Empirical decomposition of the explained variation in the variance components form of the mixed model

Nicholas Schreck

University of Mannheim, Mannheim, Germany

December 28, 2019

Abstract

The coefficient of determination is a standard characteristic in linear models. It is widely used to assess the proportion of variation explained, to determine the goodness-of-fit and to compare models with different covariates. However, there has not been an agreement on a similar quantity for the class of linear mixed models yet. We introduce a natural extension of the well-known adjusted coefficient of determination in linear models to the variance components form of the linear mixed model. We propose a novel coefficient of determination which is dimensionless, has an intuitive and simple definition in terms of variance explained, is additive for several random effects and reduces to the adjusted coefficient of determination in the linear model. To this end, we prove a full decomposition of the sum of squares of the independent variable into the explained and residual variance. Based on the restricted maximum likelihood equations, we introduce a novel measure for the explained variation which we allocate specifically to the contribution of the fixed and the random covariates of the model. We illustrate on two genomic datasets that the novel coefficient of determination can be applied as an improved estimator of the heritability of complex traits. Exemplarily, we allocate the explained variation to the five chromosomes of the model plant *Arabidopsis thaliana* and determine the contribution of each chromosome as well as cross-correlations between the chromosomes to the trait heritability.

Key words and phrases. Coefficient of determination; Explained variation; Genomic variance; Heritability; Linear mixed model; Model comparison; Variance components; Variance decomposition.

1 Introduction

Genome-wide association studies focus on finding associations between a complex phenotype and marker loci spread on the whole genome, as well as on studying the most common genetic variation in a population (Astle, 2009). Often, testing single nucleotide polymorphisms in separate ordinary regressions has proven to find useful associations and in some cases even to detect quantitative trait loci (The Wellcome Trust Consortium, 2007; Newton-Cheh et al., 2009). This approach, however, suffers from low power due to the multiple testing problem (Visscher et al., 2007), from confounding effects of genome-wide loci due to linkage disequilibrium (Segura et al., 2013), and from spurious correlation, i.e. associations that arise not at or not even near a causal locus (Astle, 2009). Therefore, the linear mixed model

$$y = \mu 1_n + X\beta + \sum_{i=1}^r Z_i u_i + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_n), \quad u_i \sim \mathcal{N}(0, \sigma_{u_i}^2 I_{p_i}) \quad (1)$$

in the traditional variance components form (Searle et al., 1992, §6, §7) is widely applied in genome-wide association studies to fit all markers simultaneously (Kang et al., 2008, 2010; Zhang et al., 2010; Zhou and Stephens, 2012). In model (1), y is the vector of observations (e.g. phenotypic values), $\mu \in \mathbb{R}$ is an intercept, 1_n is the column vector of ones of length n , I_n denotes the n -dimensional identity matrix, and $\sigma_\varepsilon^2 > 0$ and $\sigma_{u_i}^2 \geq 0$ ($i = 1, \dots, r$). We assume that the design matrix $\tilde{X} = (1_n, X)$ for the fixed effects has full rank $(k+1) < n$. The matrices Z_i ($i = 1, \dots, r$) code covariates with random effects u_i , e.g. genotypes at $p = \sum_{i=1}^r p_i$ loci.

The heritability of a trait is defined as the proportion of variability in the phenotypic values that can be explained by genomic factors, and is a main quantity of interest in many genomic studies (Falconer and Mackay, 1996). It is well known that the whole genome contributes to the genetic variance for complex traits. It is of central interest to identify genetic variants that explain variation in a phenotype (McCarthy et al., 2008), as well as to attribute the observed variation to specific covariates, in particular to certain regions of the genome such as chromosomes or even single markers (Visscher et al., 2007; Pimentel et al., 2011). In a classical regression context, the heritability equals the coefficient of determination (Falconer and Mackay, 1996), i.e. it measures the strength of the association between the phenotype and the covariates. In more detail, let us first consider the special case

$$y = \mu 1_n + X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_n) \quad (2)$$

of the linear model (1). The coefficient of determination $R^2 = 1 - \text{RSS} / \text{TSS}$ is the most popular measure of goodness-of-fit in model (2) (e.g. Kvalseth, 1985), where $\text{TSS} = y^\top C y$ is the centred total sum of squares, $C = I_n - n^{-1} 1_n 1_n^\top$ is a centring matrix, $\text{RSS} = \hat{\varepsilon}^\top \hat{\varepsilon}$ is the residual sum of squares, and $\hat{\varepsilon}$ are the estimated residuals, respectively. The centred explained sum of squares is $\text{ESS} = y^\top Q^\top C Q y$ with $Q = \tilde{X}(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top$. Because of the intercept in model (2) we have $\text{TSS} = \text{ESS} + \text{RSS}$, which leads to $R^2 = \text{ESS} / \text{TSS}$, i.e. the coefficient of determination measures the proportion of total variation that can be explained by the covariates after the least squares fit of model (2), see Kvalseth (1985) or Anderson-Sprecher (1994). We prove in the Supplementary material that $\text{ESS} = \hat{\beta}^\top X^\top C X \hat{\beta}$, where $\hat{\beta}$ is the least-squares (or maximum likelihood) estimator for β . However, the R^2 tends to overestimate the proportion of the explained variation (Kvalseth, 1985). In order to correct for the number of covariates, the adjusted coefficient of determination

