On the accuracy in high dimensional linear models and its application to genomic selection

Charles-Elie Rabier

ISEM, Institut des Sciences de l'Evolution de Montpellier
IMAG, Institut Montpelliérain Alexander Grothendieck

Brigitte Mangin

LIPM, Laboratoire des Interactions Plantes et Microorganismes, Toulouse

Simona Grusea

Institut de Mathématiques de Toulouse, INSA de Toulouse

Outline

1. Introduction
2. Prediction accuracy
3. Our new results + illustration

“On the accuracy in high dimensional linear models and its application to genomic selection”
Rabier et al., Scandinavian Journal of Statistics 2019

“Training set optimization of genomic prediction by means of Ethacc”
Mangin et al., Plos One 2019

Outline

Introduction

Genomic Selection

- Goal = to select individuals (candidates) on the basis of genomic predictions
- One advantage = we can predict the future phenotype of young candidates as soon as their DNA has been collected
- Warning = genomic predictions should be accurate ! We want to select the best candidates for the breeding program

New sequencing technologies

- Millions of markers are available $\Rightarrow$ all the QTLs highly correlated (Strong Linkage Disequilibrium) with at least one genetic marker

Statistical tool

- K markers, n training individuals
- All the markers are analyzed simultaneously $\Rightarrow$ whole genome regression

High-dimensional problem

$K >> n$

Genomic Selection (GS)

Training population

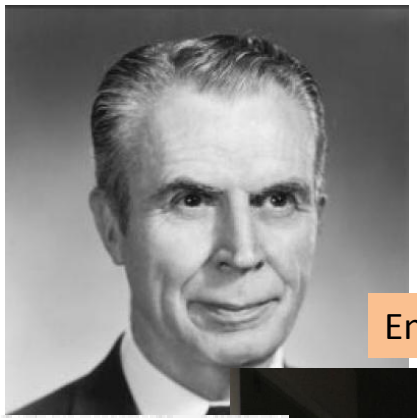
$$Y = \mu + \underbrace{f(M,\theta)}_{g} + \varepsilon$$ ← Model

$$\hat{\theta}$$ ← Estimation method

Test population (selection candidates)

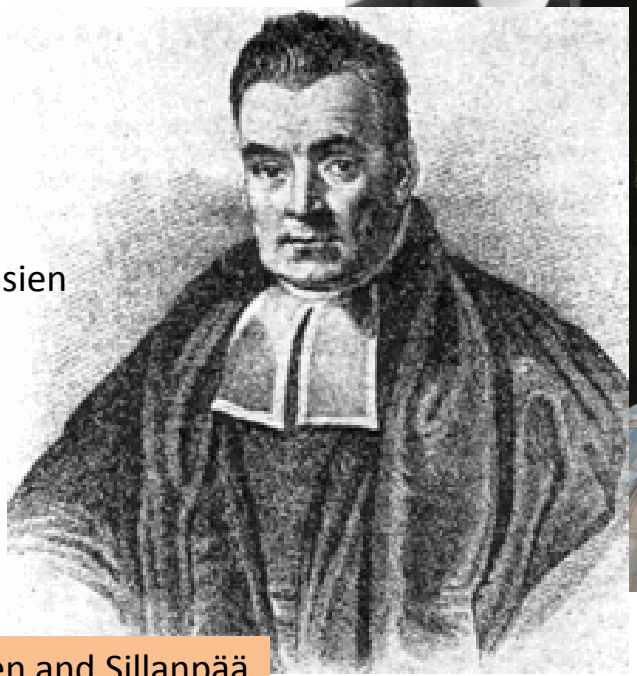
$$\underbrace{f(M,\hat{\theta})}_{\hat{g}}$$ ← Prediction

Statistical framework

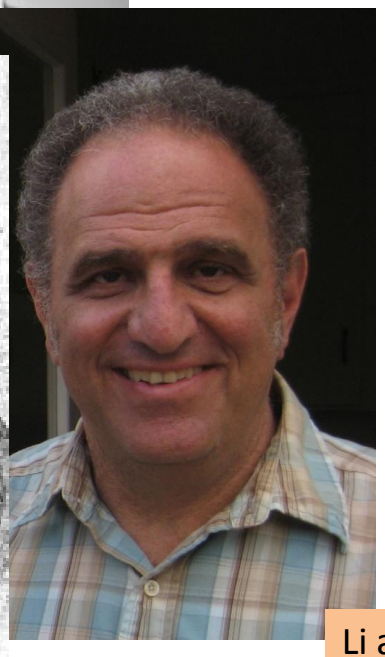


C.R. Henderson
Modèle mixte

Endelman



T. Bayes
Modèle bayésien



R. Tibshirabi
Régression pénalisée

Li and Sillanpää

Kärkkäinen and Sillanpää

Ranking

- In general, the ranking is Bayes $\geq$ Penalized regressions $>$ Mixed model
- but the methods have less influence than the marker density, the size of the Training population, the heritability (linked to the signal)
- or the distance between the Training population and the Test population

This talk : focus on GBLUP, RRBLUP, Ridge (L2 Penalty)

Outline

Goal : to predict a phenotype (continuous variable) using a large number of markers (regressors)

Causal linear model* (Q QTLs, i.e. Q true regressors)

$\theta^\star$ vector of QTL effects, $M^\star$ matrix of QTL alleles,
learning sample of size n,

$$Y = M^\star \theta^\star + e$$

where $Y = (Y_1, ..., Y_n)'$, $\theta^\star = (\theta_1^\star, ..., \theta_Q^\star)'$, $e \sim N(0, \sigma_e^2 I_n)$

