Estimation of Genetic Parameters

Palle Duun Rohde, Izel Fourie Sørensen & Peter Sørensen

2022-09-27

Contents

1 Introduction to genetic parameters

Estimation of genetic parameters is an important discipline in studying the genetics of complex traits and multifactorial diseases. First of all, estimating additive genetic variance, and possible non-additive genetic variances, contributes to a better understanding of the underlying genetic mechanisms of complex phenotypes. Secondly, estimates of phenotypic and genetic variances, and their covariances, are essential for prediction of genetic predisposition, e.g., animal breeding values or an individuals genetic liability towards a common complex disease. Parameters that are of great interest are heritability, genetic and phenotypic correlation and repeatability, which all can be computed from the estimated variance components. Genetic parameters are estimated using information on phenotypes and genetic relationships among 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 The genetic model

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

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

where μ is the population mean, a is the additive genetic effect, d is the dominance effect, i is the interaction effect (*i.e.*, epistasis), 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 offspring. In contrast, the non-additive genetic effects (i.e., dominance and epistasis) are degraded by recombination and are not transmitted from generation to generation, even though they may be important for an 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 entails that the observed phenotype (y) 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 correlations are the key genetic parameters used in estimating an individuals' genetic predisposition to a complex trait. They are defined in terms of the variance components (σ_a^2, σ_d^2 and σ_e^2) as presented in the section above.

Heritability estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation among individuals in that specific population. It quantifies how much of the phenotypic variation can be attributed to either variation in genetic factors or variation in environmental factors.

The broad sense heritability (H^2) quantifies how much of the total phenotypic variation can be explained by the overall genetic variance ($\sigma_g^2 = \sigma_a^2 + \sigma_d^2$):

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

whereas the narrow sense heritability (h^2) measures the degree of phenotypic variation that is explained by the additive genetic variation (σ_a^2):

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

A heritability of 0 implies that no genetic effects contributes to 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. Importantly, the 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. The proportion of phenotypic variance explained by additive genetic variance also sets the limit for the accuracy of genetic predisposition.

If the phenotype is binary then heritability estimates might be biased. This is because the scale at which the phenotype is measured is different from the scale at which the heritability is expressed. Also, often the proportion of cases in the study population is larger than the disease prevalence in the general population leading to ascertainment bias. Therefore, the observed genomic heritability (h_o^2) can be transformed to the liability scale (h_l^2) [Lee et al., 2011]:

$$
h_l^2 = h_o^2 \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}, \tag{4}
$$

where K is the prevalence of the disease in the population, P is the proportion of cases in the sample, and z is the height of the standard normal curve at the truncation point, thus, $z = \phi(x) = e^{-0.5x^2}/\sqrt{(2\pi)}$.

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 the genetic standard deviation for each of the traits:

$$\rho_{g_{12}} = \frac{\sigma_{g_{12}}}{\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}}}, \tag{5}$$

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 components defined above (σ^2_g and σ^2_e) for each trait in addition to the genetic covariance between traits ($\sigma_{g_{12}}$).

1.3 Data required for estimating genetic parameters

Phenotypic observations and genetic relationships among individuals in the study population are, in combination with appropriate statistical models, used for accurate estimation of genetic parameters.

Phenotypes for any trait of interest, whether it is a trait of importance in plant breeding or of importance for human health, need to be recorded accurately and for the entire study cohort. If individuals are selectively recorded (e.g., a case-control study), the estimated genetic parameters should be adjusted (see section 1.2). Data should include factors that could influence the variation in the recorded phenotype, and observations should be objectively measured, if at all possible.

Prior information about the traits is useful. Read the literature. Most likely other researchers have already done 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.

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

Additional 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.

1.4 Statistical models

For estimating genetic parameters we need to specify a statistical model that describes the genetic and non-genetic factors that may affect the phenotypic variation. Often the non-genetic factors are referred to as systematic effects such as age, sex, or year. A general representation of the model is thus:

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

Here, we make a distinction between fixed effects that determine the level (expected mean) of observations, and random effects that determine the variance. A model consists of at least one fixed effect (i.e, the mean) and one random effect (the residual error variance). If observations are also influenced by a genetic effect, 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 most of the variation in the data. 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 Considering variance components

An accurate 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 to >10,000 observations. Importantly, in cases where the data set is small, information from the literature may feature more accurate estimates of variance components. When studying new traits, we have to estimate variance components without external information, or if the study cohort is very different from what is known from the literature as variances and covariances can change over time, e.g., due to various evolutionary forces, such as genetic drift, selection, migration, or mutation.

