August Arnstad

Relative variable importance in Bayesian linear mixed models

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

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

NTNU

ABSTRACT

SAMMENDRAG

PREFACE

CONTENTS

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

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

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

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

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

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

CHAPTER

TWO

THEORY

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

2.1 Linear regression

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

2.1.1 Linear regression[1]

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

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

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

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

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

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

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

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

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

2.1.2 Qualitative covariates

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

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

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

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

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

2.1.3 Correlation among covariates in linear regression

Correlation among covariates is to be expected, as it is natural in many scenarios. However, if the correlation is very strong, this poses some serious problems when interpreting the linear regression model. The covariates $\mathbf{x}_i$ in a linear regression are assumed to be linearly independent, so that the design matrix $\mathbf{X}$ has full rank. If the design matrix is not of full rank, that is one or more covariates are perfectly correlated, the model (2.2) is said to be *multicollinear* (Poole & O'Farrell 1971). From equation (2.4) one can see that if the matrix $\mathbf{X}$ is not of full rank, the term $(\mathbf{X}^T\mathbf{X})^{-1}$ is not invertible and the MLE of $\boldsymbol{\beta}$ does not exist. Further, the variance of the MLE of $\boldsymbol{\beta}$ grows as the correlation between covariates grows (Fahrmeir et al. 2013, p. 116). A larger variance in $\hat{\boldsymbol{\beta}}$ also leads to larger standard errors and larger p-values for $\hat{\boldsymbol{\beta}}$, making it hard to assess the model. Both coefficients and covariates affect the total marginal model variance, which can be decomposed as

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

(Grömping 2007) where $\mathbf{V} = \mathrm{Cov}(\mathbf{X})$ is the $p \times p$ covariance matrix of the covariates which is assumed to be positive definite, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, v_j the regressor variances for $j = 1, ..., p$ found along the diagonal of $\mathbf{V}$ and ρ_{jk} the inter-regressor correlations between regressor j and k. The middle term in 2.7 consist of the covariance between the covariates and the variance contribution from a single covariate is not immediately clear.

2.2 Variable importance in linear regression models

In a regression setting with multiple regression coefficients, it is often desirable to be able to assign each covariate with a measure of its relative importance to the model. The relative importance of covariate X_i is defined as the contribution to explained variance in the response $\mathbf{y}$ from X_i. Assigning relative importance is no trivial task, as correlation among covariates poses a challenge in assessing the relative importance of each covariate.

2.2.1 Relative importance measures

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

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

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

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

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

(Grömping 2007). In (2.7) however, the response variance is split into three parts, the first two sums which comes from the regressors and the latter term which is the variance of the error. As mentioned, it is the middle term that poses the problem of assigning importance to each covariate, since it is not immediately clear how to distribute the contribution to variance from the covariance terms to each covariate. The literature has established some conditions that relative importance measures should fulfill, so that they can be interpreted and compared in a sensible manner (Grömping 2007). As listed in Grömping (2007), the methods should have

1. **Proper decomposition**: The model variance should be decomposed into shares for each regressor that sum up to the total variance, and the method shall allocate the shares to each regressor.
2. **Non-negativity**: Each share of the variance should be non-negative.
3. **Exclusion**: If a regressor is excluded from the model, $\beta_j = 0$, its share of the variance should be zero.
4. **Inclusion**: If a regressor is included in the model, $\beta_j \neq 0$, its share of the variance should be positive.

2.2.2 Naive decompositions

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

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

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

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

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

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

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

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

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. Conversely to the first naive approach, this decomposition assigns importances based on the full information contained in the regressor. Therefore it overestimates the importance of each variable, since the shared information is accounted for twice.

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

2.2.3 The LMG method

A method that handles correlation among covariates, and is frequently reinvented(Grömping 2007) from different approaches, is the LMG method. Therefore we shall discuss it, as it serves an important role as a leading method for assigning relative variable importance. The LMG method takes use of averaging over orders, meaning that it permutes the index set $\{1, ..., p\}$ of the regressors $(p-1)!$ times, excluding the intercept, and sequentially adds the regressors to the model for each permuted index set. By adding regressors sequentially for each permutation, one can investigate how the importance of the regressors vary depending on what other regressors are included, which is useful when they are correlated. This is justified by the assumption that there is no relevant ordering of the regressors in the index set (Kruskal 1987). For each regressor added, starting with none, it allocates a share of explained variance, or importance, and then adds a new regressor. The final allocated share to the regressor is the average of the allocated shares to that regressor for all permutations of the set of regressors indices. This would mean that for two correlated regressors whose importance share varies depending on which is added first, would receive an averaged importance. Averaging over orders is a statistical tradition (Kruskal 1987) and gives a robust assessment of each regressor's importance by considering different orderings of how they are added to the model. The iterative process for the regressors $\{X_0, X_1, X_2, X_3\}$, where X_0 is the intercept, would be

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

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

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

where $n(S)$ is the number of regressors in S. Equation (2.13) averages the increase in R^2, svar$(\{X_i\})$, when adding the covariate of interest, X_i, over all possible orderings of covariates. This mean increase over orderings is assigned as the proportion of R^2 explained by X_i. The LMG method fulfills all but the exclusion criteria described previously (Grömping 2007), but Grömping (2007) argues that this "must be seen as a natural result of model uncertainty" and therefore that this criterion is not indispensable. Therefore, we find it also suitable for our purposes to focus on the three other criteria. The setback of the LMG method is the great computational expense that the permutations require when p is large. The complexity is 2^{p-1} summations (Grömping 2007), and therefore, the LMG is not suitable for high dimensional models.

2.2.4 Relative weights method

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

The relative weights method proposes an alternative to the LMG, which is significantly less computationally expensive. Intuitively, the relative weights method projects the design matrix $\mathbf{X}$ of the fixed effects into an orthogonal column space, resulting in a matrix $\mathbf{Z}$ with orthogonal columns. The matrix $\mathbf{Z}$ is then an approximation of $\mathbf{X}$ and will be used as the design matrix in the regression. Since the columns of the design matrix $\mathbf{Z}$ are orthogonal, each covariate is uncorrelated. This allows us to decompose the variance in the straightforward manner as in equation (2.9).

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

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

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

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

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

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

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

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

and since $\mathbf{Z}$ is orthogonal, the contribution from a column of $\mathbf{z_i}$ with respect to a column $\mathbf{x}_j$ is the squared entry $\boldsymbol{\Lambda}^2_{ij}$. The contribution from a column $\mathbf{x}_j$ with respect to the response $\mathbf{y}$, *i.e.* the relative importance, is then estimated as the matrix product (Johnson 2000)

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

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

2.3 Regression models

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 include more complex data structures.

2.3.1 Generalized linear models (GLMs)

The first step in expanding the linear regression model, is to allow the responses to be non-Gaussian. Instead of considering only the normal distribution as the distribution of the response, one can consider general responses belonging to the exponential family. Assume that each we have N observations of the response y_i, where $i = 1, ..., N$, that are conditionally independent given the fixed effects. Then, y_i belongs to the univariate exponential family if

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

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

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

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

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

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

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

Some examples of link functions are the identity function for the linear regression, the logit and probit functions for the Bernoulli distribution and the log function for the Poisson distribution.