Bayesian prediction model (K markers, i.e. K regressors, with $K >> n$)

θ vector of marker effects, M matrix of marker alleles

$$Y = M\theta + \varepsilon$$

where $Y = (Y_1, ..., Y_n)'$, $\theta = (\theta_1, ..., \theta_K)' \sim N(0, \sigma_\theta^2 I_K)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$,
$\varepsilon_j \perp\!\!\!\perp \theta_k$

We suppose that the prediction model contains the Q QTLs, i.e. the Q true regressors... Then, each column of $M^\star$ is a column of M

Learning step

Joint distribution of θ and Y

$$\begin{pmatrix} \theta \\ Y \end{pmatrix}_{|M} \sim N\left(0, \begin{pmatrix} \sigma_\theta^2 I_K & \sigma_\theta^2 M' \\ \sigma_\theta^2 M & \sigma_\theta^2 M M' + \sigma_\varepsilon^2 I_n \end{pmatrix}\right)$$

Estimator $\hat{\theta}$ of θ

$$\begin{aligned} \hat{\theta} &= \mathbb{E}(\theta \mid Y) = M' \left(MM' + \lambda I_n\right)^{-1} Y \quad \text{où} \quad \lambda = \sigma_\varepsilon^2/\sigma_\theta^2 \\ &= \left(M'M + \lambda I_K\right)^{-1} M' Y \end{aligned}$$

i.e. Ridge regression (L2 Penalty) with parameter $\lambda = \sigma_\varepsilon^2/\sigma_\theta^2$

$$\hat{\theta} = \text{argmin}_\theta \quad \|Y - M\theta\|^2 + \lambda \|\theta\|^2$$

Validation set + accuracy criteria

- Let new denote an individual from the validation set

$$Y_{\text{new}} = m_{\text{new}}^{\star\prime}\,\theta^{\star} + e_{\text{new}} \quad \text{where} \quad e_{\text{new}} \sim N(0, \sigma_e^2)$$
$$\text{and} \quad m_{\text{new}}^{\star} \text{ vector of QTL alleles for ind new}$$

- Prediction of the phenotype Y_{new}

$$\hat{Y}_{\text{new}} = m'_{\text{new}}\hat{\theta} = m'_{\text{new}} M' \left(MM' + \lambda I_n\right)^{-1} Y$$
$$= m'_{\text{new}} \left(M'M + \lambda I_K\right)^{-1} M' Y$$

$\Rightarrow$ Accuracy criteria (i.e. prediction accuracy)

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}} \quad \text{with } m_{\text{new}} \text{ et } m_{\text{new}}^{\star} \text{ random, } M \text{ fixed}$$

Key criteria in genetics : it plays a role in the rate of genetic gain....

Result on the accuracy (i.e. prediction accuracy)

Theorem (Rabier Barre ... Mangin, Plos One 2016)

Let us assume that M is known, and that e, m_{new} et e_{new} are random, then

$$\rho = \frac{\theta^{\star\prime}\ Var(m_{\text{new}})\, M' V^{-1} M^{\star} \theta^{\star}}{\left\{\sigma_e^2 \mathbb{E}\left(\|m'_{\text{new}} M' V^{-1}\|^2\right) + \theta^{\star\prime} M^{\star\prime} V^{-1} M Var(m_{\text{new}})\, M' V^{-1} M^{\star} \theta^{\star}\right\}^{1/2} \Omega^{1/2}}$$

where $V = MM' + \lambda I_n$ and $\Omega = \text{Var}\,(m'_{\text{new}}\theta^{\star}) + \sigma_e^2$

One Factor affecting the accuracy :

- Column q of $M' V^{-1} M^{\star}$: LD (corrected for relatedness) between each marker and the QTL q

Existing proxies in the literature

Most of proxies are built on Daetwyler, PloS One 2008

Context of Daetwyler's study :

- locations of the Q QTLs are known
- orthogonal design (QTLs are independent)
- QTL effects are unknown
- $Q < n$

$\Rightarrow$ Y_{new} estimated by Ordinary Least Squares

$$\hat{Y}_{\text{new}}^{OLS} = m_{\text{new}}^{\star\prime}(M^{\star\prime}M^{\star})^{-1}M^{\star\prime}Y$$