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1} = 1 - \frac{\dot{\sigma}_\varepsilon^2}{\dot{\sigma}_y^2}$$

has been introduced. Here, $\dot{\sigma}_y^2$ is the empirical variance of y and the dot above a variable signifies that restricted-maximum likelihood estimators replace the variance components in the respective quantity. An unbiased estimator for the sample variance $\beta^\top \hat{\Sigma}_X \beta$ of $(X\beta)_i$ ($i = 1, \dots, n$) is

$$\dot{R}_X^2 = \hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} - \text{tr}(\hat{\Sigma}_X \hat{\Sigma}_{\hat{\beta}}),$$

where $\hat{\Sigma}_X$ is the empirical covariance matrix of X and $\Sigma_{\hat{\beta}} = \text{cov}(\hat{\beta})$. In total, we obtain the decomposition

$$\dot{\sigma}_y^2 = \dot{R}_X^2 + \dot{\sigma}_\varepsilon^2 \quad (3)$$

of the centred total sum of squares, and the alternative expression

$$\bar{R}^2 = \frac{\dot{R}_X^2}{\dot{\sigma}_y^2} = \frac{\dot{R}_X^2}{\dot{R}_X^2 + \dot{\sigma}_\varepsilon^2}. \quad (4)$$

The sample variance $\beta^\top \hat{\Sigma}_X \beta$ is induced by the empirical variation of the explanatory variables with design X , and the estimator $\dot{R}_X^2$ measures the (adjusted) empirical variation explained by the covariates and their empirical pairwise covariances. We derive these results in more detail in the Supplementary material. The $\bar{R}^2$ in equation (4) measures the goodness of the model fit, has an intuitive reasonable interpretation in terms of explained empirical variation, is independent of the units of measurement of the model variables, i.e. it is dimensionless, has values in the range of 0 and 1, and equals 1 in case of a perfect fit. These are some of the guidelines for measures of model fit introduced by Kvalseth (1985). Besides its interpretation as a measure of explained variation, the $\bar{R}^2$ is a wide-spread tool for comparisons of models with a common null model (Anderson-Sprecher, 1994), e.g. a model that contains only the intercept.

The random effects model

$$y = \mu 1_n + Zu + \varepsilon, \quad u \sim \mathcal{N}(0, \sigma_u^2 I_p), \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_n), \quad (5)$$

a special case of model (1) with only one random effect and no fixed covariates but the intercept, is used in many genomic applications. Interestingly, it has not been agreed on a unified coefficient of determination for model (5) yet. Typical approaches to define a measure of goodness-of-fit include likelihood-based methods and measures for the total model variance (e.g. Piepho, 2019). In the framework of model (5), the heritability is often defined as

$$h^2 = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}, \quad (6)$$

i.e. a ratio of the variance components of the model (e.g. Loh et al., 2015). The estimated variance component $\hat{\sigma}_u^2$ is considered as the explained variation, and the ratio (6) is structurally similar to the $\bar{R}^2$ in equation (4). However, the variance components do not completely decompose the phenotypic variation as in equation (3), and it is not possible to fully attribute the total observed variation to the covariates of the model. It has been widely stated that ratios similar to (6) should not be used to estimate the proportion of the variation explained by the genomic variance (Gianola et al., 2009; de los Campos et al., 2015). Recently, Schreck et al. (2019) have introduced the empirical best predictor

$$\dot{R}_Z^2 = \dot{u}^\top \hat{\Sigma}_Z \dot{u} + \dot{\sigma}_u^2 \text{tr}(\hat{\Sigma}_Z) - \text{tr}(\hat{\Sigma}_Z \dot{\Sigma}_{\tilde{u}}) \quad (7)$$

for the empirical genomic variance $u^\top \hat{\Sigma}_Z u$. Here, $\dot{u}$ is the empirical best linear unbiased predictor of the random effects u and $\Sigma_{\tilde{u}}$ is the covariance matrix of the best linear unbiased predictor $\tilde{u}$. The striking equality

$$\hat{\sigma}_y^2 = \dot{R}_Z^2 + \dot{\sigma}_\varepsilon^2, \quad (8)$$

which is similar to equation (3), has been observed by Schreck et al. (2019) in several applications. This suggests a connection between the estimated genomic variance in (7) and the concept of the explained variation in the random effects model similar to model (2).

In the following, we prove a decomposition of the total sum of squares for model (1). This decomposition is based on the wide-spread restricted maximum likelihood estimation of the variance components. We derive a natural and consistent extension of the concept of the coefficient of determination to the variance components form of the linear mixed model, obtaining the equalities (3) and (8) as special cases. More specifically, we decompose the empirical explained variation into separate contributions by the fixed covariates, the random covariates and empirical covariance effects between the fixed and the random covariates. We explicitly attribute the explained variation to the five chromosomes of the model plant *Arabidopsis thaliana* and quantify empirical cross-effects between the chromosomes.

2 Decomposition in the Linear Mixed Effects Model

The linear mixed model (1) has pairwise uncorrelated random effects (including the residuals). Let the random vector

$$u = (u_1^\top, \dots, u_r^\top)^\top \sim \mathcal{N}(0, D), \quad D = \text{diag}(\sigma_{u_1}^2 I_{p_1}, \dots, \sigma_{u_r}^2 I_{p_r}),$$

be of dimension p and $Z = (Z_1, \dots, Z_r)$ denote an $n \times p$ design matrix. Model (1) has the variance-covariance structure

$$V = \text{cov}(y) = ZDZ^\top + \sigma_\varepsilon^2 I_n = \sum_{i=1}^r \sigma_{u_i}^2 Z_i Z_i^\top + \sigma_\varepsilon^2 I_n,$$

such that $y \sim \mathcal{N}(1_n \mu + X\beta, V)$. Similar to (Searle et al., 1992, §6), we let $\gamma_i^2 = \sigma_{u_i}^2 / \sigma_\varepsilon^2$ ($i = 1, \dots, r$) such that $G = D / \sigma_\varepsilon^2$, i.e. G is D but with $\sigma_{u_i}^2$ replaced by γ_i^2 . Let $H = V / \sigma_\varepsilon^2 = I_n + ZGZ^\top$ and $H_C = I_n + CZGZ^\top C$, which originates from centring the columns of Z .

