Estimation of Genetic Parameters

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-03-14

Contents

1 Introduction

The estimation of genetic parameters is an important issue in human genetics. First of all, estimating additive genetic and possible non-additive genetic variances contributes to a better understanding of genetic mechanisms. Secondly, estimates of genetic and phenotypic variances and covariances are essential for the prediction of genetic predisposition. Parameters that are of interest are heritability, genetic and phenotypic correlation and repeatability, and those are computed as functions of the variance components. Genetic parameters are estimated using information on phenotypes and genetic relationships for individuals in the study population. In this section we will illustrate how different phenotypic sources and genetic relationships are used for estimating genetic parameters.

1.1 Genetic model

The phenotype for a quantitative trait is the sum of both genetic and environmental factors. In general the total genetic effect for an individual is the sum of both additive and non-additive effects:

$$y = \mu + a + d + e \tag{1}$$

where μ is the population mean, a is the additive genetic effect, d is the dominance effect, and e is the environmental deviation (or residual) not explained by the genetic effects in the model. Only the additive genetic effects are passed on to the child. In contrast non-additive genetic effects (dominance and epistasis) are degraded by recombination and are not inherited, even though they may be important for the individual's phenotype. Here we only consider the additive and dominance effects. We assume that the genetic effects (i.e., a, and d), and the residual term, e, are independent and normally distributed:

$$\begin{aligned} a &\sim N(0, \sigma_a^2), \\ d &\sim N(0, \sigma_d^2), \\ e &\sim N(0, \sigma_e^2), \end{aligned}$$

where σ_a^2 is the additive genetic variance, σ_d^2 is the dominance variance, and σ_e^2 is the residual variance. This means that the observed phenotype is also normally distributed $y \sim N(\mu, \sigma_y^2)$ with the overall phenotypic variance:

$$\sigma_y^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$$

1.2 Genetic parameters

Heritability and genetic correlation are the key genetic parameters used in estimation of genetic predisposition. They are defined in terms of the variance components (σ_a^2 and σ_e^2) presented in the previous section.

Heritability estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population. It measures how much of the variation of a trait can be attributed to variation of genetic factors, as opposed to variation of environmental factors. The narrow sense heritability is the ratio of additive genetic variance (σ_a^2) to the overall phenotypic variance:

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_d^2 + \sigma_e^2) \tag{2}$$

and the broad sense heritability is the ratio of overall genetic variance ($\sigma_g^2 = \sigma_a^2 + \sigma_d^2$) to the overall phenotypic variance:

$$H^2 = (\sigma_a^2 + \sigma_d^2)/(\sigma_a^2 + \sigma_d^2 + \sigma_e^2) \tag{3}$$

A heritability of 0 implies that no genetic effects influence the observed variation in the trait, while a heritability of 1 implies that all of the variation in the trait is explained by the genetic effects. In general the amount of information provided by the phenotype about the genetic predisposition is determined by the narrow sense heritability. Note that heritability is population-specific and a heritability of 0 does not necessarily imply that there is no genetic determinism for the trait. The trait might be highly influenced by genetic factors. Yet, the observed variation for the trait might not be due to genetic factors, because all alleles contributing to the trait are fixed, and there are no segregating causal alleles for the trait, in the population. Therefore, observed variation may only be due to environmental factors, and the heritability in that population might be 0.

Genetic correlation is the proportion of variance that two traits share due to genetic causes. Genetic correlations are not the same as heritability, as it is about the overlap between the two sets of influences and not the absolute magnitude of their respective genetic effects; two traits could be both highly heritable but

not be genetically correlated, or they could have small heritabilities and be completely correlated (as long as the heritabilities are non-zero). Genetic correlation (ρ_a) is the genetic covariance between two traits divided by the product of genetic standard deviation for each of the traits:

$$\rho_{g_{12}} = \frac{\sigma_{g_{12}}}{\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}}} \tag{4}$$

where $\sigma_{g_{12}}$ is the genetic covariance and $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ are the genetic variances for the two traits in the population. A genetic correlation of 0 implies that the genetic effects on one trait are independent of the other, while a correlation of 1 implies that all of the genetic influences on the two traits are identical. Thus in order to estimate the heritability and genetic correlation we need to estimate the variance component defined above (σ^2_g and σ^2_e), for each trait, in addition to the genetic covariance between traits.