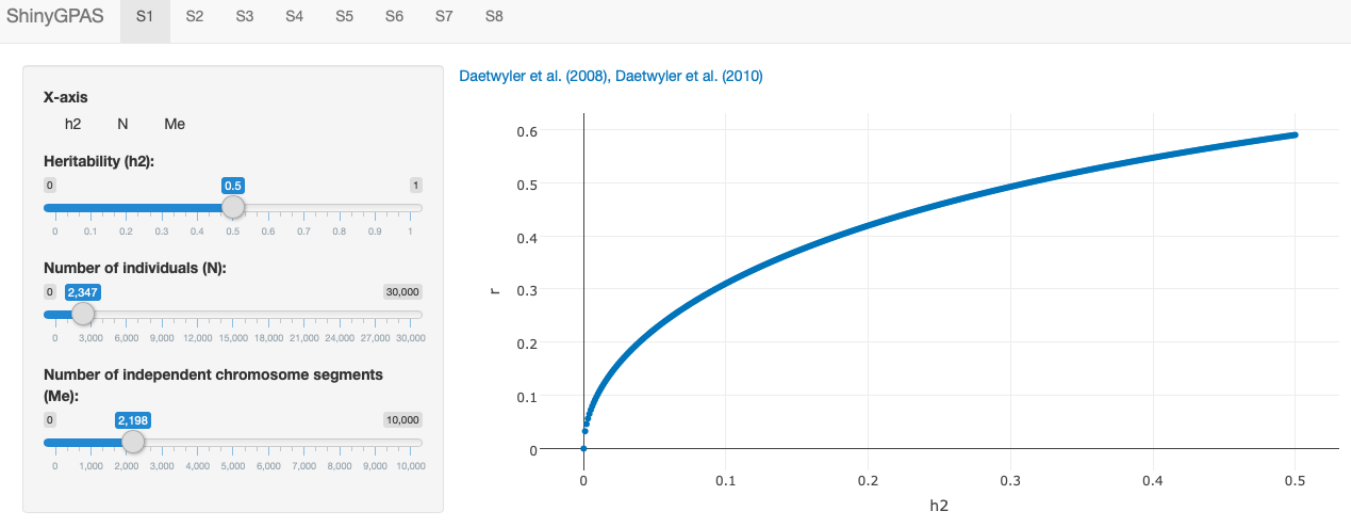
$\Rightarrow$ Daetwyler's formula (2008)

$$\rho = \frac{h\ \sqrt{h^2/(1-h^2)}}{\sqrt{\frac{Q}{n} + \frac{h^2}{1-h^2}}}$$ where h^2 is the heritability of the trait

Methods = substitute the effective number of independent M_e for Q, into Daetwyler's seminal formula (Daetwyler et al., Genetics 2010)

A software to compare several proxies for GS

ShinyGPAS by Morota (GSE, 2017)
available at https ://chikudaisei.shinyapps.io/shinygpas/



Implemented formulas :

- Daetwyler et al. (Plos One 2008, Genetics 2010) nb citations >1000
- Goddard et al. (Genetica 2009, Journal Of Animal Breeding And Genetics 2011)
- Rabier et al. (Plos One, 2016)
- de los Campos et al. (Plos Genetics, 2013)
- Karaman et al. (Plos One, 2016)
- Wientjes et al. (Genetics 2016)

Outline

Coming back to our accuracy based on Ridge regression, RRBLUP, GBLUP ...

Since the prediction model contains the true regressors (i.e. the QTLs), using a small abuse of notation

- $\theta^\star$ sparse vector of dimension K

then, the causal model can be rewritten

$$Y = M\theta^\star + e \text{ where } Y = (Y_1, ..., Y_n)' \ , \ e \sim N(0, \sigma_e^2 I_n).$$

Singular Value Decomposition (SVD)

SVD of M

$$M = U\ D\ W'$$

where

- D diagonal matrix of size $r \times r$, of full rank, with $d_1, \ldots, d_r$ diagonal elements
- U matrix of size $n \times r$, such that $U'U = I_r$
- W matrix of size $K \times r$, such that $W'W = I_r$

Before genotyping the TEST individuals

An estimation of the accuracy is

$$\hat{\rho}_{\text{before}} = \frac{\widehat{A_1}}{\left(\widehat{A_2} + \widehat{A_3}\right)^{1/2} \left(\widehat{A_4}\right)^{1/2}},$$

where

$$\widehat{A_1} = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^4}{d_s^2 + \lambda} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 \quad , \quad \widehat{A_2} = \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2 + \lambda)^2}$$

$$\widehat{A_3} = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2 + \lambda)^2} \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 \quad , \quad \widehat{A_4} = \frac{1}{n} \sum_{s=1}^{r} d_s^2 \left\| W^{(s)} W^{(s)\prime} \theta^\star \right\|^2 + \sigma_e^2.$$

We have now $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}$ in place of M_e

It is possible to evaluate the accuracy of the future prediction of TEST individuals before genotyping them

Our estimation can help geneticists to figure out whether or not their population is appropriate for GS

After genotyping the TEST individuals

M_{new} : random matrix of size $n_{\text{new}} \times K$
containing the marker alleles of the TEST individuals.

SVD of M_{new} $\quad M_{\text{new}} = WFZ'$

An estimator of the accuracy is

$$\hat{\rho}_{\text{after}} = \frac{\check{A}_1}{\left(\check{A}_2 + \check{A}_3\right)^{1/2} \left(\check{A}_4\right)^{1/2}},$$

In this case, we evaluate the accuracy of the future prediction
of TEST individuals after genotyping them

$\hat{\rho}_{\text{after}}$ relies on informations collected
on Trainings (phenotypes and markers) and on TEST (markers)

An example of application is plant breeding

Drawback of our approach

QTLs have to be known to compute our proxies !

$\Rightarrow$ estimate $\theta^\star$ by penalized regressions

LASSO (Tibshirani, JRSSB 1996)

$$\hat{\theta}^\star_{LASSO} = \text{argmin} \, \| Y - M\theta^\star \|^2 + \lambda \sum_{k=1}^{K} |\theta^\star_k|$$

Adaptative LASSO (Zou, JASA 2006)

$$\hat{\theta}^\star_{ADLASSO} = \text{argmin} \, \| Y - M\theta^\star \|^2 + \lambda \sum_{k=1}^{K} w_k \, |\theta^\star_k|$$

Group LASSO (Yuan and Lin, JRSSB 2006) ...

