On the accuracy in high dimensional linear models under imperfect linkage disequilibrium

C.E. Rabier[a,b], S. Grusea[c]

[a]ISEM, Université de Montpellier, CNRS, France
[b]LIRMM, Université de Montpellier, CNRS, France
[c]Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France

Abstract. Genomic selection (GS) consists in predicting breeding values of selection candidates, using a large number of genetic markers. An important question in GS is the determination of the number of markers required for a good prediction. Many studies show that it becomes useless to consider too many markers. In contrast, for some species, the number of markers remains too small to cover the huge genome size. Under such sparse genetic map, it is likely to observe some imperfect linkage disequilibrium: the alleles at a gene location and at a marker located nearby vary. In this context, we tackle here the problem of imperfect linkage disequilibrium in the Ridge regression framework. We present theoretical results regarding the accuracy criteria, i.e., the correlation between predicted value and true value. We show the influence of the projection of the causal regression function (i.e. at genes) on the space spanned by the columns of the design matrix (i.e. at markers). Asymptotic results, in a high dimensional framework, are given, and we prove that the convergence to an optimal accuracy depends on a few limiting factors. This study generalizes our recent results (Rabier et al. (2018)) obtained under perfect linkage disequilibrium. Last, illustrations on simulated and real data are proposed.

Keywords: Accuracy, Genomic Selection, High Dimension, Linear Model, Prediction, Ridge Regression, Singular Value Decomposition, Sparsity.

1. Introduction and background

Genomic Selection (GS), an extremely popular technique in genetics (Meuwissen et al. (2001)), consists in predicting breeding values of selection candidates using a large number of genetic markers. The goal is to predict the future phenotype of young candidates as soon as their DNA has been collected. GS was first applied to animal breeding (see Hayes et al. (2009) for a review), and GS is nowadays extensively investigated in plants. For instance, we can mention studies on apple (Muranty et al. (2015)), eucalyptus (Tan et al. (2017)), japanese pears (Minamikawa et al. (2018)), strawberry (Gezan et al. (2017)), banana (Nyine et al. (2018)), coffea (Ferrao et al. (2018)) ... Note that in medicine, the predictive ability of complexe diseases with the help of genome data, is also a topic of large interest (e.g. Lee et al. (2017), Abraham et al. (2014)). All these application fields make the topic "genomic prediction" very exciting for geneticists and statisticians, eager to propose new tools for improving the predictions (see Momen et al. (2018)).

From a methodological point of view, GS relies on the expectation that each QTL will be highly correlated with at least one marker (Schulz-Streeck et al. (2012)). In genetics, this correlation is named Linkage Disequilibrium (LD): it refers to the non independence of alleles at 2 different loci (see Durett (2008) for more details). A usual estimator of LD is the square of Pearson correlation. However, several factors are known to be responsible for artificial LD in a population (e.g. relatedness, population structure ...). In Mangin et al. (2012), the authors proposed new LD estimators (so called novel measures) that correct bias due to relatedness and population structure. These measures seem to be key elements in GS since they are also present in our formula (Rabier et al. (2016)) on prediction in GS.

The aim of this paper is to generalize our recent theoretical study on the accuracy of genomic prediction (Rabier et al. (2018)) in GS to the case of imperfect LD. Indeed, in our study, we focused only on perfect LD: QTLs were located exactly on a few markers. When QTLs do not match marker locations, we generally observe imperfect LD since the alleles generally vary at a QTL location and at a marker located nearby. Imperfect LD is a topic of interest since for some species, the number of markers remains too small to cover the huge genome size. In that sense, this density of markers is unable to perfectly tag QTL locations.

An underlying research topic in GS is the determination of the number of markers required for implementing GS. In their study on maize population, Zhang et al. (2015) showed that the prediction of a complex trait required a large number of markers (around 58000 markers thanks to Genotyping By Sequencing, and after filtering), whereas 200 markers were sufficient for predicting a simple trait. In our study on GS in raygrass (Rabier et al. (2016)), we showed that 24957 markers were unable to cover the entire genome (2.7 Gb). Furthermore, in a recent study on GS in coffea, Ferrao et al. (2018) showed that predictions relying on 4000 markers gave similar results as those based on 35000 markers. In this context, we propose to tackle here the problem of imperfect LD in GS.

In what follows, we will focus on Ridge regression since it is one of the most popular method chosen by geneticists to perform predictions.

1.1. A linear model

Let us introduce the statistical model associated to GS. The quantitative trait is observed on n training (TRN) individuals and we denote by $Y_1, \ldots, Y_n$ the observations. p markers lie on the genome. In what follows, X is a matrix of size $n \times p$, and $'$ denotes transposition. The i-th row of X, written as $x_i' = (X_{i,1}, ..., X_{i,p})$, represents the genome information at each marker available for the i-th individual.

m QTLs lie on the genome, having an effect on the quantitative trait. $\beta_j^\star$ for $1 \leq j \leq m$ refers to the j-th QTL effect, and $X^\star$ is the analogue of X at QTL locations.

We assume the following causal linear model for the quantitative trait:

$$Y = X^\star \beta^\star + \varepsilon, \tag{1}$$

where $Y = (Y_1, ..., Y_n)'$, $\beta^\star = (\beta_1^\star, ..., \beta_m^\star)'$, $\varepsilon \sim N(0, \sigma_e^2 I_n)$, I_n is the identity matrix of size n and σ_e^2 refers to the environmental variance.

In this manuscript, we will propose an analysis conditional on $x_1, \ldots, x_n$, $x_1^\star, \ldots, x_n^\star$. However, before imposing this conditioning, we have to precise that the matrix $X^\star$ is independent of ε. Besides, some correlation is present between the matrices $X^\star$ and X: for instance, due to the fixed genome size, x_i and $x_i^\star$ are necessarily correlated. Simulated data will be generated accordingly. In what follows, r (resp. $r^\star$) will denote the rank of the matrix X (resp. $X^\star$), and $\mathcal{R}_{\text{rows}}(X)$ (resp. $\mathcal{R}_{\text{rows}}(X^\star)$) will refer to the linear space generated by the rows of X (resp. $X^\star$). In the same way, $\mathcal{R}_{\text{col}}(X)$ and $\mathcal{R}_{\text{col}}(X^\star)$ will denote to the linear space spanned by the columns.

1.2. Introducing a test individual

A supplementary individual, so-called test (TST) individual (denoted *new*) is genotyped but not phenotyped. Using same notations as those used for the TRN population, $x^\star_{new}$ denotes the column vector containing the genome information at the m QTLs of the individual *new*. As a result, the quantitative trait Y_{new} can be written

$$Y_{new} = x^{\star\prime}_{new}\ \beta^\star + \varepsilon_{new},$$

where $\varepsilon_{new} \sim N(0, \sigma_e^2)$.

We suppose that $x^\star_{new}$, ε_{new} and ε are all independent. Using same notations as before, x_{new} denotes the random genome information at markers, and x_{new} and $x^\star_{new}$ are correlated because of the fixed genome size responsible for some genetic linkage.

1.3. Introducing the accuracy and the prediction model

In GS, we are interested in predicting either the genotypic value $x^{\star\prime}_{new}\ \beta^\star$, or the phenotypic value Y_{new}. In both cases, an estimator $\hat{Y}_{new}$ is constructed from a prediction model learned on n TRN individuals. $\hat{Y}_{new}$ is a function of the random variables x_{new} and ε. Then, the quality of the prediction is evaluated according to some accuracy criteria, i.e. the correlation between predicted and true values. The *phenotypic accuracy*, ρ_{ph}, also called predictive ability, is defined in the following way (e.g. Visscher et al. (2010))

$$\rho_{ph} := \frac{\operatorname{Cov}\left(\hat{Y}_{new},\ Y_{new}\right)}{\sqrt{\operatorname{Var}\left(\hat{Y}_{new}\right)\operatorname{Var}\left(Y_{new}\right)}}, \tag{2}$$

whereas the *genotypic accuracy*, ρ_g, is defined as (e.g. Daetwyler et al. (2008, 2010))

$$\rho_g := \frac{\operatorname{Cov}\left(\hat{Y}_{new},\ x^{\star\prime}_{new}\beta^\star\right)}{\sqrt{\operatorname{Var}\left(\hat{Y}_{new}\right)\operatorname{Var}\left(x^{\star\prime}_{new}\beta^\star\right)}}\ . \tag{3}$$

Note that, when $x^\star_{new}$, ε_{new} and ε are all independent, these two accuracies are linked by the relationship $\rho_{ph}/\rho_g = h$, where h is the squared root of the heritability of the trait:

$$h^2 := \frac{\operatorname{Var}\left(x^{\star\prime}_{new}\ \beta^\star\right)}{\operatorname{Var}\left(Y_{new}\right)}\ . \tag{4}$$

In what follows, we set $\sigma_G^2 = \mathrm{Var}\left(x_{new}^{\star\prime}\,\beta^\star\right)$, and as a consequence, we have the relationship $h^2 = \sigma_G^2 / \left(\sigma_G^2 + \sigma_e^2\right)$.

Besides, the *oracle situation* will denote the settings where the QTLs locations and their effects are known. Then, under the oracle situation, the natural predictor is $\hat{Y}_{new}^{oracle} = x_{new}^{\star\prime}\,\beta^\star$. As a result, according to formula (2), the oracle accuracies are the following

$$\rho_g^{oracle} = 1\ , \quad \rho_{ph}^{oracle} = h.$$

As in our previous study (Rabier et al, 2018), we will focus on Ridge regression (Tihonov (1963); Hoerl et al. (1970)), called random regression best linear unbiased predictor (RRBLUP) in genetics. It is known that RRBLUP is equivalent to genomic best linear unbiased predictor (GBLUP).

The Ridge estimator, based on genome information at markers, presents the advantage to be suitable in a high dimensional setting (i.e. $p > n$). Its expression is the following:

$$\hat{\beta} := \left(X'X + \lambda I_p\right)^{-1} X'Y, \tag{5}$$

where λ refers to a regularization (or tuning) parameter, and I_p denotes the identity matrix of size $p \times p$.

Before presenting our roadmap, let us introduce a notation regarding perfect LD.

Notation: Under perfect LD, the m QTLs are located on a few markers and β denotes the sparse vector of size p, containing the components of $\beta^\star$.

1.4. Our contributions and roadmap

Since this paper is a generalization of Rabier et al. (2018), we will follow the same outline as in our previous article. It should make the reading easier, and should help the reader to compare the different results.

Our study starts in Section 2, by recalling a recent formula on the accuracy, suitable under imperfect LD. We also introduce two singular value decompositions, the one of the design matrix (i.e. at markers), and the one of the causal matrix (i.e. at genes). Then, we state our Theorem 1, the analogue of Theorem 1 of Rabier et al. (2018), dealing here with imperfect LD. This theorem is somewhat essential since other results, appropriate under imperfect LD, are built on it.

Section 3 is devoted to the case where TRN and TST are sampled from the same probability distribution. Theorem 2 introduces an estimation $\hat{\rho}_g$ of ρ_g that does not require the genome information of TST individuals. According to this theorem, the projection of the regression function $X^\star\beta^\star$ on $\mathcal{R}_{\mathrm{col}}(X)$ is a key element for the genotypic accuracy. From Theorem 2, we can retrieve results under perfect LD: the key factor becomes the projection of the signal β on $\mathcal{R}_{\mathrm{rows}}(X)$ (as in Rabier et al. (2018)). Lemma 1 introduces, under imperfect LD, a lower bound for $\hat{\rho}_g$: it takes account a global projection (same weights on each subspace) of $X^\star\beta^\star$ on the space spanned by the columns of X. Lemma 2 assumes that the signal $\beta^\star$ is spread out uniformly on each subspace of $\mathcal{R}_{\text{rows}}(X^\star)$. The oracle accuracy is reached as soon as the limit of a loss factor (so called 1-lim $\sqrt{\xi(p,n)}$) is equal to zero. Lemma 2 relies on different assumptions than the ones assumed in Lemma 2 of Rabier et al. (2018). In that sense, our Lemma 2 is not a generalization to imperfect LD.

Section 4 investigates the case where TRN and TST are not sampled from the same probability distribution. The genome information of TST individuals needs to be known, in order to compute the estimator $\hat{\rho}_g$ of ρ_g.

Last, Section 5 introduces the modified predictor $\hat{\hat{\rho}}_g$ of Rabier et al. (2018), that may improve the quality of the prediction. Recall that it relies on the projection of Y on a well chosen subspace of $\mathcal{R}_{\text{col}}(X)$. Lemma 4 proposes an estimation of that predictor's accuracy: as expected, under imperfect LD, it depends on the projection of the regression function $X^\star\beta^\star$ on the chosen subspace. After having introduced bounds for $\hat{\hat{\rho}}_g$ in Lemma 5, we will show that Lemma 6 of Rabier et al. (2018) that compares $\hat{\hat{\rho}}_g$ and $\hat{\rho}_g$ is still suitable under imperfect LD.

To conclude, in Section 6, we will illustrate our theoretical results on simulated and real data. We propose to investigate a topic in GS that has not been studied before (as far as we know): the accuracy of the prediction when the genetic map of TRN differs from the one of TST. In particular, we suggest to consider a more dense map for TRN than for TST: the dense TRN map will help to estimate the nuisance parameters $X^\star$ and $\beta^\star$, required to compute our estimation $\hat{\rho}_g$. This concept relies on the expectation that QTLs will be in perfect LD with markers under this dense TRN map, which is not the case for the TST map (imperfect LD). Contrary to our "perfect LD" study where the Adaptive LASSO (Zou (2006)) was found to be the best substitute for β, we found here that the LASSO (Tibshirani (1996)) was the best substitute for $\beta^\star$ when a sparse TST map was considered. Moreover, the Adaptive LASSO was more appropriate for a dense TST map.

Finally, performances of the modified ridge estimator are also illustrated, and we analyzed real data of Spindel et al. (2015) on GS in rice, considering different density of markers. With the help of our "imperfect LD" proxies, we will show that geneticists can evaluate the accuracy of their prediction and figure out if they should redensify their genetic map to improve the reliability of their predictions.

2. General expression for the accuracy

2.1. An existing formula suitable under imperfect LD

Since we have the well-known relationship

$$(X'X + \lambda I_p)^{-1} X' = X' (XX' + \lambda I_n)^{-1}, \tag{6}$$

the computation of $\hat{\beta}$ only requires the inversion of a $n \times n$ matrix.

In this context, the predictor for the so-called *new* individual is the following:

$$\hat{Y}_{new} := x'_{new}\hat{\beta} = x'_{new}X'V^{-1}Y \text{ , where } V = XX' + \lambda I_n.$$

In what follows, we will assume that Y, Y_{new}, x_{new}, $x^{\star}_{new}$, the columns of X and the columns of $X^{\star}$ are centered.

Assuming that $x_1, \dots, x_n, x^{\star}_1, \dots, x^{\star}_n$ are known, and that ε, x_{new}, $x^{\star}_{new}$ and ε_{new} are random, the genotypic accuracy, according to formula (5) of Rabier et al. (2016), has the following expression:

$$\rho_g = \frac{\beta^{\star\prime}\ \mathbb{E}\left(x^{\star}_{new}x'_{new}\right) X'V^{-1}X^{\star}\beta^{\star}}{\left(\sigma_e^2\mathbb{E}\left(\left\|x'_{new}X'V^{-1}\right\|^2\right) + \beta^{\star\prime}X^{\star\prime}V^{-1}X\mathrm{Var}\left(x_{new}\right)X'V^{-1}X^{\star}\beta^{\star}\right)^{1/2}\sigma_G} \tag{7}$$

where $\|.\|$ is the L^2 norm.