Assumption 1 (Inverses of matrices). *We assume that $\tilde{X} = (1_n, X)$ has full column rank and that the inverses of H , $\tilde{X}^\top H^{-1} \tilde{X}$, H_C and $X^\top C H_C^{-1} C X$ exist.*

For notational convenience let

$$P_H = H^{-1} - H^{-1} X (X^\top H^{-1} X)^{-1} X^\top H^{-1}, \quad P_H^I = H^{-1} - H^{-1} \tilde{X} (\tilde{X}^\top H^{-1} \tilde{X})^{-1} \tilde{X}^\top H^{-1}.$$

The following lemma states the invariance of the best linear unbiased estimator $\hat{\beta}$ and predictor $\tilde{u}$ to centring the data y , X , and Z . The invariance also holds for their respective covariance matrices. In particular, the inclusion of a random intercept term in model (1) is not necessary to arrive at this result. In the framework of model (1) this is a striking property, because centring Z automatically changes the covariance structure of the model.

Lemma 2. *The best linear unbiased estimator of the intercept μ is $\hat{\mu} = b^{-1} 1_n^\top P_H y$, $b = 1_n^\top P_H 1_n$, and the best linear unbiased estimator of the effects β with its covariance matrix is*

$$\hat{\beta} = (X^\top C H_C^{-1} C X)^{-1} X^\top C H_C^{-1} C y, \quad \Sigma_{\hat{\beta}} = \sigma_\varepsilon^2 (X^\top C H_C^{-1} C X)^{-1}.$$

The best linear unbiased predictor of the random effects u is

$$\tilde{u} = GZ^\top CH_C^{-1}C(y - X\hat{\beta})$$

with covariance matrix

$$\Sigma_{\tilde{u}} = \sigma_\varepsilon^2 GZ^\top CH_C^{-1} \left\{ H_C - CX(X^\top CH_C^{-1}CX)^{-1}X^\top C \right\} H_C^{-1}CZG.$$

The following lemma introduces unbiased estimators and predictors for the empirical explained variances in model (1).

Lemma 3. Based on model (1) and the quantities introduced in Lemma 2, the estimator

$$\hat{R}_X^2 = \hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} - \text{tr}(\hat{\Sigma}_X \Sigma_{\hat{\beta}})$$

is unbiased for the sample variance $\beta^\top \hat{\Sigma}_X \beta$ of $(X\beta)_i$ ($i = 1, \dots, n$), the predictor

$$\tilde{R}_Z^2 = \tilde{u}^\top \hat{\Sigma}_Z \tilde{u} - \text{tr}(\hat{\Sigma}_Z \Sigma_{\tilde{u}}) + \text{tr}(D\hat{\Sigma}_Z)$$

is unbiased for the sample variance $u^\top \hat{\Sigma}_Z u$ of $(Zu)_i$ ($i = 1, \dots, n$), and

$$\tilde{R}_{X \times Z}^2 = 2\hat{\beta}^\top \hat{\Sigma}_{XZ} u$$

is an unbiased predictor of the sample covariance $2\beta^\top \hat{\Sigma}_{XZ} \tilde{u}$ induced by the possible non-orthogonality of the columns of CX and CZ . The quantities $\hat{R}_X^2$, $\tilde{R}_Z^2$, and $\tilde{R}_{X \times Z}^2$ are invariant to centring the data y , X , and Z .

The estimation of the variance components is a principal aspect to be considered in the process of fitting model (1). Typically, estimation procedures for the variance components are based on likelihood approaches, where the restricted maximum likelihood approach (Patterson and Thompson, 1971) is preferred in most applications. The restricted maximum likelihood estimates for the variance components are the solution to a system of non-linear equations.

Lemma 4. The set of REML equations (Searle et al., 1992, §6)

$$\dot{\sigma}_\varepsilon^2 = \frac{1}{n - k - 1} y^\top \dot{P}_H^I y = \frac{1}{\text{tr}(\dot{P}_H^I)} y^\top \dot{P}_H^I \dot{P}_H^I y \quad (9)$$

and

$$\dot{\sigma}_\varepsilon^2 \text{tr}(\dot{P}_H^I Z_i Z_i^\top) = y^\top \dot{P}_H^I Z_i Z_i^\top \dot{P}_H^I y, \quad i = 1, \dots, r, \quad (10)$$

is invariant to centring y , X and Z . In addition, the equation for the residual variance can be reformulated in the following three ways:

$$\begin{aligned} \dot{\sigma}_\varepsilon^2 &= \frac{1}{n - k - 1} (y - 1_n \dot{\mu} - X\dot{\beta})^\top \dot{H}^{-1} (y - 1_n \dot{\mu} - X\dot{\beta}) \\ &= \frac{1}{n - k - 1} (y - X\dot{\beta})^\top C \dot{H}_C^{-1} C (y - X\dot{\beta}) \\ &= \frac{1}{\text{tr}(C \dot{P}_H^I C)} (y - X\dot{\beta})^\top C \dot{H}_C^{-2} C (y - X\dot{\beta}). \end{aligned} \quad (11)$$

From now on, we assume that we have admissible solutions to the system of equations (9) and (10). We replace the variance components by their estimates and mark the resulting quantities by a dot. In case that the variable already has a superscript, we drop the first superscript for notational convenience. Here, this is only necessary when plugging estimates of the variance components into $\hat{\mu}$, $\hat{\beta}$ and $\tilde{u}$. The resulting empirical best linear unbiased estimators $\dot{\mu}$ and $\dot{\beta}$ as well as the best linear unbiased predictor $\dot{u}$ are no longer linear in y , and properties such as unbiasedness hold only under additional assumptions (Kackar and Harville, 1984; Jiang, 1999). The following theorem states the empirical decomposition of the centred total sum of squares.

