Robust estimation of heritability and predictive accuracy in plant breeding

V.M. Lourenço¹, J. Ogutu² & H.-P. Piepho²

¹ CMA & FCT – NOVA University of Lisbon, Portugal; vmml@fct.unl.pt
² Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, Germany

Supported by Portuguese National Funds

Project PTDC/MAT-STA/0568/2012
Erasmus Contract 29191(T)009/2015/STT

Statistics Seminar DM-IST-UL, 13 April, 2016
BREEDERS

What is their main goal?
To select the best plants, i.e., those which maximize the traits of interest
(e.g., flowering time; grain yield; etc.)

More precisely, in plant breeding the goal usually is
To select the best genotypes for the next season
This is called Genomic Selection (GS)

So how do they decide?
The breeder’s selection is made on the basis of genomic estimated breeding values
(GEBVs) obtained from genome-wide DNA marker information
(genotypes are usually ranked based on the GEBVs)

We talk about Genomic Prediction (GP) when
we use a genetic model to predict the unobserved true breeding value (TBV) of a genotype
Practical Motivation – Genomic Prediction

Genomic prediction (GP) is used both in animal and plant breeding to help identify the best genotypes for selection.

Here,

- Two of the most important measures of the effectiveness and/or reliability of GP in plant studies are predictive accuracy (PA) and heritability ($H^2$).

However,

- In plant breeding, as in many other applications, the models of choice for analyzing field data are regression models.

And since,

- Models that make use of the classical likelihood are known to perform poorly when their underlying assumptions are violated.

It is likely that,

- The associated biases may translate to inaccurate estimates of heritability and predictive accuracy.
In this research work, we are particularly interested in two recently proposed methods for the estimation of $H^2$ (Method 5 only) and PA (Methods 5 and 7) shown by Estaghvirou et. al. [BMCGenomics13] to consistently give the least biased, most precise and stable estimates of PA across all the simulated scenarios; Additionally Method 5 gave the most accurate estimates of $H^2$ which are founded on the linear mixed effects model as well as on ridge regression best linear unbiased prediction (RR-BLUP) through a two-stage approach used to estimate breeding values — Piepho [CropSci09]; Piepho et.al.[CropSci12]

And so we worry that inaccurate estimation of genotypic means ($\hat{\mu}$) in the 1st stage, may bias the estimation of breeding values ($\hat{g}$) in the 2nd stage and consequently bias $H^2$ and PA estimates

In previous work, Methods 5 and 7 were shown to perform well in the presence of a single outlying observation whether it was a mild, intermediate or gross outlier Estaghvirou et.al. [G3-14]; again across all simulated scenarios.
**INTRODUCTION**

**Practical Motivation – Genomic Prediction**

But of course,

one outlier does not reflect the reality of field data if we consider all the sources of possible phenotype contamination!

Remember that the process generating the outliers may also vary across locations and/or trials;
Also note that outliers may arise in plant breeding studies from measurement errors, inherent characteristics of the studied genotypes, environments or even years.

And therefore,

we now consider an extension of the work of Estaghvirou et al. [G3-14], where different percentages of phenotypic data contamination (outliers) are taken into account
Focusing on the estimation of genotypic means, breeding values and subsequent estimation of heritability through Method 5 and predictive accuracy through Methods 5 and 7.

Additionally,

At the same time we evaluate the performance of the usual classical approach in the estimation of the parameters of interest, we also asses if any gain can be achieved by using a more robust approach
In this case we propose a robust LMM approach for the first stage of the two-stage approach (phenotypic analysis) where adjusted genotypic means are computed
Why robust statistics?

- To obtain reliable results, a statistical theory is needed that accounts for the departure from particular assumptions of parametric regression models. (The classical assumptions of normality, independence, and linearity are often not fulfilled.)

- Robust statistics works in a “neighbourhood” of parametric models. It uses the advantages of parametric models but allows for deviations.

- Robust statistics relies on a theory of approximate parametric models.

Hampel et al. [NewYorkWiley86] gave the definition:

“In a broad informal sense, robust statistics is a body of knowledge, partly formalized into theories of robustness, relating to deviations from idealized assumptions in statistics.”

