Robust heritability and predictive accuracy estimation in plant breeding

Vanda M. Lourenço[1], Hans-Peter Piepho[2] and Joseph O. Ogutu[2]

[1]DM & CMA, FCT, Universidade NOVA de Lisboa; [2] Biostatistics Unit, Institute of Crop Science, Hohenheim Universität

[1]Contact Email: vmml@fct.unl.pt

Introduction & Aim

In this work, we are interested in two recently proposed methods for the estimation of heritability (H^2; Method 5 only) and predictive accuracy (PA; Methods 5 and 7; Estaghvirou et.al. [BMCGenomics13]) which are both founded on the linear mixed effects model as well as on ridge regression best linear unbiased prediction through a two-stage approach (Piepho [CropSci09]; Piepho et.al.[CropSci12]). This means, that estimates of H^2 and PA are likely to be adversely affected by the presence of outlying observations in the phenotypic data. Here, we propose a robust LMM approach for the 1st stage of the two-stage approach (phenotypic analysis) where adjusted genotypic means are computed, and compare the performances of both approaches in the estimation of the parameters of interest.

Materials & Methods

The two-stage approach of Piepho et.al. that is used to predict true breeding values ($\mathbf{g}$) that are then used to estimate H^2 and PA proceeds as follows:

1st Stage. LMM (1) is used to estimate the adjusted means, $\widehat{\boldsymbol{\mu}}$, for the testcross genotypes that will then be submitted to the 2nd stage.

2nd Stage. LMM (2) is used in a ridge-regression formulation to compute the predicted breeding values $\widehat{\mathbf{g}}$, i.e., BLUP($\mathbf{g}$) $=\widehat{\mathbf{g}}$.

$$\mathbf{y}=\mathbf{X}\boldsymbol{\mu}+\mathbf{f} \qquad (1)$$

phenotype = intercept + genotype + replicates + blocks within replicates + error

$$\widehat{\boldsymbol{\mu}}=\phi\mathbf{1}+\mathbf{g}+\mathbf{e} \qquad (2)$$

estimated adjusted means = general mean + breeding values + + error

Method 5**.** This method calculates PA as

$$E(r_{g,\widehat{g}}) \approx \frac{trace(\mathbf{P}_u\mathbf{C}\mathbf{G})}{\sqrt{trace(\mathbf{P}_u\mathbf{G})trace(\mathbf{C}^T\mathbf{P}_u\mathbf{C}\mathbf{V})}} \qquad (3)$$

and H^2 as $H^2_{m_5}=E(r_{g,\widehat{g}})^2$. Here, $\mathbf{G}=\mathbf{Z}\mathbf{Z}^T\sigma^2_u$, $\mathbf{R}=\sigma^2_e\mathbf{I}$ and $\mathbf{V}=\mathbf{G}+\mathbf{R}$, with $\mathbf{Z}$ a matrix of biallelic markers (single nucleotide polymorphisms).

Method 7**.** This method is commonly used by animal breeders to directly compute PA from LMM equations:

$$PA=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{\rho}^{\,2}_i} \text{ where } \rho^2_i=\frac{(cov(g_i,\widehat{g}_i))^2}{var(g_i)var(\widehat{g}_i)} \qquad (4)$$

The robust approach. The robust analogue of this method considers in the 1st stage that $\boldsymbol{\mu}$ are estimated via a robust LMM (Koller, M. [PhDthesis2013]). Here, a derivation of the classical log-likelihood is considered and an objective function that contains the observation level residuals and the random effects as separate terms is obtained. A system of score equations follows and bounded influence functions ψ are applied to both the residual and random effects terms. Having robustly estimated the $\widehat{\boldsymbol{\mu}}$ values, these are now carried to the 2nd stage as before and the method proceeds in the usual way with PA and H^2 estimated by Methods 5. and 7. above.

Maize Dataset & Simulation

Dataset. KWS-Synbreed maize dataset (Project 2009/15) extracted for one location, 698 genotyped testcrosses & 11646 SNP markers. Variance components estimated from this dataset were used to simulate true breeding values and phenotypic data assuming that the 698 genotypes are correlated, $\sigma^2_e=53.87$ and $\sigma^2_u=0.0059$. **Contamination settings.** (I) 1, 3 & 5% of phenotypic contamination; (II) 1 and 2 whole block contamination. Good observations are replaced by their observed value + 5-, 8- or 10- times σ_e.