2.3.2 Linear mixed models (LMMs) [2]

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

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m)^T \ , \tag{2.23}$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{2.24}$$

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.

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

Since $\boldsymbol{\alpha}$ is a random variable, the parameter to estimate is the variance of each random effect $\boldsymbol{\Sigma}_{kk} = \sigma_k^2$ and their covariance $\boldsymbol{\Sigma}_{k,l} = \sigma_{k,l}$, where $k, l = 1, ..., q$. In practice it is often easier to estimate the precision rather than the variance, so calculations often involve the precision matrix $\mathbf{Q}$ rather than the covariance matrix $\boldsymbol{\Sigma}$. In this model the independence between clusters are conserved for the response as a whole, but it expresses the correlation that observations of the same cluster have through the random effects. As for the simple linear regression it is assumed that $\mathbf{X}\boldsymbol{\beta}$ is fixed, and that $\mathbf{U}$ is given, so they do not contribute to the model's variance. Therefore, the conditional expectation $\mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \mathbf{X}\boldsymbol{\beta}$ is easily obtained, and the conditional variance can be calculated as

$$\mathrm{Var}(\mathbf{y}|\mathbf{X}, \mathbf{U}) = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) = \mathbf{U}\mathrm{Var}(\boldsymbol{\alpha})\mathbf{U}^T + \sigma^2\mathbf{I} = \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I} \ , \quad (2.25)$$

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

$$\mathbf{y}|\mathbf{X}, \mathbf{U} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{U}\mathbf{G}\mathbf{U}^T + \sigma^2\mathbf{I}) \ . \quad (2.26)$$

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

$$\theta_{i,j} = \eta_{i,j} = \mathbf{x}_{i,j}^T\boldsymbol{\beta} + \mathbf{u}_{i,j}^T\boldsymbol{\alpha}_j \ , \quad (2.27)$$

where $j = 1, ..., m$ denotes the cluster and $i = 1, ..., n_j$ denotes the observations in cluster j, $\mathbf{x}_{i,j}$ and $\mathbf{u}_{i,j}$ are the i-th columns of the submatrices $\mathbf{X}_j$ and $\mathbf{U}_j$ of the larger design matrices $\mathbf{X}$ and $\mathbf{U}$ respectively, for cluster j. The assumption of conditional independent observations $y_{i,j}$ is now conditional on the random effect as well as the covariates, and the conditional distribution of $y_{i,j}$ is still assumed to belong to the exponential family. When estimating parameters in a GLMM, the joint distribution of the observations is not generally tractable. Therefore, parameter estimation requires more sophisticated methods in both the likelihood and Bayesian framework.

2.4 Extending R^2 to GLMMs

As we generalized the linear regression to LMMs, GLMs and GLMMs, it is desirable to also generalize the concept of the R^2 to be applicable to these models. This is fundamental to be able to propose a method for decomposing the R^2 and thereby assigning relative importance to covariates. However, the task of determining the R^2, and decomposing it, is not a trivial task in the linear regression case and becomes even more complex in the case of GLMMs. Many extensions have been proposed, but due to a variety of theoretical problems and/or computational difficulties, no consensus has been reached on a framework for calculating the R^2

for GLMMs (Nakagawa & Schielzeth 2013). To get an overview of the status quo for R^2, we will follow the paper by Nakagawa & Schielzeth (2013) and go through the different components added to the linear regression to compose the GLMMs.

2.4.1 R^2 for GLMs

Recalling the definition of the R^2 from Equation (2.8), we now generalize this to the GLMs. To illustrate the generalization, consider the deviance $\mathcal{D}(\mathbf{y}|\theta)$ function which is defined as twice the difference between the log likelihood of the **saturated model** and the log-likelihood of the model of interest (McCullagh & Nelder 1989). The saturated model denotes the model of the maximum achievable log likelihood, and therefore fits the data perfectly. For a linear regression, with $\theta = (\boldsymbol{\beta}, \sigma^2)$, we would therefore obtain

$$\begin{aligned}\mathcal{D}(\mathbf{y}|\hat{\theta}) &= -2\left(\ln(\mathcal{L}(\boldsymbol{\beta}, \sigma^2|\mathbf{y})) - \ln(\hat{\mathcal{L}}(\hat{\boldsymbol{\beta}}, \widehat{\sigma^2}|\mathbf{y}))\right) = -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \widehat{\sigma^2}|\mathbf{y})\right)\\ &= -2\left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{n}{2}\ln(2\pi\sigma^2)\right)\\ &= \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\\ &= 1 - R^2\ ,\end{aligned} \tag{2.28}$$

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

$$\begin{aligned}\mathcal{D}(\mathbf{y}|\hat{\theta}) - \mathcal{D}(\mathbf{y}|\theta_0) &= -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \widehat{\sigma^2}|\mathbf{y})\right) + 2\left(l(\boldsymbol{\beta}_0, \sigma_0^2|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}, \widehat{\sigma^2})\right)\\ &= -2\left(l(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) - l(\boldsymbol{\beta}_0, \sigma_0^2|\mathbf{y})\right)\\ &= -\frac{2}{2\sigma^2}\left(-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + (\mathbf{y}-\overline{\mathbf{y}})^T(\mathbf{y}-\overline{\mathbf{y}})\right)\\ &= 1 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}\\ &= R^2\ .\end{aligned} \tag{2.29}$$

This is the basis 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 R^2 measures are proposed as

$$R_G^2 = \left[1 - \left(\frac{\mathcal{L}_0}{\mathcal{L}_M}\right)^{2/n}\right]\frac{1}{1-(\mathcal{L}_0)^{2/n}} \tag{2.30}$$

and

$$R_D^2 = 1 - \frac{-2\ln(\mathcal{L}_M)}{-2\ln(\mathcal{L}_0)} \tag{2.31}$$

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. A problem with 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 an extension to the 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^2_ε while at the same time increasing σ^2_α, 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^2_1 and R^2_2. Considering the simple random intercept model in scalar form;

$$y_{i,j} = \beta_0 + \mathbf{x}^T_{i,j}\boldsymbol{\beta} + \alpha_j + \varepsilon_{i,j} \ , \tag{2.32}$$

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

$$\begin{aligned} R^2_1 &= 1 - \frac{\mathrm{Var}(y_{i,j} - \hat{y}_{i,j})}{\mathrm{Var}(y_{i,j})} = 1 - \frac{\sigma^2_\varepsilon + \sigma^2_\alpha}{\sigma^2_{\varepsilon 0} + \sigma^2_{\alpha 0}} \\ \hat{y_{i,j}} &= \beta_0 + \mathbf{x}^T_{i,j}\beta \ , \end{aligned} \tag{2.33}$$

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

$$\begin{aligned} R^2_2 &= 1 - \frac{\mathrm{Var}(y_j - \hat{\overline{y}}_j)}{\mathrm{Var}(\overline{y_j})} = 1 - \frac{\sigma^2_\varepsilon + \sigma^2_\alpha/k}{\sigma^2_{\varepsilon 0} + \sigma^2_{\alpha 0}/k} \\ k &= \frac{M}{\sum_{j=1}^{M} \frac{1}{m_j}} \ , \end{aligned} \tag{2.34}$$

where $\overline{y_j}$ is the mean for each observed value of the jth cluster, $\hat{\overline{y}}_j$ is the mean of the fitted values for the jth cluster, k is the harmonic mean of the number of observations per cluster, m_j is the number of observations for the jth cluster and