1.3 Data required for estimating genetic parameters

Information on phenotypes and genetic relationships for individuals in a study population are, in combination with appropriate statistical models, used for accurate estimation of genetic parameters and genetic predisposition of individuals.

Phenotypes for traits of importance for human health and disease need to be recorded accurately and completely.If individuals are selectively recorded (e.g. case-control study) the genetic parameters estimated should be adjusted. Data should include factors that could influence the trait phenotype. Observations should be objectively measured, if at all possible.

Genetic relationships for the individuals in the study population are required. Genetic relationships can be inferred from a pedigree or, alternatively, computed from genetic markers. Individuals and their parents need to be uniquely identified in the data.

Information about development (e.g., birth dates), ancestry, and genotypes for various markers could also be stored. If individuals are not uniquely identified, then genetic analysis of the population may not be possible at the individual level.

Prior information about the traits is useful. Read the literature. Most likely other researchers have already made analyses of the same traits. Even though their study populations are not the same as yours, their models could be useful starting points for further analyses. Their parameter estimates could result in useful predictions. The idea is to avoid the pitfalls and problems that other researchers have already encountered.

1.4 Statistical models and variance components

For estimating genetic parameters we need to specify a statistical model that describes the genetic and non-genetic factors that may affect the trait phenotypes. Often the non-genetic factors are referred to as systematic effect such as age, sex, or year:

$$\text{phenotype} = \text{mean} + \text{systematic effect} + \text{genetic effect} + \textit{residual}$$

Here we make a distinction between fixed effects, that determine the level (expected mean) of observations, and random effects that determine variance. A model consists of at least one fixed effect (i.e. mean) and one random effect (the residual error variance). If observations also are influenced by a genetic contribution of the individuals, then a genetic variance component exists as well. In that situation, we have two components contributing to the total variance of the observations: a genetic and a residual variance component.

The statistical model is a formal representation of our quantitative genetic theory, but it is important to realize that all models are simple approximations of how genetic and non-genetic factors influence a trait. The goal of the statistical analysis is to find the best practical model that explains the most variation in the data. Statistical knowledge is required. The methods used for estimating genetic parameters is based on statistical concepts such as random variables, matrix algebra, multivariate normal theory, and linear (mixed) models. These concepts and their use will be explained in the following sections.

1.5 When to estimate variance components

In general, the estimation of variance components has to be based on a sufficient amount of data. Depending on the data structure and measurements, estimations can be based on hundreds (in selection experiments) or more than 10,000 observations (in field recorded data). Importantly, in cases where the data set is small, the information from the literature may yield more accurate estimates of variance components. In general, we have to estimate variance without external information if we study a new trait, for which no prior parameter estimates are available, or a different sample: variances and covariances might have changed over time, or due to various evolutionary forces (genetic drift, selection, migration, or mutation).

Generally, it is assumed that variances and covariances, and especially their ratio (like heritability, correlation) do not rapidly change over time. However, it is well known that the genetic variance changes as a consequence of selection or genetic drift. Changes are expected, especially when generation intervals are short, selection intensity is high, or the trait under selection is determined by few causal genes with large effects. Moreover, the circumstances under which measurements are taken can change. If measurement conditions are better controlled, and getting more uniform over time, the environmental variance decreases, and consequently the heritability increases. Finally, the biological basis of a trait may change from one environment to another. In conclusion, there are sufficient reasons for regular estimation of (co-)variance components.

1.6 Methods for estimation of genetic parameters

In general, estimation of heritability and genetic correlation are based on methods that determine resemblance between genetically related individuals. Close (compared to distant) relatives share more alleles and, if the trait is under genetic influence, they will therefore share phenotypic similarities. Methods for estimating heritability include parent-offspring regression, analysis of variance (ANOVA) for family data (e.g., half-sib/full-sib families) and restricted maximum likelihood (REML) analysis for a general pedigree. These methods are increasingly more complex, but they are also increasingly more flexible. While REML can analyze any type of relationships and structures, ANOVA can only analyze groups of individuals with similar relationships (e.g., half-sib, or full-sib families), and regression analysis can only analyze pairs of individuals with similar relationships (e.g., pairs of parent and respective offspring, or pairs of monozygotic twins).

2 Estimating genetic parameters for a general pedigree using REML

Genetic parameters are nowadays estimated using restricted maximum likelihood (REML) or Bayesian methods. These methods allow for estimation of genetic parameters using phenotypic information for individuals from a general pedigree (with arbitrary relationships among them). REML is based on linear mixed model methodology and uses a likelihood approach to estimate genetic parameters.