Notation: **I** 1_5, 1_8, 1_10, 3_5, 3_8, 3_10, 5_5, 5_8 & 5_10; **II** 1_5, 1_8, 1_10, 2_5, 2_8 & 2_10.

Results under H$_0$

Observed MSD between the estimated adjusted means using both approaches (**CLa**S**sical** & **ROB**ust)

$$\text{MSD}_\mu=\sum_{j=1}^{1000}\sum_{i=1}^{698}\frac{(\widehat{\mu}^{ROB}_{ij}-\widehat{\mu}^{CLS}_{ij})^2}{698\times 1000}\simeq 0.0616. \qquad (5)$$

Results under H$_1$

Scenarios		CLS	ROB
I (%)	1_5	3.13	3.16
	1_8	8.00	7.93
	1_10	12.48	12.32
	3_5	9.28	9.33
	3_8	23.72	23.63
	3_10	37.03	36.84
	5_5	15.54	15.65
	5_8	39.66	39.59
	5_10	61.85	61.67

Scenarios		CLS	ROB
II (block)	1_5	1.22	0.43
	1_8	1.89	0.67
	1_10	2.32	0.81
	2_5	2.10	0.56
	2_8	3.69	0.84
	2_10	5.01	0.96

Table 1. Observed MSD between the estimated $\widehat{\boldsymbol{\mu}}$ and benchmark $\boldsymbol{\mu}^{\text{CLS}}_{null}$