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \ , \tag{2.35}$$

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

$$\sigma_f^2 = \text{Var}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}^T \text{Var}(\mathbf{X})\boldsymbol{\beta} \ , \tag{2.36}$$

and further define the R^2 for the LMM as

$$R^2_{\text{LMM(m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{i=1}^q \sigma_i^2 + \sigma_\varepsilon^2} \ . \tag{2.37}$$

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

$$R^2_{\text{LMM(c)}} = \frac{\sigma_f^2 + \sum_{i=1}^q \sigma_i^2}{\sigma_f^2 + \sum_{i=1}^q \sigma_i^2 + \sigma_\varepsilon^2} \ . \tag{2.38}$$

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

$$y_{i,j} = \beta_0 + \mathbf{x}_{i,j}^T\boldsymbol{\beta} + \alpha_{0,j} + \alpha_{1,j}x_{i,j} + \varepsilon_{i,j} \ , \tag{2.39}$$

where the same notation is used as in (2.32) with $\boldsymbol{\alpha_j} = (\alpha_{0,j}, \alpha_{1,j})$ being the random effect, $\alpha_{0,j}$ denoting the random intercept and $\alpha_{1,j}$ now denoting the random deviation from the global slope β_1, for cluster j. The general assumption on the random effects are that

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} , \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\alpha_0} & \sigma_{\alpha_0,\alpha_1} \\ \sigma_{\alpha_0,\alpha_1} & \sigma^2_{\alpha_1} \end{pmatrix} \right) \ , \tag{2.40}$$

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

$$\begin{aligned}\mathrm{Var}(\alpha_{0,j}+\alpha_{1,j}x_{i,j}) &= \mathrm{Var}(\alpha_{0,j}) + 2x_{i,j}\mathrm{Cov}(\alpha_{0,j},\alpha_{1,j}) + x^2_{i,j}\mathrm{Var}(\alpha_{1,j}) \\ &= \sigma^2_{\alpha_0} + 2x_{i,j}\sigma_{\alpha_0,\alpha_1} + x^2_{i,j}\sigma^2_{\alpha_1} =: \sigma^2_{r,i,j} \ ,\end{aligned} \tag{2.41}$$

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

$$\overline{\sigma^2_r} = \frac{1}{N}\sum_{j=1}\sum_{i=1}\left(\sigma^2_{r,i,j}\right) \ . \tag{2.42}$$

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

$$\overline{\sigma^2_r} = \mathrm{Tr}(\mathbf{U_j}\mathbf{Q}\mathbf{U_j}^T), \quad r = 1, ..., q \ . \tag{2.43}$$

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

$$R^2_{\mathrm{LMM(m)}} = \frac{\sigma^2_f}{\sigma^2_f + \sum_{r=1}^{q}\overline{\sigma^2_r} + \sigma^2_\varepsilon} \ , \tag{2.44}$$

and

$$R^2_{\mathrm{LMM(c)}} = \frac{\sigma^2_f + \sum_{r=1}^{q}\overline{\sigma^2_r}}{\sigma^2_f + \sum_{i=1}^{q}\overline{\sigma^2_r} + \sigma^2_\varepsilon} \ , \tag{2.45}$$

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

$$\overline{\sigma^2_r} = \frac{1}{N}\sum_{j=1}\sum_{i=1}\left(\sigma^2_{r,i,j}\right) = \sigma^2_{r,i,j} = \sigma^2_r \ . \tag{2.46}$$

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 is inherited from the link function used, and is therefore known before analysis is done. However, the multiplicative and additive dispersion is modelled to account for the variance present that exceeds the distribution specific variance, *i.e.* overdispersion (Nakagawa & Schielzeth 2010). Therefore, one must specify upon implementation on what scale the overdispersion is to be modelled. The multiplicative dispersion, denoted by ω, is overdispersion on the response (data) scale and modelled as a distinct parameter of the assumed distribution of the response $\mathbf{y}$ (Nakagawa & Schielzeth 2010). Conversely, the additive dispersion, denoted by e, is overdispersion on the latent scale and introduced to the model as an additional random effect in the linear predictor (Nakagawa & Schielzeth 2010). Defining the residual variance now depends on the choice of dispersion modelling, and is either defined as

$$\sigma_\varepsilon^2 = \omega\sigma_d^2 \tag{2.47}$$

or

$$\sigma_\varepsilon^2 = \sigma_d^2 + \sigma_e^2 \ , \tag{2.48}$$

for multiplicative and additive dispersion respectively. With the residual variance defined, the generalization to of the R^2 to GLMMs (thereby also the GLMs) follows the same logic as the LMMs, and $R^2_{\text{GLMM (m)}}$ is defined as

$$R^2_{\text{GLMM(m, m)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \omega\sigma_d^2} \ , \tag{2.49}$$

and

$$R^2_{\text{GLMM(m, a)}} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \sigma_\varepsilon^2} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \sigma_d^2 + \sigma_e^2} \ , \tag{2.50}$$

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

$$R^2_{\text{GLMM(c, m)}} = \frac{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2}}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \omega\sigma_d^2} \ , \tag{2.51}$$

and

$$R^2_{\text{GLMM(c, a)}} = \frac{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2}}{\sigma_f^2 + \sum_{r=1}^{q}\overline{\sigma_r^2} + \sigma_d^2 + \sigma_e^2} \ , \tag{2.52}$$

completing the generalization.

2.5 The Bayesian framework

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

2.5.1 General idea

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

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \ . \tag{2.53}$$

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

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})} \ , \tag{2.54}$$

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 $\pi(\mathbf{y}) = \int \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ is the marginal distribution of the data (Gelman et al. 2015). In practice, the marginal distribution is often omitted and one only consider the proportionality of (2.54), i.e.

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ . \tag{2.55}$$

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

2.5.2 Prior and posterior distributions

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

2.5.3 Penalising complexity (PC) priors

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

1. **Occams razor** - If there is no evidence for a complex mode, a base model should be preferred.
2. **Measure of complexity** - The measure of model complexity is deifned as $d(f||g) = \sqrt{2\text{KLD}(f||g)}$ where $\text{KLD}(f||g)$ denotes the Kullback-Leibler divergence (Simpson et al. 2017, for more information).
3. **Constant rate penalisation** - The penalisation, i.e. the decay of prior mass, grows as the complexity grows, but it is desirable that this growth is constant.
4. **User defined scaling** - Assuming that the user has an idea of the magnitude of the parameter of interest, the user should be able to scale the prior accordingly.

