels. Consequently, we argue that the BVI method allocates the covariates with a plausible relative importance, both for Gaussian and non-Gaussian models. The non-negativity criterion 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 criterion will not be considered in our assessment, as Grömping (2007) argues this is not in general reasonable. Lastly, violating the inclusion criterion 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 criterion. It has not yet been properly assessed how the inclusion criterion 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. The Bayesian framework is designed to provide uncertainty, and therefore subjecting its result against the rigid thresholds of the frequentist framework is not necessarily reasonable. In the case of the inclusion criterion, we interpret this to mean that if the posterior relative importances of a non-zero regressor contains zero, this is a violation the criterion. By not considering the inclusion criterion, 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 criterion might not pose a problem at all. With this in mind, the results from both the Gaussian and non-Gaussian simulation studies 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.

Another part of validating the BVI method is to assess how well the methodology performs on real data. To investigate this, we applied the BVI method to an LMM modelling three phenotypic traits of a house sparrow population. The modelling of phenotypic traits included complex correlation structures between related birds, so the model formulation and pedigree was constructed with the help of domain experts. For each trait, the BVI method estimated the posterior relative importance distributions for all covariates in the model, with the heritability of each trait being of particular interest. The heritability estimates from the BVI method were compared to those of Silva et al. (2017) and Muff et al. (2019). For all traits the posterior distribution of the heritability from the BVI method covers the estimates from the domain experts, and places them close to the mean. We observed that the average heritability estimate for body mass from the BVI method was narrowly smaller than both estimates from Silva et al. (2017) and Muff et al. (2019). Investigating the wing length, the average heritability estimated by the BVI method was marginally larger than that of Muff et al. (2019) and a bit smaller than the estimate from Silva et al. (2017). The average heritability estimate for tarsus length was very close to the estimates from Silva et al. (2017), and in this case Muff et al. (2019) had no estimate. All posterior distributions of heritability were seemingly normally distributed, with some varying spread and kurtosis. That the BVI method is in such close agreement with estimates from published papers by domain experts is very promising, and strengthens our belief that the methodology and implementation can be used in practice.

While the house sparrow study was particularly interested in heritability values, our methodology’s true strength lies in its ability to extend this analysis to the broader concept of relative variable importance. The BVI method allows researchers to evaluate the relative variable importance of all covariates in the models, not limited to a specific measure. This comprehensive approach provides additional information that can help researchers gain a deeper understanding of the statistical models they apply. Importantly, this capability is not limited to quantitative genetics. The BVI method presents a general framework for relative importance of covariates in generalized linear mixed models (GLMMs), regardless of the context. Given the promising results observed so far, we believe that the methodology is robust, which supports its potential for application in many disciplines.

It was difficult to find a real world example to compare the binomial and Poisson GLMMs to. The solution was to compare the BVI method to the `rptR` method in a case study on repeatability as well as in the non-Gaussian simulation study where the package was applicable. The case study was created by the authors of the `rptR` package, which suggests a repeatability measure for GLMMs. As repeatability is closely linked to relative variable importance, it was possible to use the package in such a way that it could be compared to the BVI method. In the case study, the BVI method and the `rptR` package closely agreed for both the binomial and Poisson models. The spread of the posterior repeatability from the BVI method was more narrow than the confidence interval from the `rptR`. In terms of computational efficiency, the BVI method was significantly faster than

all the models from **rptR** as it does not need to bootstrap to quantify the uncertainty. When using the **rptR** package for comparisons in the simulation study, the overall results aligned well with the BVI method, with some small exceptions. For the random effects, we observed some differences that were relatively large compared to the average allocated relative importance for positive correlations. Similar observations were made for the R^2 , however these differences were smaller relative to the average R^2 estimates. Overall, the BVI method was more in line with the expected results than the **rptR** package, and demonstrated significantly faster computational performance.

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. The almost uniform distribution may not be reasonable for relative variable importance, but it is sensible for cases where there is little or no prior information available. The shrinkage prior methods generally do not follow the patterns we see from the BVI and the LMG methods, and yield estimates with large uncertainty. We argue that for the specific task of assigning relative variable importance, the BVI method is more suitable and more reliable than the R2D2 and GDR2 methods. However, the use of these shrinkage priors are primarily not focused on calculating the specific variable importance. Shrinkage prior methods could perhaps be developed further, with an emphasis on variable importance, to yield more suitable estimates for posterior relative variable importance distributions. 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). Considering 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 penalising 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 author's 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 optimise the use of shrinkage priors for variable importance by further studying the topic.