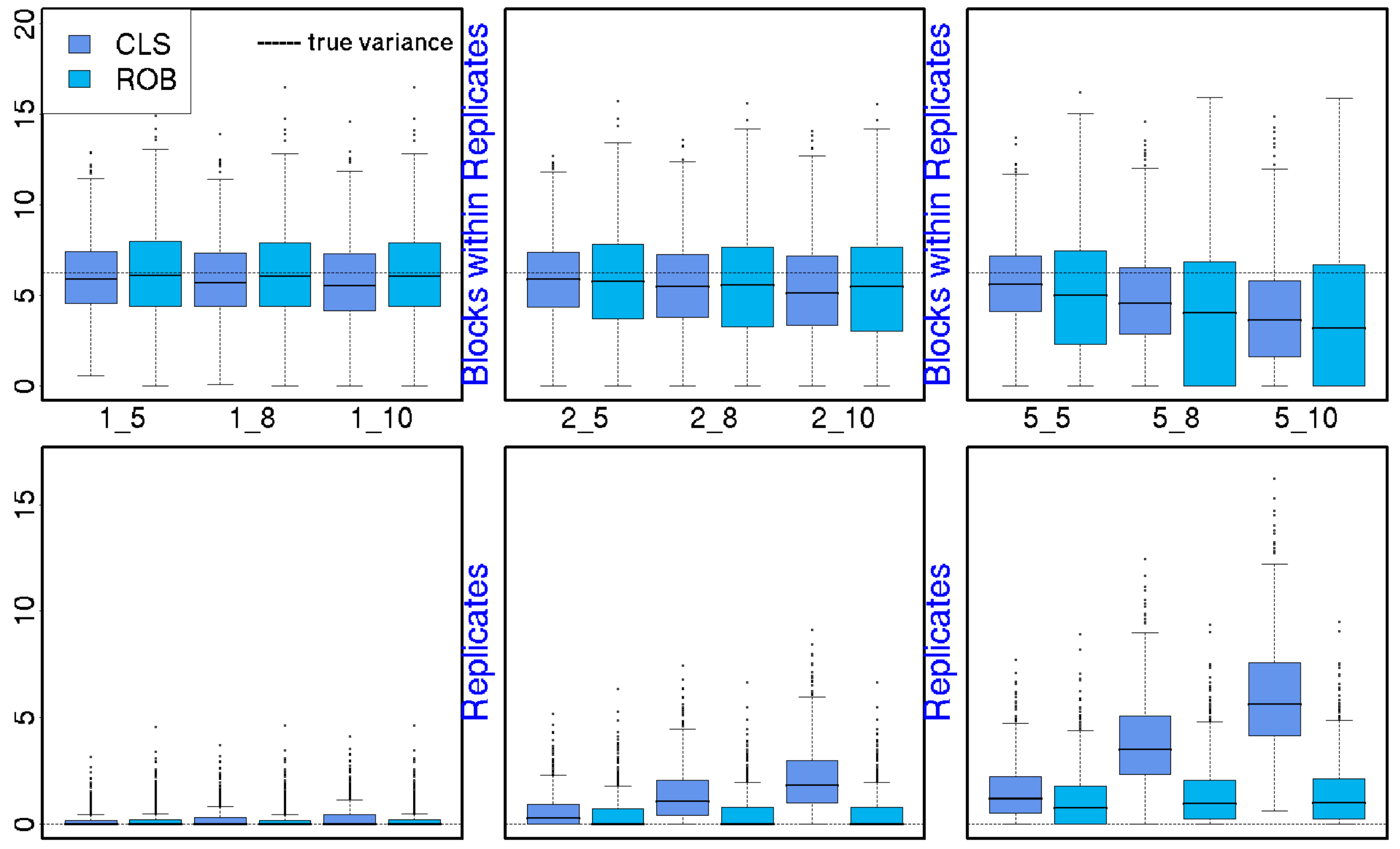

Figure 1. Variance components – Scenarios **I**

Results under H$_1$ (cont.)