We introduce the following notations

$$A_1 := \beta^{\star\prime}\ \mathbb{E}\left(x^{\star}_{new}x'_{new}\right) X'V^{-1}X^{\star}\beta^{\star} \quad , \quad A_2 := \sigma_e^2\mathbb{E}\left(\left\|x'_{new}X'V^{-1}\right\|^2\right)$$
$$A_3 := \beta^{\star\prime}X^{\star\prime}V^{-1}X\mathrm{Var}\left(x_{new}\right)X'V^{-1}X^{\star}\beta^{\star} \quad , \quad A_4 := \sigma_G^2.$$

2.2. SVD decomposition

Following Shao and Deng (2012) and Bühlmann (2013), let us consider the singular value decomposition of X:

$$X = PDQ', \tag{8}$$

where P is a $n \times r$ matrix satisfying $P'P = I_r$, Q is a $p \times r$ matrix satisfying $Q'Q = I_r$, and $D = Diag\left(d_1, \dots, d_r\right)$ with $d_1 \geq \ldots \geq d_r > 0$. The columns of Q (resp. P) constitute an orthogonal basis of the space spanned by the rows (resp. columns) of X. In what follows, $Q^{(s)}$ will denote the s-th column of Q, and as a consequence $\mathcal{R}(X) = Span\left\{Q^{(1)}, \dots, Q^{(r)}\right\}$. By construction QQ' is an idempotent matrix, and as mentioned in Shao and Deng (2012), we have the relationship

$$QQ'\hat{\beta} = \hat{\beta}.$$

In other words, since the projection of $\hat{\beta}$ onto $\mathcal{R}(X)$ is still $\hat{\beta}$, the ridge estimator is always in $\mathcal{R}(X)$.

In the same way, let us introduce the singular value decomposition of $X^{\star}$:

$$X^{\star} = P^{\star}D^{\star}Q^{\star\prime}, \tag{9}$$

where $P^{\star}$ is a $n \times r^{\star}$ matrix satisfying $P^{\star\prime}P^{\star} = I_{r^{\star}}$, $Q^{\star}$ is a $m \times r^{\star}$ matrix satisfying $Q^{\star\prime}Q^{\star} = I_{r^{\star}}$, and $D^{\star} = Diag\left(d^{\star}_1, \dots, d^{\star}_{r^{\star}}\right)$ with $d^{\star}_1 \geq \ldots \geq d^{\star}_{r^{\star}} > 0$.

2.3. Results

Our first theorem can be viewed as the analogue of Theorem 1 of Rabier et al. (2018), dealing with imperfect LD.

Theorem 1. *Let us assume that X and $X^\star$ are known, and that ε, x_{new}, $x^\star_{new}$ and ε_{new} are random. Then, the genotypic accuracy has the following expression*

$$\rho_g = \frac{A_1}{\left(A_2 + A_3\right)^{1/2} \left(A_4\right)^{1/2}}$$

where

$$\begin{aligned}
A_1 &= \sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda}\ \beta^{\star\prime}\ \mathbb{E}\left(x^\star_{new} x'_{new}\right)\ Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \quad , \\
A_2 &= \sigma_e^2 \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2+\lambda)^2}\ \mathbb{E}\left(\left\| Q^{(s)} Q^{(s)\prime} x_{new} \right\|^2\right) \quad , \\
A_3 &= \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda}\ Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)' \ \mathbb{E}\left(x_{new} x'_{new}\right) \\
&\quad \times \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda}\ Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right) \quad , \\
A_4 &= \beta^{\star\prime} \mathbb{E}\left(x^\star_{new} x^{\star\prime}_{new}\right) \beta^\star .
\end{aligned}$$

The proof is given in Section 7.1. The phenotypic accuracy is obtained by replacing the term A_4 at the denominator by $A_4 + \sigma_e^2$.

Note that another expression of A_1 is the following:

$$A_1 = \sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda}\ \beta^{\star\prime}\ \mathbb{E}\left(x^\star_{new} x'_{new}\right)\ Q^{(s)} Q^{(s)\prime} Q^{(s)} P^{(s)\prime} X^\star \beta^\star . \tag{10}$$

Recall that under perfect LD, the QTLs are located on a few markers and β denotes the sparse vector of size p, containing the components of $\beta^\star$. According to the above formula (10), we can notice that the term $d_s Q^{(s)} Q^{(s)\prime} \beta$ from Theorem 1 of Rabier et al. (2018) has been replaced here by the quantity $Q^{(s)} Q^{(s)\prime} Q^{(s)} P^{(s)\prime} X^\star \beta^\star$. In other words, under imperfect LD, we have to consider the projection of the vector $P^{(s)\prime} X^\star \beta^\star Q^{(s)}$ on $Span\left\{Q^{(s)}\right\}$, whereas under perfect LD, the projection of $d_s \beta$ on $Span\left\{Q^{(s)}\right\}$ is taken into account. Same remark holds for A_3 at the denominator.

Last, since formulas obtained under imperfect LD are more general, we can easily find formulas suitable under perfect LD from formulas obtained under imperfect LD. We just have to consider that the regression function is the same (i.e. $X^\star \beta^\star = X\beta$), and $X^\star \beta^\star Q^{(s)}$ is obviously equal to $d_s \beta$.

3. Estimation when TRN and TST samples come from the same probability distribution

In this section, let us consider the case where the TRN and TST samples are from the same probability distribution. In this context, using the empirical covariances $X^{\star\prime}X/n$, $X'X/n$ and $X^{\star\prime}X^{\star}/n$ as estimates of the covariances $\mathbb{E}\left(x^{\star}_{new}x'_{new}\right)$, $\mathbb{E}\left(x_{new}x'_{new}\right)$ and $\mathbb{E}\left(x^{\star}_{new}x^{\star\prime}_{new}\right)$ appearing in Theorem 1, we obtain the following theorem.

Theorem 2. *Let us assume that $x_1, \ldots, x_n$ and x_{new} are independent and identically distributed (i.i.d.). In the same way, let us assume that $x^{\star}_1, \ldots, x^{\star}_n$ and $x^{\star}_{new}$ are i.i.d. Besides, let us consider that X and $X^{\star}$ are known, and that ε, x_{new} and ε_{new} are random. Then, an estimation of the genotypic accuracy is*

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

where

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^2}{d_s^2+\lambda}\left\|P^{(s)}P^{(s)\prime}X^{\star}\beta^{\star}\right\|^2 \quad ,\\
\hat{A}_2 &= \frac{\sigma_e^2}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2} \quad , \quad \hat{A}_3 = \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\left\|P^{(s)}P^{(s)\prime}X^{\star}\beta^{\star}\right\|^2\\
\hat{A}_4 &= \frac{1}{n}\sum_{\ell=1}^{r^{\star}} d_\ell^{\star 2}\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^{\star}\right\|^2 .
\end{aligned}$$

The proof is given in Section 7.2.

We can see that the term $d_s^2\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2$ from Theorem 2 of Rabier et al. (2018) has been replaced by the quantity $\left\|P^{(s)}P^{(s)\prime}X^{\star}\beta^{\star}\right\|^2$, in the expressions of $\hat{A}_1$ and $\hat{A}_3$. As said before, this theorem is more general than Theorem 2 of Rabier et al. (2018): we can easily switch from imperfect LD formulas to perfect LD formulas as soon as we impose $X^{\star}\beta^{\star} = X\beta$.

Let us now give bounds for the quantity $\hat{\rho}_g$.

Lemma 1 (Bounds on $\hat{\rho}_g$). *Using same assumptions as in Theorem 2, we always have*

$$\frac{\left\|PP'X^{\star}\beta^{\star}\right\|^2 \ \min_s \frac{d_s^2}{d_s^2+\lambda}}{\sqrt{\sigma_e^2\ r + \left\|PP'X^{\star}\beta^{\star}\right\|^2 \ \max_s \frac{d_s^4}{(d_s^2+\lambda)^2}}\ \sqrt{\left\|Q^{\star}Q^{\star\prime}\beta^{\star}\right\|^2 \ \max_\ell d_\ell^{\star 2}}} \leq \hat{\rho}_g \leq \rho_g^{\text{oracle}}.$$

The proof is given in Section 1 of the Supplementary material.

Note that $\hat{A}_1$ and $\hat{A}_3$ can be rewritten in the following way:

$$\hat{A}_1 = \frac{1}{n}\sum_{s=1}^{r} \beta^{\star\prime} \frac{d_s^2}{d_s^2+\lambda} \sum_{\ell=1}^{r^\star} Q^{\star(\ell)} d_\ell^\star P^{\star(\ell)\prime} P^{(s)} \sum_{j=1}^{r^\star} d_j^\star P^{(s)\prime} P^{\star(j)} Q^{\star(j)\prime} \beta^\star \ ,$$

$$\hat{A}_3 = \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \left(\sum_{\ell=1}^{r^\star} d_\ell^\star P^{(s)\prime} P^{\star(\ell)} Q^{\star(\ell)\prime} \beta^\star \right)^2 .$$

3.1. Asymptotic study of $\hat{\rho}_g$ when $n \to +\infty$ and $p \to +\infty$, m bounded

Recall that $d_1^\star \geq d_2^\star \geq \ldots \geq d_{r^\star}^\star > 0$ are the singular values of $X^\star$, and that $d_1 \geq d_2 \geq \ldots \geq d_r > 0$ are the singular values of X. Note that since the number of QTLs m is bounded, the rank $r^\star$ is bounded. In contrast, the rank r may diverge because we let p and n tend to $+\infty$ in our high dimensional setting.

To study asymptotic properties of $\hat{\rho}_g$, we consider that

$$d_1^{\star 2} \sim n^\psi \ \text{ with } 0 < \psi \leq 1,$$
$$d_{r^\star}^{\star 2} \sim n^\eta \ \text{ with } \eta \leq \psi \leq 1 \ \text{ and } \ \eta \text{ and } \psi \ \text{ do not depend on } n.$$

Recall that $u_n \sim v_n$ means that $\frac{u_n}{v_n} \longrightarrow 1$ when $n \to \infty$. Besides, we will assume that

$$\left\| Q^\star Q^{\star\prime} \beta^\star \right\|^2 \sim n^{2\tau} \ \text{ with } \tau < \eta \ \text{ and } \ \tau \ \text{ not depending on } n.$$

Although that $r^\star$ is bounded in our study, these conditions are somewhat inspired from Shao and Deng (2012) and Fan and Lv (2008). Let us consider a regularization parameter λ such as $\lambda \to \infty$ and $\lambda = o\left(d_1^{\star 2}\right)$. Besides, $\Omega_1^\star$, $\Omega_2^\star$, $\Omega_3^\star$ and their analogues Ω_1, Ω_2, Ω_3, will denote the following sets:

$$\Omega_1 = \left\{ s \,\middle|\, \lambda = o(d_s^2) \right\},\ \Omega_2 = \left\{ s \,\middle|\, d_s^2 \ \sim \frac{1}{C_s}\lambda \text{ with } C_s > 0 \right\},\ \Omega_3 = \left\{ s \,\middle|\, d_s^2 = o(\lambda) \right\},$$

$$\Omega_1^\star = \left\{ \ell \,\middle|\, \lambda = o(d_\ell^{\star 2}) \right\},\ \Omega_2^\star = \left\{ \ell \,\middle|\, d_\ell^{\star 2} \ \sim \frac{1}{C_\ell^\star}\lambda \text{ with } C_\ell^\star > 0 \right\},\ \Omega_3^\star = \left\{ \ell \,\middle|\, d_\ell^{\star 2} = o(\lambda) \right\}.$$

Note that $\Omega_1^\star$ contains at least the index 1. This study is singular because of the 6 sets considered instead of the 3 sets from our previous "perfect LD" study.

3.1.1. The projected signal is spread out uniformly on each subspace

Let us define the following sets Ω_k^ℓ associated to ℓ:

$$\forall k \in \{1,2,3\}\,, \quad \Omega_k^\ell := \left\{ s \in \Omega_k \mid \left\| P^{(s)} P^{(s)\prime} P^{\star(\ell)} \right\|^2 \neq 0 \right\}.$$

We impose all the sets Ω_k^ℓ to be disjoints ($\bigcap_{1 \leq k \leq 3} \bigcap_{1 \leq \ell \leq r^\star} \Omega_k^\ell = \emptyset$). In other words, a given "s" can not tag different ℓ.

Besides, $\forall \ell \in \Omega_1^\star$, we will impose the corresponding set Ω_1^ℓ to be non empty: each "ℓ" associated to a large singular value of $X^\star$, is tagged by a few "s" associated to large singular values of X. It implies that $\#\Omega_1^\star \leq \#\Omega_1$. Note that this condition is not required for the other sets associated to ℓ: Ω_2^ℓ and Ω_3^ℓ may be empty or not. In that sense, each $\ell \in \Omega_1^\star$ can also be tagged by some "s" that belong to Ω_2 or Ω_3.

Furthermore, let us consider a more general ℓ, i.e. $1 \leq \ell \leq r^\star$, and taking into account that $\left\|P^{\star(\ell)}\right\|^2 = 1$, let us define the quantities ξ_1^ℓ, ξ_2^ℓ and ξ_3^ℓ such as:

$$
\begin{aligned}
&\xi_1^\ell 1_{\#\Omega_1^\ell \neq 0} + \xi_2^\ell 1_{\#\Omega_2^\ell \neq 0} + \xi_3^\ell 1_{\#\Omega_3^\ell \neq 0} \leq 1 \text{ with } \left(\xi_1^\ell, \xi_2^\ell, \xi_3^\ell\right) \in]0,1]^3 \ , \\
&\forall s \in \Omega_1^\ell \ , \ \left\|P^{(s)} P^{(s)\prime} P^{\star(\ell)}\right\|^2 \sim \frac{\xi_1^\ell}{\#\Omega_1^\ell} \text{ if } \#\Omega_1^\ell \neq 0 \ , \\
&\forall s \in \Omega_2^\ell \ , \ \left\|P^{(s)} P^{(s)\prime} P^{\star(\ell)}\right\|^2 \sim \frac{\xi_2^\ell}{\#\Omega_2^\ell} \text{ if } \#\Omega_2^\ell \neq 0 \ , \\
&\forall s \in \Omega_3^\ell \ , \ \left\|P^{(s)} P^{(s)\prime} P^{\star(\ell)}\right\|^2 \sim \frac{\xi_3^\ell}{\#\Omega_3^\ell} \text{ if } \#\Omega_3^\ell \neq 0
\end{aligned}
$$

where $\#\Omega$ refers to the cardinal of the set Ω.

In other words, the projection of $P^{\star(\ell)}$ on $\mathit{Span}\left\{P^{(1)}, \ldots, P^{(r)}\right\}$ is spread out on the subspaces $\underset{s \in \Omega_1^\ell}{\mathit{Span}}\left\{P^{(s)}\right\}$, $\underset{s \in \Omega_2^\ell}{\mathit{Span}}\left\{P^{(s)}\right\}$, and $\underset{s \in \Omega_3^\ell}{\mathit{Span}}\left\{P^{(s)}\right\}$. Note that within each subspace $\underset{s \in \Omega_k^\ell}{\mathit{Span}}\left\{P^{(s)}\right\}$, the projection is spread out uniformly on each component $P^{(s)}$.

Let us consider a few extra conditions. In what follows, conditions denoted with a star are specific to this paper, whereas others were already present in Rabier et al. (2018).