It is well known that 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. For example, if the conditions under which the measurements are taken are better controlled, and getting more uniform over time, the environmental variance decreases, and consequently the heritability increases. On a similar note, the genetic predisposition to a multifactorial disease is stable through life, however, the environmental exposure varies and may accumulate risk effects as we age, consequently meaning that the heritability for a trait may decrease with age. 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

Estimating heritability and genetic correlations 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 relationship and structure, 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 mono- and dizygotic twins).

2 Estimating genetic parameters for a general pedigree using REML

Genetic parameters are nowadays estimated using REML or Bayesian inference. 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 is:

$$y = Xb + Za + e, \tag{6}$$

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 variables, 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 variable 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 are 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.

2.1.1 Model parameters and assumptions

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 defined as:

$$Var(y) = Var(a) + Var(e)$$
$$Var(a) = A\sigma_a^2$$
$$Var(e) = I\sigma_e^2$$

where A is the additive genetic relationship matrix, and I is an identity matrix. The matrices $Var(y)$, $Var(a)$ and $Var(e)$ are square matrices of the phenotypic, genetic, residual (co)variances among the individuals, respectively.

2.1.2 Genetic relationship matrices

The genetic relationship can be (1) inferred from general pedigrees, e.g., a numerator relationship matrix, A, or (2) estimated from genetic markers e.g., a genomic relationship matrix, G. The main difference between the two types of genetic relationship matrices, A and G, is that A is based on the concept of identity by descent (sharing of the same alleles, transmitted from common ancestors) whereas G is based on the concept of identity by state (sharing of the same alleles, regardless of their origin).

2.1.3 Advanced linear mixed models

The linear mixed model framework is very flexible and can be expanded in a number of ways. First, additional effects can be included to describe the data more accurately: maternal, permanent environmental, or QTL effects. These effects may be fitted as additional random effects. Second, non-additive genetic effects can be included in the model such as dominance (i.e., modeled by a dominance relationship matrix, D) or additive by additive interaction effects (i.e., modeled by an epistatic relationship matrix, $A \circ A$). Third, the multiple genetic variance components may be defined by partitioning the total genetic variance using genomic relationship matrices computed from different marker sets defined by different functional marker categories. Finally, multiple trait models can be fitted:

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

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

2.2 Estimating variance components using REML

Restricted Maximum Likelihood, or REML, is a method that is used to estimate the variance components (σ_a^2 and σ_e^2) in the linear mixed model specified in Section 2.1. The general principle used in maximum likelihood methods is to find the set of parameters ($\hat{\theta}$) 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 parameter estimates can be specified as $\hat{\theta} = \hat{b}$, $\hat{\sigma}_a^2$, $\hat{\sigma}_e^2$ (where the *hat* refers to that these are estimates of parameters).

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 false in most cases, and as a 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 component's estimates are non-linear functions of the variance components.] It is therefore necessary to apply an iterative method to obtain the estimates, $\hat{\theta}$.

2.2.1 Estimates of genetic parameters

From the REML estimate of the variance components the narrow sense heritability can easily be estimated:

$$\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2), \tag{7}$$

where the hat ($\hat{}$) indicates that the value is an estimate, and therefore not the true value. 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:

$$\hat{\rho}_{a_{12}} = \frac{\hat{\sigma}_{a_{12}}}{\sqrt{\hat{\sigma}_{a_1}^2 \hat{\sigma}_{a_2}^2}} \tag{8}$$

2.2.2 Statistical test of variance components

The concept of likelihood also provides a framework for testing hypotheses regarding the variance components in the models. In particular, the so-called likelihood ratio (LR) test is 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 (T_{LRT}) can be derived by using the following formula:

$$T_{\mathrm{LRT}} = 2 \ln \left[\frac{L(\hat{\theta}|\mathbf{y})}{L(\hat{\theta}_r|\mathbf{y})} \right] = -2 \left[l(\hat{\theta}_r|\mathbf{y}) - l(\hat{\theta}|\mathbf{y}) \right], \tag{9}$$

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

For example, 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 within 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.

2.2.3 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 could be "dangerous" in the sense that 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, and therefore 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).

References

S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, 88:294–305, 2011.

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.