# Robust heritability estimation in plant studies

Vanda M. Lourenço*
FCT & CMA, NOVA University of Lisbon, Portugal - vmml@fct.unl.pt

Miguel Fonseca
CMA, NOVA University of Lisbon, Portugal - fonsecamig@fct.unl.pt

Ana M. Pires
IST & CEMAT, University of Lisbon, Portugal - apires@math.ist.utl.pt

Paulo C. Rodrigues
Federal University of Bahia, Brazil - paulocanas@gmail.com

## Abstract

Heritability is of major importance in plant studies to help achieve better yield and other agronomic traits of interest. In candidate gene studies regression models are used to test for associations between phenotype and candidate single nucleotide polymorphisms (SNPs). SNP imputation guarantees that marker information is complete and so both the coefficient of determination, $R^2$, and broad-sense heritability are equivalent. However, when the normality assumption is violated, the classical $R^2$ may be seriously affected. Recently two $R^2$ alternatives with good properties were proposed for the linear mixed model: a marginal $R^2_m$ for the variance explained by the fixed factors and a conditional $R^2_c$ for the variance explained by both the fixed and random factors. In this work we step forward a robust version of $R^2_c$ and assess the adequacy of both classical and robust counterparts in the estimation of true broad-sense heritability via simulation, where a particular contamination scenario is considered. An example of application with a real maize data set is also presented.

**Keywords**: Robust linear mixed model; Coefficient of determination; Single nucleotide polymorphism (SNP); Heritability estimation.

## 1. Introduction

The extent to how much a certain phenotype is genetically determined, i.e., its heritability, can be inferred from traditional crossing experiments, either breeding trials or pedigree-based approaches, or more recently, from genomic-based approaches (Staton-Geddes et al., 2013). Once a trait is known to be high heritable then association studies are performed genome-wide so that the single nucleotide polymorphisms (SNPs) underlying those traits variation may be found. Candidate gene studies have a useful role in validating the results obtained from genome-wide association studies. Here, regression models are used to test for associations between phenotype and every candidate SNP via single-SNP and/or multiple-SNP analysis. These data are collected population wide and missing SNP values are usually imputed so that marker information is complete. Hence, the usual coefficient of determination and broad-sense heritability are equivalent (Piepho & Möhring, 2007). However, when data violates the normality assumption, results from the association analysis may be biased. Moreover, the classical coefficient of determination can be seriously affected.

Robust procedures are designed so that the effects of outlying observations on parameter estimation, testing procedures and goodness-of-fit measures may be minimized (Lourenço et al., 2011; Lourenço & Pires, 2014) but although throughout the years several robust alternatives to the classical $R^2$ have been placed in the literature in the context of the linear regression model (Rousseeuw, 1984; Anderson, 1994; Croux & Dehon, 2003; Maronna et al., 2006), in the context of the linear mixed model there hasn't been a consensus as to what a "good" coefficient of determination should be. Recently Nakagawa & Schielzeth (2013) proposed a marginal coefficient of determination, $R^2_m$, for the variance explained by the fixed factors and a conditional one, $R^2_c$, for the variance explained by both the fixed and random factors, that were shown to have good desirable properties. One of those properties is that these coefficients may be computed even if REML estimation has been used.

In this work, after a brief description of the methods (Section 2) we assess the performance of $\mathrm{R}_c^2$ in the estimation of true broad-sense heritability when there is violation of the normality premise and step forward a robust counterpart of this coefficient. Both approaches are compared via simulation where a particular contamination scenario is considered (Section 3). Finally, we present an example of application with a real maize data set (Section 4) and discuss the usefulness of the proposed robust methodology (Section 5).