- $(C1^\star) \dfrac{n^{2\tau}}{r^\star} \sum_{\ell \in \Omega_1^\star} d_\ell^{\star 2} \to +\infty$
- $(C2) \sum_{s \in \Omega_3} d_s^2 = o(\lambda)$
- $(C3) \sum_{s \in \Omega_3} d_s^4 = o(\lambda^2)$
- $(C4^\star) \dfrac{n^{2\tau}}{r^\star} = o(1/\lambda)$
- $(C5) \#\Omega_1 = O(1)$
- $(C6) \#\Omega_2 = O(1)$
- $(C7^\star) \dfrac{n^{2\tau}}{r^\star} \sum_{\ell \in \Omega_1^\star} \xi_2^\ell d_\ell^{\star 2} = o(1)$
- $(C8^\star) \dfrac{n^{2\tau}}{r^\star} \sum_{\ell \in \Omega_1^\star} \xi_3^\ell d_\ell^{\star 2} = o(1)$

Because of conditions $(C5)$ and $(C6)$, when $p > n$, the rank r of the matrix X which is bounded by n, will diverge to $+\infty$ if and only if the number of elements of Ω_3 diverges. On the other hand, since the number m of QTLs is bounded, the rank $r^\star$ of the matrix $X^\star$ is bounded and $\Omega_1^\star$, $\Omega_2^\star$ and $\Omega_3^\star$ are finite sets. Some intuition and explanations on these conditions are given in Section 4 of the Supplementary material.

The following lemma, so-called Lemma 2, assumes imperfect LD and that the signal is spread out uniformly on each subspace of $\mathcal{R}_{\text{rows}}(X^\star)$. This lemma is different from Lemma 2 of Rabier et al. (2018), which was restricted to perfect LD, and supposed that the signal was spread out uniformly on the subspaces of $\mathcal{R}_{\text{rows}}(X)$.

Lemma 2 (Convergence to the oracle accuracy). *Let us consider same assumptions as in Theorem 2. Besides, let us suppose that the projected signal is spread out uniformly on each subspace Span* $\left\{Q^{\star(\ell)}\right\}$, *i.e.*

$$\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 \sim \frac{n^{2\tau}}{r^\star}, \ell = 1, \ldots, r^\star. \tag{11}$$

Moreover, let us assume that $\forall \ell \in \Omega_1^\star\ \Omega_1^\ell \neq \oslash$, *and let us set* $\xi_1^\ell = \xi(p,n)$ $\forall \ell \in \Omega_1^\star$, *with* $0 < \xi(p,n) \leq 1$. *Then, assuming that* $\bigcap\limits_{1\leq k\leq 3}\ \bigcap\limits_{1\leq \ell\leq r^\star} \Omega_k^\ell = \oslash$ *and conditions* $(C1^\star - C2 - C3 - C4^\star - C5 - C6 - C7^\star - C8^\star)$,

- *for large* p *and* n, *we have* $\hat{\rho}_g \sim \sqrt{\xi(p,n)}\ \rho_g^{oracle}$
- *when* $p \to +\infty$ *and* $n \to +\infty$, *we have* $\hat{\rho}_g \longrightarrow \left(\lim \sqrt{\xi(p,n)}\right) \rho_g^{oracle}$
- *if* $\forall \ell \in \Omega_1^\star$, $\xi_2^\ell = 1/n^{\theta_1}$ *and* $\xi_3^\ell = 1/n^{\theta_2}$ *with* $\theta_1 > \psi$ *and* $\theta_2 > \psi$, *then we have* $\xi(p,n) \longrightarrow 1$ *and* $\hat{\rho}_g \longrightarrow \rho_g^{oracle}$.

The proof is given in Section 5 of the Supplementary material (see also Section 4 for some intuition). For each $\ell \in \Omega_1^\star$, $\xi(p,n)$ is the percentage of the L2 norm of $P^{\star(\ell)}$ represented on $\underset{s\in\Omega_1^\ell}{Span}\left\{P^{(s)}\right\}$. Note that under our conditions, we are only able to capture this percentage of the L2 norm of $P^{\star(\ell)}$ (see Sections 4 and 5 of the Supplementary material). $1 - \xi(p,n)$ can be viewed as a loss coefficient: it is the percentage of the L2 norm of $P^{\star(\ell)}$ that is unable to be captured (from either $\underset{s\in\Omega_2^\ell}{Span}\left\{P^{(s)}\right\}$, either $\underset{s\in\Omega_3^\ell}{Span}\left\{P^{(s)}\right\}$ or from the complementary subspace).

As a consequence, according to Lemma 2, $\hat{\rho}_g$ converges to $\left(\lim \sqrt{\xi(p,n)}\right) \rho_g^{oracle}$. The oracle accuracy is reached as soon as $\lim \sqrt{\xi(p,n)}$ is equal to one. Typically, this is the case when we set $\xi_2^\ell = 1/n^{\theta_1}$ and $\xi_3^\ell = 1/n^{\theta_2}$.

3.1.2. The projected signal belongs only to one component

Let us come back to the assumptions given at the beginning of Section 3.1 (before paragraph 3.1.1). In this context, we propose to study here the asymptotic behavior of our estimate $\hat{\rho}_g$ when the projected signal belongs only to one component.

Lemma 3 (Extreme cases). *Let us consider same assumptions as in Theorem 2. Besides, let us suppose that the projected signal belongs only to Span* $\left\{Q^{\star(1)}\right\}$ *that is to say*

$$\left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^{\star}\right\|^{2} \sim n^{2\tau}, \quad \left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^{\star}\right\|^{2} = 0, \text{ for } 1 < \ell \leq r^{\star}.$$

Moreover, let us assume that $\ell = 1$ *is tagged only by one* s *such as* $\left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^{2} \sim \xi(p,n)$ *with* $0 < \xi(p,n) \leq 1$, *and* $\left\|P^{(u)}P^{(u)\prime}P^{\star(1)}\right\|^{2} = 0\ \forall u \neq s$. *Then*

- *If* $s \in \Omega_1 \cup \Omega_2$, *and*
  - *if* $2\tau + \psi > 1$, *then* $\hat{\rho}_g \to \left(\lim \sqrt{\xi(p,n)}\right)\ \rho_g^{oracle}$.
  - *if* $2\tau + \psi < 1$, *then*
    - *if* $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{2\tau+\psi}\right)$, *then* $\hat{\rho}_g \to \left(\lim \sqrt{\xi(p,n)}\right)\ \rho_g^{oracle}$
    - *if* $n^{2\tau+\psi} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right)$, *then* $\hat{\rho}_g \to 0$.
- *If* $s \in \Omega_3$, $\lambda \sim Cn^{\kappa+\eta}$, $d_s \sim n^{\gamma}$ *with* $\kappa > \max(0, -\eta)$, $\gamma < (\kappa+\eta)/2$
  - *if* $4\gamma - 2\kappa - 2\eta + 2\tau + \psi > 1$, *then* $\hat{\rho}_g \to \left(\lim \sqrt{\xi(p,n)}\right)\ \rho_g^{oracle}$
  - *if* $4\gamma - 2\kappa - 2\eta + 2\tau + \psi < 1$, *then*
    - *if* $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{4\gamma-2\kappa-2\eta+2\tau+\psi}\right)$, *then* $\hat{\rho}_g \to \left(\lim \sqrt{\xi(p,n)}\right)\ \rho_g^{oracle}$
    - *if* $n^{4\gamma-2\kappa-2\eta+2\tau+\psi} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right)$, *then* $\hat{\rho}_g \to 0$.

The proof is given in Section 6 of Supplementary material.

The analogue of this lemma, dealing with a projected signal that belongs only to *Span* $\left\{Q^{\star(r^{\star})}\right\}$ is also given in Supplementary material.

4. Estimation when TRN and TST samples do not come from the same probability distribution

In this section, we will consider the general case when the TRN and TST samples do not come necessarily from the same probability distribution. Furthermore, let us assume that n_{new} *new* individuals are available, and that we are willing to predict the phenotypes of those individuals. X_{new} will be a random matrix of size $n_{new} \times p$ containing the genomic markers of the new individuals. The singular value decomposition of X_{new} is the following:

$$X_{new} = WFZ',$$

where W is a $n_{new} \times r_{new}$ matrix satisfying $W'W = I_{r_{new}}$, Z is a $p \times r_{new}$ matrix satisfying $Z'Z = I_{r_{new}}$, and F is $r_{new} \times r_{new}$ diagonal matrix of full rank.

In the same way, $X^\star_{new}$ is the random matrix at gene locations, and we consider $W^\star F^\star Z^{\star\prime}$ the SVD decomposition of $X^\star_{new}$. $r^\star_{new}$ denotes the rank of $X^\star_{new}$.

Using $X^{\star\prime}_{new} X_{new}/n_{new}$, $X'_{new} X_{new}/n_{new}$ and $X^{\star\prime}_{new} X^\star_{new}/n_{new}$ to estimate the covariances $\mathbb{E}\left(x^\star_{new} x'_{new}\right)$, $\mathbb{E}\left(x_{new} x'_{new}\right)$ and $\mathbb{E}\left(x^\star_{new} x^{\star\prime}_{new}\right)$, we obtain the following Theorem 3, a random version of Theorem 2.

Theorem 3. *Let us assume that X and $X^\star$ are given and that X_{new} and $X^\star_{new}$ are random. Besides, we suppose that the rows of X_{new} are i.i.d., and also that the rows of $X^\star_{new}$ are i.i.d. Then, an estimator of the genotypic accuracy is.*

$$\check{\rho}_g = \frac{\check{A}_1}{\left(\check{A}_2 + \check{A}_3\right)^{1/2} \left(\check{A}_4\right)^{1/2}}, \tag{12}$$

where

$$\check{A}_1 = \frac{1}{n_{new}} \sum_{s=1}^{r} \frac{d_s}{d_s^2 + \lambda} \sum_{\ell=1}^{r_{new}} \sum_{k=1}^{r^\star_{new}} f^\star_k f_\ell < W^{\star(k)}, W^{(\ell)} > \sum_{j=1}^{r^\star} d^\star_j < P^{(s)}, P^{\star(j)} >< Z^{(\ell)} Z^{\star(k)\prime} \beta^\star, Q^{(s)} Q^{\star(j)\prime} \beta^\star > ,$$

$$\check{A}_2 = \frac{\sigma_e^2}{n_{new}} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \sum_{i=1}^{n_{new}} \left(\sum_{\alpha=1}^{r_{new}} f_\alpha \, Q^{(s)\prime} Z^{(\alpha)} W_i^{(\alpha)} \right)^2 ,$$

$$\check{A}_3 = \frac{1}{n_{new}} \sum_{s=1}^{r} \sum_{\ell=1}^{r} \frac{d_s}{d_s^2 + \lambda} \frac{d_\ell}{d_\ell^2 + \lambda} \sum_{\alpha=1}^{r_{new}} f_\alpha^2 \; < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)}, Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} > \sum_{j=1}^{r^\star} d^\star_j < P^{(s)}, P^{\star(j)} > Q^{\star(j)\prime} \beta^\star$$
$$\times \sum_{k=1}^{r^\star} d^\star_k < P^{(\ell)}, P^{\star(k)} > Q^{\star(k)\prime} \beta^\star ,$$

$$\check{A}_4 = \frac{1}{n_{new}} \sum_{\alpha=1}^{r^\star_{new}} f_\alpha^{\star 2} \left\| Z^{\star(\alpha)} Z^{\star(\alpha)\prime} \beta^\star \right\|^2 .$$

The proof is given in Section 8.

5. How to improve the quality of the prediction

As in Rabier et al. (2018), we propose to project the vector Y on a well chosen subspace of the space spanned by the columns of X, in order to improve the quality of the prediction. Let $1 \le \tilde{r} \le r$ and $\sigma(.)$ a one-to-one map $\sigma : \{1, \dots, \tilde{r}\} \to \{1, \dots, r\}$. We thus have $\sigma(k) \neq \sigma(k')$ for $k \neq k'$.

Let us consider the estimator

$$\tilde{\beta} = X' V^{-1} \tilde{P} \tilde{P}' Y \text{ where } \tilde{P} = \left(P^{\sigma(1)}, \dots, P^{\sigma(\tilde{r})} \right).$$

Note that $\tilde{P}\tilde{P}'Y$ is the projection of Y on *Span*$\left\{ P^{\sigma(1)}, \dots, P^{\sigma(\tilde{r})} \right\}$. Besides, we set $\tilde{Q} = \left(Q^{\sigma(1)}, \dots, Q^{\sigma(\tilde{r})} \right)$. Then, the corresponding prediction for the so-called *new* individual is the following:

$$\tilde{Y}_{new} = x'_{new}\tilde{\beta} = x'_{new}X'V^{-1}\tilde{P}\tilde{P}'Y\ .$$

Let $\tilde{\rho}_g$ be the analogue of ρ_g, with $\hat{Y}_{new}$ replaced by $\tilde{Y}_{new}$ (cf. formula (3)):

$$\tilde{\rho}_g := \frac{\mathrm{Cov}\left(\tilde{Y}_{new},\ x'_{new}\beta\right)}{\sqrt{\mathrm{Var}\left(\tilde{Y}_{new}\right)\mathrm{Var}\left(x'_{new}\beta\right)}}. \tag{13}$$

A more explicit formula for $\tilde{\rho}_g$ is given in Lemma 1 of Section 2 of the Supplementary material. This lemma can be viewed as a version of Theorem 1 based on this new estimator. Let us now present a lemma which is the analogue of Theorem 2.

Lemma 4. *Let us consider same hypotheses as in Theorem 2. Then, an estimation of the quantity $\tilde{\rho}_g$ is*

$$\hat{\tilde{\rho}}_g = \frac{\hat{\tilde{A}}_1}{\left(\hat{\tilde{A}}_2 + \hat{\tilde{A}}_3\right)^{1/2}\left(\hat{\tilde{A}}_4\right)^{1/2}},$$

where

$$\hat{\tilde{A}}_1 = \frac{1}{n}\sum_{s=1}^{\tilde{r}}\frac{d^2_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}\left\|P^{(\sigma(s))}P^{(\sigma(s))\prime}X^\star\beta^\star\right\|^2\ ,\ \hat{\tilde{A}}_2 = \frac{\sigma_e^2}{n}\sum_{s=1}^{\tilde{r}}\frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}$$

$$\hat{\tilde{A}}_3 = \frac{1}{n}\sum_{s=1}^{\tilde{r}}\frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}\left\|P^{(\sigma(s))}P^{(\sigma(s))\prime}X^\star\beta^\star\right\|^2\ ,\ \hat{\tilde{A}}_4 := \hat{A}_4.$$

The proof is given in Section 3 of the Supplementary material.

Let us now give bounds for the quantity $\hat{\tilde{\rho}}_g$.

Lemma 5 (Bounds on $\hat{\tilde{\rho}}_g$). *Using same assumptions as in Theorem 2, we always have*

$$\frac{\left\|\tilde{P}\tilde{P}'X^\star\beta^\star\right\|^2\ \min_{1\le s\le\tilde{r}}\frac{d^2_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}}{\sqrt{\sigma_e^2\ \tilde{r} + \left\|\tilde{P}\tilde{P}'X^\star\beta^\star\right\|^2\ \max_{1\le s\le\tilde{r}}\frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}}\ \sqrt{\left\|Q^\star Q^{\star\prime}\beta^\star\right\|^2\ \max_\ell d_\ell^{\star 2}}} \le \hat{\tilde{\rho}}_g \le \rho_g^{\mathrm{oracle}}.$$

The proof relies heavily on the proof of Lemma 1, using the expressions of $\hat{\tilde{A}}_1$, $\hat{\tilde{A}}_2$ ans $\hat{\tilde{A}}_3$ given in Lemma 4. We can notice that at the denominator, the quantities $\tilde{r}$ and $\left\|\tilde{P}\tilde{P}'X^\star\beta^\star\right\|^2$ replace now the quantities r and $\|PP'X^\star\beta^\star\|^2$ of Lemma 1. This decrease at the denominator will be profitable provided that the numerator does not decrease too much.