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

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right), \quad \tau, \lambda > 0 \ . \tag{2.56}$$

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

2.5.4 Hierarchical Bayesian modelling

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

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}, \phi)\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\phi})\pi(\boldsymbol{\phi}) \ , \tag{2.57}$$

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

$$y_i|\theta_i \sim \text{Po}(\theta_i) \ , i = 1, ..., n \ , \tag{2.58}$$

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

$$\theta_i|\phi \sim \text{Gamma}(\alpha, \beta) \ , \tag{2.59}$$

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

$$\alpha \sim \text{Exp}(a) \text{ and } \beta \sim \text{Gamma}(b, c) \ , \tag{2.60}$$

The full posterior density now reads

$$\pi(\boldsymbol{\theta}, \alpha, \beta|\mathbf{y}) \propto \underbrace{\prod_{i=1}^{n} \theta_i^{y_i} e^{-\theta_i}}_{\text{Po}(\theta_i)} \underbrace{\prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\beta)} \theta_i^{\alpha-1} e^{-\beta\theta_i}}_{\text{Gamma}(\alpha,\beta)} \underbrace{\alpha^{a-1} e^{-\alpha}}_{\text{Exp}(a)} \underbrace{\beta^{b-1} e^{-c\beta}}_{\beta\sim\text{Gamma}(b,c)} \ , \tag{2.61}$$

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 the model in (2.19) with dispersion parameter ϕ and linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} \ , \tag{2.62}$$

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

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho}|\mathbf{y}) &\propto \left(\prod_{j=1}^{m} \pi(\mathbf{y}_j|\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \boldsymbol{\rho})\right) \pi(\boldsymbol{\alpha}|\boldsymbol{\rho})\pi(\boldsymbol{\beta})\pi(\phi)\pi(\boldsymbol{\rho}) \ , \\ &\propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}(\boldsymbol{\rho})\boldsymbol{\alpha} + \sum_{j=1}^{m} \ln \pi(\mathbf{y}_j|\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi)\right) |\mathbf{Q}(\boldsymbol{\rho})|^{1/2} \pi(\boldsymbol{\beta})\pi(\phi)\pi(\boldsymbol{\rho}) \ ,\end{aligned} \tag{2.63}$$

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

2.5.5 R^2 in the Bayesian framework[3]

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

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

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

$$R_s^2 = \frac{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\mathbf{T}\mathbf{X}} \boldsymbol{\beta}_s}{\boldsymbol{\beta}_s^T \boldsymbol{\Sigma}_{\mathbf{X}^\mathbf{T}\mathbf{X}} \boldsymbol{\beta}_s + \sigma_s^2} \ , \tag{2.64}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}^\mathbf{T}\mathbf{X}}$ 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 a posterior distribution. Drawing enough samples from (2.64) one would eventually obtain also a distribution for the R^2 value.

2.6 The INLA framework[4]

As we have seen, the analytical posterior is possible to obtain for some hierarchical structures. However, in the case of GLMMs, the posterior distribution is not in general analytically tractable (Fong et al. 2010). This calls for the use of numerical methods, such as Markov Chain Monte Carlo (MCMC) methods, to be able to sample from the posterior distribution. Such methods are computationally expensive, and require careful analysis to justify convergence and mixing of the Markov chains to the posterior distribution. Therefore it is desirable, under certain conditions, to look at other methods that are more computationally efficient. In this thesis we will consider the Integrated Nested Laplace Approximation (INLA) method (Gómez-Rubio 2020).

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

$$\eta_i = \alpha + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}(u_{ki}) + \varepsilon_i \ , \qquad i = 1, ..., n \ , \tag{2.65}$$

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

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

term. This gives rise to the key assumption that the INLA method needs in order to be applicable, namely that the latent field $\mathbf{x}$, denoted as

$$\mathbf{x} = (\eta_1, ..., \eta_n, \alpha, \beta_1, ..., \beta_n) \ , \tag{2.66}$$

is a Gaussian Markov Random Field (GMRF). Further, it is assumed that observations are independent given this latent field and the latent field is distributed according to some hyperparameters $\boldsymbol{\theta}$. The structure of the GMRF is given by a precision matrix $\mathbf{Q}(\theta)$, which is sparse and can be represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (see Section 2.8 for more details). This along with the assumed conditional independence makes computations very fast and is why INLA is effective. Now, the posterior distribution of the latent field $\boldsymbol{x}$ is given by

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ , \tag{2.67}$$

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

$$\pi(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}) \ , \tag{2.68}$$

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

$$\pi(x_l|\mathbf{y}) = \int \pi(x_l|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ , \tag{2.69}$$

and

$$\pi(\theta_k|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-k} \ , \tag{2.70}$$

(Gómez-Rubio 2020) respectively, $\boldsymbol{\theta}_{-k}$ is the vector of hyperparameters excluding element θ_k and the latter integral is possible to integrate numerically due to the low dimension of $\boldsymbol{\theta}$ (Rue et al. 2009). The approximations of these integrals are omitted, see Rue et al. (2009) for the full details. Lastly, the joint posterior distribution can be approximated from the so-called Skew Gaussian Copula class, as specified in Chiuchiolo et al. (2021), and allows for sampling from the joint distribution. The INLA method is implemented in the R-package `R-INLA` (Gómez-Rubio 2020) and is used in this thesis to fit the models and draw from the obtained posteriors. We note that for the random effects INLA outputs the precision for the parameters involved, which is defined as the inverse covariance matrix. For the posterior marginal distribution of variance for the random effects the package has a function for transforming the precision marginal to the variance marginal. The priors used for the models in this thesis follow the recommendations of penalizing priors by Simpson et al. (2017). EXPAND ON INLA? SINCE SOME OF THE FEEDBACK STATED THAT IT WAS A BIT UNCLEAR.

2.7 The Animal Model and quantitative genetics

An important application of GLMMs, which we will later analyse, is in the context of evolutionary biology and quantitative genetics. To introduce the animal model and biological terminology, the section will rely heavily on the work of Kruuk (2004) and Conner & Hartl (2004). The animal model is a mathematical model, used as a tool for quantitative genetic analysis in evolutionary biology where the aim is to explain the phenotypic variation in a population. A phenotype is defined as *the outward appearance of an organism for a given characteristic* (Conner & Hartl 2004), such as eye color, height or behavior. In an organism, the observed phenotypic trait is a result of the complex interaction between environment 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 (Kruuk 2004), a decomposition of the phenotypic variance is of great interest. The simplest partition is to define the phenotypic variance as the sum of the genetic variance and environmental variance (Conner & Hartl 2004). However, for species that mate with other individuals in the population rather than self-fertilize, it is common to further decompose the genetic variance into three parts. The **total phenotypic variance** can therefore be partitioned as

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2 \ , \tag{2.71}$$

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

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\alpha} \ , \tag{2.72}$$

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. As in Section 2.3.2, we let $\mathbf{G}$ denote the covariance matrix of the random effect $\boldsymbol{\alpha}_A$, which in the animal model can be derived from the expected covariance between relatives (Kruuk 2004). This derivation can be done by considering the coefficient of coancestry, $\Theta_{i,j}$, defined as *the probability that an allele drawn at random from an individual i will be identical by descent to an allele drawn at random from individual j* (Kruuk 2004). We use the coefficient of coancestry to define the expected covariance between relatives, or additive relationship matrix, as $\mathbf{A}_{i,j} = 2\Theta_{i,j}$ and consequently $\mathbf{G} = \sigma_A^2\mathbf{A}$ (Kruuk 2004). The parameters of interest are now the additive genetic variance σ_A^2 and the **heritability**, defined as the proportion of the total phenotypic variance that is present due to the additive genetic variance, σ_A^2/σ_P^2 (Wilson 2008*a*).