Theorem 5. *Given restricted maximum likelihood estimators $\dot{\sigma}_\varepsilon^2$ for the residual variance σ_ε^2 and $\dot{\sigma}_{u_i}^2$ for the variance components $\sigma_{u_i}^2$ ($i = 1, \dots, r$) in model (1), we find*

$$\dot{\sigma}_y^2 = \dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2 + \dot{\sigma}_\varepsilon^2. \quad (12)$$

An intuitive reasoning for variables $\dot{R}_X^2$, $\dot{R}_Z^2$, and $\dot{R}_{X \times Z}^2$ is given by Lemma 3. A proof of this theorem can be found in the Appendix. Similar to $\bar{R}^2$ in equation (4), let $\dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2$ be the empirical explained variation and

$$\dot{R}^2 = \frac{1}{\dot{\sigma}_y^2} (\dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2) \quad (13)$$

be a coefficient of determination for the linear mixed model (1). The additive decomposition derived in (12) gives

$$\dot{R}^2 = \frac{\dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2}{\dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2 + \dot{\sigma}_\varepsilon^2} = 1 - \frac{\dot{\sigma}_\varepsilon^2}{\dot{\sigma}_y^2},$$

i.e. the possible range of values of $\dot{R}^2$ is $[0, 1]$ and it equals 1 if the residual variance equals 0. The quantity $\dot{R}^2$ is dimensionless, i.e. it is invariant to scaling and centring the data y , X , and Z . The $\dot{R}^2$ behaves additively for several random effects because

$$\dot{R}_Z^2 = \sum_{i=1}^r \dot{\sigma}_{u_i}^2 \text{tr}(\hat{\Sigma}_{Z_i}) + \sum_{i=1}^r \sum_{j=1}^r \dot{u}_i^\top \hat{\Sigma}_{Z_i Z_j} \dot{u}_j - \sum_{i=1}^r \sum_{j=1}^r \text{tr}(\hat{\Sigma}_{Z_i Z_j} \dot{\tilde{u}}_j \dot{\tilde{u}}_i) \quad (14)$$

and

$$\dot{R}_{X \times Z}^2 = 2\dot{\beta}^\top \sum_{i=1}^r \hat{\Sigma}_{X Z_i} \dot{u}_i, \quad (15)$$

i.e. we can allocate the explained variation to specific contributions of covariates and empirical covariances between covariates. The definition of $\dot{R}^2$ in (13) reduces to $\bar{R}^2$ in the linear model (2), i.e. it is a natural extension of the adjusted coefficient of determination from model (2) to model (1). We obtain the variance decomposition (8) in case of the random effects model (5), see also Schreck et al. (2019). The $\dot{R}^2$ is an analytic function of the estimates returned after a model fit with restricted maximum likelihood so that its computation is straightforward and fast. Maybe even more importantly, due to its explicit nature, an interpretation of the origin and the nature of the explained variation is possible. In particular, the explained variation is caused

by the sample variation of the covariates, no matter their specific nature. The contribution of each covariate to the explained variation is clearly visible. Equation (12) implies that we can additively attribute the explained variance within model (1) to the fixed effects, the random effects, and a contribution caused by potential sample correlations between the fixed and random effects co-variables.

3 Data Analysis

3.1 Datasets

We consider the mice dataset that comes with the R package BGLR (Perez and de los Campos, 2014) and originates from experiments by Valdar et al. (2006a,b). The dataset includes $n = 1814$ mice. We consider the phenotype “Obesity Bodyweight” as the dependent variable. In the framework of model (1), we will use “Obesity Bodylength” as a fixed covariate, and the full set of $p = 10,346$ biallelic genotyped loci as random effects with a common variance component, i.e. we set $r = 1$.

In addition, we consider a dataset for the model organism “*Arabidopsis thaliana*” which is publicly available (The 1001 Genomes Consortium, 2016). The dataset has $n = 1057$ lines which were genotyped for $p = 193,697$ single nucleotide polymorphism markers. We relate the phenotype flowering time at $10^\circ C$ to the fixed effect covariates, in this case the longitude and the latitude at which the flowers were picked. We considered the marker genotypes on the $r = 5$ chromosomes of the organism as random variables with common variance components. The five chromosomes were genotyped for $p_1 = 47,518$, $p_2 = 25,550$, $p_3 = 38,813$, $p_4 = 33,240$ and $p_5 = 48,576$ markers, respectively.

3.2 Settings

We conduct the following analysis with the free software R (R Development Core Team, 2018). We use the package sommer (Covarrubias-Pazarán, 2017) to fit the linear mixed model defined in (1). We use the option restricted maximum likelihood in sommer to estimate the variance components of the model, see also Lemma 4. Besides the estimated variance components, the package also provides the quantities defined in Lemma 2. We use this outcome to calculate the variance decomposition introduced in Theorem 5. For simplicity and comparability we scale the sample variance of the dependent variable y to be equal to 100, i.e. all the variables that describe empirical variation are automatically relative to the empirical variation in the phenotypes.