We can notice that Lemma 6 of Rabier et al. (2018) is still suitable in our study. Then, for fixed n, we can easily compare the quantities $\hat{\tilde{\rho}}_g$ and $\hat{\rho}_g$. Let us recall this lemma.

Lemma 6 (Rabier et al. (2018)). *Let us suppose that $\hat{A}_1 - \hat{\hat{A}}_1 \neq 0$. Then, we have $\hat{\hat{\rho}}_g \geq \hat{\rho}_g$ if and only if the following relation holds:*

$$\frac{\hat{\hat{A}}_1}{\hat{A}_1 - \hat{\hat{A}}_1} \geq \frac{(\hat{\hat{A}}_2 + \hat{\hat{A}}_3)}{\hat{A}_2 + \hat{A}_3 - (\hat{\hat{A}}_2 + \hat{\hat{A}}_3)} \left(1 + \sqrt{\frac{\hat{A}_2 + \hat{A}_3}{\hat{\hat{A}}_2 + \hat{\hat{A}}_3}}\right).$$

Let us briefly recall explanation given in Rabier et al. (2018). $\vec{\beta}$ denotes the estimator, such as

$$\vec{\beta} := X'V^{-1}\vec{P}\vec{P}'Y$$

where $\vec{P}$ denotes the matrix obtained from P by removing the column vectors $P^{\sigma(1)}, \ldots, P^{\sigma(\tilde{r})}$. $\vec{Y}_{\text{new}}$ denotes the prediction:

$$\vec{Y}_{\text{new}} = x'_{\text{new}}\vec{\beta}.$$

We have the relationships $\hat{\beta} = \tilde{\beta} + \vec{\beta}$ and $\hat{Y}_{\text{new}} = \tilde{Y}_{\text{new}} + \vec{Y}_{\text{new}}$.

Then, the different terms of the lemma can be rewritten in the following way:

$$\hat{\hat{A}}_1 = \widehat{\text{Cov}}(\tilde{Y}_{\text{new}}, Y_{\text{new}}) \ , \ \ \hat{A}_1 - \hat{\hat{A}}_1 = \widehat{\text{Cov}}(\vec{Y}_{\text{new}}, Y_{\text{new}})$$
$$\hat{\hat{A}}_2 + \hat{\hat{A}}_3 = \widehat{\text{Var}}(\tilde{Y}_{\text{new}}) \ , \ \ \hat{A}_2 + \hat{A}_3 = \widehat{\text{Var}}(\hat{Y}_{\text{new}})$$
$$\hat{A}_2 + \hat{A}_3 - (\hat{\hat{A}}_2 + \hat{\hat{A}}_3) = \widehat{\text{Var}}(\vec{Y}_{\text{new}}).$$

Let us introduce a new lemma:

Lemma 7 (Extreme cases). *Let us consider same assumptions as in Theorem 2. Besides, let us suppose that the projected signal belongs only to Span $\left\{Q^{\star(1)}\right\}$, that is to say*

$$\left\|Q^{\star(1)}Q^{\star(1)\prime}\beta\right\|^2 \sim n^{2\tau}, \quad \left\|Q^{\star(s)}Q^{\star(s)\prime}\beta\right\|^2 = 0, \text{ for } 1 < s \leq r^\star.$$

Moreover, let us assume that $\ell = 1$ is tagged only by one s such as $\left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \sim \xi(p,n)$ with $0 < \xi(p,n) \leq 1$, and $\left\|P^{(u)}P^{(u)\prime}P^{\star(1)}\right\|^2 = 0 \ \forall u \neq s$. If we suppose that $s \in \{\sigma(1), \ldots, \sigma(\tilde{r})\}$, then

1. *If $s \in \Omega_1 \cup \Omega_2$ and if $2\tau + \psi < 1$ and the following two conditions hold*
   - $\sum_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(u)}}{\left(d^2_{\sigma(u)} + \lambda\right)^2} = o\left(n^{2\tau+\psi}\right)$;
   - $n^{2\tau+\psi} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right)$,

   we have $\hat{\hat{\rho}}_g \longrightarrow \left(\lim \sqrt{\xi(p,n)}\right) \rho_g^{oracle}$, whereas $\hat{\rho}_g \longrightarrow 0$.
2. *If $s \in \Omega_3$, $\lambda \sim Cn^{\kappa+\eta}$, $d_s \sim n^\gamma$ with $\kappa > \max(0, -\eta)$, $\gamma < (\kappa+\eta)/2$, $4\gamma - 2\kappa - 2\eta + 2\tau + \psi < 1$, and the following two conditions hold*

- $\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(u)}^4}{\left(d_{\sigma(u)}^2+\lambda\right)^2} = o\left(n^{4\gamma-2\kappa-2\eta+2\tau+\psi}\right);$
- $n^{4\gamma-2\kappa-2\eta+2\tau+\psi} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right),$

we have $\hat{\tilde{\rho}}_g \longrightarrow \left(\lim \sqrt{\xi(p,n)}\right) \rho_g^{oracle}$, *whereas* $\hat{\rho}_g \longrightarrow 0$.

The proof is largely inspired from the proof of Lemma 3. According to this lemma, there are a few cases where at the same time, the new accuracy $\hat{\tilde{\rho}}_g$ is not negligible (equal to $\lim \sqrt{\xi(p,n)}\rho_g^{oracle}$) and the classical accuracy $\hat{\rho}_g$ is null. Note that the analogue of this lemma, for a projected signal belonging only to $Span\left\{Q^{\star(r^\star)}\right\}$ can be deduced easily.

6. Applications under imperfect LD

In this section, we propose to illustrate our theoretical results, with the help of simulated data. We refer to Rabier et al. (2018) and Rabier et al. (2016) for a more precise description of the simulation framework. Populations were simulated by random mating between haploid individuals (i.e. with only one copy of each chromosome), during (a) 30, (b) 50, (c) 70 generations, or (d) 100 generations. In generation zero, either two or eight haploid founder lines were crossed. The eight founder set up, was suppose to introduce less LD due to relatedness. We focused on one chromosome of length 1 Morgan and also on a genome of length 4 Morgan or 6 Morgan. We considered 4 different densities of genetic markers equally spaced on the chromosome: (a) 100, (b) 500, (c) 1,000, or (d) 2,000 SNPs. We studied different configurations for the phenotypic model and the environmental variance σ_e^2 was set to 1.

The prediction model was learnt using 500 TRN individuals and the prediction model was evaluated on 100 TST (in all cases) produced in the last generation. We did not consider any full sib. Note also that all the quantities presented in the different tables are averages based on 100 simulations. Since we analyze the case where X and $X^\star$ do not vary across replicates, one simulation consists (a) in regenerating 100 TST individuals by random mating between individuals from the penultimate generation, and (b) in regenerating new phenotypes (TRN+TST).

The empirical accuracy was computed with the R software, using the empirical correlation between the predicted values and the true values. The regularization parameter λ was chosen by Restricted Maximum Likelihood (Corbeil and Searle (1976)) using the matrix X.

In what follows, to make the reading easier, we will adopt the notation $\hat{\rho}_{ph}(X^\star, \beta^\star)$ and $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$ for $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$ respectively. It will help for enumerating the nuisance parameters that have to be estimated.

6.1. Comparison between $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$ (Tables 1 and 2)

Tables 1 and 2 investigate the differences between $\hat{\rho}_{ph}(X^\star, \beta^\star)$ and $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$, when 2 large QTLs lie on a chromosome of size T=1M. Table 1 focuses on the case where TRN and TST may be sampled from different probability distributions. The TRN sample was always based on 30 generations, whereas the TST sample was based on either 30, either 50 or 70 generations. Different density of markers were considered. According to Table 1, the estimation $\hat{\rho}_{ph}(X^\star, \beta^\star)$ and the estimator $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$ gave fair results when TRN and TST relied on 30 generations. However, when the number of generations for TEST was greater than the one for TRN, the performances of $\hat{\rho}_{ph}(X^\star, \beta^\star)$ deteriorated heavily: we did not observe an agreement with the empirical accuracy. In contrast, $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$ was still a good estimator.

Table 2 investigates in details the formulas given in Theorems 2 and 3. TRN and TST were based on the same number of generations and the number of markers was set to 1000. We focused not only on $\hat{\rho}_{ph}(X^\star, \beta^\star)$ and $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$ but also on extra estimators (or estimations for $\hat{\rho}_{ph}$) relying on partial sums: these estimators consist in replacing the true quantities r, r_{new} with smaller values. In other words, we studied partial estimators relying on a few subspaces of $\mathcal{R}_{\text{rows}}(X)$, $\mathcal{R}_{\text{rows}}(X_{new})$, $\mathcal{R}_{\text{col}}(X)$, $\mathcal{R}_{\text{col}}(X_{new})$. Recall that under our simulation set up, n=500 and n_{new}=100: as a consequence, r is bounded by $\min(500, 1000)$ (i.e. 500), whereas r_{new} is bounded by $\min(100, 1000)$ (i.e. 100).

According to Table 2, when r and r_{new} were set to 25, we observed a non-negligible difference between the corresponding estimators and the empirical accuracy, specially for a larger number of generations (50 and 70). The number of considered subspaces was probably too low. In contrast, for higher values of r and r_{new}, the estimators relying on more subspaces gave estimates close to the empirical accuracy. Note that since $\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$ handles explicitly the TST sample, it suffered from a decrease of the size of the underlying subspace, which was not the case of $\hat{\rho}_{ph}(X^\star, \beta^\star)$ exclusively dedicated to TRN individuals.

6.2. TRN and TST do not share the same genetic map (Tables 3, 4, 5, 6, 7)

Tables 3, 4 and 5 deal with 25 QTLs lying either on a chromosome of size 1M, or a genome of size 4M, or 6M.

We propose here to study a new topic in GS: the accuracy of the prediction when the genetic map of TRN differs from the one of TST. In this context, let us consider a more dense map for TRN than for TST. Since the estimation $\hat{\rho}_{ph}(X^\star, \beta^\star)$ depends on nuisance parameters $X^\star$ and $\beta^\star$, we propose to estimate the parameters using the dense TRN map. This concept relies on the expectation that QTLs will be in perfect LD with markers under this dense TRN map, which is not the case for the TST map (imperfect LD).

The key point is that the dense TRN map is only used to estimate the nuisance parameters. The predictor for the so-called *new* individual is still $\hat{Y}_{new} = x'_{new}\hat{\beta} = x'_{new}X'V^{-1}Y$, where X denotes the design matrix (of size $n \times p$) for TRN (the columns of X match exactly marker locations of TST). In the same way, the estimation $\hat{\rho}_{ph}(X^\star, \beta^\star)$, built on Theorem 2, relies on that X design matrix. In this context, using the same number of generations for TRN and TST, boths samples (TRN and TST) share the same probability distribution, and it is reasonable to consider the estimation $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star)$ as a proxy for the predictive ability. In order to estimate $\beta^\star$ in a high-dimensional setting, we will concentrate on the LASSO (Tibshirani (1996)), the Adaptive LASSO (Zou (2006)) and on the Group LASSO (Yuan and Lin (2006)) estimators, as in Rabier et al. (2018).

Tables 3–7 compare the performances of our new proxies, that handle imperfect LD, with proxies suggested in Rabier et al. (2018) under perfect LD assumptions. We chose the Adaptive LASSO as a substitute for β, as advised by the authors. In what follows, $\hat{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$ (resp. $\check{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$) will refer to the "perfect LD" proxies available before (resp. after) genotyping TST individuals.

Tables 3, 4 and 5, deal respectively with 250 markers, 500 markers and 1000 markers, equally spaced on $[0,T]$. The dense TRN map contains twice the number of markers. According to Tables 3, 4 and 5, there is a clear advantage to handle explicitly imperfect LD for T=4 and T=6, whatever the density of markers: the proxies $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$ and $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$ gave always better performances than the quantities $\hat{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$ and $\check{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$ relying on perfect LD.

In contrast, when a chrosomome of length 1M was studied, $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$ was the only proxy found to be more accurate than "perfect LD" proxies. Indeed, when p was set to 500 or 1000, $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$ was outperformed by "perfect LD" proxies. This result is not so surprising since this genetic map is close to micmick perfect LD situation, and the Adaptive Lasso was the best substitute for β according to Rabier et al. (2018).

Same conclusions hold for the 100 QTLs scenario (cf. Tables 6 and 7).

To sum up, the best proxy (see gray underlined in each table) for each simulation set up, was found to be $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$ for T=1, and in most cases, $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$ for T=4 and T=6.

6.3. The quality of the prediction can be improved (Tables 8 and 9)

We propose to illustrate here the performances of the estimator $\tilde{\beta}$ that relies on the projection of Y on a well chosen subspace of $\mathcal{R}_{\text{rows}}(X)$.

In order to find an appropriate subspace, we used the same kind of procedure as in Rabier et al. (2018). We decided that $\frac{d^2_{\sigma(k)}}{d^2_{\sigma(k)}+\lambda} \left\| P^{(\sigma(k))} P^{(\sigma(k))\prime} X^\star \beta^\star \right\|^2$ was the k-th largest term of the sequence $\frac{d_s^2}{d_s^2+\lambda} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2_{s=1,\dots,r}$.

The value of $\tilde{r}$ was chosen as the largest value satisfying the condition $\hat{\tilde{A}}_1 / \hat{A}_1 \leq \upsilon$, where υ denotes a tuning parameter. The corresponding accuracy was then computed for a given value of υ. Since υ was unknown, we performed an optimization over the grid
$\{0.7, 0.8, 0.9, 0.925, 0.95, 0.975, 0.99\}$ and kept the highest accuracy.

During this procedure, $\beta^\star$ was estimated with the help of a penalized likelihood method. Table 8 compares the empirical correlations $\hat{cor}\left(\tilde{Y}_{\text{new}}, Y_{\text{new}}\right)$ when subspaces were chosen according to the Adaptive LASSO or according to the LASSO. It reports also the empirical accuracy, relying on the classical Ridge estimator.

In all the cases studied in Table 8, the empirical accuracy associated to the new estimator was always greater than the classical empirical accuracy based on the Ridge estimator. Moreover, for choosing the subspaces, we could not establish the superiority of one penalized likelihood method over an other.

Last, Table 8 investigates the case where the vector $\beta^\star$ belongs to $\mathcal{R}_{\text{rows}}(X^\star)$. As expected, we observe a significant increase in terms of accuracy when the "modified predictor" is adopted.

6.4. Real data: GS in rice

We propose to illustrate our theoretical results on real data of Spindel et al. (2015), regarding GS in rice. An important research topic in GS is to determine the number of markers required for implementing GS. We focused on the rice flowering time (days to 50% flowering) collected in Los Banos, Philippines, during the dry season 2012.

Of the observations, 80% were chosen for the TRN set, and the remaining 20% were affected to the TST set. According to the data, the number of TRN individuals was 252 whereas the number of TEST individuals was 63.

We considered 4 subset sizes, 448, 781, 1553 and 3076, chosen by the authors from their 73147 SNPs. For each subset size, we considered exactly the 10 random sets provided by the authors. Recall that these random sets contain SNPs located at random position along the rice genome.