Limitations

It could be questioned if our investigations of the Bayesian Variable Importance method has been sufficient. In the Gaussian simulation study, we chose to investigate covariates that were highly correlated, whereas in the non-Gaussian we looked into more moderate correlation levels, including negative correlation. Arguably, we should have looked into negative correlation for the Gaussian simulation study, and more extreme correlation levels for the non-Gaussian study as well. However, with limited time and resources, we had to make some choices. We believe that the correlation levels used in the simulation studies were sufficient to show the general performance of the BVI method.

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 research field in itself, 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 penalising 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 mostly 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 rely on approximations and sampling. The relative weights method can be viewed an approximation of the Lindeman, 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. Further, the integration strategy used to compute the marginal posterior distributions of covariates, which again is used to approximate the joint posterior, can affect the sampled values. The samples may either be compromised by poor numerical integration due to a high dimensional hyperparameter vector,

or perhaps the assumption of the latent layer being Gaussian is not met, causing the samples to not be representative of the true posterior. During the development of the BVI package, we experienced that the choice of integration strategy could have an impact on the results. For instance, in the house sparrow study, the grid and CCD integration strategies yielded different shapes of the posterior distributions of relative importance for some traits (Appendix C.1). Although the distributions were centered around the same mean, one might interpret the results differently as a consequence of the different shapes. This exemplified the sensitivity of the BVI method to the chosen strategy, and is something the researcher should be aware of when applying the method.

Despite these limitations, we argue that the results obtained from the BVI method are satisfactory. For the simulation studies, the results align well with our expectation in cases where an expectation can be given. Additionally, the patterns for varying correlation levels seem to be logical and the results plausible, even though a true value is hard to obtain. Furthermore, the BVI method is in close alignment with comparable methods and performs well on real data. Based on this, we believe that the BVI method can be a useful tool, which is accurate enough for moderate correlation levels and does not require extensive prior tuning.

Future work

We are not 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.

Although not developed for relative variable importance, the shrinkage priors R2D2 and GDR2 could be further explored to see if they can pose as viable variable importance measures. Recently, the author was also introduced to the article *Intuitive Joint Priors for Variance Parameters* by Fuglstad et al. (2020). In the article, a framework for selecting priors based on a hierarchical decomposition of total model variance is proposed (Fuglstad et al. 2020). When prior knowledge is not available, the authors use the Dirichlet decomposition as the R2D2 method, however penalising complexity (PC) priors are used if the user has a logical idea of how to decompose the variance. In addition to using PC priors, the method proposed is designed for latent Gaussian models, which are both features of the BVI method when using INLA to fit the Bayesian GLMM. Therefore, the method by Fuglstad et al. (2020) could be of interest as a possible bridge between the BVI method and the discussed shrinkage prior methods. Due to time constraints, this

was not investigated further in this thesis, but could be of interest to explore in the future.

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 full theoretical investigations of the BVI method. More theoretical investigations 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.

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 chosen 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 being 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 very valuable practical insight 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*".

CHAPTER
SIX

CONCLUSIONS

The goal of this thesis was to provide a novel variable importance measure in the Bayesian framework for generalized linear mixed models. To do so, we applied the relative weights method and fit a Bayesian GLMM. Then, we extended a simple definition of the R^2 for GLMMs into the Bayesian framework to obtain our proposed definition. The posterior distribution of the Bayesian GLMM is sampled, before the R^2 is decomposed and distributed to the covariates, to allocate them a relative importance. The methodology is named the Bayesian Variable Importance (BVI) method and wrapped in an R package.

From simulation studies, case studies and real world applications, it has been shown that the BVI method is capable of providing plausible and robust estimates. The uncertainty in estimates is quantified, and the method allows researchers to carry out comprehensive inference. Being a general method, the BVI method can be applied to a wide range of regression models, and has proven to be computationally efficient. It is available to any reader with access to the statistical software R, and has many areas of applications across sciences. There is much potential for further augmentation of the method, both theoretically and practically. It is our aspiration, that the BVI method provides a useful tool, and that it can provide researchers with more inference on their statistical models.