3.3 Illustration of the empirical decomposition

In Table 1 we illustrate the empirical variance decomposition of the dependent variable and the attribution of the empirical variance to the components of model (1). For both datasets the empirical variance decomposition holds up to at least the sixth digit which is not illustrated for reasons of clarity. The remaining difference is due to the lack of computational accuracy. Compared to *Arabidopsis thaliana*, the residual variance for the mice dataset is considerably large, i.e. the model for *Arabidopsis thaliana* explains a larger part of the variability in the phenotypes. We

Table 1: Empirical explained variation; each entry is multiplied by 100; The total sum of squares was scaled to 100. All numbers have been rounded to two digits.

	$\dot{R}_X^2$	$\dot{R}_Z^2$	$\dot{R}_{X \times Z}^2$	$\dot{\sigma}_\varepsilon^2$	Σ
Mice	16.90	30.21	-8.22	61.11	100.00
<i>Arabidopsis thaliana</i>	1.23	85.73	5.44	7.59	100.00

Table 2: Contribution of the five chromosomes of *Arabidopsis thaliana* to the explained variation; first, we depict the 5×5 matrix of empirical explained correlations; the last column contains the explained variation by the respective chromosome; All numbers have been rounded to two digits.

Chromosomes	1	2	3	4	5	$(\dot{R}_Z^2)_i$
1	1.00	0.12	0.06	0.13	0.17	0.04
2	0.12	1.00	0.06	0.16	0.17	0.06
3	0.06	0.06	1.00	0.08	0.11	0.03
4	0.13	0.16	0.08	1.00	0.23	0.14
5	0.17	0.17	0.11	0.23	1.00	0.35

are able to attribute most of the explained variance to the random effects covariates. The additive genomic variance, and because of the standardized phenotypes also the heritability, is about 30% for the mice dataset and about 86% for the model plant *Arabidopsis thaliana*. We conclude that the phenotype flowering time is highly heritable, whereas the variation in the bodyweight of the mice can be roughly attributed to the variation of body length, heritage, and unknown (environmental) factors in the relation 1 : 2 : 4. Also, the variance explained by the only fixed effect variable, body length, is about half of the variance explained by the 10,346 random genomic covariates. This is somewhat intuitive, because the body mass index depends on the body length. In both datasets, a minor, but not negligible, part of the variation in the observations stems from sample covariation between the fixed and random covariates.

3.4 Attribution of empirical explained variation

In Table 2 we illustrate the decomposition of $\dot{R}_Z^2$ into the contributions of the five chromosomes of *Arabidopsis thaliana* and their cross-effects. No fixed effects but the intercept have been included in the model fit. The formula for the decomposition is given in equation (14), and we attribute

$$(\dot{R}_Z^2)_i = \dot{\sigma}_{u_i}^2 \text{tr}(\hat{\Sigma}_{Z_i}) + \dot{u}_i^\top \hat{\Sigma}_{Z_i} \dot{u}_i - \text{tr}(\hat{\Sigma}_{Z_i} \dot{\Sigma}_{\tilde{u}_i}), \quad i = 1, \dots, 5,$$

to the single chromosomes. We define the empirical explained covariances analogously by

$$(\dot{R}_Z^2)_{ij} = \dot{u}_i^\top \hat{\Sigma}_{Z_i Z_j} \dot{u}_j - \text{tr}(\hat{\Sigma}_{Z_i Z_j} \dot{\Sigma}_{\tilde{u}_j \tilde{u}_i}), \quad i = 1, \dots, 5, \quad j = 1, \dots, 5,$$

and the correlations by division with the corresponding square roots of the empirical variances.

Chromosome five alone contributes the largest amount to the explained variation, whereas the contributions of the first three chromosomes are comparably small. The empirical explained

Figure 1: Illustration of the contribution of each marker to the explained variation.

correlations are the largest between the fifth and the other chromosomes. The magnitude of the correlations indicates the necessity to include them in the analysis. The empirical covariances between the chromosomes contribute about 22% to the observed phenotypic variation.

We infer the direct and indirect contribution of single (fixed and random) components to the explained variation for the mice dataset. We allocate the contribution of each single random effect to the explained variation by setting $r = 1$:

$$\dot{R}_Z^2 = \dot{\sigma}_u^2 \sum_{j=1}^{10,346} (\hat{\Sigma}_Z)_{jj} + \sum_{j=1}^{10,346} \sum_{i=1}^{10,346} \left(\dot{u}_i \dot{u}_j - (\dot{\Sigma}_{\tilde{u}})_{ji} \right) (\hat{\Sigma}_Z)_{ij},$$

i.e. the explained variation by the random effect covariates can be decomposed into the contribution of the p markers. In Figure 1 we illustrate the spread of the explained variation across the genome of the mice. This illustration is similar to the well-known Manhattan plots, i.e. a graph depicting the p -values of the single markers in ordinary regressions. However, Figure 1 is an holistic and relative analysis, i.e. the contributions of each marker are always with respect to the other markers, which contrasts the classical single marker analysis in genome-wide association studies. We obtain a map of the genotyped markers and are able to specify regions of lower and larger contributions to the explained variation. This allows for an overview of the genome with respect to the specific trait, and in a next step the further analysis of certain regions of the genome.