Illustration on rice data from Spindel et al. (Plos Genetics, 2015)

- Two traits of interest : Flowering and Yield (dry season 2012)
- Flowering : $h^2 = 0.4378$, Emp Acc=0.5576
- Yield : $h^2 = 0.3213$, Emp Acc=0.3361
- $K = 13101$ markers, $n = 252$ for Flowering, $n = 248$ for Yield
- $n_{\text{new}} = 63$ in both cases

TABLE – Mean squared error (with respect to the Empirical accuracy)

MSE based on	Flowering	Yield
$\hat{\rho}_{\text{after}}(\hat{\theta}^{\star}_{\text{ADLASSO}})$	1.6248×10^{-2}	2.807×10^{-2}
$\hat{\rho}_{\text{after}}(\hat{\theta}^{\star}_{\text{ADLASSO}})$	2.41×10^{-2}	4.85×10^{-2}
Rabier et al. (2016)	7.08×10^{-2}	1.25×10^{-1}
M_{e1} *Goddard (2009)*	4.49×10^{-2}	5.70×10^{-2}
M_{e2} *Goddard et al (2011)*	4.18×10^{-2}	5.10×10^{-2}
M_{e3} *Goddard et al (2011)*	3.83×10^{-2}	4.43×10^{-2}
M_{LJ} *Li and Ji (2005)*	4.71×10^{-2}	6.27×10^{-2}

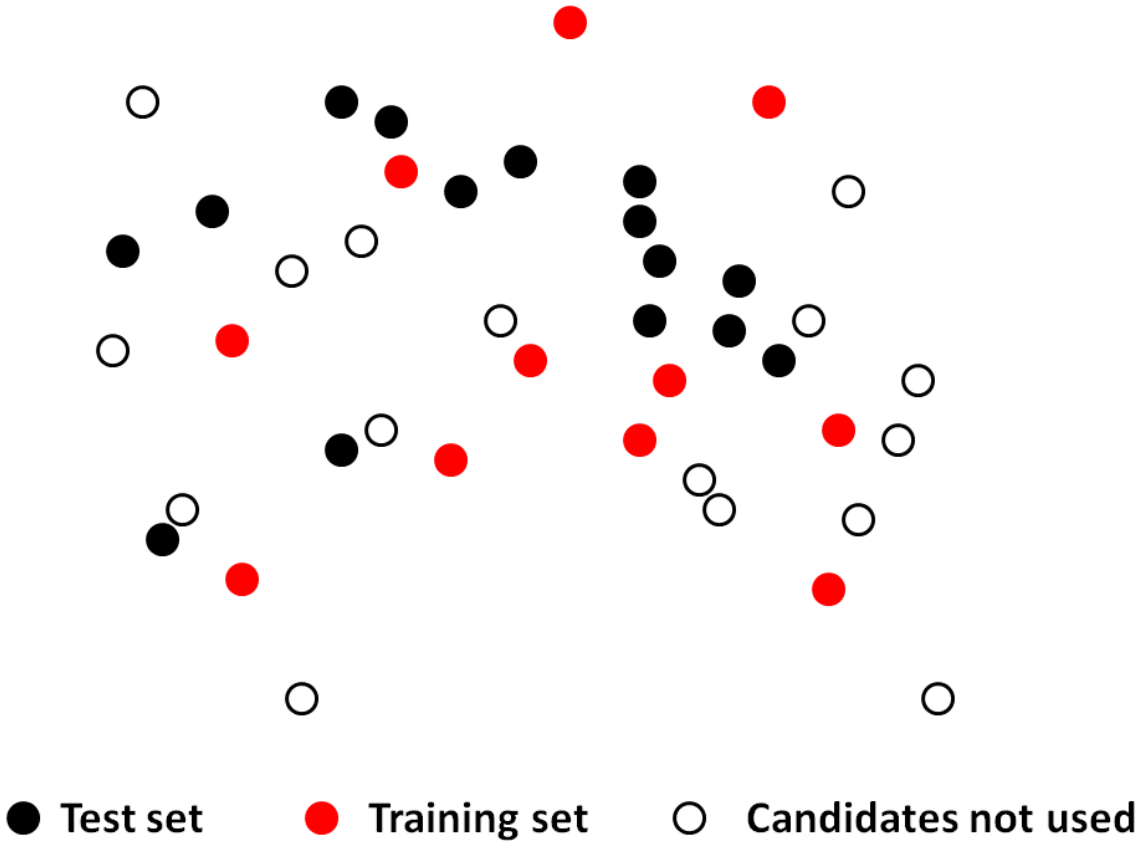
Training set optimization of genomic prediction by means of Ethacc (Mangin et al., Plos One 2019)

- Comparison of several multi-locus GWAS methods to locate QTLs
  - penalized regressions (Lasso, EN05.1se, EN01.1se) Waldmann et al., Frontiers in Genetics 2013 (EN05.FDR) Yi et al., Genetics 2015
  - MLMM Segura et al., Nature Genetics 2012
- Once located, QTL effects are estimated by OLS