2.8 The Animal Model as a GMRF

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. If a Gaussian random variable $\mathbf{X} = (X_1, ..., X_n)$ possesses the Markov property it means that for some $i \neq j$, X_i is independent of X_j conditioned $X_{-i,j}$, where $X_{-i,j}$ denotes all other elements of $\mathbf{X}$ except X_i and X_j (Rue et al. 2009). This property readily visualized in a conditional independence graph, and for the animal model the pedigree structure 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. FIGUR AV DETTE? The pedigree can also be used to construct the relatedness matrix $\mathbf{A}$, previously defined as the expected covariance between relatives, and the gives rise to the sparse precision matrix $\mathbf{Q} := \mathbf{A}^{-1}$ which is needed for calculations. As we consider each node as an individual, the corresponding variable of that node is its breeding value $\boldsymbol{\alpha}$ (Steinsland & Jensen 2010).

2.8.1 Single and Multitrait Animal Model

When modelling the observed phenotypic trait values of individuals in a population, caution must be made when modelling the genetic component of the trait using INLA. If only one trait is considered, The breeding values $\boldsymbol{\alpha}$ of an individual represents the genetic component of an observed phenotypic trait. FINSIH THIS PART

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 PLACE NAME HERE. The method applies to GLMMs modelled with Binomial, Poisson and Gaussian responses, and assumes the distinct random effects to be independent. Random slopes are not included in the method.

If categorical covariates with more than two levels are contained in the fixed effects, they should be encoded using distinct names in order to make sure the code runs. Place this in the documentation of the code!!!

3.1 Variable importance in the Bayesian framework

There are a few considerations necessary in order to calculate variable importance on GLMMs in a Bayesian framework. First of all, the characteristics of the Bayesian framework must be considered. When fitting a GLMM in the frequentist framework, point estimates of the coefficients of the fixed effects as well as point estimates of the variance from 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 will be used 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 with additive overdispersion in Nakagawa & Schielzeth (2013). This definition makes GLMMs analogous to LMMs, thus supporting a unified approach to both types of models. Therefore, we consider only GLMMs modeled with additive overdispersion, although we believe the method could be extended to handle multiplicative overdispersion as well. These considerations are the basis for our proposed model for calculating relative variable importance in Bayesian GLMMs. The presented method does not include categorical variables with more than two categories or random slopes. The exclusion of categorical variables are seen as outside the scope of this thesis, whereas the random slopes are omitted 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 model are handled in our method.

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

The basis of our 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

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

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

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

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

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

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

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

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

3.3 Decomposing the R^2 value

We now seek to decomposed the proposed R^2 value and assign each covariate with a proportion of the variance explained, i.e. assign each covariate with a *relative*

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

Distribution	**Link Function**	σ_d^2
Binomial	Logit	$\pi^2/3$
Binomial	Probit	1
Poisson	Log	$\ln\left(1 + 1/\exp\left(\beta_0 + 0.5(\sum_{k=1}^{q} \sigma_{\alpha_k}^2 + \sigma_e^2)\right)\right)$
Poisson	Square Root	0.25

Table 1: Distribution-specific variance σ_d^2 for the Binomial and Poisson distributions for the link functions implemented in this thesis. The variances correspond to the values in Nakagawa & Schielzeth (2013) and the calculation for the log-link Poisson follow the recommendations of Nakagawa et al. (2017).

variable importance. Recall that the fixed and random effects are assumed to be independent, so that one can consider the variances of the fixed and random effects separately. Further, the residual variance is also considered as independent of both fixed and random effects.

3.3.1 Applying the relative weights method in the Bayesian framework

To remedy the problems of calculating importance of correlated covariates, we will apply the relative weights method to the fixed effects before fitting the model. Following Section 2.2.4, we project the design matrix $\mathbf{X}$ of the fixed effects to obtain the matrix $\mathbf{Z}$. The model is fit using $\mathbf{Z}$ as an approximated design matrix of fixed effects, and from the joint posterior distribution samples of the coefficients $\boldsymbol{\beta}_{\mathbf{Z}}$ can be drawn. Each sample $\boldsymbol{\beta}_{\mathbf{Z},s}$, $s = 1, ..., S$ can be used to approximate a sample of the importance of the columns $\mathbf{X}$. Using equations (2.17) and (2.18), we calculate this sample as

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

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

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

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

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

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

3.3.2 Random effects

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

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

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

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

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

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

and obtain an approximate posterior distribution of the ICC.

3.4 Heritability of phenotypic traits

A particularly interesting application of variable importance, and an area of much active research, is estimating the heritability of phenotypic traits. As mentioned in Section 2.7, heritability is defined as the ratio of additive genetic variance to total phenotypic variance (Wilson 2008*b*). When modeling a phenotypic trait as the response, the variable importance of the random effect accounting for additive genetic variance can be interpreted as the heritability of the phenotypic trait. Therefore, this is a useful application of our variable importance method, and has been the motivation for the development of the method.

3.4.1 House sparrow study

We now apply our 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 traits using a Gaussian LMM, namely the body mass, wing length and tarsus length. The fixed effects in the model consist of observations of *sex*, a standardized inbreeding coefficient denoted *FGRM*, the standardized *month* of the year (measurements were made during May-August), the *age* of each bird, and dummy variables encoding the location of the *native island* group of the bird (three levels, outer islands, inner islands or other islands). In addition, we model the *hatchyear* as an independent and identically distributed (i.i.d.) random intercept. To account for the correlation between relatives, we include a random effect for the pedigree structure, specifying the relatedness matrix as the correlation structure. Lastly, to account for individual differences we add an i.i.d. random intercept for the individual bird. We prefer to use the INLA framework, described in Section 2.6, to fit our LMM as it is computationally efficient and easy to use.

3.5 Non-Gaussian case studies

3.5.1 Binomial and Poisson case studies

To investigate how well our method generalizes to non-Gaussian responses, we perform a case study using the setup described in the vignette of the R-package `rptR` (Stoffel et al. 2017). The dataset, introduced for a different purpose, is simulated to replicate a study on twelve different beetle larvae populations (Stoffel et al. 2017). It contains the covariates *population*, the discrete *habitat* of the larvae, the dietary *treatment* of the larvae, the *sex* and *container* of which the larvae was contained in. The phenotypes to be modeled by the Binomial and Poisson distributions are the two distinct male colour morph and the number of eggs laid by female larvae respectively. Both models use treatment as the only fixed effect and place i.i.d. random intercepts on the population and container covariates. As before, our modelling is carried out using INLA, whereas the models in the vignette are calculated from functions in the `rptR` package. For more details on the model formulation used by our method, see Appendix A and Appendix B.

3.5.2 Binomial and Poisson simulation studies

As the case studies previously described all consider the heritability or repeatability as the quantity of interest, we simulate data to be able to evaluate the performance of our method when calculating relative variable importance for all parameters. We simulate $N = 10000$ responses from a Poisson distribution with log-link, a Binomial distribution with a logit link and a Binomial distribution with a probit link. The linear predictor contains three fixed effects and one random intercept. The fixed effects $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0, 0, 0)^T$, $\Sigma_{i,i} = 1$ and $\Sigma_{i,k} = \rho, k \neq i$. The true regression coefficient are set to be $\boldsymbol{\beta} = (1, \sqrt{2}, \sqrt{3})^T$. The random effect comes from $m = 100$ clusters, each with $n_j = 100$ observations for $j = 1, ..., m$. Further, we draw the random effect from a normal distribution with mean zero and variance $\sigma^2_{\alpha_1} = 1$. All data is standardized before fitting the models and we fit four different models letting ρ vary for each model by taking on the values $0, 0.1, 0.5$ and 0.9. The INLA framework is used to fit the GLMMs and the methodology described used to calculate the relative importance.