BIBLIOGRAPHY

- Abramowitz, M. & Stegun, I. A. (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 10th edn, National Bureau of Standards, Washington, D.C, USA.
- Aguilar, J. E. & Bürkner, P.-C. (2024), ‘Generalized decomposition priors on R^2 ’, *arXiv preprint arXiv:2007.04558*. Accessed on 2024-05-02.
URL: <https://arxiv.org/html/2401.10180v1>
- Akaike, H. (1974), ‘A New Look at the Statistical Model Identification’, *IEEE Transactions on Automatic Control* **19**(6), 716–723. Accessed on 2023-12-03.
URL: <https://doi.org/10.1109/TAC.1974.1100705>
- 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. Available upon request.
- 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. Accessed on 2023-11-14.
URL: <https://doi.org/10.1098/rstl.1763.0053>
- 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. Accessed on 2023-12-01.
URL: <https://www.nature.com/articles/s41562-017-0189-z>
- Blakeley B. McShane, Christian Robert, J. L. T. & Gelman, A. (2019), ‘Abandon statistical significance’, *The American Statistician* **73**(sup1), 235–245. Accessed on 2023-12-01.
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. Accessed on 2023-11-20.
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. Accessed on 2024-01-19.
URL: <https://www.sciencedirect.com/science/article/pii/S0304407696018180>
- Chevan, A. & Sutherland, M. (1991), ‘Hierarchical partitioning’, *The American Statistician* 45(2), 90–96. Accessed on 2023-14-11.
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*. Accessed on 2023-10-05.
URL: <https://arxiv.org/pdf/2112.02861.pdf>
- Conner, J. K. & Hartl, D. L. (2004), *Primer of Ecological Genetics*, Harvard University Press, Cambridge, MA, USA.
- Fabbris, L. (1980), ‘Measures of predictor variable importance in multiple regression: An additional suggestion’, *Quality and Quantity* 14, 787–792. Accessed on 2023-12-03.
URL: <https://doi.org/10.1007/BF00145808>
- Fahrmeir, L., Lang, S., Kneib, T. & Marx, B. (2013), *Regression - Models, Methods and Applications*, Springer Berlin, Heidelberg, Germany.
- 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. Accessed on 2023-10-17.
URL: <https://doi.org/10.1093/biostatistics/kxp053>
- Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H. & Riebler, A. (2020), ‘Intuitive Joint Priors for Variance Parameters’, *Bayesian Analysis* 15(4), 1109 – 1137. Accessed: 2024-05-21.
URL: <https://doi.org/10.1214/19-BA1185>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2015), *Bayesian Data Analysis*, 3 edn, Chapman & Hall/CRC, New York.
- Gelman, A., Goodrich, B., Gabry, J. & Ali, I. (2017), R-squared for bayesian regression models, Technical report, Linköping. Accessed on 2023-09-03.
URL: https://www.ida.liu.se/~732G43/bayes_R2.pdf
- Genizi, A. (1993), ‘Decomposition of R^2 in multiple regression with correlated regressors’, *Statistica Sinica* 3, 407–420. Accessed on 2023-12-07.
URL: <https://www3.stat.sinica.edu.tw/statistica/oldpdf/A3n210.pdf>
- Grömping, U. (2007), ‘Estimators of Relative Importance in Linear Regression Based on Variance Decomposition’, *The American Statistician* 61, 139–147. Accessed on 2023-08-31.
URL: <https://www.jstor.org/stable/27643865>

- Grömping, U. (2015), ‘Variable importance in regression models’, *WIREs Computational Statistics* **7**(2), 137–152. Accessed on 2023-10-15.
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. Accessed on 2023-12-14.
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. Accessed on 2023-10-16.
URL: https://doi.org/10.1207/S15327906MBR3501_1
- Johnson, P. C. (2014), ‘Extension of nakagawa & schielzeth’s R^2 GLMM to random slopes models’, *Methods in Ecology and Evolution* **5**, 944–946. Accessed on 2024-03-12.
URL: <https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.12225>
- Johnson, R. (1966), ‘The minimal transformation to orthonormality’, *Psychometrika* **31**, 61–66. Accessed on 2023-10-14.
URL: <https://doi.org/10.1007/BF02289457>
- Kruskal, W. (1987), ‘Relative importance by averaging over orderings’, *The American Statistician* **41**(1), 6–10. Accessed on 2023-11-14.
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. Accessed on 2023-09-27.
URL: <http://rstb.royalsocietypublishing.org/>
- Kullback, S. & Leibler, R. A. (1951), ‘On Information and Sufficiency’, *The Annals of Mathematical Statistics* **22**(1), 79–86. Accessed on 2024-05-07.
URL: <https://doi.org/10.1214/aoms/1177729694>
- Laplace, P. S. (1774), ‘Mémoire sur la probabilité des causes par les événements’, *Mémoires de l’Académie royale des Sciences de MI (Savants étrangers)* **4**, 621–656. Reprinted in: Laplace, *Oeuvres complètes* (Paris, France: Gauthier-Villars et fils, 1841), vol. 8, pp. 27–65, Accessed on 2024-05-20.
URL: <https://gallica.bnf.fr/ark:/12148/bpt6k77596b/f32.image>
- Lindeman, R. H., Merenda, P. F. & Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis*, Scott, Foresman and Company, Glenview, IL, USA.