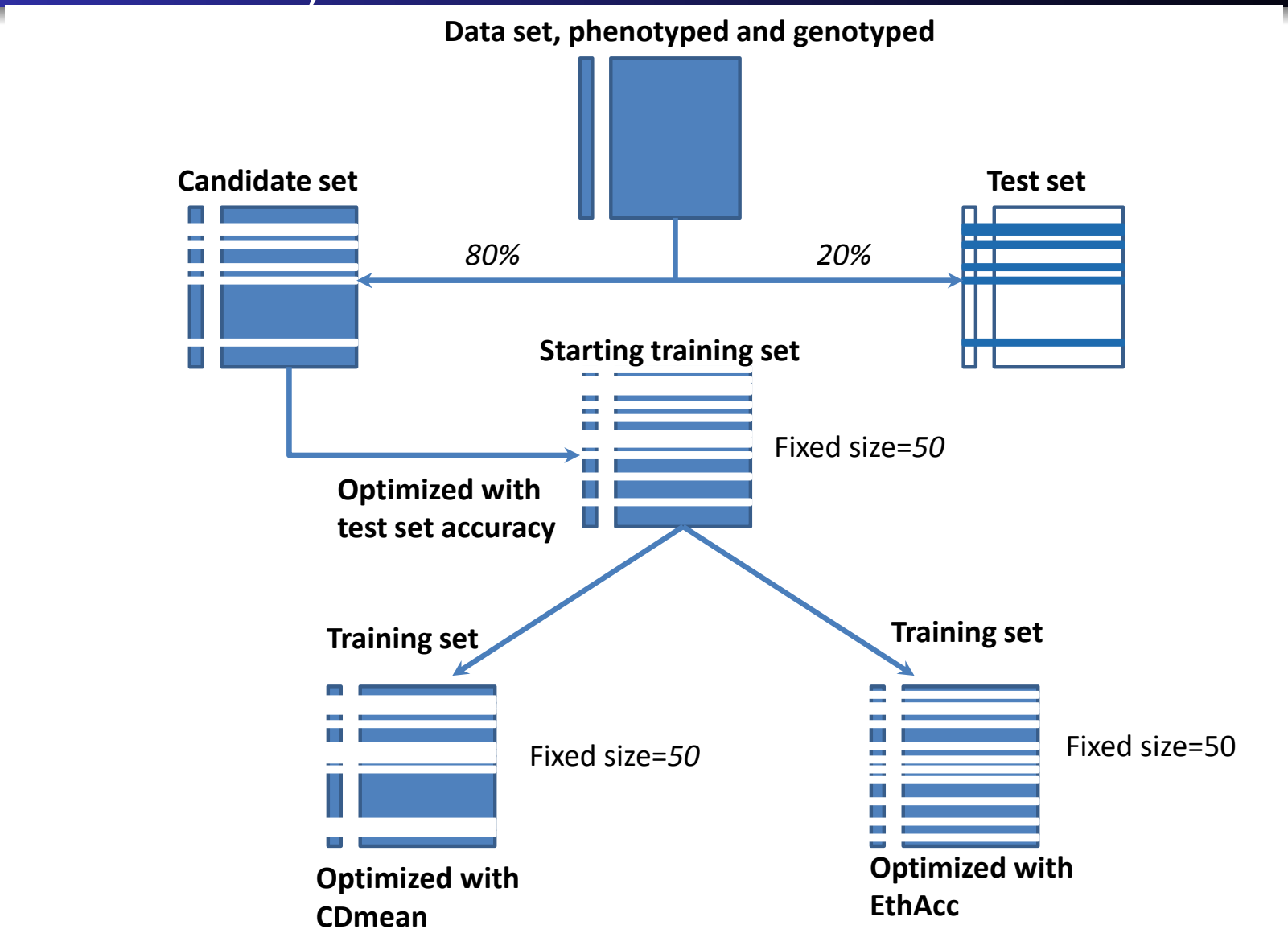
TABLE – Mean squared error on 7 traits on sugar beet

Method	MSE
MLMM	1.22×10^{-3}
LASSO.min	3.25×10^{-3}
LASSO.1se	1.60×10^{-3}
EN05.1se	1.65×10^{-3}
EN01.1se	1.78×10^{-3}
EN05.FDR	8.54×10^{-3}