For each subset size, Table 10 reports the average performance of different GS proxies over the 10 random sets. In contrast, Table 11 is dedicated to the configuration with 448 SNPs, and provides results regarding each random set. Note that Tables 1, 2 and 3, that handle 781, 1553 and 3076 SNPs respectively, are included in supplementary material.

According to Table 10, $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$ was the most interesting proxy (combining all SNPs scenarios). In particular, a small density of markers deteriorated "perfect LD" proxies: the phenotypic accuracy was underestimated when $p = 448$ or $p = 781$. For instance, for $p = 448$, $\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$ and $\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$ were equal on average to 0.3168 and 0.3662 respectively, instead of 0.4789 (see also sets 3, 6, 8 and 10 in Table 11). In contrast, the "imperfect LD" proxy $\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$ was satisfactory for all density of markers. This proxy did not suffer from the lack of markers, since the nuisance parameters were learned using a TRN map based on 73147 SNPs. Moreover, as observed before (cf. our simulation study, section 6.2), for such large genome size ($T = 13.101$M in rice), it seems that we should choose the LASSO and not the Adaptive LASSO, as a substitute for $\beta^\star$, when computing our "imperfect LD" proxies. Last, as expected, the more markers there are, the more similar the behavior of perfect and "imperfect LD" proxies is.

To conclude, in this study, we proposed an "imperfect LD" proxy that should be of interest for geneticists. Predictions of TST individuals can be performed using only a few markers, and the accuracy of these predictions can be evaluated according to our reliable proxy, as soon as a large number of markers is available for the TRN map. If this accuracy is found to be too low, geneticists should reconsider the density of markers used for their TST map.

Table 1: Comparison among different estimators of the phenotypic accuracy as a function of the number of generations during which the TST sample evolved (TRN sample is always based on 30 generations). The chromosome is of length 1M and 2 QTLs are located at 2cM and 80cM with effects +1 and −2, respectively ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$, 2 founders). Emp. Acc. refers to the empirical *phenotypic accuracy.* The causal SNPs (i.e. the QTLs) were not observed in the TRN and TST samples.

Nb Markers	Nb TST generations	Emp. Acc.	$\hat{\rho}_{ph}(X^\star, \beta^\star)$	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$
100	30	0.5118	0.5679	0.5124
	50	0.3726	0.5665	0.3789
	70	0.2354	0.5691	0.2411
1000	30	0.6956	0.7155	0.6878
	50	0.6506	0.7145	0.6429
	70	0.6189	0.7143	0.6105
2000	30	0.6881	0.6919	0.6830
	50	0.6003	0.6922	0.5911
	70	0.3864	0.6895	0.3904

Table 2: Comparison among different estimators of the phenotypic accuracy, as a function of the number of considered subspaces. The chromosome is of length 1M and 2 QTLs are located at 3cM and 80cM with effects +1 and −2, respectively ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$, 2 founders). TRN and TST evolved during the same number generations (30, 50 or 70). Emp. Acc. refers to the empirical *phenotypic accuracy.* 1000 markers are equally spaced on the chromosome. The causal SNPs (i.e. the QTLs) were not observed in the TRN and TST samples.

Method	Subspaces	30 generations	50 generations	70 generations
Emp. Acc.	All	0.6956	0.6596	0.6722
$\hat{\rho}_{ph}(X^\star, \beta^\star)$	$(r, r^\star) = (100, 2)$	0.7157	0.6579	0.6843
	$(r, r^\star) = (75, 2)$	0.7151	0.6548	0.6752
	$(r, r^\star) = (50, 2)$	0.7109	0.6432	0.6536
	$(r, r^\star) = (25, 2)$	0.6889	0.6113	0.5998
$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	All	0.6878	0.6563	0.6659
	$(r, r^\star, r_{new}, r^\star_{new}) = (100, 2, 100, 2)$	0.6867	0.6547	0.6626
	$(r, r^\star, r_{new}, r^\star_{new}) = (75, 2, 75, 2)$	0.6923	0.6579	0.6593
	$(r, r^\star, r_{new}, r^\star_{new}) = (50, 2, 50, 2)$	0.6903	0.6462	0.6408
	$(r, r^\star, r_{new}, r^\star_{new}) = (25, 2, 25, 2)$	0.6594	0.6136	0.5687

7. Proofs

7.1. Proof of Theorem 1

By definition,

$$A_1 = \beta^{\star\prime}\, \mathbb{E}\left(x^\star_{new} x'_{new}\right) X' V^{-1} X^\star \beta^\star.$$

Table 3: Comparison among different estimators of the phenotypic accuracy, when the causal SNPs (i.e. the QTLs) are not observed in the TST samples ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$, 8 founders). The nuisance parameters are estimated thanks to a TRN map containing 500 markers equally spaced on the chromosome on [0,T]. In contrast, the TST map contains only 250 markers equally spaced on [0,T]. For both maps, the first marker is located respectively at 0.002M, 0.008M, and 0.012M, when T=1, T=4, and T=6. 25 QTLs with effects 0.45 are located respectively every 0.04M, 0.16M, and 0.24M when T=1, T=4, and T=6. Emp. Acc. refers to the empirical phenotypic accuracy, whereas $\hat{\rho}_{ph}^{scand}$ and $\check{\rho}_{ph}^{scand}$ refer to complete LD proxies from Rabier et al. (2018). The Mean Squared Error (MSE) with respect to the Empirical Accuracy is given in brackets. $\overline{\text{MSE}}$ refers to the average over the 3 number of generations. For each chromosome length T, the proxy with the smallest $\overline{\text{MSE}}$ is highlighted in gray.

	Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
$T = 1$	Emp. Acc.	0.2925	0.2976	0.3224	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.2833 (0.0070)	0.3099 (0.0078)	0.3221 (0.0068)	0.0072
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.1241 (0.0397)	0.1312 (0.0380)	0.1767 (0.0336)	0.0371
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.08366 (0.0561)	0.0998 (0.0501)	0.1393 (0.0464)	0.0509
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.2947 (0.0108)	0.3129 (0.0107)	0.3521 (0.0110)	0.0108
	$\check{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.1762 (0.0324)	0.2179 (0.0238)	0.2708 (0.0159)	0.0240
	$\hat{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.1955 (0.0302)	0.2361 (0.0222)	0.3086 (0.0149)	0.0224
$T = 4$	Emp. Acc.	0.3021	0.2671	0.2043	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.3021 (0.0057)	0.2670 (0.0067)	0.2088 (0.0056)	0.0060
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.2848 (0.0102)	0.3042 (0.0111)	0.2591 (0.0114)	0.0109
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.2549 (0.0133)	0.2677 (0.0108)	0.2370 (0.0107)	0.0116
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4029 (0.0199)	0.4197 (0.0316)	0.3708 (0.0362)	0.0292
	$\check{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.1669 (0.0438)	0.1240 (0.0457)	0.0283 (0.0416)	0.0437
	$\hat{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.1878 (0.0416)	0.1446 (0.0453)	0.0312 (0.0413)	0.0427
$T = 6$	Emp. Acc.	0.2284	0.2441	0.2331	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.2212 (0.0064)	0.2433 (0.0067)	0.2327 (0.0075)	0.0069
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.2832 (0.0141)	0.2870 (0.012)	0.2529 (0.0118)	0.0126
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.2624 (0.0127)	0.2600 (0.0126)	0.2336 (0.0121)	0.0125
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.3907 (0.0366)	0.4109 (0.0379)	0.3836 (0.0339)	0.0361
	$\check{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.0742 (0.0387)	0.0817 (0.0483)	0.0841 (0.0449)	0.0439
	$\hat{\rho}_{ph}^{scand}(\hat{\beta}_{ADLASSO})$	0.0848 (0.0374)	0.0931 (0.0477)	0.0991 (0.0449)	0.0433

We set $\overline{D} = Diag\left(\frac{d_1}{d_1^2+\lambda}, \ldots, \frac{d_r}{d_r^2+\lambda}\right)$. With this notation, we have the relation:

$$X'V^{-1} = Q\overline{D}P'. \tag{14}$$

Recall that $X^\star = P^\star D^\star Q^{\star\prime}$. After easy calculations, we obtain

$$X'V^{-1}P^\star D^\star Q^{\star\prime}\beta^\star = \sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \tag{15}$$

Table 4: Same as Table 3 except that more markers are considered. The nuisance parameters are estimated thanks to a TRN map containing 1000 markers on [0,T]. The TST map contains only 500 markers on [0,T]. For both maps, the first marker is located respectively at 0.001M, 0.004M, and 0.006M, when T=1, T=4, and T=6. QTL locations are the same as in Table 3.

	Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
$T=1$	Emp. Acc.	0.5287	0.5396	0.5173	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.5152 (0.0043)	0.5412 (0.0043)	0.5176 (0.0029)	0.0038
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4370 (0.0175)	0.4638 (0.0013)	0.4642 (0.0092)	0.0093
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.4033 (0.0239)	0.4469 (0.0163)	0.4471 (0.0115)	0.0172
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5371 (0.0073)	0.5691 (0.0063)	0.5589 (0.0069)	0.0068
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5011 (0.0098)	0.5324 (0.0079)	0.5172 (0.0049)	0.0075
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5411 (0.0099)	0.5758 (0.0094)	0.5690 (0.0087)	0.0093
$T=4$	Emp. Acc.	0.3909	0.3772	0.3217	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.3795 (0.0055)	0.3759 (0.0075)	0.3266 (0.0064)	0.0065
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.3397 (0.0112)	0.3436 (0.0132)	0.2629 (0.0146)	0.0130
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.2413 (0.0334)	0.3059 (0.0179)	0.2178 (0.0228)	0.0247
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4677 (0.01293)	0.4821 (0.0222)	0.4093 (0.0164)	0.0172
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2599 (0.0389)	0.2647 (0.0355)	0.0846 (0.0722)	0.0489
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2970 (0.0336)	0.3182 (0.0306)	0.0986 (0.0693)	0.0445
$T=6$	Emp. Acc.	0.3749	0.3319	0.3155	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.3751 (0.0052)	0.3339 (0.0054)	0.3206 (0.0045)	0.0050
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.37 (0.0034)	0.3548 (0.0094)	0.3415 (0.0093)	0.0074
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.3395 (0.01132)	0.3259 (0.0093)	0.3048 (0.0094)	0.0100
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5045 (0.02488)	0.4981 (0.0355)	0.4703 (0.0317)	0.0307
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2351 (0.0436)	0.2383 (0.0358)	0.2423 (0.0307)	0.0367
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.1929 (0.0519)	0.1906 (0.0397)	0.2045 (0.0319)	0.0412

Then,

$$A_1 = \sum_{s=1}^{r} \frac{d_s}{d_s^2 + \lambda} \, \beta^{\star\prime} \, \mathbb{E}\left(x^\star_{new} x'_{new}\right) \, Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star .$$

By definition,

$$A_2 = \sigma_e^2 \mathbb{E}\left(\left\| x'_{new} X' V^{-1} \right\|^2 \right).$$

According to formula (14), we have

$$\begin{aligned} \left\| x'_{new} X' V^{-1} \right\|^2 &= x'_{new} X' V^{-1} \left(X' V^{-1} \right)' x_{new} \\ &= x'_{new} Q \overline{D} P' P \overline{D} Q' x_{new} \\ &= x'_{new} Q \overline{D}^2 Q' x_{new}. \end{aligned}$$

Furthermore, we have

$$Q \overline{D}^2 Q' = \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} Q^{(s)} Q^{(s)\prime}.$$

Table 5: Same as Table 3 except that more markers are considered. The nuisance parameters are estimated thanks to a TRN map containing 2000 markers on [0,T]. The TST map contains only 1000 markers on [0,T]. For both maps, the first marker is located respectively at 0.0005M, 0.002M, and 0.003M, when T=1, T=4, and T=6. QTL locations are the same as in Table 3.

	Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
$T=1$	Emp. Acc.	0.5239	0.5561	0.5907	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.5224 (0.0036)	0.5441 (0.0030)	0.5853 (0.0033)	0.0033
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4218 (0.0181)	0.4213 (0.0224)	0.4676 (0.0220)	0.0208
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.3856 (0.0269)	0.3949 (0.0309)	0.4546 (0.0247)	0.0275
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5261 (0.0061)	0.5298 (0.0043)	0.5709 (0.0057)	0.0054
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4624 (0.0096)	0.4734 (0.0114)	0.5241 (0.0092)	0.0101
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5107 (0.0068)	0.5153 (0.0062)	0.5641 (0.0065)	0.0065
$T=4$	Emp. Acc.	0.4244	0.4027	0.4162	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.4315 (0.0046)	0.3935 (0.0055)	0.4093 (0.0053)	0.0051
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.3614 (0.013)	0.3224 (0.0193)	0.3478 (0.0156)	0.0159
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.2974 (0.0260)	0.2521 (0.0403)	0.2929 (0.0256)	0.0306
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5063 (0.0147)	0.4642 (0.0146)	0.5001 (0.0152)	0.0148
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3037 (0.0291)	0.2441 (0.0414)	0.2906 (0.0328)	0.0344
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3612 (0.0226)	0.3205 (0.0305)	0.3483 (0.0259)	0.0263
$T=6$	Emp. Acc.	0.3724	0.4037	0.3477	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.3814 (0.0052)	0.3959 (0.0041)	0.3435 (0.0057)	0.0050
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.3215 (0.0127)	0.3325 (0.0135)	0.2709 (0.0167)	0.0143
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.2619 (0.0236)	0.2799 (0.0240)	0.2071 (0.0299)	0.0258
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4863 (0.0212)	0.4966 (0.0144)	0.4401 (0.0167)	0.0174
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2024 (0.0478)	0.2309 (0.0499)	0.1844 (0.0413)	0.0463
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2510 (0.0399)	0.2935 (0.0397)	0.2347 (0.0324)	0.0373

Since $Q^{(s)}Q^{(s)\prime}$ is an idempotent matrix, we obtain

$$\begin{aligned}
\left\|x_{new}^\prime X^\prime V^{-1}\right\|^2 &= \sum_{s=1}^r \frac{d_s^2}{(d_s^2+\lambda)^2} x_{new}^\prime Q^{(s)}Q^{(s)\prime} x_{new} \\
&= \sum_{s=1}^r \frac{d_s^2}{(d_s^2+\lambda)^2} x_{new}^\prime Q^{(s)}Q^{(s)\prime}Q^{(s)}Q^{(s)\prime} x_{new} \\
&= \sum_{s=1}^r \frac{d_s^2}{(d_s^2+\lambda)^2} \left\|Q^{(s)}Q^{(s)\prime} x_{new}\right\|^2.
\end{aligned}$$

Finally,

$$A_2 = \sum_{s=1}^r \frac{d_s^2}{(d_s^2+\lambda)^2} \mathbb{E}\left(\left\|Q^{(s)}Q^{(s)\prime} x_{new}\right\|^2\right).$$

By definition,

$$A_3 = \beta^{\star\prime} X^{\star\prime} V^{-1} X \operatorname{Var}(x_{new}) X^\prime V^{-1} X^\star \beta^\star.$$

Table 6: Same as Table 4 except that more QTLs are considered on [0,T]. 100 QTLs with effects 0.30 are located respectively every 0.01M, 0.04M, and 0.06M when T=1, T=4, and T=6. Recall that the nuisance parameters are estimated thanks to a TRN map containing 1000 markers on [0,T]. The TST map contains only 500 markers on [0,T].

	Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
$T=1$	Emp. Acc.	0.6489	0.6499	0.6872	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.6383 (0.0027)	0.6451 (0.0028)	0.6846 (0.0018)	0.0024
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.6102 (0.0059)	0.6793 (0.0050)	0.6978 (0.0027)	0.0045
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.5909 (0.0075)	0.6451 (0.0044)	0.6916 (0.0026)	0.0048
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.6433 (0.0039)	0.6793 (0.0050)	0.7069 (0.0027)	0.0039
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.6578 (0.0044)	0.6667 (0.0044)	0.7156 (0.0029)	0.0039
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.6839 (0.0058)	0.7163 (0.0092)	0.7598 (0.0074)	0.0075
$T=4$	Emp. Acc.	0.4451	0.4821	0.4053	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.4450 (0.0039)	0.4634 (0.0039)	0.4095 (0.0068)	0.0049
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4652 (0.0094)	0.4234 (0.0138)	0.4326 (0.0136)	0.0123
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.4264 (0.0118)	0.3610 (0.0257)	0.3872 (0.0152)	0.0176
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5551 (0.0192)	0.5103 (0.0108)	0.5273 (0.0252)	0.0184
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3603 (0.0245)	0.3602 (0.0326)	0.2866 (0.04651)	0.0345
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4414 (0.0212)	0.4104 (0.0243)	0.3371 (0.0419)	0.0291
$T=6$	Emp. Acc.	0.3895	0.3666	0.3599	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.3861 (0.0049)	0.3575 (0.0045)	0.3507 (0.0042)	0.0045
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.3983 (0.0123)	0.4171 (0.0131)	0.3774 (0.0121)	0.0125
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.3403 (0.01824)	0.3575 (0.0116)	0.3312 (0.0137)	0.0145
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5007 (0.0233)	0.5085 (0.0294)	0.4894 (0.0247)	0.0258
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.1124 (0.0995)	0.2016 (0.0569)	0.1847 (0.0545)	0.0703
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.1415 (0.0926)	0.2556 (0.0546)	0.2293 (0.0493)	0.0655

According to formula (15), we obtain the desired result

$$A_3 = \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2 + \lambda} \; Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)^{\prime} \; \mathbb{E}\left(x_{new} x_{new}^{\prime}\right)$$
$$\times \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2 + \lambda} \; Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right).$$

Last, since $A_4 = \sigma_G^2$, we have the relationship

$$A_4 = \beta^{\star\prime} \mathbb{E}\left(x^\star_{new} x^{\star\prime}_{new}\right) \beta^\star.$$

7.2. Proof of Theorem 2

Let us define $\hat{A}_1$ in the following way:

$$\hat{A}_1 = \sum_{s=1}^{r} \frac{d_s}{d_s^2 + \lambda} \; \beta^{\star\prime} \; \hat{\Sigma} \; Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star,$$

Table 7: Same as Table 6 except that more markers are considered. The nuisance parameters are estimated thanks to a TRN map containing 2000 markers on [0,T]. The TST map contains only 1000 markers on [0,T]. QTL locations are the same as in Table 6.

	Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
$T=1$	Emp. Acc.	0.6612	0.6484	0.6831	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.6616 (0.0023)	0.6396 (0.0023)	0.6787 (0.0020)	0.0022
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.5935 (0.0098)	0.5855 (0.0079)	0.6333 (0.0066)	0.0081
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.5722 (0.0131)	0.5665 (0.0115)	0.6180 (0.0082)	0.0109
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.6477 (0.0033)	0.6213 (0.0042)	0.6676 (0.0035)	0.0037
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.6149 (0.0054)	0.5825 (0.0077)	0.6676 (0.0035)	0.0055
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.6449 (0.0037)	0.6291 (0.0037)	0.6636 (0.0036)	0.0037
$T=4$	Emp. Acc.	0.5047	0.4723	0.4760	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.5118 (0.0039)	0.4596 (0.0039)	0.4727 (0.0039)	0.0039
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.5157 (0.0083)	0.4574 (0.0122)	0.4201 (0.0153)	0.0119
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.4547 (0.0123)	0.4078 (0.0189)	0.3663 (0.0227)	0.0179
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5986 (0.0163)	0.5477 (0.0180)	0.5420 (0.0128)	0.0157
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4366 (0.0166)	0.3639 (0.0294)	0.3416 (0.0409)	0.0289
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5206 (0.0197)	0.4567 (0.0219)	0.4171 (0.0327)	0.0247
$T=6$	Emp. Acc.	0.4306	0.4870	0.4384	
	$\check{\rho}_{ph}(X^\star, X^\star_{new}, \beta^\star)$	0.4244 (0.0039)	0.4805 (0.0029)	0.4342 (0.0042)	0.0037
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4205 (0.0173)	0.4529 (0.0155)	0.3733 (0.0194)	0.0174
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{GPLASSO})$	0.3429 (0.0229)	0.4009 (0.0192)	0.3279 (0.0267)	0.0229
	$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.5307 (0.0241)	0.5582 (0.0146)	0.4994 (0.0178)	0.0188
	$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2890 (0.0476)	0.3424 (0.0419)	0.2581 (0.0650)	0.0515
	$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3611 (0.0415)	0.4269 (0.0313)	0.3156 (0.0597)	0.0442

where $\hat{\Sigma} := X^{\star\prime}X/n$.

We have the relationship $XQ^{(s)} = d_s P^{(s)}$. As a consequence, after some straightforward matrix algebra, we obtain

$$X^{\star\prime}XQ^{(s)} = d_s \sum_{\ell=1}^{r^\star} Q^{\star(\ell)} d^\star_\ell P^{\star(\ell)\prime} P^{(s)}.$$

We deduce

$$\hat{A}_1 = \frac{1}{n}\sum_{s=1}^{r} \beta^{\star\prime} \frac{d_s^2}{d_s^2+\lambda} \sum_{\ell=1}^{r^\star} Q^{\star(\ell)} d^\star_\ell P^{\star(\ell)\prime} P^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star.$$

Table 8: Illustration of the predictions based on $\tilde{\beta}$. $\hat{cor}\left(\hat{Y}_{\text{new}}, Y_{\text{new}}\right)$ refers to the empirical correlation between $\hat{Y}_{\text{new}}$ and Y_{new}. $\hat{cor}\left(\tilde{Y}_{\text{new}}^{ADLASSO}, Y_{\text{new}}\right)$ (resp. $\hat{cor}\left(\tilde{Y}_{\text{new}}^{LASSO}, Y_{\text{new}}\right)$) refers to the empirical correlation between $\tilde{Y}_{\text{new}}$ and Y_{new}, with the help of the Adaptive Lasso (resp. Lasso) for the choice of the subspace. The chromosome is of length T and 2 QTLs located at 3cM and 80cM with effects $+2$ and -4 respectively ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 4$, 8 founders). For TRN, p markers are equally spaced on the chromosome on [0,T], whereas for TST $p/(2T)$ markers are equally spaced on [0,1], and the same map (as TRN) is kept on $[1, T]$. The QTLs were observed only in the TRN sample (i.e. not observed in the TST sample).

	Generations	$\hat{cor}\left(\hat{Y}_{\text{new}}, Y_{\text{new}}\right)$	$\hat{cor}\left(\tilde{Y}_{\text{new}}^{LASSO}, Y_{\text{new}}\right)$	$\hat{cor}\left(\tilde{Y}_{\text{new}}^{ADLASSO}, Y_{\text{new}}\right)$
$T = 4$	50	0.4537	0.4625	0.4668
$p = 4000$	100	0.4051	0.4059	0.4126
$T = 6$	50	0.3171	0.3174	0.3246
$p = 6000$	100	0.3468	0.3536	0.3527
$T = 4$	50	0.2975	0.2985	0.3094
$p = 8000$	100	0.2642	0.2726	0.2741
$T = 6$	50	0.3510	0.3578	0.3604
$p = 12000$	100	0.3563	0.3604	0.3655

Table 9: Comparisons among predictions based on $\hat{\beta}$ and $\tilde{\beta}$ when the vector $\beta^\star$ belongs to $\mathcal{R}_{\text{rows}}(X^\star)$. $\hat{cor}\left(\hat{Y}_{\text{new}}, Y_{\text{new}}\right)$ refers to the empirical correlation between $\hat{Y}_{\text{new}}$ and Y_{new}. $\hat{cor}\left(\tilde{Y}_{\text{new}}^{ADLASSO}, Y_{\text{new}}\right)$ (resp. $\hat{cor}\left(\tilde{Y}_{\text{new}}^{LASSO}, Y_{\text{new}}\right)$) refers to the empirical correlation between $\tilde{Y}_{\text{new}}$ and Y_{new}, with the help of the Adaptive Lasso (resp. Lasso) for the choice of the subspace. The chromosome is of length T ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$, 8 founders). For TRN, p markers are equally spaced on the chromosome on [0,T], whereas for TST $p/(2T)$ markers are equally spaced on [0,1], and the same map (as TRN) is kept on $[1, T]$. QTLs are located at marker locations of the TRN map on [0, 1]. The vector $\beta^\star$ is such that $\beta^\star = \omega Q^{\star(1)} + \omega Q^{\star(2)} + \omega Q^{\star(3)}$.

	Generations	ω	$\hat{cor}\left(\hat{Y}_{\text{new}}, Y_{\text{new}}\right)$	$\hat{cor}\left(\tilde{Y}_{\text{new}}^{LASSO}, Y_{\text{new}}\right)$	$\hat{cor}\left(\tilde{Y}_{\text{new}}^{ADLASSO}, Y_{\text{new}}\right)$
$T = 4$	50	0.3	0.5660	0.5791	0.5845
$p = 8000$	100	0.3	0.5561	0.5644	0.5691
$T = 6$	50	0.3	0.4769	0.4815	0.4824
$p = 12000$	100	0.3	0.4649	0.4834	0.4834
$T = 4$	50	0.6	0.7978	0.8115	0.8078
$p = 8000$	100	0.6	0.7912	0.8067	0.8019
$T = 6$	50	0.6	0.7244	0.7371	0.7273
$p = 12000$	100	0.6	0.7127	0.7324	0.7247

A natural estimation of A_2 is

$$\begin{aligned}
\hat{A}_2 &= \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \sum_{i=1}^{n} \left\| Q^{(s)} Q^{(s)\prime} x_i \right\|^2 \\
&= \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \operatorname{Tr}\left(X Q^{(s)} Q^{(s)\prime} Q^{(s)} Q^{(s)\prime} X' \right) \\
&= \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \operatorname{Tr}\left(X Q^{(s)} Q^{(s)\prime} X' \right) \\
&= \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \operatorname{Tr}\left(P D Q' Q^{(s)} Q^{(s)\prime} Q D P' \right).
\end{aligned}$$

Table 10: Comparison among different estimators of the phenotypic accuracy on rice data from Spindel et al. (2015). The trait considered is the flowering time during the dry season 2012. Different density of markers for the TST samples are studied, and the nuisance parameters are estimated thanks to a TRN map containing 73147 markers. Each computed accuracy relies on 1000 data sets: sets 1 to 10 of Spindel et al. (2015) are studied, and 100 draws are considered for each set (with random individuals in TRN and TST sets, $n = 252$, $n_{new} = 63$). The Mean Squared Error (MSE) with respect to the Empirical Accuracy is given in brackets. For each density of markers, the proxy with the smallest MSE is highlighted in gray. $\overline{\text{MSE}}$ refers to the average over the 4 density of markers.

Method	448 SNPs	781 SNPs	1553 SNPs	3076 SNPs	$\overline{\text{MSE}}$
Emp. Acc.	0.4789	0.4919	0.5275	0.5237	
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4269 (0.0355)	0.4379 (0.0376)	0.4520 (0.0419)	0.4466 (0.0428)	0.03945
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4621 (0.0244)	0.4653 (0.0226)	0.4737 (0.0254)	0.4732 (0.0261)	0.0246
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3168 (0.0529)	0.3571 (0.0364)	0.4233 (0.0264)	0.413 (0.0299)	0.0364
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3662 (0.0454)	0.4202 (0.0281)	0.4919 (0.0215)	0.4905 (0.0363)	0.0328

Table 11: Same as Table 10 except that only 448 SNPs are used for the TST sample. Moreover, the results according to each set Spindel et al. (2015) are fully described here. The nuisance parameters are still estimated thanks to a TRN map containing 73147 markers. Each computed accuracy relies on 100 data sets: for each set of Spindel et al. (2015), 100 draws are considered (with random individuals in TRN and TST sets, $n = 252$, $n_{new} = 63$).

Dataset ID	Set 1	Set 2	Set 3	Set 4
Emp. Acc.	0.5993	0.5445	0.4117	0.5054
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4764 (0.0429)	0.4441 (0.0441)	0.4053 (0.0322)	0.4358 (0.0356)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.5125 (0.0271)	0.4847 (0.02486)	0.4380 (0.0236)	0.4808 (0.0207)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5065 (0.0171)	0.4712 (0.0154)	0.1580 (0.0959)	0.4222 (0.0176)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5404 (0.0124)	0.5128 (0.0128)	0.2059 (0.0867)	0.4663 (0.0153)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.4676	0.4081	0.4878	0.4455
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4309 (0.0353)	0.4070 (0.0348)	0.4362 (0.0373)	0.4214 (0.0348)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4653 (0.0233)	0.4207 (0.0232)	0.4676 (0.0227)	0.4508 (0.0244)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3251 (0.0398)	0.1774 (0.0907)	0.3732 (0.0286)	0.2726 (0.0668)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3953 (0.0298)	0.2179 (0.0823)	0.4343 (0.0211)	0.3274 (0.0586)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4427	0.4696		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4117 (0.0218)	0.4130 (0.0366)		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4622 (0.0316)	0.4382 (0.0229)		
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2789 (0.0404)	0.1829 (0.1179)		
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3255 (0.0314)	0.2366 (0.1036)		

Note that

$$DQ'Q^{(s)} = d_s e_s,$$

where e_s denotes the s-th vector of the canonical basis of $\mathbb{R}^r$.

$$\begin{aligned}\hat{A}_2 &= \frac{\sigma_e^2}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\operatorname{Tr}\left(Pe_se_s'P'\right)\\ &= \frac{\sigma_e^2}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\operatorname{Tr}\left(P'Pe_se_s'\right)\\ &= \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}.\end{aligned}$$

Let us consider the following estimation of A_3

$$\hat{A}_3 = \frac{1}{n}\left(\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\; Q^{(s)}P^{(s)\prime}\sum_{j=1}^{r^\star}d_j^\star P^{\star(j)}Q^{\star(j)\prime}\beta^\star\right)'\; X'X$$
$$\times\left(\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\; Q^{(s)}P^{(s)\prime}\sum_{j=1}^{r^\star}d_j^\star P^{\star(j)}Q^{\star(j)\prime}\beta^\star\right).$$

We have

$$\hat{A}_3 = \frac{1}{n}\left(\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\; XQ^{(s)}\sum_{j=1}^{r^\star}d_j^\star Q^{\star(j)\prime}\beta^\star P^{(s)\prime}P^{\star(j)}\right)'$$
$$\times\left(\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\; XQ^{(s)}\sum_{j=1}^{r^\star}d_j^\star Q^{\star(j)\prime}\beta^\star P^{(s)\prime}P^{\star(j)}\right).$$

Note that

$$XQ^{(s)}Q^{\star(j)\prime}\beta = PDQ'Q^{(s)}Q^{\star(j)\prime}\beta = d_sPe_sQ^{\star(j)\prime}\beta^\star = d_sP^{(s)}Q^{\star(j)\prime}\beta^\star.$$