4 Discussion

We have proven a novel partitioning of the centred sum of squares in the components form of the linear mixed model, where we have used only estimates that result from the model fit using the restricted maximum likelihood approach. The coefficient of determination $\hat{R}^2$ satisfies the most important criteria for appropriate measures of fit, and it is a natural extension of the coefficient of determination from the linear model (2). The decomposition of the total sum of squares derived in Theorem 5 and the associated natural extension of the classical coefficient of determination to the components form of the linear mixed model is in particular important in genomic applications.

We have extended the best prediction approach introduced by Schreck et al. (2019) from the random effects model to the linear mixed effects model. This enables an improved estimation of the additive genomic variance and the heritability for model (1). We have attributed the explained variation, and in particular the genomic variance, to specific components and empirical covariances between covariates of model (1). This decomposition results in graphs that illustrate how the estimated genomic variance is spread across certain regions of the genome. This enables an overview over the relative explained variation by specific (groups of) covariates in a model where these variables are fitted simultaneously. Hence, we have introduced an alternative to the common Manhattan plots, where the variables are tested in separate ordinary regressions.

Variable or model selection is becoming more important with the advent of high-dimensional data. Classically, model comparisons based on (2) are based on the concept of the coefficient of determination, and the model with the larger explained variation is typically chosen. Finding parsimonious models in the framework of model (1) is, however, not straightforward, and there has not been a consensus on how to approach that subject yet (Müller et al., 2013). Approaches include to drop out random covariates with zero variance components (Chen and Brunson, 2003) and a simultaneous selection of fixed and random effects based on a re-parameterized likelihood (Bondell et al., 2010). We investigate the contribution of covariates to the novel coefficient of determination $\hat{R}^2$. These contributions are not based on the variance components alone, and the sample covariances between different covariates also influence the explained variation. Including these covariances is vital, and can only be neglected if the independent variables are orthogonal to each other. Based on $\hat{R}^2$ we expect new strategies for variable selection in the framework of model (1).

Acknowledgement

Most sincere thanks are due to Martin Schlather for valuable comments and conversations on this manuscript. I am grateful to Jonas Brehmer and Christopher Dörr for proofreading the manuscript. I was supported by the Deutsche Forschungsgemeinschaft and the Research Training Group 1953.

Supplementary material

Supplementary Material available online includes the technical details for the empirical variance decomposition in §1, additional technical lemmata for the variance components model and proofs of the lemmata in §2.

Appendix 1

Proof of the main theorem

Proof. (Theorem 5.) First, we make the preliminary considerations that

$$X^\top C \dot{P}_H^C = X^\top C \dot{H}_C^{-1} - X^\top C \dot{H}_C^{-1} C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1} = 0$$

and

$$y - X\dot{\beta} = (I_n - X(X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1})y.$$

Based on that, we find

$$X^\top C \dot{H}_C^{-1} C (y - X\dot{\beta}) = X^\top C \dot{P}_H^C y = 0,$$

which implies that $\dot{H}_C^{-1} C (y - X\dot{\beta})$ is orthogonal to the column space of X . We use this result together with Lemma 2 and the definition of H_C to calculate

$$\begin{aligned} 2\dot{\beta}^\top \hat{\Sigma}_{XZ} \dot{u} &= \frac{2}{n-1} \dot{\beta}^\top X^\top C Z \dot{G} Z^\top C \dot{H}_C^{-1} C (y - X\dot{\beta}) \\ &= \frac{2}{n-1} \dot{\beta}^\top X^\top (I_n - \dot{H}_C^{-1}) C (y - X\dot{\beta}) \\ &= \frac{2}{n-1} \dot{\beta}^\top X^\top C (y - X\dot{\beta}) - \frac{2}{n-1} \dot{\beta}^\top X^\top \dot{H}_C^{-1} C (y - X\dot{\beta}) \\ &= \frac{2}{n-1} \dot{\beta}^\top X^\top C (y - X\dot{\beta}). \end{aligned} \tag{16}$$

Next, we consider the quadratic form in the empirical best linear unbiased estimators

$$\begin{aligned} \dot{u}^\top \hat{\Sigma}_Z \dot{u} &= \frac{1}{n-1} (y - X\dot{\beta})^\top C \dot{H}_C^{-1} C Z \dot{G} Z^\top C Z \dot{G} Z^\top C \dot{H}_C^{-1} C (y - X\dot{\beta}) \\ &= \frac{1}{n-1} (y - X\dot{\beta})^\top C (I_n - \dot{H}_C^{-1})^2 C (y - X\dot{\beta}) \\ &= \frac{1}{n-1} (y - X\dot{\beta})^\top C (y - X\dot{\beta}) \\ &\quad + \frac{1}{n-1} (y - X\dot{\beta})^\top C \dot{H}_C^{-2} C (y - X\dot{\beta}) \\ &\quad - 2 \frac{1}{n-1} (y - X\dot{\beta})^\top C \dot{H}_C^{-1} C (y - X\dot{\beta}). \end{aligned}$$

We use equation (11) from Lemma 4 to reformulate

$$(y - X\dot{\beta})^\top C\dot{H}_C^{-2}C(y - X\dot{\beta}) = \dot{\sigma}_\varepsilon^2 \text{tr}(C\dot{P}_H^C C)$$

and

$$(y - X\dot{\beta})^\top C\dot{H}_C^{-1}C(y - X\dot{\beta}) = \dot{\sigma}_\varepsilon^2(n - k - 1).$$