## 2. Methods

**The linear mixed model (LMM).**
We consider the linear mixed effects model proposed by Yu et al. (2006) for genetic association studies of quantitative traits

$$\mathbf{y} = X\boldsymbol{\beta} + I\mathbf{u} + \boldsymbol{\varepsilon} \tag{1}$$

where: $\mathbf{y} \in \mathcal{M}_{n\times 1}$ is the vector of observed values; $X \in \mathcal{M}_{n\times p}$ with full column rank $p$; $\boldsymbol{\beta} = (\beta_0, ..., \beta_{p-1}) \in \mathcal{M}_{p\times 1}$ and $\mathbf{u} \in \mathcal{M}_{n\times 1}$ are the correspondent fixed and random effects with $\mathbf{u} \sim N(\mathbf{0}, 2\sigma^2 K)$, $K$ is a kinship matrix; $\boldsymbol{\varepsilon} \in \mathcal{M}_{n\times 1}$ is a vector of residual deviations with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_e^2 I)$. Both $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent and so $cov(\mathbf{u}, \boldsymbol{\varepsilon}) = \mathbf{0}$ and, taking $\sigma_*^2 = \frac{\sigma^2}{\sigma_e^2}$,

$$var(\mathbf{y}) = \sigma^2 A + \sigma_e^2 I = \Phi = \sigma_e^2\Big(\frac{\sigma^2}{\sigma_e^2} 2K + I\Big) = \sigma_e^2\Big(\sigma_*^2 2K + I\Big) = \sigma_e^2 \Phi^*.$$

The vector of unknown variance parameters of $\Phi$ thus becomes $(\sigma_*^2, \sigma_e^2)$ and hence $\mathbf{y} \sim N(X\beta, \sigma_e^2\Phi^*)$. Additionally, $X\beta = \text{intercept} + P\alpha + S\gamma$ where $P$ is a matrix derived from principal component analysis (to help correct for population structure) and $S$ is a matrix of genetic markers, in this case, SNPs.

**Robust estimation in the LMM.**
Estimation in the LMM ($\Phi^*$ unknown; $\Phi^*$ semi-definite positive) can be done via derivative free methods, e.g., by maximizing the profile restricted maximum log-likelihood (REML). The maximization of the REML profile-log-likelihood, or minimization of minus the profile-log-likelihood, is done by an iterative process and as such, in each step of the process $\Phi^*$ is fixed. The robust estimators will be plugged-in in the derivative-free REML estimation of the parameters of model (1), which we call robust-derivative-free REML estimation, in the following way:

1. We begin by considering a transformation of model (1) via a $n \times \big(n-p\big)$ matrix $U$ such that $U^T X\beta = 0$[1] and thus
$$U^T\mathbf{y} = U^T(Iu + \boldsymbol{\varepsilon}) \tag{2}$$
where $var(U^T\mathbf{y}) = var(U^T(Iu+\boldsymbol{\varepsilon})) = \sigma_e^2 U^T\Phi^* U = \sigma_e^2\Phi^\circ$; $E(U^T\mathbf{y}) = \mathbf{0}$; $\Phi^\circ = U^T\Phi^* U$ is semi-definite positive. Hence $U^T\mathbf{y} \sim N(\mathbf{0}, \sigma_e^2\Phi^\circ)$.

2. Taking the inverse of the Cholesky decomposition of $\Phi^\circ$ as $\Delta = (chol(\Phi^\circ))^{-1}$, equation (2) re-writes to
$$Y^* = \boldsymbol{\varepsilon}^*$$
where now $Y^* = \Delta U^T\mathbf{y}$, $\boldsymbol{\varepsilon}^* = \Delta U^T(Iu + \varepsilon)$; $var(Y^*) = var(\boldsymbol{\varepsilon}^*) = \sigma_e^2 I$ and $E(Y^*) = \mathbf{0}$. Thus, $Y^* \sim N(\mathbf{0}, \sigma_e^2 I)$.

3. At this point, $\sigma_e^2$ may be estimated as some robust estimate of scale, e.g., through $Q_n(\Delta U^T Y)^2$, $MAD(\Delta U^T Y)^2$, $hubers(\Delta U^T Y)^2$ or other (see Huber, 1981; Venables & Ripley, 2002). Noting that the profile log-likelihood writes as
$$l_P(\sigma_*^2|\sigma_e^2 = \hat{\sigma}_e^2, \mathbf{y}) = \mathbf{c} - \frac{n-p}{2}\log(\hat{\sigma}_e^2) - \frac{1}{2}\log|U^T\Phi^* U|, \tag{3}$$
the robust estimate of scale $\widehat{\sigma}_e^2$ is plugged in (3) and an estimate $\widehat{\sigma}_*^2$ obtained through an optimization process.