As a consequence,

$$\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\; XQ^{(s)}\sum_{j=1}^{r^\star}d_j^\star Q^{\star(j)\prime}\beta^\star P^{(s)\prime}P^{\star(j)} = \sum_{s=1}^{r}\frac{d_s^2}{d_s^2+\lambda}\; P^{(s)}\sum_{j=1}^{r^\star}d_j^\star P^{(s)\prime}P^{\star(j)}Q^{\star(j)\prime}\beta^\star.$$

Last, we obtain

$$\hat{A}_3 = \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\left(\sum_{\ell=1}^{r^\star}d_\ell^\star P^{(s)\prime}P^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right)^2.$$

Finally, let us consider the following estimation of A_4:

$$\hat{A}_4 = \frac{1}{n}\beta^{\star\prime}X^{\star\prime}X^\star\beta^\star.$$

We have

$$\hat{A}_4 = \frac{1}{n}\beta^{\star\prime}Q^{\star}D^{\star 2}Q^{\star\prime}\beta^{\star} = \frac{1}{n}\sum_{s=1}^{r} d_s^{\star 2}\beta^{\star\prime}Q^{\star(s)}Q^{\star(s)\prime}\beta^{\star}$$
$$= \frac{1}{n}\sum_{s=1}^{r} d_s^{\star 2}\beta^{\star\prime}Q^{\star(s)}Q^{\star(s)\prime}Q^{\star(s)}Q^{\star(s)\prime}\beta^{\star} = \frac{1}{n}\sum_{s=1}^{r} d_s^{\star 2}\left\|Q^{\star(s)}Q^{\star(s)\prime}\beta^{\star}\right\|^2.$$

8. Proof of Theorem 3

Let us define $\check{A}_1$ in the following way:

$$\check{A}_1 = \frac{1}{n_{new}}\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\;\beta^{\star\prime}\;X_{new}^{\star\prime}X_{new}\;Q^{(s)}P^{(s)\prime}\sum_{j=1}^{r^\star} d_j^{\star}P^{\star(j)}Q^{\star(j)\prime}\beta^{\star}.$$

We have:

$$X_{new}^{\star\prime}X_{new} = Z^{\star}F^{\star}W^{\star\prime}WFZ'$$
$$= \sum_{s=1}^{r_{new}}\sum_{k=1}^{r_{new}^{\star}} Z^{\star(k)}f_k^{\star}f_s W^{\star(k)\prime}W^{(s)}Z'^{(s)}.$$

Then,

$$\check{A}_1 = \frac{1}{n_{new}}\sum_{s=1}^{r}\frac{d_s}{d_s^2+\lambda}\sum_{\ell=1}^{r_{new}}\sum_{k=1}^{r_{new}^{\star}} f_k^{\star}f_\ell < W^{\star(k)}, W^{(\ell)} > \sum_{j=1}^{r^\star} d_j^{\star} < P^{(s)}, P^{\star(j)} >< Z^{(\ell)}Z^{\star(k)\prime}\beta^{\star}, Q^{(s)}Q^{\star(j)\prime}\beta^{\star} > .$$

Further, a natural estimator of A_2 is

$$\check{A}_2 = \frac{\sigma_e^2}{n_{new}}\sum_{s=1}^{r}\frac{d_s^2}{(d_s^2+\lambda)^2}\mathrm{Tr}\left(X_{new}Q^{(s)}Q^{(s)\prime}Q^{(s)}Q^{(s)\prime}X_{new}'\right)$$
$$= \frac{\sigma_e^2}{n_{new}}\sum_{s=1}^{r}\frac{d_s^2}{(d_s^2+\lambda)^2}\mathrm{Tr}\left(WFZ'Q^{(s)}Q^{(s)\prime}ZFW'\right).$$

We can easily see that

$$\mathrm{Tr}\left(WFZ'Q^{(s)}Q^{(s)\prime}ZFW'\right) = \sum_{i=1}^{n_{new}}\left(\sum_{\alpha=1}^{r_{new}} f_\alpha Q^{(s)\prime}Z^{(\alpha)}W_i^{(\alpha)}\right)^2,$$

which gives

$$\check{A}_2 = \frac{\sigma_e^2}{n_{new}}\sum_{s=1}^{r}\frac{d_s^2}{(d_s^2+\lambda)^2}\sum_{i=1}^{n_{new}}\left(\sum_{\alpha=1}^{r_{new}} f_\alpha Q^{(s)\prime}Z^{(\alpha)}W_i^{(\alpha)}\right)^2.$$

A natural estimator of A_3 is

$$\check{A}_3 = \frac{1}{n_{new}} \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)' \; X_{new}' X_{new}$$
$$\times \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right).$$

We have the relationship

$$\check{A}_3 = \frac{1}{n_{new}} \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, X_{new} Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)'$$
$$\times \left(\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, X_{new} Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right).$$

Using the fact that

$$X_{new} Q^{(s)} = W F Z' Q^{(s)} = \sum_{\alpha=1}^{r_{new}} f_\alpha Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)},$$

we deduce

$$\sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, X_{new} Q^{(s)} P^{(s)\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)} Q^{\star(j)\prime} \beta^\star$$
$$= \sum_{s=1}^{r} \frac{d_s}{d_s^2+\lambda} \, \sum_{\alpha=1}^{r_{new}} f_\alpha Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \sum_{j=1}^{r^\star} d_j^\star P^{(s)\prime} P^{\star(j)} Q^{\star(j)\prime} \beta^\star.$$

Consequently,

$$\check{A}_3 = \frac{1}{n_{new}} \sum_{s=1}^{r} \sum_{\ell=1}^{r} \frac{d_s}{d_s^2+\lambda} \frac{d_\ell}{d_\ell^2+\lambda} \sum_{\alpha=1}^{r_{new}} f_\alpha^2 \; < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)}, Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} > \sum_{j=1}^{r^\star} d_j^\star < P^{(s)}, P^{\star(j)} > Q^{\star(j)\prime} \beta^\star$$
$$\times \sum_{k=1}^{r^\star} d_k^\star < P^{(\ell)}, P^{\star(k)} > Q^{\star(k)\prime} \beta^\star.$$

References

Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS genetics*. **10**, (2), e1004137.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*. **19**, (4), 1212-1242.

Corbeil, R.R., & Searle, S.R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*. **18**, (1), 31-38.

Daetwyler, H.D., Villanueva, B. & Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*. **3**, (10), e3395.

Daetwyler, H.D., Villanueva, B. & Woolliams, J.A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*. **185**, (3), 1021-1031.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*. **22**, (1), 1-37.

Durrett, R. (2008). *Probability models for DNA sequence evolution.* Springer Science & Business Media.

Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*. **4**, (3), 250-255.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B*. **70**, (5), 849-911.

Ferrao, L. F. V., Ferrao, R. G., Ferrao, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., Garcia, A. A. F. (2018). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*.

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning*, Springer series in statistics Springer, Berlin.

Gezan, S. A., Osorio, L. F., Verma, S., Whitaker, V. M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*. **4**, 16070.

Goddard, M. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*. **136**, (2), 245-257.

Hayes, B., Bowman, P., Chamberlain, A. & Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*. **92**, (2), 433-443.

Haldane J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet.*. **8**, (29), 299-309.

Hoerl, A.E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. **12**, (1), 55-67.

Lee, S. H., Weerasinghe, W. S. P., Wray, N. R., Goddard, M. E., Van Der Werf, J. H. (2017). Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific reports.* **7**, 42091.

Lian, L., Jacobson, A., Zhong, S., Bernardo, R. (2014). Genomewide prediction accuracy within 969 maize biparental populations. *Crop Science.* **54**, (4), 1514-1522.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity.* 108(3), 285.

Meuwissen, T.H., Hayes, B. & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* **157**, (4), 1819-1829.

Minamikawa, M. F., Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H., ... Iwata, H. (2018). Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). *Scientific reports.* **8(1)**, 11994.

Momen, M., Mehrgardi, A. A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., ... Gianola, D. (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific reports.* **8**.

Muranty, H., Troggio, M., Sadok, I. B., Al Rifaï¿½, M., Auwerkerken, A., Banchi, E., ... Kumar, S. (2015). Accuracy and responses of genomic selection on key traits in apple breeding. *Horticulture research.* **2**, 15060.

Nyine, M., Uwimana, B., Blavet, N., HÅibovÃ¡, E., Vanrespaille, H., Batte, M., ... DoleÅ¾el, J. (2018). Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. *The Plant Genome.* **11(2)**, 170090.

Rabier, C.E., Barre, P., Asp, T., Charmet, G. & Mangin, B. (2016). On the Accuracy of Genomic Selection. *PloS One.* **11**, (6), e0156086. doi:10.1371/journal.pone.0156086.

Rabier, C.E., Mangin, B. & Grusea, S. (2018). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics.* https://doi.org/10.1111/sjos.12352.

Schulz-Streeck, T., Ogutu, J., Karaman, Z., Knaak, C. & Piepho, H. (2012). Genomic selection using multiple populations. *Crop Science.* **52**, (6), 2453-2461.

Shao, J. & Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics.* **40**, (2), 812-831.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al (2015). Genomic Selection and Association Mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics.* **11**, (2), e1004982.

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B., Ingvarsson, P. K. (2017). Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology.* **17**, (1),110.

Technow, F. (2014). *R Package hypred: Simulation of Genomic Data in Applied Genetics.* Available from: ??? 06/12/2015].

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological).* 267-288.

Tikhonov, A.N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk..* **SSSR 151**, 501-504.

Visscher, P.M., Yang, J. & Goddard, M.E. (2010). A commentary on "common SNPs explain a large proportion of the heritability for human height" by Yang et al.(2010). *Twin Research and Human Genetics.* **13**, (06), 517-524.

Wu, R., Ma, C. & Casella, G. (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL.* Springer Science & Business Media; 2007.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B.* **68**, (1), 49-67.

Zhang, X., Perez-Rodriguez, P., Semagn, K., Beyene, Y., Babu, R., Lopez-Cruz, M. A., ... Prasanna, B. M. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity.* **114**, (3), 291.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association.* **101**, (476), 1418-1429.

Charles-Elie Rabier (ce.rabier@gmail.com)
ISEM, Université de Montpellier, CNRS, France.
Simona Grusea (grusea@insa-toulouse.fr)
INSA de Toulouse, Institut de Mathématiques de Toulouse, Université de Toulouse, France.

Text S1: Supplementary material of "On the accuracy in high dimensional linear models under imperfect linkage disequilibrium"

C.E. Rabier[a,b], S. Grusea[c]

[a] *ISEM, Université de Montpellier, CNRS, France*
[b] *LIRMM, Université de Montpellier, CNRS, France*
[c] *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France*

---

1. Proof of Lemma 1 of the main manuscript

To begin with, we have to notice that

$$\|PP'\beta^\star\|^2 = \sum_{s=1}^{r} \left\| P^{(s)} P^{(s)\prime} \beta^\star \right\|^2 .$$

Then, using the Cauchy-Schwartz inequality and the fact that $X^\star \beta^\star$ belongs to $\mathit{Span}\left(P^{\star(1)}, \ldots, P^{\star(r^\star)}\right)$, we have

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2 \\
&= \frac{1}{n} \sum_{s=1}^{r} \left(\frac{d_s^2}{d_s^2+\lambda} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\| \right) \left(\left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\| \right) \\
&\leq \frac{1}{n} \left(\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2 \right)^{1/2} \left(\sum_{s=1}^{r} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2 \right)^{1/2} \\
&= \frac{1}{n} \left(\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2 \right)^{1/2} \|PP'X^\star\beta^\star\| \\
&\leq \frac{1}{n} \left(\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \left\| P^{(s)} P^{(s)\prime} X^\star \beta^\star \right\|^2 \right)^{1/2} \|P^\star P^{\star\prime} X^\star\beta^\star\| \\
&= \hat{A}_3^{1/2} \left(\sum_{\ell=1}^{r^\star} d_\ell^{\star 2} \left\| Q^{\star(\ell)} Q^{\star(\ell)\prime} \beta^\star \right\|^2 \right)^{1/2} \\
&= \hat{A}_3^{1/2} \hat{A}_4^{1/2} .
\end{aligned}$$

Besides, since $\hat{A}_2 \geq 0$ and $\rho_g^{oracle} = 1$, we obtain

$$\hat{\rho}_g \leq \frac{\hat{A}_1}{\hat{A}_3^{1/2}\hat{A}_4^{1/2}} \leq \rho_g^{oracle}.$$

In order to obtain the lower bound, we just have to notice that

$$n\hat{A}_1 = \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} \left\| P^{(s)}P^{(s)\prime}X^\star\beta^\star \right\|^2 \geq \|PP'X^\star\beta^\star\|^2 \min_s \frac{d_s^2}{d_s^2+\lambda} \ ,$$

$$n\hat{A}_3 = \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \left\| P^{(s)}P^{(s)\prime}X^\star\beta^\star \right\|^2 \leq \max_s \frac{d_s^4}{(d_s^2+\lambda)^2} \|PP'X^\star\beta^\star\|^2 \ ,$$

$$n\hat{A}_4 = \sum_{\ell=1}^{r^\star} d_\ell^{\star 2} \left\| Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star \right\|^2 \leq \|Q^\star Q^{\star\prime}\beta^\star\|^2 \ \max_\ell d_\ell^{\star 2}.$$

Since $\frac{d_s^4}{(d_s^2+\lambda)^2}$ is bounded by one, we also have $n\hat{A}_2 = \sigma_e^2 \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \leq \sigma_e^2 r$. As a consequence, we have:

$$\frac{\|PP'X^\star\beta^\star\|^2 \ \min_s \frac{d_s^2}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \|PP'X^\star\beta^\star\|^2 \ \max_s \frac{d_s^4}{(d_s^2+\lambda)^2}} \ \sqrt{\|Q^\star Q^{\star\prime}\beta^\star\|^2 \ \max_\ell d_\ell^{\star 2}}} \leq \hat{\rho}_g.$$

2. Introduction

Lemma 1. *Let us consider same hypotheses as in Theorem 1 of the main manuscript. Then, the quantity $\tilde{\rho}_g$ defined in Section 5 of the main manuscript has the following expression*

$$\tilde{\rho}_g = \frac{\tilde{A}_1}{\left(\tilde{A}_2+\tilde{A}_3\right)^{1/2}\left(\tilde{A}_4\right)^{1/2}},$$

where

$$\tilde{A}_1 = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2+\lambda} \ \beta^{\star\prime} \ \mathbb{E}\left(x_{new}^\star x_{new}'\right) \ Q^{(\sigma(s)}P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)}Q^{\star(j)\prime}\beta^\star \ ,$$

$$\tilde{A}_2 = \sigma_e^2 \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{(d_s^2+\lambda)^2} \ \mathbb{E}\left(\left\| Q^{(\sigma(s))}Q^{(\sigma(s))\prime}x_{new} \right\|^2\right)$$

$$\tilde{A}_3 = \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2+\lambda} \ Q^{(\sigma(s))}P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)}Q^{\star(j)\prime}\beta^\star\right)' \ \mathbb{E}\left(x_{new}x_{new}'\right)$$
$$\times \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2+\lambda} \ Q^{(\sigma(s))}P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d_j^\star P^{\star(j)}Q^{\star(j)\prime}\beta^\star\right) \ ,$$

$$\tilde{A}_4 = A_4.$$

Proof. After having replaced the quantity $X'V^{-1}$ by $X'V^{-1}\tilde{P}\tilde{P}'$, formula (5) of Rabier et al. (2016) becomes

$$\rho_g = \frac{\beta^{\star\prime}\,\mathbb{E}\left(x^{\star}_{new}x'_{new}\right)X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}}{\left(\sigma_e^2\mathbb{E}\left(\left\|x'_{new}X'V^{-1}\tilde{P}\tilde{P}'\right\|^2\right)+\beta^{\star\prime}X^{\star\prime}\tilde{P}\tilde{P}'V^{-1}X\mathrm{Var}\left(x_{new}\right)X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}\right)^{1/2}\sigma_G}.$$