Introduction ends.
The two-stage approach of Piepho et al. [BiomJou12] that is used to predict true breeding values \( (g) \) that are then used to estimate heritability and predictive accuracy proceeds as follows:

**1\textsuperscript{st} Stage.** Model

\[
y = X\mu + f
\]

\( y \) is used to estimate the adjusted means for the testcross genotypes, where

- \( y \) is the vector of the observed phenotypic plot values \( (\text{yield}) \)
- \( X \) is the design matrix of fixed effects of genotypes, testers, etc
- \( \mu \) is the vector of unknown genotypic means
- \( f \) is a vector that combines all the other fixed, random design and error effects (replicates, blocks, etc.).

Estimated adjusted genotypic means \( \hat{\mu} \) are submitted to the \( 2\textsuperscript{nd} \) stage of the two-stage approach. Also, weights computed from \( \Sigma\mu \) will also be used in the \( 2\textsuperscript{nd} \) stage.
2nd Stage. The linear mixed model

\[ \hat{\mu} = \phi 1 + g + e \]  

\( \hat{\mu} = \text{general mean} + \text{breeding values} + \text{error} \)

is used in a ridge-regression formulation to compute the predicted breeding values \( \hat{g} \), i.e., \( \text{BLUP}(g) = \hat{g} \), where

- \( \hat{\mu} \) are the estimated adjusted genotypic means from the 1\textsuperscript{st} stage
- \( \phi \) is the general mean
- \( g \) are the random effects, modeled through a linear regression on the SNP markers \( Z \) as \( g = Zu \), where
  - \( Z \) is the matrix of biallelic markers of the single nucleotide polymorphisms (SNPs), coded as \(-1\) for genotypes AA, \( 1 \) for BB and \( 0 \) for AB, BA or missing values;
  - \( u \sim N(0, \sigma_u^2 I) \) with \( \sigma_u^2 \) being the variance of the marker effects; thus \( \text{var}(g = Zu) = \sigma_u^2 ZZ' \)
  - \( e \sim N(0, R) \)
Method 5.

This method calculates predictive accuracy as

\[
E(r_{g,\hat{g}}) \approx \frac{\text{trace}(P_uCG)}{\sqrt{\text{trace}(P_uG)\text{trace}(C^TP_uCV)}}
\]

(3)

where

- \(G = ZZ^T\sigma_u^2\) and \(V = G + R\) are the variance-covariance matrices for genotypes and phenotypes, respectively;
- \(P_u = \frac{1}{n-1}(I - \frac{1}{n}J_n)\), with \(J_n\) a \(n \times n\) matrix of ones;
- \(C = GV^{-1}Q\), with \(Q = I - 1(1^T V^{-1} 1)^{-1}1^T V^{-1}\), and \(1\) denoting a vector of ones.

Heritability can then be computed from (3) as

\[
H_{m5}^2 = E(r_{g,\hat{g}})^2.
\]
**Method 7.**

This method first computes the reliability for each genotype as

$$\hat{\rho}_i^2 = \frac{(\text{cov}(g_i, \hat{g}_i))^2}{\text{var}(g_i) \text{var}(\hat{g}_i)}.$$  \hspace{1cm} (4)

The reliability of all the genotypes in each dataset is then estimated by

$$\hat{\rho}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_i^2$$  \hspace{1cm} (5)

where $n$ is the total number of genotypes in the dataset.

Predictive accuracy is computed as

$$E(r_{\hat{g}, \hat{g}}) \approx \sqrt{\hat{\rho}^2}.$$
The robust approach.

The robust analogue of the two-stage method estimates $\mu$ in the 1$^{st}$-stage via a robust LMM (Koller, M.).

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 $\psi$ are applied to both the residual and random effects terms.

Having robustly estimated the $\hat{\mu}$ values, these are now carried to the 2$^{nd}$ stage as before and the method proceeds in the usual way with PA and $H^2$ estimated by Methods 5 and 7 above.