[1] $U$ is such that: $UU^T = I - P$, $P$ the orthogonal projection matrix on the range space of X; $U^T U = I$.

4. Having $\widehat{\sigma}^2_*$, an estimate of the variance-covariance matrix $\Phi^*$ is computed and a final estimate of the residual variance, $\widehat{\sigma}^2_e$, taken through $Q_n(\widehat{\Delta}U^TY)^2$, $MAD(\widehat{\Delta}U^TY)^2$ or $hubers(\widehat{\Delta}U^TY)^2$. The genetic unknown variance is then estimated by $\widehat{\sigma}^2 = \widehat{\sigma}^2_* \times \widehat{\sigma}^2_e$.
5. Once the variance components have been robustly estimated, the parameters $\beta$ are then estimated by a robust fit of model
$$\Delta^*Y = \Delta^*X\beta + \boldsymbol{\varepsilon}_0$$
where $\Delta^*$ is a matrix such that $\Delta^*\Phi^*(\Delta^*)^T = I$ and $\boldsymbol{\varepsilon}_0 \sim N(\mathbf{0}, \sigma^2_e I)$.

**A robust coefficient of determination for the LMM.**
The conditional coefficient of determination defined by Nakagawa & Schielzeth (2013) rewrites in the case of model (1) to

$$R^2_c = \frac{\sigma^2_f + \sigma^2}{\sigma^2_f + \sigma^2 + \sigma^2_e}$$

with $\sigma^2_f$ to be estimated by $\hat{\sigma}^2_f = var(X\hat{\beta})$. In the case of the robust fit of model (1) we may consider $\hat{\sigma}^2_f = MAD(X\hat{\beta})^2$, $\hat{\sigma}^2_f = Q_n(X\hat{\beta})^2$, $\hat{\sigma}^2_f = hubers(X\hat{\beta})^2$ or other, depending on the robust estimate of scale already used in (4.). $R^2_c$ can then be adjusted in the usual way.

**Broad-sense heritability and the coefficient of determination.**
Heritability is the proportion of phenotypic variation in a population that is explained by the genetic variation among individuals. Although there are two formulae for this concept we will focus on broad-sense heritability only:

$$H^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e}$$

where $\sigma^2_g$ and $\sigma^2_e$ are the variance components associated with the genetic effects and the residual error, respectively. The residual error may include undetected genetic effects, environmental effects, gene-environment interaction effects and the random effects. $H^2$ relates to $R^2_c$ in the sense that $\sigma^2_g = \sigma^2_f + \sigma^2$.

## 3. Simulation

**Simulation settings and model.**
The following data and settings are considered for the simulation study:

- the maize data of Zhang *et al.* (2010): one SNP data set(matrix $S$); 5 Principal Components derived for this SNP data set (matrix $P$); a Kinship matrix $K$ also derived for these data;
- 10 SNPs contributing equally to the quantitative trait/phenotype;
- trait heritabilities $H^2$ of $0.15,\ 0.3,\ 0.5$ & $0.75$ (assuming no epistatic effects; $\sigma^2 = 1$);
- 1% 5-unit shift-outlier contamination taken from a normal distribution;
- 1000 replications.

The phenotypic values are generated according to model (no intercept included)

$$Y \sim P\alpha + S\beta + I\mathbf{u} + \varepsilon$$

where $\varepsilon_i \sim N(0, (\sigma^2_f + \sigma^2)\frac{1-H^2}{H^2})$ with $\sigma^2_f = var(P\alpha + S\beta)$ and $\sigma^2 = 1$; $\mathbf{u} = Z\mathbf{u}^*$ with $u^*_i \sim N(0, 1)$ and $Z = PD^{1/2}$ with $P$ and $D$ the singular value decomposition matrices of $K$ (i.e, $K = PDP^T$ with $P$ orthogonal matrix and $D$ diagonal matrix of K's eigenvalues). $Y$ is then standardized so that contamination may be drawn from a $N(5, 1)$ distribution.