2.1 Linear mixed model:

The linear mixed model contains the observation vector for the trait(s) of interest (y), the 'fixed effects' that explain systematic differences in y, and the 'random effects' which capture unidentified factors affecting y, e.g., random genetic effects and random residual effects.

A matrix formulation of a general model equation is:

$$y = Xb + Za + e$$

where

y : is the vector of observed values of the trait(s),

b : is a vector of factors, collectively known as fixed effects,

a : is a vector of factors known as random additive genetic effects,

e : is a vector of residual terms, also random,

X : is a known design matrix that relates the elements of b to their corresponding element in y.

Z : is a known design matrix that relates the elements of a to their corresponding element in y.

The factors (or 'variables') which describe fixed and random effects, may be either continuous or categorical.

Continuous variables have (theoretically) an infinite range of possible values (e.g., body weight or height in humans).

Categorical variables fall in distinct categories (e.g., different years). These variables do not describe a gradient of values along a single axis, like height of individuals (values between 0 and "infinity"). Instead, they have distinct classes (or 'levels'), each of which has its own estimated effect.

In addition to continuous or categorical (factor), it is necessary to distinguish between **fixed** and **random** effects in the linear mixed model.

Fixed effect: If the number of levels of a categorical variables is small or limited to a fixed number, and inferences about that factor are going to be limited to that set of levels, and to no others, then its effects is usually fixed. In other words, if a new sample of observations is made (from a new experiment), and the same levels of that factor are in both samples, then the factor is usually fixed. Continuous variables are usually fixed too (but not always).

Random effect: If the number of levels of a categorical variable is large, then that factor may be random. If the inferences about that factor are going to be made for an entire population of levels, and if the levels of the factor are a sample from an infinitely large population, then that factor is usually random. In other words, if a new sample of observations are made (from a new experiment), and the levels are completely different between the two samples, then the factor is usually random.

2.2 Expectation and variance of variables in the linear mixed model:

In the statistical model (specified above) the random effects (a and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal distribution. In general terms the expectations of these random variables are:

$$\begin{aligned} E(y) &= E(Xb) + E(Za) + E(e) \\ &= Xb + 0 + 0 \\ &= Xb \end{aligned}$$

and the variance-covariance matrices are:

$$\begin{aligned} Var(a) &= G \\ Var(e) &= R \\ Var(y) &= G + R = V \end{aligned}$$

where G, R and V are square matrices of genetic, residual and phenotypic (co)variances among the individuals, respectively. For a single trait model these covariances can be written as:

$$\begin{aligned} Var(a) &= A\sigma_a^2 \\ Var(e) &= I\sigma_e^2 \end{aligned}$$

where A is the additive genetic relationship matrix. For a multiple trait model:

$$Var(a) = A \otimes V_a$$
$$Var(e) = I \otimes V_e$$

where V_a and V_e are square matrcies of genetic and residual (co)variances among traits, respectively. To simplify we have assumed one record per individual per trait and therefore the Z matrix reduces to an identity matrix which can be left out of the equation system.

2.3 Genetic relationships among individuals

Estimating narrow sense heritability using REML (similar to the parent-offspring regression and ANOVA methods) requires that the phenotypic covariance between related individuals can be expressed by their additive genetic relationship and the additive genetic variance (σ_a^2). Related individuals share more alleles and thus resemble each other (have correlated phenotypes, to an extent that depends on additive genetic relationships).

In general, the genetic covariance between individuals depends on the additive genetic relationship. Examples of different types of additive genetic relationships can be found in the table below. The additive genetic relationship (A_{ij}) between the various sources (j) and the individual itself (i) can be seen in the table below.

Relative	A_{ij}
Self	1.0
Unrelated	0
Mother	0.5
Father	0.5
Grandparent	0.25
Half-sib	0.25
Full-sib	0.5
Cousin	0.0625
Child	0.5
Twin(MZ/DZ)	1/0.5

The A matrix expresses the additive genetic relationship among individuals in a population, and is called the **numerator relationship matrix** A. The matrix A is symmetric and its diagonal elements A_{ii} are equal to $1 + F_i$ where F_i is the **coefficient of inbreeding** of individual i. F_i is defined as the probability that two alleles taken at random from individual i are identical by descent. As such, F_i is also the kinship coefficient of its parents (half their genetic relationship).