Statistical methods end.
**Simulation**

**Real dataset, phenotypic data simulation and data contamination**

Real dataset and simulation of 1000 phenotypic datasets
We consider a real maize dataset from the KWS-Synbreed Project (2009-15), extracted for one location from a larger dataset consisting, in particular, of

- 698 genotyped testcrosses (four testers)
- 11646 SNP markers

Variance components estimated from this dataset were used to simulate true breeding values and phenotypic data assuming that all the 698 genotypes are correlated. It was also assumed that: ● $\sigma^2_e \simeq 53.9$ and ● $\sigma^2_u \simeq 0.006$ (marker-effect variance)

<table>
<thead>
<tr>
<th>sim</th>
<th>rep</th>
<th>block</th>
<th>geno_rnd</th>
<th>yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-7.416505</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1.945098</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>698</td>
<td>1</td>
<td>1</td>
<td>39</td>
<td>25.097810</td>
</tr>
<tr>
<td>699</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>12.640605</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1396</td>
<td>1</td>
<td>2</td>
<td>39</td>
<td>18.859413</td>
</tr>
</tbody>
</table>

Data contamination settings

Type of outliers. good observations are replaced by their observed value plus 5, 8 or 10 times $\sigma_e$

Scenarios I. 1, 3, 5, 7 and 10% of randomly assigned phenotypic data contamination on 1 replicate (overall contamination of $\simeq 0.5, 1.5, 2.5, 3.5$ and 5%)

Scenarios II. 1, 2, 3, 4 and 5 whole block contamination on 1 replicate (overall contamination of $\simeq 1.3, 2.6, 3.9, 5.2$ and 6.5%)
We compare the robust & classical approaches via the raw deviations and the following mean square deviations (MSD):

1\textsuperscript{st} Stage.

$$\text{MSD}_\mu = \sum_{j=1}^{N} \sum_{i=1}^{n} \frac{ (\hat{\mu}_{ij} - \mu_{ij})^2 }{ n \times N }$$  \hspace{1cm} (6)

2\textsuperscript{nd} Stage.

$$\text{MSD}_g = \sum_{j=1}^{N} \sum_{i=1}^{n} \frac{ (\hat{g}_{ij} - g_{ij})^2 }{ n \times N }$$ \hspace{1cm} (7)

3\textsuperscript{rd} Stage.

$$\text{MSD}_h = \sum_{j=1}^{N} \frac{ (\hat{H}_j^2 - (r_{g,\hat{g},j})^2)^2 }{ N }$$ \hspace{1cm} (8)

$$\text{MSD}_{PA} = \sum_{j=1}^{N} \frac{ (r_{g,\hat{g},j} - r_{g,\hat{g},j})^2 }{ N }$$ \hspace{1cm} (9)

where $n$ is the total number of observations with $i = 1, \ldots, n$ denoting the $i$-th observation and $N$ is the total number of simulated datasets, with $j = 1, \ldots, N$ denoting the $j$-th dataset.
**H₀: No Contamination Scenario**

1st Stage.

\[
\text{MSD}_\mu = \frac{1000 \sum_{j=1}^{698} \sum_{i=1}^{698} (\hat{\mu}_{ij}^R - \hat{\mu}_{ij}^C)^2}{698 \times 1000} \approx 0.0616; \\
\text{MSD}^C_\mu = \frac{1000 \sum_{j=1}^{698} \sum_{i=1}^{698} (\hat{\mu}_{ij}^C - \mu_{ij})^2}{698 \times 1000} \approx 28.97; \\
\text{MSD}^R_\mu = \frac{1000 \sum_{j=1}^{698} \sum_{i=1}^{698} (\hat{\mu}_{ij}^R - \mu_{ij})^2}{698 \times 1000} \approx 29.08
\]

**Figure 1.** Plot of \(\text{MSD}^k_\mu = \frac{1000 \sum_{j=1}^{698} (\hat{\mu}_{ijk} - \mu_{ijk})^2}{1000}\) for each of the \(i = 1, \ldots, 698\) genotypes and \(k = \text{rob, cls}\) methods
**Simulation**

Classical vs Robust results — under the null

**H₀**: No Contamination Scenario

1ˢᵗ Stage.

<table>
<thead>
<tr>
<th>true mean=8.923</th>
<th>ROB</th>
<th>CLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>overall mean $\hat{\mu}$</td>
<td>8.906</td>
<td>8.908</td>
</tr>
<tr>
<td>Spearman Corr</td>
<td>0.747</td>
<td>0.747</td>
</tr>
</tbody>
</table>

Figure 2. Classical and robust $\text{var}(\hat{\mu}_{ij})$, $i = 1, \ldots, 698$, $j = 1, \ldots, 1000$ against the true LSMEANS variances
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE NULL**

$H_0$: No Contamination Scenario

1$^{\text{st}}$ Stage.