For the binomial simulation with distributional variance σ^2_d independent of the fitted model, we can empirically calculate the relative importance of the parameters when they are not correlated. When uncorrelated, the proportion of variance explained by each covariate in the linear predictor is equal to the square of the true coefficient. By defining $\sigma^2_{x_k}$ as the variance contribution to the linear predictor for covariate k, we then have

$$\sigma^2_{x_1} = \sigma^2_{\alpha_1} = 1 \quad \text{and} \quad \sigma^2_{x_2} = 2 \quad \text{and} \quad \sigma^2_{x_3} = 3 \ . \tag{3.11}$$

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

$$\begin{aligned} \text{RI}(\mathbf{X})_1 = \text{RI}(\alpha_1) &= \frac{\sigma^2_{x_1}}{\sum_{i=1}^{3} \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d}, \\ \text{RI}(\mathbf{X}_2) &= \frac{\sigma^2_{x_3}}{\sum_{i=1}^{3} \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d}, \\ \text{RI}(\mathbf{X}_3) &= \frac{\sigma^2_{x_3}}{\sum_{i=1}^{3} \sigma^2_{x_i} + \sigma^2_{\alpha_1} + \sigma^2_d} \ . \end{aligned} \tag{3.12}$$

This means that the expected relative importance in the binomial model with probit link of X_1 and α_1 is 1/8 and the expected relative importance of X_2 and X_3 is 2/8 and 3/8 respectively. For the logit link we have an expected relative importance of 0.0972 for X_1 and α_1, 0.194 for X_2 and 0.292 for X_3. For the Poisson simulation however, the distributional variance is approximated by $\sigma^2_d = \ln(1 + 1/\mathbb{E}[\lambda])$ with $\mathbb{E}[\lambda] = \exp\big(\beta_0 + 0.5(\sum_{k=1}^{q} \sigma^2_{\alpha_k} + \sigma^2_e)\big)$, and is therefore dependent on the fitted model (Nakagawa et al. 2017).

CHAPTER

FOUR

RESULTS

4.1 Heritability of house sparrow traits

We now investigate the results of applying our method to the house sparrow dataset. As mentioned, estimating the heritability of phenotypic traits can be seen as a special case of relative variable importance. The findings we present here are direct calculations obtained by our method. We present the samples of relative variable importance obtained of the variance component that captures additive genetic variance, and use the results from Silva et al. (2017) and Muff et al. (2019) to compare with. In Table 1 the mean of sampled heritability along with confidence intervals is presented, as well as the corresponding measures from the comparable studies.

	Silva et al. (2017)		Muff et al. (2019)		BVI	
	Estimate	CI	Estimate	CI	Mean	CI
h^2_{mass}	0.300	[0.2314, 0.3686]	0.2875	[0.2188, 0.3707]	0.2841	[0.2324 0.3442]
h^2_{wing}	0.388	[0.3525, 0.4605]	0.3438	[0.2939, 0.4085]	0.3543	[0.3208 0.3906]
h^2_{tarsus}	0.415	[0.3327, 0.4973]	-	-	0.4016	[0.3305 0.4690]

Table 1: Heritability estimates and confidence interval from Silva et al. (2017), posterior means of additive genetic variance divided by the posterior means of total phenotypic variance in Muff et al. (2019) with corresponding confidence interval and the mean and confidence interval of the heritability samples obtained from the BVI method for the phenotypic traits; body mass, wing length and tarsus length.

For the sampled heritability of body mass (Figure 1), we have a mean of 0.2841 and a distribution which seems slightly skewed to the left. The mean and mode seem to be approximately the same.

The samples of wing length heritability form a smooth bell curve with a mean of 0.3543 (Figure 2). Also here the mean and mode seem to be roughly the same, but the distribution is more symmetric than for body mass.

The heritability samples of tarsus length (Figure 3) has a mean of 0.4016 and a distribution that has three distinct peaks. The center peak is the highest, and the peaks on each side seem to be of equal height. We suspect that this pattern is a result of how INLA performs the sampling of the LMM. The mean and mode are approximately the same.