Each off-diagonal elements (A_{ij}) is the genetic relationship between individuals i and j. Multiplying the matrix A by the additive genetic variance σ_a^2 leads to the covariance among breeding values. Thus if a_i is the breeding value of individual i then

$$var(a_i) = A_{ii}\sigma_a^2 = (1 + F_i)\sigma_a^2 \tag{5}$$

2.3.1 Algorithm to compute the numerator relationship matrix A

The matrix A can be computed using a recursive method. In what follows the recursive method to compute the components of A is described. Initially, individuals in a pedigree are numbered from 1 to n and ordered such that parents precede their children. The following rules are then used to compute the components of A.

- If both parents s and d of individual i are known, then

  - the diagonal element A_{ii} corresponds to: $A_{ii} = 1 + F_i = 1 + \frac{1}{2}A_{sd}$ and
  - the off-diagonal element A_{ji} is computed as: $A_{ji} = \frac{1}{2}(A_{js} + A_{jd})$
  - because A is symmetric $A_{ji} = A_{ij}$

- If only one parent s of individual i is known and assumed unrelated to the mate
  - $A_{ii} = 1$
  - $A_{ij} = A_{ji} = \frac{1}{2}(A_{js})$

- If both parents are unknown
  - $A_{ii} = 1$
  - $A_{ij} = A_{ji} = 0$

2.3.1.1 Numeric Example We are given the following pedigree and we want to compute the matrix A using the recursive method described in 2.3.1.

Table 1: Example Pedigree To Compute Additive Genetic Relationship Matrix

Child	Father	Mother
3	1	2
4	1	NA
5	4	3
6	5	2

The first step of the computations of A are the numbering and the ordering of all the individuals. This is already done in the pedigree shown in Table 1. The components of A are computed row-by-row starting with A_{11}.

$$
\begin{aligned}
A_{11} &= 1 + F_1 = 1 + 0 = 1 \\
A_{12} &= 0 = A_{21} \\
A_{13} &= \frac{1}{2}(A_{11} + A_{12}) = 0.5 = A_{31} \\
A_{14} &= \frac{1}{2}A_{11} = 0.5 = A_{41} \\
A_{15} &= \frac{1}{2}(A_{14} + A_{13}) = 0.5 = A_{51} \\
A_{16} &= \frac{1}{2}(A_{15} + A_{12}) = 0.25
\end{aligned}
$$

The same computations are also done for all the other components of the matrix A. The final result for the matrix looks as follows

$$
A = \begin{bmatrix}
1 & 0 & 0.5 & 0.5 & 0.5 & 0.25 \\
0 & 1 & 0.5 & 0 & 0.25 & 0.625 \\
0.5 & 0.5 & 1 & 0.25 & 0.625 & 0.5625 \\
0.5 & 0 & 0.25 & 1 & 0.625 & 0.3125 \\
0.5 & 0.25 & 0.625 & 0.625 & 1.125 & 0.6875 \\
0.25 & 0.625 & 0.5625 & 0.3125 & 0.6875 & 1.125
\end{bmatrix}
$$

As a result, we can see from the components of the above shown matrix A that individuals 1 and 2 are not related to each other. Furthermore from the diagonal elements of A, it follows that individuals 5 and 6 are

inbred while individuals 1 to 4 are not inbred. Finally, we can see that different types of relationships were included in this data. In comparison, only two types of relationships could exist in regression and ANOVA analyses: unrelated (e.g., A_{ij}=0 between individuals from different families) or not (e.g., A_{ij}=0.5 between individuals from the same full-sib family).

2.4 Restricted Maximum Likelihood approach for variance component estimation

Restricted Maximum Likelihood is a method that is used to estimate the parameters (i.e. variance components σ_a^2 and σ_e^2) in the linear mixed model specified above. The general principle used in maximum likelihood methods is to find the set of parameters which maximizes the likelihood of the data, i.e., the probability of observations given the model and its parameter estimates $p(y|\hat{\theta})$. For a single trait model including additive genetic effects the vector of parameters can be specified as $\hat{\theta} = \hat{b}, \hat{\sigma}_a^2, \hat{\sigma}_e^2$.