Classical analysis is performed using the R routine `lmekin()` from the R package *coxme* where the derivative-free REML method is implemented. The robust derivative-free plug-in REML method was programmed by the authors, also in the R language. We also consider the *hubers* robust estimate of scale, available in the R package *MASS* via routine `hubers()`.

**Simulation results.**

Figure 1 shows that under the null hypothesis of no data contamination and as expected, the performance of the classical methodology produces conditional $R_c^2$ boxplots symmetric and centered on the mean for all the heritabilities considered in the simulations. Also as expected, the robust approach produces results that are close to the classical ones with boxplots that are also symmetric and more-or-less centered on the mean.

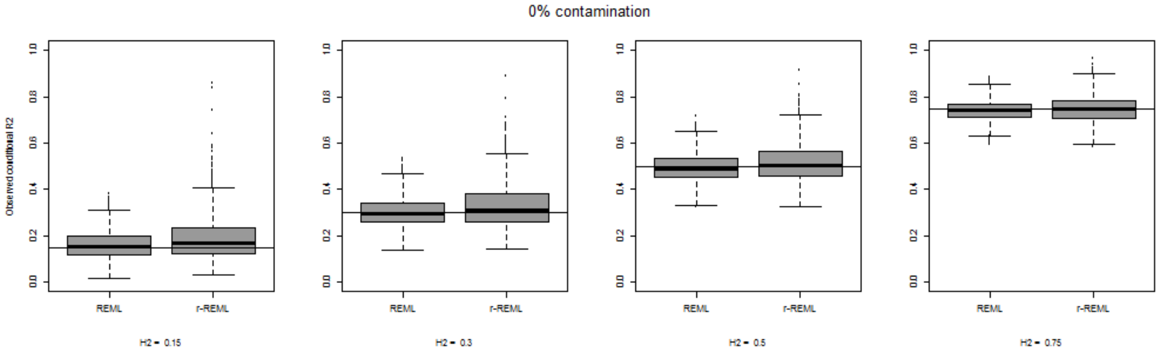

Figure 1: Boxplots for 1000 observed classical (REML) and robust (r-REML) r-squared conditional values.

Under the alternative hypothesis 1% 5-unit shift-outlier contamination was placed in the data by replacing 1% of good observations. Here, Figure 2 shows that the classical approach is no longer able to accurately estimate true broad sense heritability (0.15 case excluded) whereas the robust counterpart is still able to deliver good approximations although some underestimation is already seen in the case of $H^2 = 0.75$.

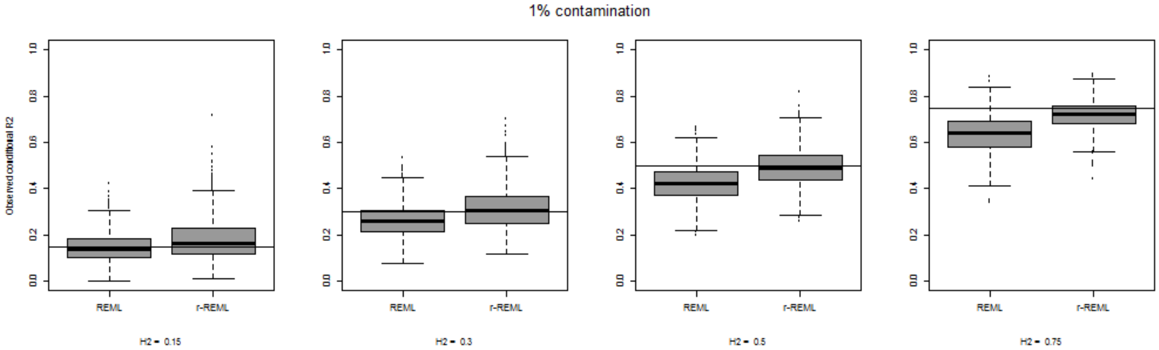

Figure 2: Boxplots for 1000 observed classical (REML) and robust (r-REML) r-squared conditional values.