![Figure 3](image.png)

**Figure 3.** Variance component and residual variance estimation
**Simulation**

**Classical vs Robust Results — Under the Null**

**H₀: No Contamination Scenario**

2ⁿᵈ Stage. Smith weighting scheme used

\[
\text{MSD}_g = \frac{1000 \sum_{j=1}^{698} \sum_{i=1}^{1000} (\hat{g}_{ij}^{R} - \hat{g}_{ij}^{C})^2}{698 \times 1000} \approx 0.031; \quad \text{MSD}^R = \frac{1000 \sum_{i=1}^{698} \sum_{j=1}^{1000} (\hat{g}_{ij}^{R} - g_{ij})^2}{698 \times 1000} \approx 25.48; \quad \text{MSD}^C = \frac{1000 \sum_{i=1}^{698} \sum_{j=1}^{1000} (\hat{g}_{ij}^{C} - g_{ij})^2}{698 \times 1000} \approx 25.18
\]

Figure 4. Plot of \( \text{MSD}_g^{ik} = \sum_{j=1}^{1000} \frac{(\hat{g}_{ijk} - g_{ijk})^2}{1000} \) for each of the \( i = 1, \ldots, 698 \) genotypes and \( k = \text{rob, cls} \) methods.
Simulation

Classical vs Robust results — under the null

$H_0$: No Contamination Scenario

2nd Stage. Smith weighting scheme used

<table>
<thead>
<tr>
<th>ROB</th>
<th>CLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman Corr</td>
<td>0.875</td>
</tr>
</tbody>
</table>

Figure 5. Estimated of classical and robust $\text{var}(\hat{g}_{ij})$, with $i = 1, \ldots, 698, j = 1, \ldots, 1000$ and under $H_0$ (2nd stage)
**Simulation**

Classical vs Robust results — under the null

$H_0$: No Contamination Scenario

$H^2$ and PA Estimation

not done yet
**H₀: No Contamination Scenario**

**Results summary**

- Both methods perform similarly, with the classical method coming slightly better than the robust.

- They both methods provide moderate (overall MSD$_μ$) to high (some per-genotype MSD$_μ$) MSD values.

- These observed MSD values have no effect on the ranking of the genotypes (high SCC) in the 2$^{nd}$-stage, the quantity of primary interest to plant breeders doing GS.
Simulation

Classical vs Robust results — under the alternative hypothesis

$H_1$: 1st Stage

— Contamination Scenarios I (RANDOM) —
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**Table 1. Observed MSD between the estimated \( \hat{\mu} \) and true \( \mu \)**

<table>
<thead>
<tr>
<th>Scenarios I</th>
<th>% cont</th>
<th>sdt</th>
<th>CLS</th>
<th>ROB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>-</td>
<td>28.97</td>
<td>29.08</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>32.44</td>
<td>32.51</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8</td>
<td>37.83</td>
<td>37.78</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>10</td>
<td>42.80</td>
<td>42.64</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5</td>
<td>39.36</td>
<td>39.43</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8</td>
<td>55.46</td>
<td>55.27</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10</td>
<td>70.23</td>
<td>69.89</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>46.26</td>
<td>46.39</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>72.98</td>
<td>72.87</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10</td>
<td>97.50</td>
<td>97.25</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>5</td>
<td>53.12</td>
<td>53.34</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8</td>
<td>90.42</td>
<td>90.44</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>10</td>
<td>124.68</td>
<td>124.55</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5</td>
<td>63.30</td>
<td>63.65</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>8</td>
<td>116.40</td>
<td>116.67</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>165.22</td>
<td>165.30</td>
</tr>
</tbody>
</table>