Combining these equations leads to

$$\begin{aligned} \dot{u}^\top \hat{\Sigma}_Z \dot{u} &= \frac{1}{n-1} y^\top C y + \dot{\beta}^\top \hat{\Sigma}_X \dot{\beta} - \frac{2}{n-1} \dot{\beta}^\top X^\top C y \\ &\quad + \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 (\text{tr}(C\dot{P}_H^C C) - 2n + 2k + 2). \end{aligned} \quad (17)$$

We use Lemma 2 and the cycle property of the trace to calculate

$$\begin{aligned} \text{tr}(\hat{\Sigma}_Z \dot{\Sigma}_{\tilde{u}}) &= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}(Z^\top C Z \dot{G} Z^\top C \dot{H}_C^{-1} C Z \dot{G}) \\ &\quad - \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}(Z^\top C Z \dot{G} Z^\top C \dot{H}_C^{-1} C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1} C Z \dot{G}) \\ &= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}\{(\dot{H}_C - I_n) \dot{H}_C^{-1} (\dot{H}_C - I_n)\} \\ &\quad - \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}\{(\dot{H}_C - I_n) \dot{H}_C^{-1} C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1} (\dot{H}_C - I_n)\} \\ &= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 (\text{tr}(\dot{H}_C) + \text{tr}(\dot{H}_C^{-1}) - 2n) - \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}\{(\dot{H}_C - I_n)^2 (\dot{H}_C^{-1} - \dot{P}_H^C)\} \end{aligned}$$

We use

$$\text{tr}(\dot{H}_C \dot{P}_H^C) = \text{tr}(I_n - C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1}) = n - k$$

because of the full rank of k of X to simplify

$$\begin{aligned} \text{tr}\{(\dot{H}_C - I_n)^2 (\dot{H}_C^{-1} - \dot{P}_H^C)\} &= \text{tr}(\dot{H}_C - 2I_n + \dot{H}_C^{-1} - \dot{P}_H^C - \dot{H}_C^2 \dot{P}_H^C + 2\dot{H}_C \dot{P}_H^C) \\ &= \text{tr}(\dot{H}_C) - 2k + \text{tr}(\dot{H}_C^{-1}) - \text{tr}(\dot{P}_H^C) + \text{tr}(\dot{H}_C \dot{P}_H^C \dot{H}_C). \end{aligned}$$

In total, we obtain

$$\text{tr}(\hat{\Sigma}_Z \dot{\Sigma}_{\tilde{u}}) = \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \{2k - 2n + \text{tr}(\dot{P}_H^C) + \text{tr}(\dot{H}_C \dot{P}_H^C \dot{H}_C)\}. \quad (18)$$

The analogous expression for the fixed effects is

$$\begin{aligned} \text{tr}(\hat{\Sigma}_X \dot{\Sigma}_{\tilde{\beta}}) &= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}(X^\top C X (X^\top C \dot{H}_C^{-1} C X)^{-1}) \\ &= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr}(\dot{H}_C \dot{H}_C^{-1} C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1} \dot{H}_C) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr} (\dot{H}_C (\dot{H}_C^{-1} - \dot{P}_H^C) \dot{H}_C) \\
&= \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \{ \text{tr} (\dot{H}_C) - \text{tr} (\dot{H}_C \dot{P}_H^C \dot{H}_C) \}.
\end{aligned} \tag{19}$$

In addition, we consider

$$\text{tr} (\dot{D} \hat{\Sigma}_Z) = \frac{1}{n-1} \text{tr} (\dot{D} Z^\top C Z) = \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \text{tr} (C Z \dot{G} Z^\top C) = \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 (\text{tr} (\dot{H}_C) - n). \tag{20}$$

Finally, we use (16), (17), (18) (19), and (20) to calculate the right-hand-side of Lemma 5:

$$\begin{aligned}
\dot{R}_X^2 + \dot{R}_Z^2 + \dot{R}_{X \times Z}^2 + \dot{\sigma}_\varepsilon^2 &= \dot{\beta}^\top \hat{\Sigma}_X \dot{\beta}^\top - \text{tr} (\hat{\Sigma}_X \dot{\Sigma}_{\hat{\beta}}) + \text{tr} (\dot{D} \hat{\Sigma}_Z) + \dot{u}^\top \hat{\Sigma}_Z \dot{u} - \text{tr} (\hat{\Sigma}_Z \dot{\Sigma}_{\hat{u}}) \\
&\quad + 2 \dot{\beta}^\top \hat{\Sigma}_{XZ} \dot{u} + \dot{\sigma}_\varepsilon^2 \\
&= \dot{\beta}^\top \hat{\Sigma}_X \dot{\beta} - \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \{ \text{tr} (\dot{H}_C) - \text{tr} (\dot{H}_C \dot{P}_H^C \dot{H}_C) \} \\
&\quad + \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 (\text{tr} (\dot{H}_C) - n) + \frac{1}{n-1} y^\top C y \\
&\quad + \dot{\beta}^\top \hat{\Sigma}_X \dot{\beta} - \frac{2}{n-1} \dot{\beta} X^\top C y \\
&\quad + \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 (\text{tr} (C \dot{P}_H^C C) - 2n + 2k + 2) \\
&\quad - \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \{ 2k - 2n + \text{tr} (\dot{P}_H^C) + \text{tr} (\dot{H}_C \dot{P}_H^C \dot{H}_C) \} \\
&\quad + \frac{2}{n-1} \dot{\beta}^\top X^\top C (y - X \dot{\beta}) + \frac{n-1}{n-1} \dot{\sigma}_\varepsilon^2 \\
&= \frac{1}{n-1} y^\top C y + \frac{1}{n-1} \dot{\sigma}_\varepsilon^2 \{ \text{tr} (C \dot{P}_H^C C) - \text{tr} (\dot{P}_H^C) + 1 \}.
\end{aligned}$$