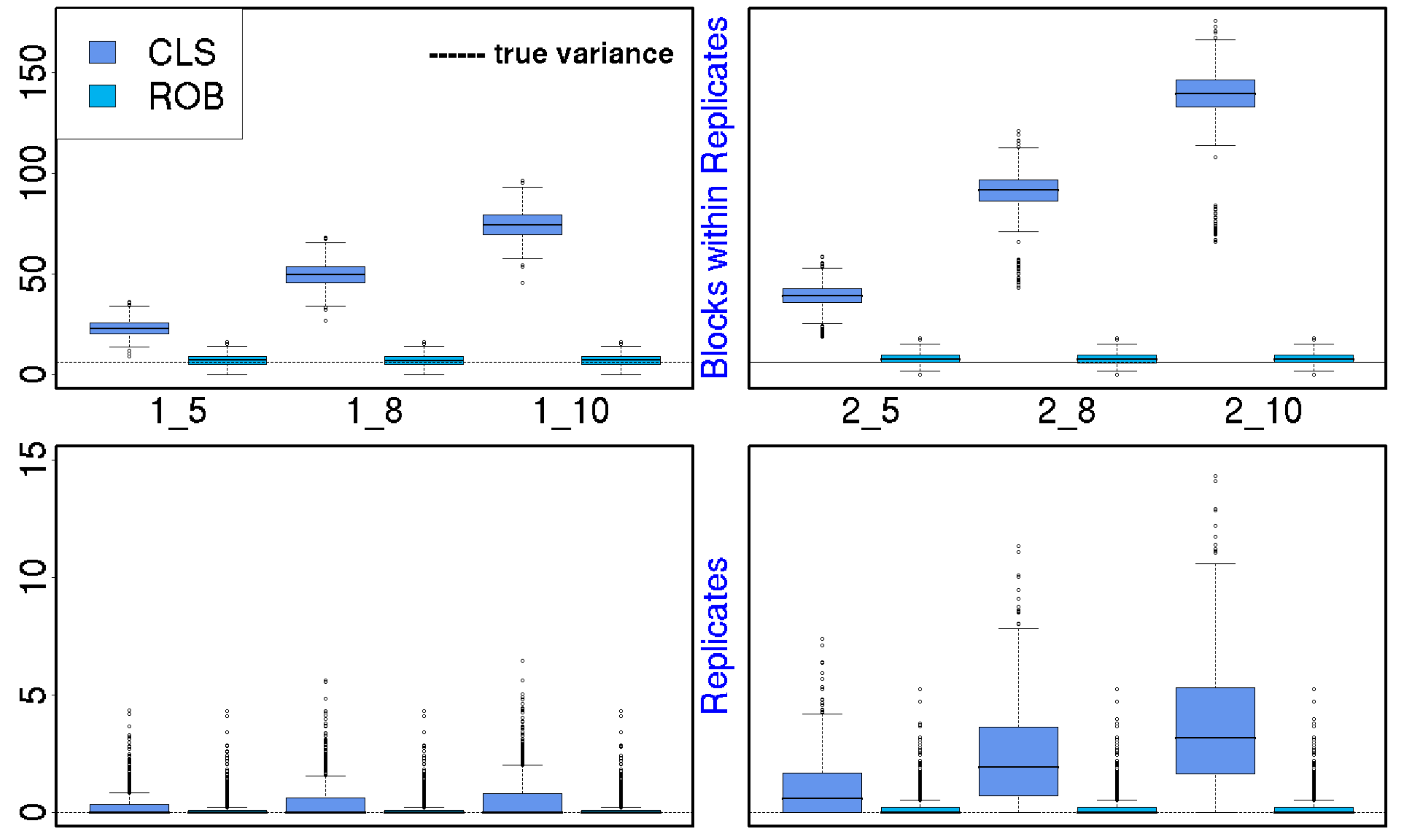

Figure 2. Variance components – Scenarios **II**

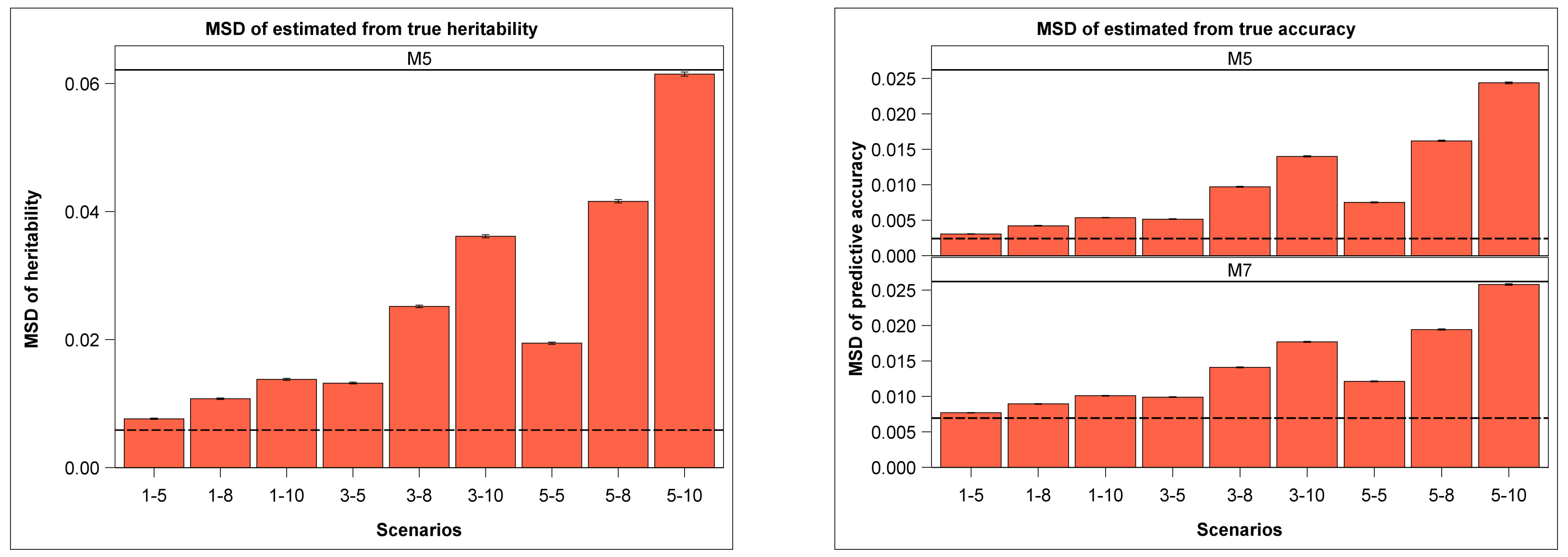

Figure 3. H^2 & PA MSDs – CLS results **I**

Discussion

Under H$_0$: MSD$_{\boldsymbol{\mu}}\simeq 0$ which is desirable for any alternative method. **Under H$_1$: Table 1.** MDS$_{\boldsymbol{\mu}}$ values increase with the % of contamination and also with the increase of the shift outliers. In the **II** scenarios the ROB approach presents 2.8 to 5.2 times smaller MSDs than CLS. The ROB estimated random effects variances are more accurate (**Figures 1. & 2.; I & II Scenarios**). The biased parameter and variance estimation of the CLS method translates in the underestimation of both H^2 and PA (**Figure 3.**). The biases from the CLS approach are also seen in the **II** scenarios (not shown). The robust approach is expected to present better results in terms of the final estimation and more simulations are underway to better assess its usefulness.

Acknowledgements

V. M. Lourenço acknowledges partial financial support by the Fundação para a Ciência e a Tecnologia through projects PTDC/MAT-STA/0568/2012 and UID/MAT/00297/2013 and grant SFRH/BSAB/105935/2014.