It is useful to recall that the likelihood $L(\theta|y)$ may be any function of the parameters (θ) that is proportional to $p(y|\theta)$. Maximizing $L(\theta|y)$ leads to obtaining the most likely value of θ ($\hat{\theta}$) given the data y. The REML method was developed by Patterson and Thompson [1971] as an improvement of the standard Maximum Likelihood (ML). ML assumes that fixed effects are known without error which is in most cases false and, as consequence, it produces biased estimates of variance components (usually, the residual variance is biased downward). As a solution to this problem, REML estimators maximize only the part of the likelihood which does not depend on the fixed effects, and REML, by itself, does not estimate the fixed effects. There are no simple one-step solutions for estimating the variance components based on REML [Lynch and Walsh, 1998]. Instead, the partial derivatives of the likelihoods are inferred with respect to the variance components. The solutions to these involve the inverse of the variance-covariance matrix, which themselves includes the variance components, so the variance components estimates are non-linear functions of the variance components. It is therefore necessary to apply iterative methods to obtain the estimates.

2.4.1 Estimates of genetic parameters

From the REML estimate of the variance components estimates of the narrow sense heritability can easily be computed by

$$\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) \tag{6}$$

where the the hat ($\hat{}$) refers to estimators. Similar estimates for the genetic correlation (ρ_a) is the genetic covariance between two traits divided by the product of genetic standard deviation for each of the traits:

$$\hat{\rho}_{a_{12}} = \frac{\hat{\sigma}_{a_{12}}}{\sqrt{\hat{\sigma}_{a_1}^2 \hat{\sigma}_{a_2}^2}} \tag{7}$$

2.4.2 Statistical test of variance components and genetic parameters

The concept of likelihood also provides a framework for testing hypotheses regarding, for example, the variance components in the models. In particular, the so-called likelihood ratio tests are used to assess whether a reduced model fits the data better than a full model by comparing the likelihoods of the two models.

The LR test statistic can be derived by using the following formula:

$$T_{\mathrm{LRT}} = 2 \ln \left[\frac{L(\hat{\theta}|\mathbf{y})}{L(\hat{\theta}_r|\mathbf{y})} \right] = -2 \left[l(\hat{\theta}_r|\mathbf{y}) - l(\hat{\theta}|\mathbf{y}) \right] \tag{8}$$

where $l(\hat{\theta}|y)$ is the log-likelihood for the full model, and $l(\hat{\theta}_r|y)$ is the log-likelihood for the reduced model.

When the sample size is sufficiently large, the LR statistic is χ^2 distributed with κ degrees of freedom, where κ parameters that were free in the full model, have been assigned fixed values in the reduced.

A high likelihood ratio shows that the full model with two (different) variance components is better at explaining the observed phenotypic variance than the reduced model with only one variance component.

It is fundamental for the reduced model to be nested in the full model, otherwise this approach does not make any sense. When the REML procedure is used, it is also important for the two models being compared to have the same fixed effects, otherwise the two likelihoods are not comparable, as can be easily understood by looking at the concept of restricted likelihood.

2.5 Advantages of using REML for estimating genetic parameters

Although REML does not produce unbiased estimates, it is still the method of choice due to the fact that this source of bias is also present but higher in ML estimates [Lynch and Walsh, 1998].

REML requires that y have a multivariate normal distribution although various authors have indicated that ML or REML estimators may be an appropriate choice even if normality does not hold (Meyer, 1990).

REML can account for selection when the complete mixed model is used with all genetic relationships and all data used for selection is included (Sorensen and Kennedy, 1984; Van der Werf and De Boer, 1990).

There is obviously an advantage in using (RE)ML methods that are more flexible in handling several (overlapping) generations (and possibly several random effects). However, the use of such methods are "dangerous" in the sense we no longer need to think explicitly about the data structure. For example, to estimate additive genetic variance, we need to have a data set that contains a certain family structure which allows us to separate differences between families from differences within families. Or in other words, we need to differentiate genetic and residual effects, so the structure due to genetic relationships must be different from the structure due to residual effects (i.e., the G and R matrices must be different enough).

Today, heritability can be estimated based on genetic relationships, inferred from general pedigrees or estimated from genetic markers. Linear mixed models are also used in genetic evaluation, allowing information on all known relationships between individuals to be incorporated simultaneously in the analysis. Linear mixed models can include additional effects to describe the data more accurately: maternal, permanent environmental, cytoplasmic or dominance effects and QTL effects. These effects may be fitted as additional random effects.

References

Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, USA, 1998. ISBN 0-87893-481-2.

H. D. Patterson and R. Thompson. Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3):545–554, December 1971. ISSN 00063444. doi: 10.2307/2334389. URL http://www.jstor.org/stable/2334389.