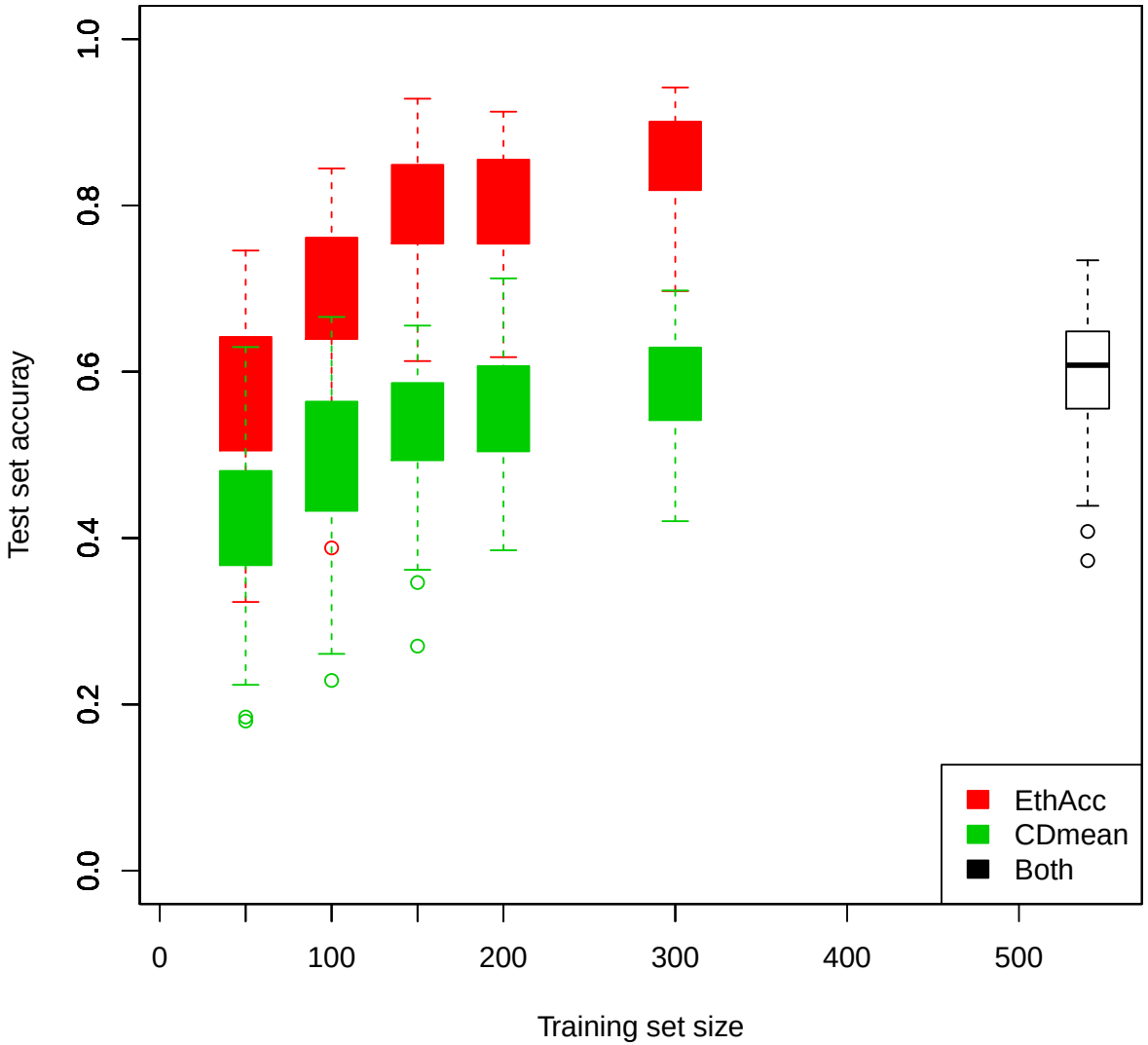
Choice of the Training individuals



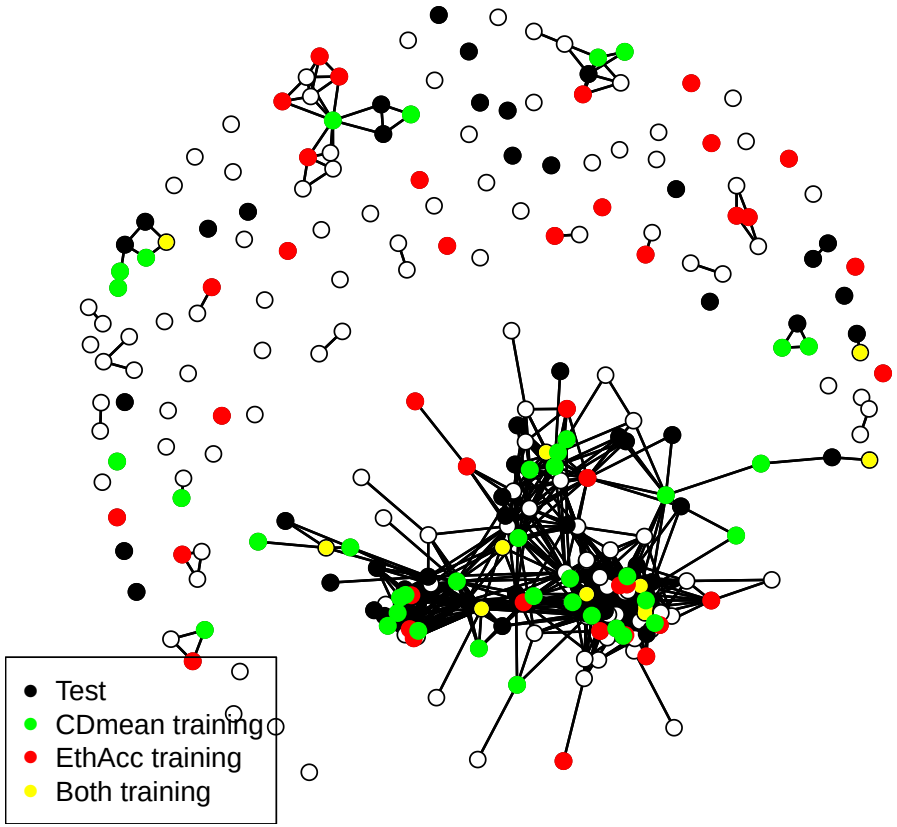
Choice of the Training individuals : Our approach (EthAcc) versus Mixed Model (CDmean, Rincent et al. Genetics 2012)



Sugar beet : EthAcc vs CDmean ($K = 692$, $n_{\text{new}} = 420$, $n = 50, \ldots, 500$)



An extreme case on maize ($K = 25682$, $n = 50$)



- EthAcc optimization gave an accuracy of 0.76
- CDmean optimization gave an accuracy of 0.07

How to improve the quality of the prediction

Idea : consider a subspace of lower dimension

Reminder : $U = \left(U^{(1)}, \ldots, U^{(r)}\right)$ orthonormal basis for the space spanned by the columns of M

Let us choose $\tilde{r}$ columns of U and define $\sigma : \{1, \ldots, \tilde{r}\} \to \{1, \ldots, r\}$