**Table 2. Overall Mean value of \( \hat{\mu} \) & Spearman correlation coefficient**

<table>
<thead>
<tr>
<th>Scenarios I</th>
<th>true mean = 8.923/SCC</th>
<th>% cont</th>
<th>sdt</th>
<th>CLS</th>
<th>ROB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>-</td>
<td>8.908/0.75</td>
<td>8.906/0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
<td>9.092/0.74</td>
<td>9.090/0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>8</td>
<td>9.202/0.73</td>
<td>9.200/0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>10</td>
<td>9.276/0.73</td>
<td>9.274/0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>5</td>
<td>9.460/0.71</td>
<td>9.458/0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>8</td>
<td>9.792/0.70</td>
<td>9.789/0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>10</td>
<td>10.013/0.70</td>
<td>10.010/0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
<td>9.829/0.69</td>
<td>9.827/0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td>10.381/0.68</td>
<td>10.379/0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>10</td>
<td>10.749/0.67</td>
<td>10.747/0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>5</td>
<td>10.196/0.68</td>
<td>10.193/0.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>10.969/0.65</td>
<td>10.967/0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>10</td>
<td>11.484/0.65</td>
<td>11.483/0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>5</td>
<td>10.748/0.65</td>
<td>10.748/0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>8</td>
<td>11.853/0.61</td>
<td>11.853/0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>10</td>
<td>12.589/0.61</td>
<td>12.589/0.61</td>
</tr>
</tbody>
</table>
SIMULATION

CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS

H₁: 1ˢᵗ Stage

— Contamination Scenarios I —

Computing the MSD for each genotype

Across the 1000 simulations

And across all the RANDOM scenarios
SIMULATION

CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS

**H$_1$: 1$^{st}$ Stage — Contamination Scenarios I (RAND)**

**Figure 6.** Plot of $MSD_{jk}^\mu = \sum_{j=1}^{1000} \frac{(\hat{\mu}_{ijk} - \mu_{ijk})^2}{1000}$ for each of the $i = 1, \ldots, 698$ genotypes and $k = rob, cls$ methods RANDOM cont.
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**H₁: 1st Stage — Contamination Scenarios I (RAND)**

**Figure 7.** Plot of $\text{MSD}_{\mu}^{ik} = \frac{1000}{1000} \sum_{j=1}^{1000} (\hat{\mu}_{ijk} - \mu_{ijk})^2$ for each of the $i = 1, \ldots, 698$ genotypes and $k = \text{rob, cls}$ methods RANDOM cont.
**H₁: 1st Stage**

— Contamination Scenarios I —

Computing the overall mean of $\hat{\mu}$ for each genotype

Across the 1000 simulations

And across all the RANDOM scenarios

Against the true values of $\mu$
Simulation
Classical vs Robust results — under the alternative hypothesis

$H_1$: 1st Stage — Contamination Scenarios I (RAND)

Figure 8. Estimated per-genotype overall mean value of $\hat{\mu}$ computed across the 1000 simulations, for both classical and robust methods against the true overall mean value of $\mu$. 
**SIMULATION**

**CLASSICAL vs ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**H₁: 1st Stage — Contamination Scenarios I (RAND)**

*Figure 9.* Estimated per-genotype overall mean value of $\hat{\mu}$ computed across the 1000 simulations, for both classical and robust methods against the true overall mean value of $\mu$. 
**SIMULATION**

**CLASSICAL vs ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**H$_1$: 1$^{st}$ Stage — Contamination Scenarios I (RAND)**

*Figure 10. Histograms of classical and robust $\text{var}(\hat{\mu}_{ij})$, with $i = 1, \ldots, 698$, $j = 1, \ldots, 1000$*
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

\[ H_1: \text{1st Stage — Contamination Scenarios I (RAND)} \]

**Figure 11.** Histograms of classical and robust \( \hat{\text{var}}(\hat{\mu}_{ij}) \), with \( i = 1, \ldots, 698, j = 1, \ldots, 1000 \)
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

$H_1$: 1st Stage — Contamination Scenarios I (RAND)