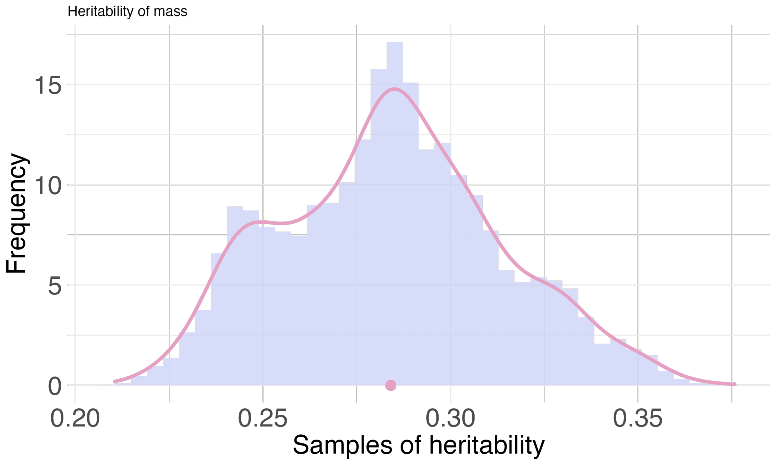

Figure 1: Heritability of body mass

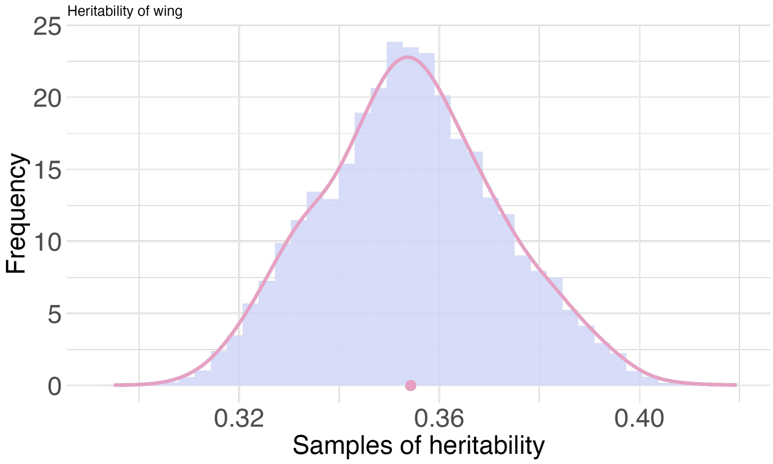

Figure 2: Heritability of wing length

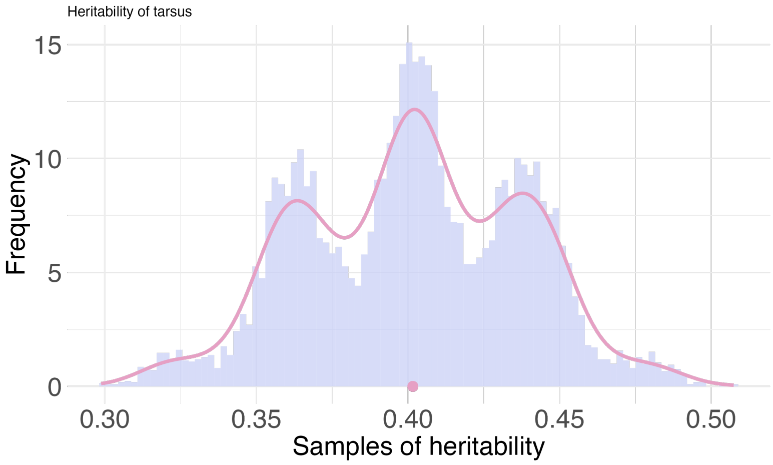

Figure 3: Heritability of tarsus length

4.2 Comparison with `rptR` package

To further assess our method, a comparison to the vignette for the **rptR** was made. No expected results were available, and so we can only compare our method to the results made by the authors of the vignette. It should however be noted, that the **rptR** package only returns the marginal R^2, whereas our method automatically decomposes this value and assigns a share to each fixed effect.

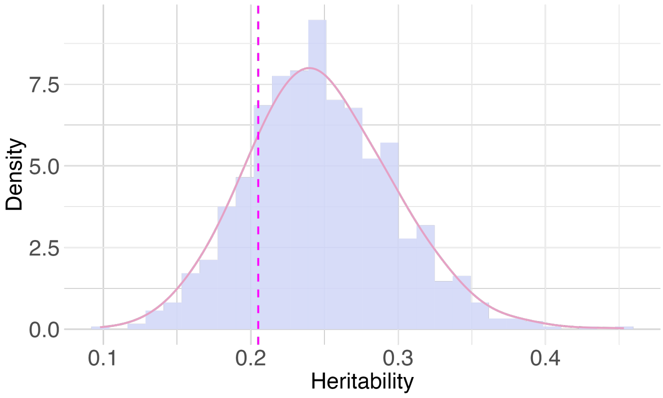

Figure 4: Heritability of color of male beetles from BVI method (top) and Stoffel (bottom)

The heritability of the color of male beetles is modelled by a binomial GLMM with binary outcome. We see that the approximated distribution from the BVI method is centered a bit to the right of the point estimate from **rptR**. The stochasticity of the BVI method makes it so that the distributions return will vary based on what seed was made to generate the results. Therefore, we expect to see a difference between the centering of the distribution and point estimates.

The case for number of eggs laid by female larvae is centered a bit more to the left than the Stoffel estimate. We notice that the heritability of eggs laid from the BVI method seems to be roughly twice that of the color of male beetles, whereas the **rptR** package estimates the heritability of eggs laid to be more than twice that of color.

4.3 Simulation study

	Binomial (logit)	Binomial (probit)	Poisson (log)
Random	0.0954 [0.0698, 0.1272]	INSERT	0.1338 [0.1039, 0.1700]
Fixed 1	0.0976 [0.0822, 0.1154]	INSERT	0.1336 [0.1254, 0.1412]
Fixed 2	0.1951 [0.1773, 0.2132]	INSERT	0.2675 [0.2552, 0.2816]
Fixed 3	0.2916 [0.2673, 0.3174]	INSERT	0.4013 [0.3852, 0.4216]
R_m^2	0.5843 [0.5545, 0.6157]	INSERT	0.8024 [0.7697, 0.8340]
R_c^2	0.6797 [0.6527, 0.7037]	INSERT	0.9361 [0.9270, 0.9472]

Table 2: Average relative importance across simulations using the BVI method, Stoffels results with 10 bootstrap samples and the average expected importance (dependent on the fitted model for the Poisson model)

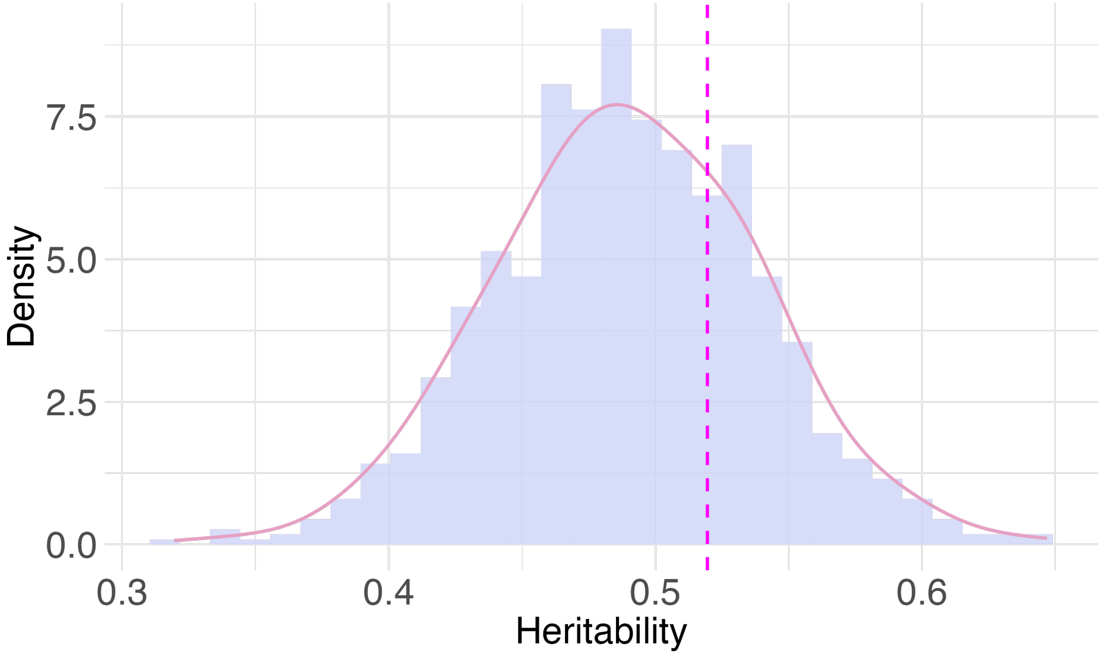

Figure 5: Heritability of eggs laid by female beetles from BVI method (top) and Stoffel (bottom)

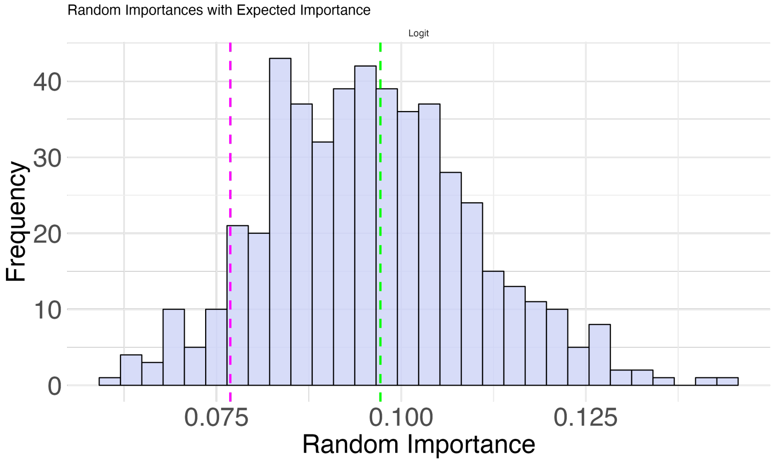

Figure 6: Relative importance of the random effect for binomial GLMM with logit link and Poisson GLMM with log link. The magenta line is Stoffel, green line is expected importance