Lipovetsky, S. & Conklin, M. (2001), ‘Analysis of regression in game theory approach’, *Applied Stochastic Models in Business and Industry* **17**, 319 – 330. Accessed on 2023-12-06.

URL: <https://doi.org/10.1002/asmb.446>

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*. Accessed on 2024-04-28.

URL: <https://doi.org/10.1002/9781118445112.stat08212>

Matre, A. (2022), *Relative Variable Importance Approaches for Linear Models with Random Intercepts*, Master’s thesis, Department of Mathematical Sciences NTNU – Norwegian University of Science and Technology.

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

Menard, S. (2000), ‘Coefficients of determination for multiple logistic regression analysis’, *American Statistician* **54**(1), 17–24. Accessed on 2024-01-12.

Mirsky, L. (1960), ‘SYMMETRIC GAUGE FUNCTIONS AND UNITARILY INVARIANT NORMS’, *The Quarterly Journal of Mathematics* **11**, 50–59. Accessed on 2023-10-13.

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). Accessed on 2024-03-03.

URL: <https://doi.org/10.1186/s12711-019-0449-7>

Murtaugh, P. A. (2014), ‘In defense of p values’, *Ecology* **95**(3), 611–617. Accessed on 2024-05-27.

URL: <https://doi.org/10.1890/13-0590.1>

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. Accessed on 2024-03-14.

URL: <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, Accessed on 2024-03-15.

URL: <https://doi.org/10.1111/j.1469-185X.2010.00141.x>

Nakagawa, S. & Schielzeth, H. (2013), ‘A general and simple method for obtaining R^2 from generalized linear mixed-effects models’, *Methods in Ecology and Evolution* **4**, 133–142. Accessed on 2023-09-30.

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. Accessed on 2023-09-07.
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. Accessed on 2023-11-14.
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.
- Romero, J. E. A. & Bürkner, P.-C. (2024), ‘Generalized decomposition priors on R^2 ’, 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)* **71**, 319—392. Accessed on 2023-09-03.
URL: <https://doi.org/10.1111/j.1467-9868.2008.00700.x>
- Ryan, T. P. (1989), *Statistical Methods for Quality Improvement*, 1st edn, John Wiley & Sons, New York, New York, USA.
- Salvesen, N. (2023), ‘Thesis template NTNU’. Latex thesis template, Accessed: 2023-08-31.
URL: <https://www.overleaf.com/latex/templates/thesis-template-ntnu/kpybjxchxnmc>
- Schwarz, G. (1978), ‘Estimating the Dimension of a Model’, *The Annals of Statistics* **6**(2), 461–464. Accessed on 2023-12-03.
URL: <https://doi.org/10.1214/aos/1176344136>
- Shapley, L. S. (1953), ‘Stochastic games’, *Proceedings of the National Academy of Sciences* **39**, 1095 – 1100. Accessed on 2023-12-06.
URL: <https://api.semanticscholar.org/CorpusID:263414073>
- Silva, C., McFarlane, S., Hagen, I., Rønnegård, L., Billing, A., Kvalnes, T., Kempainen, 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. Accessed on 2024-04-20.
URL: <https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2463307>
- 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. Accessed on 2023-10-12.
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. Accessed on 2024-02-08.
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. Accessed on 2024-03-07.
URL: <https://doi.org/10.1111/2041-210X.12797>
- 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, Accessed on 2024-02-13.
URL: <https://doi.org/10.2307/2336490>
- Wilson, A. J. (2008), ‘Why h^2 does not always equal $\frac{V_a}{V_p}?$ ’, *Journal of Evolutionary Biology* **21**(3), 647–650. Accessed on 2024-02-14.
URL: <https://doi.org/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. Accessed on 2024-02-13.
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’. GitHub repository, 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*. Accessed on 2024-05-02.
URL: <https://arxiv.org/abs/2007.04558>

APPENDIX
A

GITHUB REPOSITORIES

All code and data used to produce results and all latex files used to produce this document are included in the GitHub repositories linked below. Please note that the package developed for the master's thesis encapsulates the package developed for the project thesis. Further explanations are given in the README files.