Let $\widetilde{\theta}$ be the new estimator

$$\widetilde{\theta} = M' V^{-1} \widetilde{U} \widetilde{U}' Y \quad \text{where} \quad \widetilde{U} = \left(U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right)$$

$\widetilde{U}\widetilde{U}' Y$ is the projection of Y on $Span\left\{U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right\}$

$\Rightarrow$ Prediction and accuracy built on the new estimator $\widetilde{\theta}$

$$\widetilde{Y}_{\text{new}} = m'_{\text{new}} \widetilde{\theta} \quad , \qquad \widetilde{\rho} = Cor\left(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}}\right) = \frac{\text{Cov}(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}})}{\sqrt{\text{Var}(\widetilde{Y}_{\text{new}})\text{Var}(Y_{\text{new}})}}$$

$$\widehat{\widetilde{\rho}}_{\text{before}} := \widehat{Cor}\left(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}}\right) \ldots$$

Thanks to

Simona Grusea (INSA), Brigitte Mangin (LIPM)

INSA TOULOUSE

Lipm Laboratoire Interactions Plantes Micro-organismes

isem Institut des Sciences de l'Evolution-Montpellier

IMAG INSTITUT MONTPELLIERAIN ALEXANDER GROTHENDIECK

Laurence Moreau, Renaud Rincent, Ellen Goudemand, Philippe Barre
Gilles Charmet, Muriel Tavaud, Jacques David ...

INRA SCIENCE & IMPACT SelGen

AKER Betterave2020

aMaizing

SUNRISE UNE CULTURE POUR LE FUTUR

Coming back to our accuracy based on Ridge regression, RRBLUP, GBLUP ...

Accuracy criteria (i.e. prediction accuracy)

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\hat{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}}$$

In our study on Ridge regression,

$$\rho = \frac{A_1}{(A_2 + A_3)^{1/2}\,(A_4)^{1/2}}.$$

where

$$A_1 := \theta^{\star\prime}\ \text{Var}\left(m_{\text{new}}\right) M' V^{-1} M\theta^{\star} \quad , \quad A_2 := \sigma_e^2 \mathbb{E}\left(\left\| m'_{\text{new}} M' V^{-1} \right\|^2\right)$$

$$A_3 := \theta^{\star\prime} M' V^{-1} M \text{Var}\left(m_{\text{new}}\right) M' V^{-1} M\theta^{\star} \quad , \quad A_4 := \text{Var}\left(m'_{\text{new}}\theta^{\star}\right) + \sigma_e^2.$$

We have $n\ \mathbb{E}\left(\left\| m'_{\text{new}} M' V^{-1} \right\|^2\right)$ in place of M_e

Illustration on simulated data

- 15 architectures , K=100, 1000, or 2000
- either a) two large QTLs, b) 100 small QTLs or c) a mixture between major genes and multiple small QTLs
- $n = 500$ and $n_{\text{new}} = 100$
- Population simulated by random mating between haploid individuals during a few generations

TABLE – Mean Squared Error as a function of the chosen method

MSE based on	50 generations for TEST	70 generations for TEST
$\hat{\rho}_{\text{after}}(\theta^{\star})$	5.9685×10^{-5}	3.8455×10^{-5}
$\hat{\rho}_{\text{after}}(\hat{\theta}^{\star}_{\text{ADLASSO}})$	1.2108×10^{-3}	1.2118×10^{-3}
$\hat{\rho}_{\text{before}}(\hat{\theta}^{\star}_{\text{ADLASSO}})$	2.2677×10^{-3}	1.5168×10^{-3}
Plos One (2016)	3.3056×10^{-3}	1.007×10^{-2}
M_{e1}	3.7936×10^{-3}	1.3779×10^{-2}
M_{e2}	3.7508×10^{-3}	1.3518×10^{-2}
M_{e3}	3.6970×10^{-3}	1.3165×10^{-2}
M_{LJ}	5.5578×10^{-3}	6.1021×10^{-3}

How to improve the quality of the prediction

Idea : consider a subspace of lower dimension

Reminder : $U = \left(U^{(1)}, \ldots, U^{(r)}\right)$ orthonormal basis for the space spanned by the columns of M

Let us choose $\tilde{r}$ columns of U. Let us define $\sigma : \{1, \ldots, \tilde{r}\} \to \{1, \ldots, r\}$

Let $\widetilde{\theta}$ be the new estimator

$$\widetilde{\theta} = M' V^{-1} \widetilde{U} \widetilde{U}' Y \quad \text{where} \quad \widetilde{U} = \left(U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right)$$

where $\widetilde{U}\widetilde{U}' Y$ is the projection of Y on $Span\left\{U^{\sigma(1)}, \ldots, U^{\sigma(\tilde{r})}\right\}$.

Let us note $\widetilde{W} = \left(W^{\sigma(1)}, \ldots, W^{\sigma(\tilde{r})}\right)$

$\Rightarrow$ Prediction and accuracy built on the new estimator $\widetilde{\theta}$

$$\widetilde{Y}_{\text{new}} = m'_{\text{new}} \widetilde{\theta} \quad , \qquad \widetilde{\rho} = Cor\left(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}}\right) = \frac{\text{Cov}\left(\widetilde{Y}_{\text{new}}, \ Y_{\text{new}}\right)}{\sqrt{\text{Var}\left(\widetilde{Y}_{\text{new}}\right)\text{Var}\left(Y_{\text{new}}\right)}}$$

When does the accuracy increase ?