**Figure 12.** Variance components estimation
**SIMULATION**

**CLASSICAL vs ROBUST results — UNDER THE ALTERNATIVE HYPOTHESIS**

$H_1$: 1$st$ Stage

— Contamination Scenarios II (BLOCK) —
## Simulation

Classical vs Robust results — under the alternative hypothesis

### $H_1$: 1st Stage — Contamination Scenarios II (BLOCK)

**Table 3.** Observed MSD between the estimated $\hat{\mu}$ and true $\mu$

<table>
<thead>
<tr>
<th>Scenarios II</th>
<th>CLS</th>
<th>ROB</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. blocks</td>
<td>sdrt</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>28.97</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>30.90</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>30.75</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>31.16</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>30.93</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>32.49</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>33.79</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>32.06</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>35.22</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>38.02</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>33.48</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>38.74</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>43.51</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>35.31</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>43.33</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>50.66</td>
</tr>
</tbody>
</table>

### Table 4. Overall Mean value of $\hat{\mu}$ & Spearman correlation coefficient

<table>
<thead>
<tr>
<th>Scenarios II</th>
<th>true mean =8.923/Spear CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>% cont</td>
<td>sdrt</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
Simulation

Classical vs Robust results — under the alternative hypothesis

$H_1$: 1st Stage

— Contamination Scenarios II —

Computing the MSD for each genotype

Across the 1000 simulations

And across all the BLOCK scenarios
**Simulation**

Classical vs Robust results — under the alternative hypothesis

**H₁: 1st Stage — Contamination Scenarios II (BLOCK)**

![Box plots](image)

**Figure 13.** Plot of $\text{MSD}_{\mu}^k = \frac{1}{1000} \sum_{j=1}^{1000} (\hat{\mu}_{ijk} - \mu_{ijk})^2$ for each of the $i = 1, \ldots, 698$ genotypes and $k = rob, cls$ methods BLOCK cont.
**SIMULATION**

**CLASSICAL vs ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**H₁: 1st Stage — Contamination Scenarios II (BLOCK)**

Figure 14. Plot of $\text{MSD}_{\mu}^{jk} = \frac{1000}{\sum_{j=1}^{1000} (\hat{\mu}_{ijk} - \mu_{ijk})^2}$ for each of the $i = 1, ..., 698$ genotypes and $k = \text{rob, cls}$ methods
**SIMULATION**

**CLASSICAL vs ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

\[ H_1: \text{1st Stage} \]

— Contamination Scenarios II —

Computing the overall mean of \( \hat{\mu} \) for each genotype

Across the 1000 simulations

And across all the BLOCK scenarios

Against the true values of \( \mu \)
**Simulation**

Classical vs Robust results — Under the alternative hypothesis

$H_1$: 1st Stage — Contamination Scenarios II (BLOCK)

**Figure 15.** Estimated per-genotype overall mean value of $\hat{\mu}$ computed across the 1000 simulations, for both classical and robust methods against the true overall mean value of $\mu$. 
**Simulation**

Classical vs Robust results — Under the alternative hypothesis

$H_1$: 1st Stage — Contamination Scenarios II (BLOCK)

**Figure 16.** Estimated per-genotype overall mean value of $\hat{\mu}$ computed across the 1000 simulations, for both classical and robust methods against the true overall mean value of $\mu$. 
**Simulation**

**Classical vs Robust results — under the alternative hypothesis**

**$H_1$: 1st Stage — Contamination Scenarios II (BLOCK)**

**Figure 17.** Histograms of classical and robust $\text{var}(\hat{\mu}_{ij})$, with $i = 1, \ldots, 698$, $j = 1, \ldots, 1000$
**SIMULATION**

**CLASSICAL VS ROBUST RESULTS — UNDER THE ALTERNATIVE HYPOTHESIS**

**H₁: 1st Stage — Contamination Scenarios II (BLOCK)**

*Figure 18. Histograms of classical and robust $\text{var}(\hat{\mu}_{ij})$, with $i = 1, \ldots, 698, j = 1, \ldots, 1000$*
**Simulation**

**Classical vs Robust results — under the alternative hypothesis**

\[ H_1: \text{1st Stage — Contamination Scenarios II (BLOCK)} \]

---

**Figure 19.** Variance components estimation
**Simulation**

Classical vs Robust results — under the alternative hypothesis

\[ H_1: 2^{nd \text{st} \text{ Stage}} \]

— Contamination Scenarios I —

RANDOM

not done yet
H$_1$: 2$^{nd}$ Stage

— Contamination Scenarios II —

BLOCK

- For both classical and robust methods, and across all scenarios, the overall

\[
\text{MSD}_g = \sum_{j=1}^{1000} \sum_{i=1}^{698} \frac{(\hat{g}_{ij} - g_{ij})^2}{698 \times 1000} \approx 25
\]

a value close to the MSD observed under the null

- The per-genotype MSD$_g$ values do not show differences between the methods and are therefore not shown