GitHub repositories

- Package developed for master's thesis:
<https://github.com/AugustArnstad/BayesianVariableImportance>
- Package developed for project thesis:
<https://github.com/AugustArnstad/BayesianImportance>
- Full master's thesis:
<https://github.com/AugustArnstad/TMA4900-Master-Thesis>
- Full project thesis:
<https://github.com/AugustArnstad/TMA4500-Specialization-Project>

APPENDIX

B

BAYESIAN VARIABLE IMPORTANCE USAGE

```
1 ## GENERAL SETUP
2 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.
3
4 ``{r setup, input=FALSE, echo=FALSE}
5 library(formatr)
6 showsol <- FALSE
7 library(knitr)
8 library(devtools)
9 knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 68),
10                      tidy = TRUE,
11                      warning = FALSE,
12                      error = FALSE,
13                      message = FALSE,
14                      echo = TRUE,
15                      fig.width=7,
16                      fig.height=5,
17                      fig.align="center")
18 library(remote)
19 library(INLA)
20 library(mnormt)
21 library(ggplot2)
22 library(reshape2)
23 library(RColorBrewer)
24 library(tidyr)
25 library(dplyr)
26
27
28 ## INSTALLING THE PACKAGE
29 This section ensures the devtools package is installed, which
   is required for installing packages from GitHub. We then
   install the BayesianVariableImportance package directly
   from GitHub using devtools::install_github(). In the
   package under the Hello.R file, all functions are defined
```

```

      with corresponding documentation.

30  '''{r}
31 # If not already installed, install the 'devtools' package
32 if(!require(devtools)) install.packages("devtools")
33 devtools::install_github("AugustArnstad/
34   BayesianVariableImportance")
35 library(BayesianVariableImportance)
36 '''
37
38 ## SIMULATE DATA
39 In this part, we simulate data to demonstrate the
40   functionality of the BayesianVariableImportance package.
41   We generate random variables used as fixed effects with
42   different correlation structures and random effects. Note
43   that the coefficients used here are a bit large for the
44   Poisson model, consider lowering them. The data is then
45   structured into data frames for further analysis. If you
46   have a suitable dataset you can use this instead.
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67

```

30 '''{r}

31 # If not already installed, install the 'devtools' package

32 if(!require(devtools)) install.packages("devtools")

33 devtools::install_github("AugustArnstad/

34 BayesianVariableImportance")

35 library(BayesianVariableImportance)

36 '''

37 ## SIMULATE DATA

38 In this part, we simulate `data` to demonstrate the

39 functionality of the `BayesianVariableImportance` package.

40 We generate random variables used as fixed `effects` with

41 different correlation structures and random `effects`. Note

42 that the `coefficients` used here are a bit large for the

43 Poisson `model`, consider lowering them. The `data` is then

44 structured into `data frames` for further analysis. If you

45 have a suitable dataset you can use this instead.