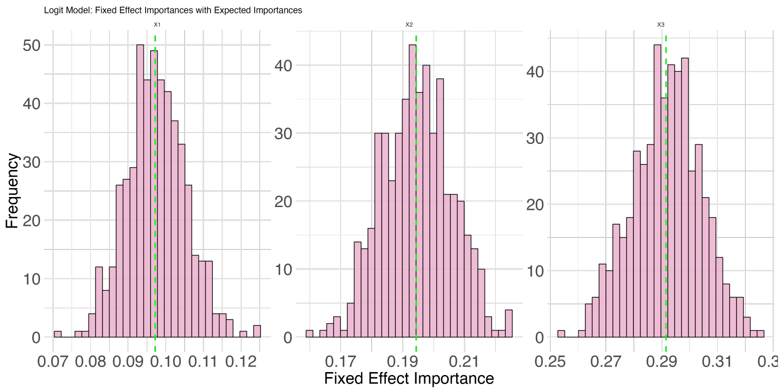

Figure 7: Relative importance of fixed effects for binomial GLMM with logit link. Blue line is expected importance

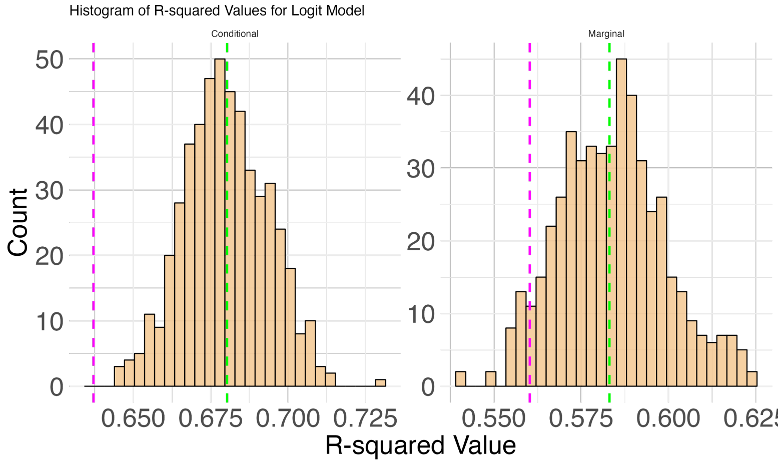

Figure 7: Conditional and marginal R^2 for binomial GLMM with logit link. The magenta line is Stoffel, green line is expected importance

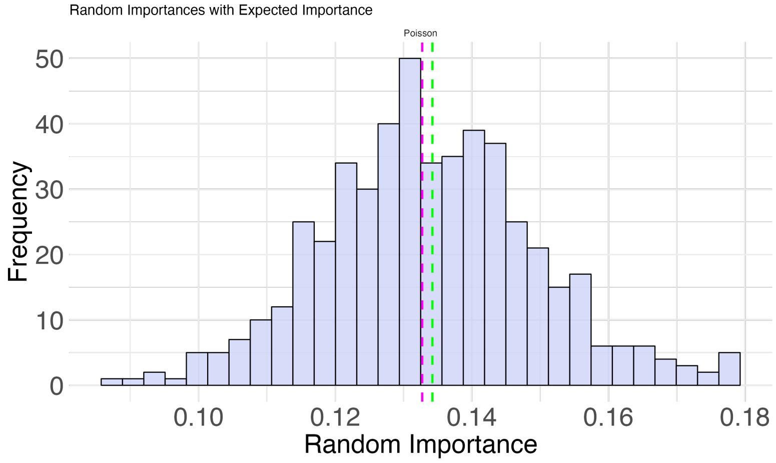

Figure 8: Relative importance of the random effect for binomial GLMM with logit link and Poisson GLMM with log link. The magenta line is Stoffel, green line is expected importance

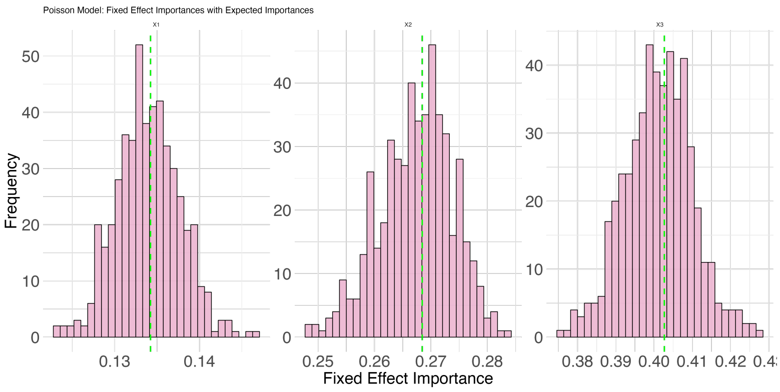

Figure 9: Relative importance of fixed effects for poisson GLMM

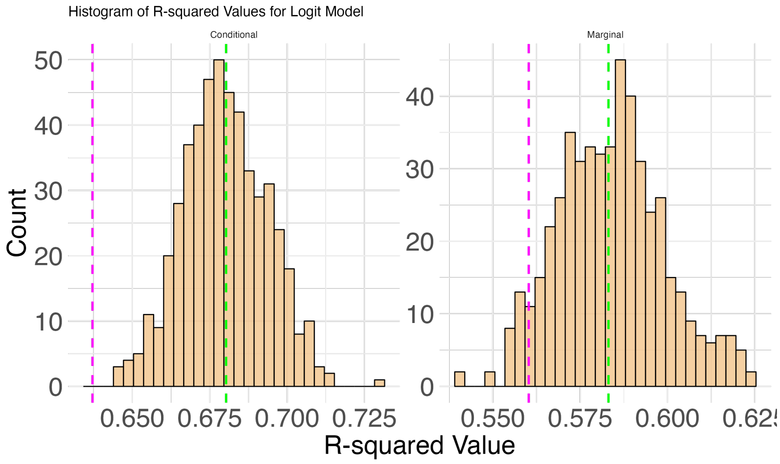

Figure 9: Conditional and marginal R^2 for poisson GLMM. The magenta line is Stoffel, green line is expected importance

CHAPTER FIVE

DISCUSSION & FURTHER WORK

CHAPTER

SIX

CONCLUSIONS

BIBLIOGRAPHY

Arnstad, A. (2024), 'Relative variable importance in bayesian linear mixed models'.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Wilson, A. J. (2008*a*), 'Why h2 does not always equal va/vp?', *Journal of Evolutionary Biology* **21**(5), 647–650.

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

APPENDIX

A

GITHUB REPOSITORY

APPENDIX

B

BAYESIAN VARIABLE IMPORTANCE USAGE

APPENDIX

C

MISCELLANEOUS PROOFS

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

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

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

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

This can be shown by considering the following:

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

and

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

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

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

and

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

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

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

and

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