- The observed values of the Spearman correlation coefficient were $\approx 0.86$ for both approaches, therefore showing an increase of $\approx 16\%$ from the correlation observed in the 1$^{st}$ stage ($\approx 0.74$)
**SIMULATION**

**CLASSICAL vs ROBUST results — UNDER THE ALTERNATIVE HYPOTHESIS**

\[ H_1: \text{Contamination Scenarios I & II} \]

\[ H^2 \text{ and PA Estimation} \]

not done yet
**H₁: Contamination Scenarios I (RANDOM)**

**Results summary / 1ˢᵗ stage**

Both methods have similar poor performance in estimating the genotypic means increasing overall and per-genotype MSDs, increasing overall genotypic mean and decreasing Spearman correlation coefficient across % of contamination and outlier sizes.

In terms of variance components estimation the robust method performed better than the classical (except regarding the block within replicates effect variance component).

The robust method provided smaller and more stable \( \text{var}(\hat{\mu}_{ij}) \), with the classical estimates growing further apart from the robust.

- As far as the 1ˢᵗ stage is concerned, the robust method does not seem to bring added value to the estimation of \( \mu \) in plant breeding in this type of contamination scenarios,

- The somewhat poor performance of the methods might be of concern for breeders who are primarily concerned with the predicted mean value for a genotype.
**H₁: Contamination Scenarios I (RANDOM)**

**Results summary / 2\textsuperscript{nd} stage**

not done yet but

- the huge differences observed between the methods in the estimation of $\text{var}(\hat{\mu}_{ij})$ in the first-stage, which directly impact the Smith/Standard weights that are used in the second-stage, are likely to produce differences between the methods in the second-stage

- since heritability uses the variance components directly, if the genetic variance increases with increasing contamination but the residual variance does not, one might expect the classical method to give a more biased estimate of heritability than the robust
**H₁: Contamination Scenarios II (BLOCK)**

**Results summary / 1st stage**

the performance of the robust method in terms of both the observed overall and per-genotype MSD_μ as well as in terms of the observed overall mean value of ̂μ was better than the classical

both methods produced equivalent Spearman correlation coefficients across % of contamination and outlier sizes

regarding var(̂μ_ij), the robust method showed more stable results, with the classical ones increasing but now in a more modest way than seen in the RANDOM scenarios

in terms of variance components estimation the robust method performed better than the classical (likewise performance in the estimation of the variance of the residual errors)

- as far as the 1st stage is concerned, the robust method thus seems to bring some added value to this stage of the two-stage approach when complete BLOCK contamination occurs
  (which is good news for breeders doing variety testing)
**H$_1$: Contamination Scenarios II (BLOCK)**

*Results summary / 2$^{nd}$ stage*

despite the differences observed between the methods in the 1$^{st}$ stage, in the 2$^{nd}$ stage, no major differences were observed between the methods regarding the overall genotypic MSD$_g$ and the Spearman correlation coefficients computed across scenarios

in turns out that the block effect is completely confounded with the effect of the contaminated observations within the block; as a result, including the block effect in the model likely accounts for the effect of contamination within the block by increasing the block variance

- despite the differences previously observed in the 1$^{st}$ stage, both methods performed in a similar way, ending with a good ranking of the genotypes *(which is good news for breeders doing genomic selection)*

- regarding heritability and predictive accuracy estimation we expect no major differences will be seen between approaches
**SO, is this the end of the line for the robust method?**

Not exactly!!

- The breakdown point of the robust method we are using is about 30% for models with continuous variables.
- The problem in this study is that, not only we are dealing with categorical variables but also we only have two replicates in our study, which means only two observations per genotype.
- And this means that for the random contamination scenarios the method is unable to determine which of the two replicate observations per genotype is the outlier.

We thus believe that by considering studies where more than 3 replicates are available (e.g., variety testing studies), it will be possible to see the real added value of the robust method, specially in the RANDOM contamination scenarios.

*To be continued...*
References


Thank you!