39 '''{r}

40

41

42 `set.seed(1)`

43

44 `simulate_data <- function(n = 10000, n_groups = 100,`

45 `covariance_level=0) {`

46 `# Simulate fixed effects`

47

48 `sigma <- matrix(c(1, covariance_level, covariance_level,`

49 `covariance_level, 1, covariance_level,`

50 `covariance_level, covariance_level, 1),`

51 `3, 3)`

52

53 `X <- MASS::mvrnorm(n = n, mu = c(0, 0, 0), Sigma = sigma)`

54 `X1 <- X[, 1]`

55 `X2 <- X[, 2]`

56 `X3 <- X[, 3]`

57

58 `# Simulate random effects groups`

59 `Z1 <- sample(1:n_groups, n, replace = TRUE)`

60 `random_effect_contributions_z1 <- rnorm(n_groups, mean = 0,`

61 `sd = 1)[Z1]`

62

63 `# Coefficients for fixed effects`

64 `beta1 <- 1`

65 `beta2 <- sqrt(2)`

66 `beta3 <- sqrt(3)`

67

68 `# Linear predictor`

69 `eta <- beta1*X1 + beta2*X2 + beta3*X3 + random_effect_`

70 `contributions_z1`

```

68 # Binomial with logit link
69 p_logit <- exp(eta) / (1 + exp(eta))
70 y_logit_bin <- rbinom(n, size = 1, prob = p_logit)
71 data_logit <- data.frame(y_logit_bin, X1, X2, X3, Z1)
72
73 # Binomial with probit link
74 p_probit <- pnorm(eta)
75 y_probit_bin <- rbinom(n, size = 1, prob = p_probit)
76 data_probit <- data.frame(y_probit_bin, X1, X2, X3, Z1)
77
78 # Poisson with log link
79 lambda <- exp(eta)
80 y_pois <- rpois(n, lambda = lambda)
81 data_poisson <- data.frame(y_pois, X1, X2, X3, Z1)
82
83 epsilon = rnorm(n, mean=0, sd=sqrt(1))
84 y_normal <- beta1*X[, 1] + beta2*X[, 2] + beta3*X[, 3] +
  random_effect_contributions_z1 + epsilon
85 data_normal <- data.frame(y_normal, X1, X2, X3, Z1)
86
87
88 list(binomial_logit = data_logit,
89       binomial_probit = data_probit,
90       poisson = data_poisson,
91       normal = data_normal)
92 }
93
94
95
96 /**
97
98
99 ## USAGE
100 Here we demonstrate the usage of the
  BayesianVariableImportance package. We fit Bayesian
  binomial, Poisson and gaussian models and sample posterior
  distributions for different simulated datasets using
  functions from the package. Then, plots are made to
  display the results.
101 ``{r}
102 set.seed(1234)
103
104 datasets <- simulate_data()
105
106 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)
  )))
107
108
109
110
111 )

```



```

145 imp_lmm <- BayesianVariableImportance::extract_importances(
  model_normal, datasets$normal)
146
147
148 samples_logit <- BayesianVariableImportance::sample_posterior_
  _count(model_logit, glmm_logit, datasets$binomial_logit, n_
  _samp=5000, additive_param = "Z1")
149 samples_pois <- BayesianVariableImportance::sample_posterior_
  _count(model_pois, glmm_pois, datasets$poisson, n_samp
  =5000, additive_param = "Z1")
150 samples_lmm <- BayesianVariableImportance::sample_posterior_
  gaussian(model_normal, lmm, datasets$normal, n_samp=5000,
  additive_param = "Z1")
151
152 plots_logit <- BayesianVariableImportance::plot_samples(
  samples_logit)
153 plots_pois <- BayesianVariableImportance::plot_samples(
  samples_pois)
154 plots_lmm <- BayesianVariableImportance::plot_samples(samples_
  _lmm)
155 ''
156
157 ## IMPORTANCES
158 The simplest way of obtaining the importances can be done by
  looking at these objects. Note that these are sampled, so
  they do not represent the mean of the samples used for
  plotting further down.
159 '''{r}
160 imp_logit
161
162 imp_pois
163
164 imp_lmm
165 ''
166
167 ## PLOTS
168 These are the default plots that are implemented in the
  package, displaying the importance of all effects and $R^2
  $ metrics.
169 '''{r}
170 plots_logit$fixed_effects
171 plots_logit$random_effects
172 plots_logit$heritability
173 plots_logit$R2
174
175 plots_pois$fixed_effects
176 plots_pois$random_effects
177 plots_pois$heritability
178 plots_pois$R2
179
180 plots_lmm$fixed_effects

```

```

181 plots_lmm$random_effects
182 plots_lmm$heritability
183 plots_lmm$R2
184
185 """
186
187
188 ## CUSTOM PLOT
189 Cutsomizing plots is often very nice to display information
   in the way you want it. Therefore, we show how one can
   customize the plots using ggplot2 based on the samples
   drawn.
190 """
191 {r}
192 random <- "Z1"
193
194 random_plot <- ggplot(samples_pois$scaled_random_samples, aes
   (x = !!sym(random))) +
   geom_histogram(aes(y = after_stat(density)), fill = "#"
      C6CDF7", alpha = 0.7, bins = 40, color = "black") +
   geom_density(color = "#E6C6DF", adjust = 1.5, linewidth
      =1.5) +
   geom_point(aes(x = mean(samples_pois$scaled_random_samples$
      Z1), y = 0), color = "#E6C6DF", size = 4) +
   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)
   )
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220

```

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

APPENDIX
C

SUPPLEMENTARY MATERIAL

C.1 Supplementary figures for the house sparrow study

We include figures of the heritability estimates obtained when using the grid and CCD integration strategies for body mass, wing length and tarsus length, in the house sparrow study (Section 3.5). The figures are presented in the same order as in the main text, starting with the body mass model, followed by the wing length model, and finally the tarsus length model.