- Ridge estimator $\hat{\theta}$ based on all the columns of U
  - accuracy $\hat{\rho}$, prediction $\hat{Y}_{\text{new}}$
- New estimator $\widetilde{\theta}$ based on $\tilde{r}$ columns of U
  - accuracy $\tilde{\rho}$, prediction $\widetilde{Y}_{\text{new}}$
- Complementary estimator $\vec{\theta}$ of the new estimator, based on the $r - \tilde{r}$ remaining columns of U
  - accuracy $\vec{\rho}$, prediction $\vec{Y}_{\text{new}}$

Notations :

$$\widehat{A_1} = \widehat{\text{Cov}}\left(\hat{Y}_{\text{new}},\ Y_{\text{new}}\right)\ ,\ \widehat{A_2} + \widehat{A_3} = \widehat{\text{Var}}\left(\hat{Y}_{\text{new}}\right)\ ,\ \widehat{A_4} = \widehat{\text{Var}}\left(Y_{\text{new}}\right)$$

$$\widehat{\widetilde{A_1}} = \widehat{\text{Cov}}\left(\widetilde{Y}_{\text{new}},\ Y_{\text{new}}\right)\ ,\ \widehat{\widetilde{A_2}} + \widehat{\widetilde{A_3}} = \widehat{\text{Var}}\left(\widetilde{Y}_{\text{new}}\right)\ ,\ \widehat{\widetilde{A_4}} = \widehat{A_4} = \widehat{\text{Var}}\left(Y_{\text{new}}\right)$$

...

Before genotyping the TEST individuals

Accuracy based on our new estimator $\tilde{\theta}$

$$\widehat{\widetilde{\rho}}_{\text{before}} = \frac{\widehat{\widetilde{A_1}}}{\left(\widehat{\widetilde{A_2}} + \widehat{\widetilde{A_3}}\right)^{1/2} \left(\widehat{\widetilde{A_4}}\right)^{1/2}},$$

where

$$\widehat{\widetilde{A_1}} = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2 + \lambda} \left\| W^{(\sigma(s))} W^{(\sigma(s))\prime} \theta^\star \right\|^2 \quad , \quad \widehat{\widetilde{A_2}} = \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2 + \lambda)^2}$$

$$\widehat{\widetilde{A_3}} = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^6}{(d_{\sigma(s)}^2 + \lambda)^2} \left\| W^{(\sigma(s))} W^{(\sigma(s))\prime} \theta^\star \right\|^2 \quad , \quad \widehat{\widetilde{A_4}} = \widehat{A_4}.$$

The Me part, $\widehat{\widetilde{A_2}}$, is smaller than previously ...

The 3 possible situations (non asymptotic result)

1. We have $\hat{\tilde{\rho}}_{\text{before}} \geq \hat{\rho}_{\text{before}}$ if and only if

$$\frac{\widehat{\text{Cov}}\left(\tilde{Y}_{\text{new}}, Y_{\text{new}}\right)}{\widehat{\text{Cov}}\left(\vec{Y}_{\text{new}}, Y_{\text{new}}\right)} \geq \frac{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}}\right).$$

   In this case, we also have $\hat{\tilde{\rho}}_{\text{before}} \geq \hat{\vec{\rho}}_{\text{before}}$.

2. We have $\hat{\vec{\rho}}_{\text{before}} \geq \hat{\rho}_{\text{before}}$ if and only if

$$\frac{\widehat{\text{Cov}}\left(\tilde{Y}_{\text{new}}, Y_{\text{new}}\right)}{\widehat{\text{Cov}}\left(\vec{Y}_{\text{new}}, Y_{\text{new}}\right)} \leq \sqrt{1 + \frac{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)}} - 1.$$

   In this case, we also have $\hat{\vec{\rho}}_{\text{before}} \geq \hat{\tilde{\rho}}_{\text{before}}$.

3. We have $\hat{\rho}_{\text{before}} \geq \hat{\tilde{\rho}}_{\text{before}}$ and $\hat{\rho}_{\text{before}} \geq \hat{\vec{\rho}}_{\text{before}}$ if and only if

$$\sqrt{1 + \frac{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)}} - 1 \leq \frac{\widehat{\text{Cov}}\left(\tilde{Y}_{\text{new}}, Y_{\text{new}}\right)}{\widehat{\text{Cov}}\left(\vec{Y}_{\text{new}}, Y_{\text{new}}\right)} \leq \frac{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)} \left(1 + \sqrt{1 + \frac{\widehat{\text{Var}}\left(\vec{Y}_{\text{new}}\right)}{\widehat{\text{Var}}\left(\tilde{Y}_{\text{new}}\right)}}\right).$$

An example where $\theta^\star$ belongs to the subspace spanned by the rows of M

- $\theta^\star = 0.3W^{(1)} + 0.3W^{(2)} + 0.3W^{(3)}$
- $\tilde{r}$ and the columns of U chosen by cross validation
- $K = 1000$ markers, $n = 500$ or 800
- $n_{\text{new}} = 100$

σ_e^2	n	Method	200 generations	400 generations
1	500	$\hat{\rho}$	0.7550	0.6721
		$\hat{\hat{\rho}}$	0.7810	0.7041
	800	$\hat{\rho}$	0.7487	0.7037
		$\hat{\hat{\rho}}$	0.7728	0.7312
9	500	$\hat{\rho}$	0.3370	0.2623
		$\hat{\hat{\rho}}$	0.3809	0.3079
	800	$\hat{\rho}$	0.3317	0.2904
		$\hat{\hat{\rho}}$	0.3734	0.3330