## 4. Real data example.

We consider the maize data of Weber *et al.* (2008) on

- 493 plants
- 1 quantitative trait (FERL: female hear length)
- 61 candidate SNPs
- a kinship matrix $K$
- 10 PCs

Both classical and robust fits of model (1) are adjusted for these data one SNP at a time (Single-SNP analysis). Significant SNPs from the Single-SNP analysis are then considered in the final multiple model (Multiple-SNP analysis). See Lourenço *et al.* (2011) for more detail on association testing.

In the classical analysis, 8 SNPs are declared significant associations with FERL in the single-SNP analysis (6 agreeing with the ones indicated by Weber *et al.*) from which 6 remain significant in the multiple-SNP analysis (4 of the 6 signalized by Weber *et al.*). Here, we have an adjusted $R_c^2 = 0.34$.

In the robust approach, 7 SNPs are declared significant associations with FERL in the single-SNP analysis (4 agreeing with the ones indicated by Weber *et al.*) from which 3 remain significant in the multiple-SNP analysis (1 of the 6 signalized by Weber *et al.*). Here, we have an adjusted $R_c^2 = 0.39$.

The proximity between both coefficients indicates that the data is not highly contaminated, which is in agreement with the conclusions of Lourenço & Pires (2014) where one to two outliers were detected, amounting to 0.5% contamination (369 observations used in the multiple regression model), which is also in line with the results from the simulations where we can see, for heritabilities between 0.3 and 0.5, a tendency of the classical $R_c^2$ to underestimate the true value when contamination is present (Figure 2).

## 5. Conclusions

In this work we have presented an alternative plug-in robust derivative-free REML alternative to the classical counterpart. Preliminary simulation results, where a particular scenario of mild non-severe contamination was considered, showed that the classical REML completely derails in the estimation of true broad-sense heritability when contamination is present, whereas the robust approach is still able to produce reasonable results. We therefore conclude that, despite this robust proposal is not yet optimal due to some underestimation for higher levels of trait heritability, it is undeniably better than the classical derivative-free REML approach. Notwithstanding, other robust approaches will have to be explored since it is expected that for higher percentages of contamination the robust plug-in method will also fail in accurately estimating $H^2$.

### Acknowledgements

The authors acknowledge financial support from the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects PTDC/MAT-STA/0568/2012 and PEst-OE/MAT/UI0297/2014 (CMA - Centro de Matemática e Aplicações).

### References

Croux, C. and Dehon, C. (2003). Estimators of the multiple correlation coefficient: local robustness and confidence intervals. *Statistical Papers* **44(3)**, 315–334.

Huber, P. J. (1981). Robust Statistics. Wiley.

Lourenço, V. M., Pires, A. M. and Kirst, M. (2011). Robust linear regression methods in association studies. *Bioinformatics* **27(6)**, 815–821.

Lourenço, V. M. and Pires, A. M. (2014). M-regression, false discovery rates and outlier detection with application to genetic association studies. *Journal of Computational Statistics and Data Analysis* **78**, 33–42.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust statistics, theory and methods.* Chichester: Wiley.

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**, 133–142.

Piepho, A.H. and Möhring, J. (2007). Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials *Genetics* **177**, 1881–1888.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880.

Staton-Geddes, J., Yoder, J. B., Briskine, R. and *et al.* (2013). Estimating heritability using genomic data. *Methods in Ecology and Evolution* **4**, 1151–1158.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Fourth edition. Springer.

Weber, A. L., Briggs, W. H., Rucker, J. *et al.* (2008). The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. *Genetics* **(180)**, 1221–1232.

Yu, J., Pressoir, G., Briggs, W. H. *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **(38)**, 203–208.

Zhang, Z., Ersoz, E., Lai, C. Q. *et al.* (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–360.