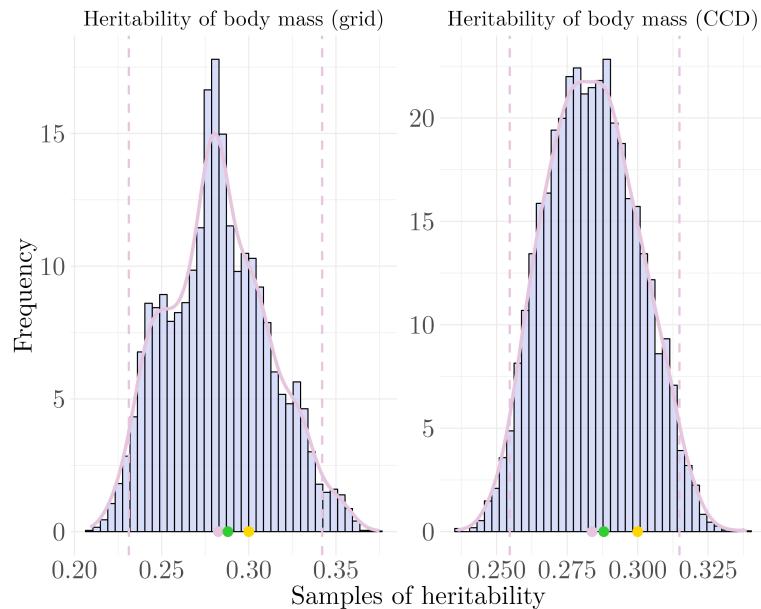


Figure C.1: Histogram depicting the estimated heritability values of body mass by the BVI method for the grid integration strategy(left) and CCD integration strategy (right) for the house sparrow dataset. The mean of the samples is marked as a pink circle at the bottom of the histogram, with the lower 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.

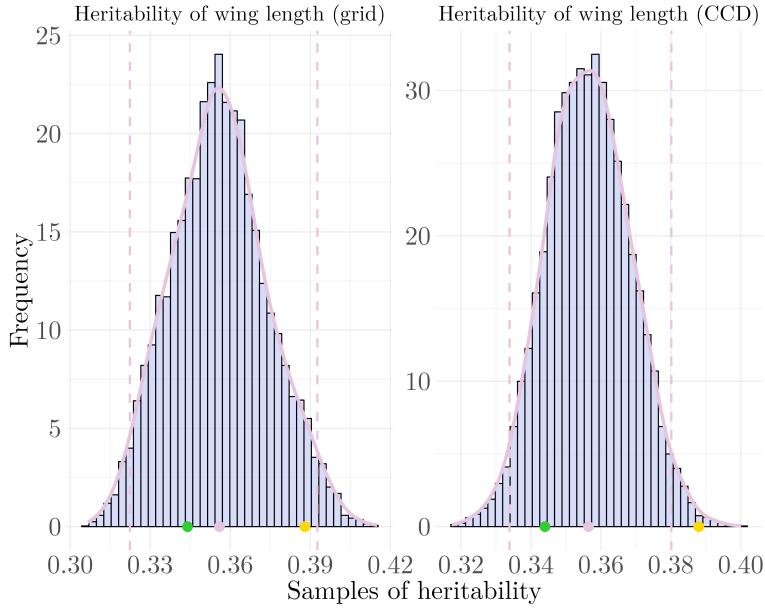


Figure C.2: Histogram of heritability values for wing length of the house sparrows estimated by the BVI method for the grid integration strategy (left) and the CCD integration strategy (right). The mean of the samples is marked as a pink circle at the bottom of the histogram, and the lower and upper 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.

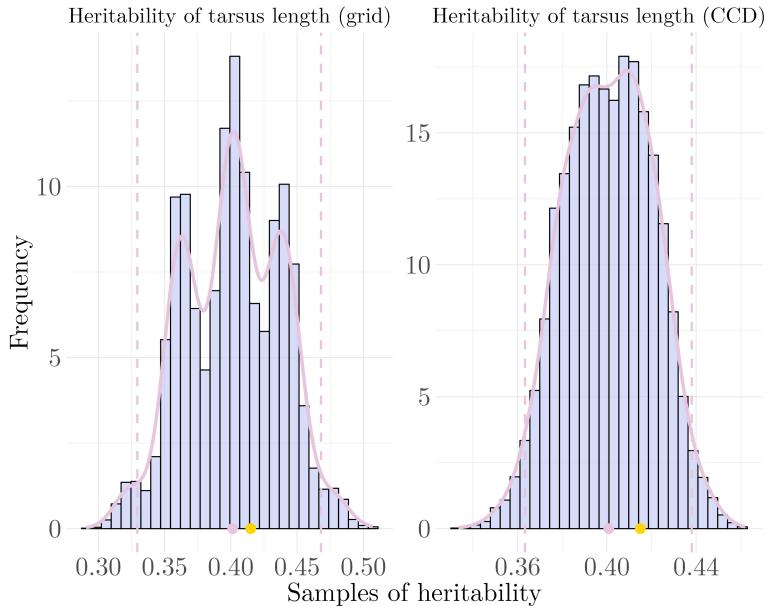


Figure C.3: Histogram showing estimated heritability values for tarsus length of the house sparrows from the BVI method for the grid integration strategy (left) and the CCD integration strategy (right). The two dots at the bottom represent the mean of the samples (pink) and the estimate from (Silva et al. 2017) (gold). The dashed lines represent the lower and upper value for the 95% percentile.