We show in the Supplementary material that $\dot{H}_C^{-1} 1_n = 0$. Therefore, we have

$$\begin{aligned}
\text{tr} (C \dot{P}_H^C C) &= \text{tr} (\dot{P}_H^C) - n^{-1} 1_n^\top \dot{P}_H^C 1_n \\
&= \text{tr} (\dot{P}_H^C) - n^{-1} 1_n^\top \dot{H}_C^{-1} 1_n + n^{-1} 1_n^\top \dot{H}_C^{-1} C X (X^\top C \dot{H}_C^{-1} C X)^{-1} X^\top C \dot{H}_C^{-1} 1_n \\
&= \text{tr} (\dot{P}_H^C) - 1,
\end{aligned}$$

which proves the empirical variance decomposition given in equation (12). The invariance of $\dot{R}_X^2$, $\dot{R}_Z^2$ and $\dot{R}_{X \times Z}^2$ (and with it of $\dot{R}^2$) to centering and scaling the columns of y , X , and Z follows by applying Lemmata 2 and 4. $\square$

References

- ANDERSON-SPRECHER, R. (1994). Model comparisons and R^2 . *Am Stat.* **48**, 113–117.
- ASTLE, W. & BALDING, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Stat Sci.* **24**, 451–471.

- BONDELL, H. D., KRISHNA, A. & GOSH, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–1077.
- CHEN, Z. & BRUNSON, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.
- COVARRUBIAS-PAZARAN, G. (2016). Genome assisted prediction of quantitative traits using the R package sommer. *PLoS ONE* **11**, 1–15
- DE LOS CAMPOS, G., SORENSEN, D. & GIANOLA, D. (2015). Genomic heritability: what is it?. *PLoS Genet.* **11**, e1005048.
- FALCONER, D. & MACKAY, T. (1996). *Introduction into Quantitative Genetics*. Essex: Prentice Hall.
- GIANOLA, D., DE LOS CAMPOS, G., HILL, W. G., MANFREDI, E. & FERNANDO, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–363.
- GREVEN, S. & KNEIB, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 773–789.
- JIANG, J. (1999). On unbiasedness of the empirical BLUE and BLUP. *Statistics and Probability Letters* **41**, 19–24.
- KACKAR, R. N. & HARVILLE, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Am. Stat. Assoc.* **79**, 853–862.
- KANG, H. M., ZAITLEN, N. A., WADE, C. M., KIRBY, A., HECKERMAN, D., DALY, M. J. & ESKIN, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723.
- KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S., FREIMER, N. B., SABATTI, C. & ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies *Nat. Genet.* **42**, 348–354.
- KVALSETH, T. O. (1985). A cautionary note about R^2 . *The American Statistician* **39**, 279–285.
- LOH, P.-R., BHATIA, G., GUSEV, A. & ET AL.(2013). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nat Genetics* **47**, 1385–1392.
- MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. & HIRSCHHORN, J. N.(2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* **9**, 356–369.
- MÜLLER, S., SEALY, J. L. & WELSH, A. H. (2013). Model selection in linear mixed models. *Statistical Science* **28**, 135–167.

- NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.
- NEWTON-CHEH, C., JOHNSON, T., GATEVA, V. & ET AL.(2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet.* **41**, 666–676.
- PATTERSON, H. D. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.
- PEREZ, P. & DE LOS CAMPOS, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **198**, 483–495.
- PIEPHO, H.-P. (2019). A coefficient of determination (R^2) for linear mixed models. *Biometrical Journal* **61**, 860–872.
- PIMENTEL, E., ERBE, M., KÖNIG, S. & SIMIANER, H. (2013). Genome partitioning of genetic variation for milk production and composition traits in Holstein cattle. *Frontiers in Genetics* **28**, 135–167.
- R DEVELOPMENT CORE TEAM (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- SCHRECK, N., PIEPHO, H.-P. & SCHLATHER, M. (2019). Best prediction of the additive genomic variance in random-effects models. *Genetics* **213**, 379–394.
- SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. (1992). *Variance Components*. New Jersey: Wiley-Interscience.
- SEGURA, V., VILHJALMSSON, B. J., PLATT, A., KORTE, A., SEREN, Ü., QUANG, L. & NORDBORG, M. (2013). An efficient multi-locus mixed model approach for genome-wide association studies in structured population. *Nat Genetics* **44**, 825–830.
- THE 1001 GENOMES CONSORTIUM (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491.
- THE WELCOME TRUST CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **477**, 661–683.
- VALDAR, W., SOLBERG, L. C., GAUGUIER, D., BURNETT, S., KLENERMAN, P., COOKSON, W. O., TAYLOR, M. S., RAWLINS, J. N. P., MOTT, R. & FLINT, J. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet.* **38**, 879–887.
- VALDAR, W., SOLBERG, L. C., GAUGUIER, D., COOKSON, W. O., RAWLINS, J. N. P., MOTT, R. & FLINT, J. (2006). Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959–984.
- VISSCHER, P. M., MACGREGOR, S., BENYAMIN, B. & ET AL.(2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am. J. Hum. Genet.* **81**, 1104–1110.

- ZHANG, Z., ERSOZ, E., LAI, C.-Q. & ET AL.(2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* **42**, 355–360.
- ZHOU, X. & STEPHENS, M. (2012). Recovery of inter-block information when block sizes are unequal. *Nat Genet.* **44**, 821–824.