C.2 Supplementary tables for the non-Gaussian simulation study

We attach the summarizing tables for the Binomial and Poisson simulation studies here, as they were referred to throughout the thesis. Firstly, Table C.1 contains summary statistics of the distribution obtained from the Binomial model, while Table C.2 contains the same for the Poisson model.

Measure		$\rho = \mathbf{0}$	$\rho = \mathbf{0.1}$	$\rho = -\mathbf{0.1}$	$\rho = \mathbf{0.4}$	$\rho = -\mathbf{0.4}$
Relative Importance of Fixed effect X1	Average	0.098	0.118	0.077	0.173	0.020
	2.5%	0.086	0.106	0.067	0.161	0.017
	97.5%	0.108	0.130	0.088	0.185	0.023
Relative Importance of Fixed effect X2	Average	0.194	0.208	0.176	0.239	0.077
	2.5%	0.177	0.194	0.161	0.226	0.066
	97.5%	0.210	0.223	0.193	0.251	0.087
Relative Importance of Fixed effect X3	Average	0.292	0.298	0.280	0.299	0.166
	2.5%	0.272	0.280	0.258	0.285	0.150
	97.5%	0.310	0.317	0.300	0.315	0.182
Relative Importance of Random effect	Average	0.097	0.087	0.107	0.067	0.171
	2.5%	0.073	0.066	0.081	0.050	0.131
	97.5%	0.121	0.117	0.139	0.086	0.211
R_m^2	Average	0.583	0.624	0.532	0.710	0.262
	2.5%	0.558	0.600	0.507	0.690	0.241
	97.5%	0.607	0.646	0.557	0.730	0.284
R_c^2	Average	0.680	0.712	0.640	0.777	0.434
	2.5%	0.659	0.695	0.617	0.763	0.401
	97.5%	0.700	0.731	0.660	0.792	0.468

Table C.1: Summary of simulation study results for the quantiles of relative importance estimates of the Logit model across different correlation levels. For $\rho = 0$ the expected values are given in Table 3.3.

Measure		$\rho = \mathbf{0}$	$\rho = \mathbf{0.1}$	$\rho = -\mathbf{0.1}$	$\rho = \mathbf{0.4}$	$\rho = -\mathbf{0.4}$
Relative Importance of Fixed effect X1	Average	0.097	0.119	0.075	0.181	0.017
	2.5%	0.088	0.109	0.066	0.172	0.015
	97.5%	0.107	0.128	0.084	0.189	0.020
Relative Importance of Fixed effect X2	Average	0.193	0.212	0.171	0.251	0.066
	2.5%	0.181	0.200	0.158	0.241	0.057
	97.5%	0.205	0.224	0.184	0.260	0.075
Relative Importance of Fixed effect X3	Average	0.290	0.302	0.272	0.314	0.143
	2.5%	0.276	0.288	0.256	0.302	0.130
	97.5%	0.307	0.316	0.289	0.326	0.156
Relative Importance of Random effect	Average	0.098	0.090	0.106	0.072	0.149
	2.5%	0.074	0.068	0.081	0.054	0.111
	97.5%	0.127	0.112	0.136	0.093	0.190
R_m^2	Average	0.580	0.633	0.519	0.746	0.225
	2.5%	0.558	0.613	0.496	0.727	0.207
	97.5%	0.601	0.654	0.541	0.763	0.244
R_c^2	Average	0.679	0.723	0.625	0.817	0.374
	2.5%	0.658	0.704	0.602	0.804	0.337
	97.5%	0.700	0.741	0.648	0.830	0.413

Table C.2: Summary of simulation study results for the quantiles of relative importance estimates the Poisson model across different correlation levels. For $\rho = 0$ the expected values are given in Table 3.3.

APPENDIX
D

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 normalised probability density function $f(y|\theta, \phi)$ on the form

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

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

$$\begin{aligned} \mathbb{E}(Y|\theta) &= b'(\theta) \\ \text{Var}(Y|\theta) &= b''(\theta) \end{aligned} \quad (\text{D.2})$$

This can be proven by considering the following:

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

and

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

Now, assuming mild regularity to interchange derivation and integration and noting that $\int_{\mathbb{R}} f(y|\theta) dy = 1$, we have

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

and

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

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

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

and

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