As a result, let us define

$$\tilde{A}_1 := \beta^{\star\prime}\,\mathbb{E}\left(x^{\star}_{new}x'_{new}\right)X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}\ ,\quad \tilde{A}_2 := \sigma_e^2\mathbb{E}\left(\left\|x'_{new}X'V^{-1}\tilde{P}\tilde{P}'\right\|^2\right),$$
$$\tilde{A}_3 := \beta^{\star\prime}X^{\star\prime}\tilde{P}\tilde{P}'V^{-1}X\mathrm{Var}\left(x_{new}\right)X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}\ ,\quad \tilde{A}_4 := A_4.$$

Using the fact that $X'V^{-1} = Q\overline{D}P'$, we have

$$\tilde{A}_1 = \beta^{\star\prime}\,\mathbb{E}\left(x^{\star}_{new}x'_{new}\right)Q\overline{D}P'\tilde{P}\tilde{P}'X^{\star}\beta^{\star}.$$

After some simple algebra, we obtain

$$Q\overline{D}P'\tilde{P} = \left(\frac{d_{\sigma(1)}}{d^2_{\sigma(1)}+\lambda}Q^{(\sigma(1))},\ldots,\frac{d_{\sigma(\tilde{r})}}{d^2_{\sigma(\tilde{r})}+\lambda}Q^{(\sigma(\tilde{r}))}\right). \tag{1}$$

Then,

$$\tilde{A}_1 = \beta^{\star\prime}\,\mathbb{E}\left(x^{\star}_{new}x'_{new}\right)\left(\sum_{s=1}^{\tilde{r}}\frac{d_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}Q^{(\sigma(s))}P^{(\sigma(s))\prime}\right)\left(\sum_{s=1}^{r^{\star}}d^{\star}_sP^{\star(s)}Q^{\star(s)\prime}\right)\beta^{\star}.$$

Let us now consider $\tilde{A}_2$. According to Rabier et al. (2018), we have

$$\tilde{A}_2 = \sigma_e^2\sum_{s=1}^{\tilde{r}}\frac{d^2_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}\;\mathbb{E}\left(\left\|Q^{(\sigma(s))}Q^{(\sigma(s))\prime}x_{new}\right\|^2\right).$$

Furthermore, recall that

$$\tilde{A}_3 = \beta^{\star\prime}X^{\star\prime}\tilde{P}\tilde{P}'V^{-1}X\mathrm{Var}\left(x_{new}\right)X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}.$$

Since the expression of $X'V^{-1}\tilde{P}\tilde{P}'X^{\star}\beta^{\star}$ is also present in $\tilde{A}_1$, we easily obtain

$$\tilde{A}_3 = \left(\sum_{s=1}^{\tilde{r}}\frac{d_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}\;Q^{(\sigma(s))}P^{(\sigma(s))\prime}\sum_{j=1}^{r^{\star}}d^{\star}_jP^{\star(j)}Q^{\star(j)\prime}\beta^{\star}\right)'\;\mathbb{E}\left(x_{new}x'_{new}\right)$$
$$\times\left(\sum_{s=1}^{\tilde{r}}\frac{d_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}\;Q^{(\sigma(s))}P^{(\sigma(s))\prime}\sum_{j=1}^{r^{\star}}d^{\star}_jP^{\star(j)}Q^{\star(j)\prime}\beta^{\star}\right).$$

□

3. Proof of Lemma 3 of the main manuscript

To begin with, let us recall the expression $\tilde{A}_1$ given in Lemma 1 above:

$$\tilde{A}_1 = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d^2_{\sigma(s)} + \lambda} \; \beta^{\star\prime} \, \mathbb{E}\left(x^{\star}_{new} x'_{new}\right) \; Q^{(\sigma(s)} P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \; . \quad (2)$$

Let us consider the following natural estimation $\hat{\tilde{A}}_1$:

$$\hat{\tilde{A}}_1 := \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d^2_{\sigma(s)} + \lambda} \; \beta^{\star\prime} \; X^{\star\prime} X \; Q^{(\sigma(s)} P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star.$$

We have the relationship $XQ^{(\sigma(s))} = d_{\sigma(s)} P^{(\sigma(s))}$. As a consequence, after some straightforward matrix algebra, we obtain:

$$X^{\star\prime} X Q^{(\sigma(s))} = d_{\sigma(s)} \sum_{\ell=1}^{r^\star} d^\star_\ell Q^{\star(\ell)} P^{\star(\ell)\prime} P^{(\sigma(s))}.$$

Then,

$$\hat{\tilde{A}}_1 = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \beta^{\star\prime} \frac{d^2_{\sigma(s)}}{d^2_{\sigma(s)} + \lambda} \sum_{\ell=1}^{r^\star} Q^{\star(\ell)} d^\star_\ell P^{\star(\ell)\prime} P^{(\sigma(s))} \sum_{j=1}^{r^\star} d^\star_j P^{(\sigma(s))\prime} P^{\star(j)} Q^{\star(j)\prime} \beta^\star.$$

According to Rabier et al. (2018),

$$\hat{\tilde{A}}_2 = \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)} + \lambda)^2}.$$

An estimation for the quantity $\tilde{A}_3$ is the following

$$\begin{aligned}\hat{\tilde{A}}_3 = \frac{1}{n} & \left(\sum_{s=1}^{\tilde{r}} X \frac{d_{\sigma(s)}}{d^2_{\sigma(s)} + \lambda} \; Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)' \\ \times & \left(\sum_{s=1}^{\tilde{r}} X \frac{d_{\sigma(s)}}{d^2_{\sigma(s)} + \lambda} \; Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^\star} d^\star_j P^{\star(j)} Q^{\star(j)\prime} \beta^\star \right)\end{aligned}$$

Using the fact that $XQ^{(\sigma(s))} = d_{\sigma(s)} P^{(\sigma(s))}$ and after some straightforward matrix algebra, we obtain:

$$\hat{\tilde{A}}_3 = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d^4_{\sigma(s)}}{(d^2_{\sigma(s)} + \lambda)^2} \left(\sum_{\ell=1}^{r^\star} d^\star_\ell P^{(\sigma(s))\prime} P^{\star(\ell)} Q^{\star(\ell)\prime} \beta^\star \right)^2 .$$

4. Some intuition on the different conditions and some explanations on the proof of Lemma 1 of the main manuscript

First, we have to highlight the fact that the shrinkage will potentially have an impact on the singular values d_s of X (e.g. see the terms $d_s^2/(d_s^2+\lambda)$ in $\hat{A}_1$). In contrast, the singular values $d_\ell^\star$ of $X^\star$ are not directly affected by the shrinkage. Recall that the shrinkage parameter λ is necessary in order to handle the high dimensional setting $p >> n$.

Let us consider a "ℓ" that belongs to $\Omega_1^\star$. The key point is the following. When "ℓ" is tagged by a "s" that belongs to Ω_1, the shrinkage does not have any impact since λ is negligible compared to d_s. As soon as "ℓ" is tagged by a "s" that belongs to either Ω_2 or Ω_3, there is a loss due to shrinkage, since λ is not negligible compared to d_s. Condition $(C7^\star)$ (resp. $(C8^\star)$) will ensure that the projection ξ_2^ℓ (resp. ξ_3^ℓ) of $P^{\star(\ell)}$ on $\underset{s\in\Omega_2^\ell}{Span}\left\{P^{(s)}\right\}$ (resp. $\underset{s\in\Omega_3^\ell}{Span}\left\{P^{(s)}\right\}$) is small enough. In that sense, the loss due to the shrinkage will have no impact. In contrast, the projection ξ_1^ℓ of $P^{\star(\ell)}$ on $\underset{s\in\Omega_1^\ell}{Span}\left\{P^{(s)}\right\}$ has to be the largest possible.

On the other hand, let us consider a "s" belonging to Ω_1, that is to say associated to large singular values of X. This "s", not impacted by shrinkage, may tag a "ℓ" belonging to $\Omega_2^\star$ and $\Omega_3^\star$. However, the related terms will be negligible because of conditions $(C4^\star)$ and because of the order of $d_\ell^\star$ compared to λ. We refer to the proof of Lemma 2 in supplementary material for more details (see below).

5. Proof of Lemma 2 of the main manuscript

According to Rabier et al. (2018) (proof relying on Condition C3), we have:

$$n\hat{A}_2 \sim \sigma_e^2 \#\Omega_1 + \sigma_e^2 \sum_{s\in\Omega_2} \frac{1}{(1+C_s)^2}.$$

On the other hand, recall that $\hat{A}_3 = \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}\left(\sum_{\ell=1}^{r^\star} d_\ell^\star P^{(s)\prime}P^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right)^2$.

Then,

$$n\hat{A}_3 \sim \sum_{s\in\Omega_1}\left(\sum_{\ell=1}^{r^\star} d_\ell^\star P^{(s)\prime}P^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right)^2 + \sum_{s\in\Omega_2}\frac{1}{(1+C_s)^2}\left(\sum_{\ell=1}^{r^\star} d_\ell^\star P^{(s)\prime}P^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right)^2$$
$$+ \sum_{s\in\Omega_3}\frac{d_s^4}{\lambda^2}\left(\sum_{\ell=1}^{r^\star} d_\ell^\star P^{(s)\prime}P^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right)^2.$$

Since each "s" is allowed to tag only one ℓ, we have (cf. assumptions in Section 3.1.1 of the main manuscript)

$$
\begin{aligned}
n\hat{A}_3 \sim & \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 \qquad (3)\\
& + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\\
& + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 .
\end{aligned}
$$

From now, let us set $\xi_1^\ell = \xi(p,n)\ \ \forall\ell\in\Omega_1^\star$ with $0<\xi(p,n)\leq 1$. To begin with, let us focus on the first term of formula (3). We have:

$$
\begin{aligned}
\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 &\sim \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi(p,n)}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star}\\
&\sim \sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\xi(p,n)\frac{n^{2\tau}}{r^\star} .
\end{aligned}
$$

Let us focus on the second term of formula (3). We have the relationship

$$
\begin{aligned}
\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 &\sim \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\\
&\sim \sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell .
\end{aligned}
$$

Besides, $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell \leq \sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}$. Since by definition the cardinality of $\Omega_2^\star$ is bounded, and since $\lambda\frac{n^{2\tau}}{r^\star} = o(1)$ (Condition $C4^\star$), we have $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} = o(1)$, that implies $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell = o(1)$.
Let us consider the third term of formula (3):

$$\begin{aligned}\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 &\sim \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star}\\ &\sim \sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell\ .\end{aligned}$$

We have $\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell \leq \sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}$.

Since $\Omega_3^\star$ is bounded, $\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2} = o(\lambda)$. Then, according to $(C4^\star)$, $\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} = o(1)$. As a consequence, $\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell = o(1)$.

Let us move on to the fourth term of formula (3):

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 \sim \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\ .$$

We have:

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} \leq \sum_{\ell\in\Omega_1^\star}\xi_2^\ell d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}.$$

According to Condition $(C7^\star)$, $\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_1^\star}\xi_2^\ell d_\ell^{\star 2} = o(1)$, that implies $\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} = o(1)$.

Let us focus on the fifth term of formula (3):

$$\begin{aligned}\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 &\sim \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\\ &\sim \sum_{\ell\in\Omega_2^\star}\frac{\xi_2^\ell d_\ell^{\star 2}}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\sum_{s\in\Omega_2^\ell}\frac{1}{(1+C_s)^2}.\end{aligned}$$

We have $\sum_{\ell\in\Omega_2^\star}\frac{\xi_2^\ell d_\ell^{\star 2}}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\sum_{s\in\Omega_2^\ell}\frac{1}{(1+C_s)^2} \leq \sum_{\ell\in\Omega_2^\star}\xi_2^\ell\, d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} \leq \sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}$. Since $\#\Omega_2^\star = O(1)$ and $\lambda\frac{n^{2\tau}}{r^\star} = o(1)$ (Condition $C4^\star$) , we have $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} = o(1)$.

As a consequence, $\sum\limits_{\ell\in\Omega_2^\star}\sum\limits_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}=o(1)$.
Let us consider the sixth term of formula (3):

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\sim\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\,.$$

We have

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\le\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star}d_\ell^{\star 2}\,.$$

Since $\Omega_3^\star$ is bounded, $\sum\limits_{\ell\in\Omega_3^\star}d_\ell^{\star 2}=o(\lambda)$. Then, according to $(C4^\star)$, we have $\frac{n^{2\tau}}{r^\star}\sum\limits_{\ell\in\Omega_3^\star}d_\ell^{\star 2}=o(1)$. It implies $\sum\limits_{\ell\in\Omega_3^\star}\sum\limits_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{(1+C_s)^2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}=o(1)$.

Let us study the seventh term of formula (3):

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_3^\ell}d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2\sim\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_1^\star}d_\ell^{\star 2}\sum_{s\in\Omega_3^\ell}\frac{d_s^4}{\lambda^2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\,.$$

We have,

$$\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_1^\star}d_\ell^{\star 2}\sum_{s\in\Omega_3^\ell}\frac{d_s^4}{\lambda^2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\le\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_1^\star}\xi_3^\ell d_\ell^{\star 2}\right)\left(\sum_{s\in\Omega_3}\frac{d_s^4}{\lambda^2}\right)\,.$$

According to $(C3)$ and $(C8^\star)$, the right is term is equal to $o(1)$. As a result, the left term is also negligible.

Let us focus on the eighth term of formula (3):

$$\begin{aligned}\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell}d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2&\sim\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell}d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\\&\sim\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell}\frac{\lambda}{C_\ell^\star}\frac{d_s^4}{\lambda^2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\,.\end{aligned}$$

We have

$$\begin{aligned}\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell}\frac{1}{C_\ell^\star}\frac{d_s^4}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}&\le\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star}\frac{1}{\lambda\,C_\ell^\star\,\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell}d_s^4\\&\le\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_2^\star}\frac{1}{\lambda\,C_\ell^\star\,\#\Omega_3^\ell}\right)\left(\sum_{s\in\Omega_3}d_s^4\right)\,.\end{aligned}$$

Using $(C4^\star)$, $(C3)$ and the fact that $\#\Omega_2^\star$ is bounded, we obtain that the right term of the inequality is equal to $o(1)$. Then, the left term is negligible.

Last, let us study the last (i.e. ninth) term of formula (3):

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 .$$

We have:

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^4}{\lambda^2}\left\|P^{(s)}P^{(s)\prime}P^{\star(\ell)}\right\|^2\left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 \sim \frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell}\frac{d_s^4}{\lambda^2} .$$

Besides,

$$\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell}\frac{d_s^4}{\lambda^2} \le \frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\right)\left(\sum_{s\in\Omega_3}\frac{d_s^4}{\lambda^2}\right).$$

We have already proved that $\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\right) = o(1)$. So, using $(C3)$, the right term is equal to $o(1)$. Then,

$$\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell}\frac{d_s^4}{\lambda^2} = o(1).$$

As a result, all the terms of formula (3) are negligible except the first one. It leads to the relationship:

$$n\hat{A}_3 \sim \xi(p,n)\sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} .$$

Conditions $(C5)$, $(C6)$, and $(C1^\star)$ ensure that

$$\begin{aligned} n\hat{A}_2 + n\hat{A}_3 &\sim \sigma_e^2\#\Omega_1 + \sigma_e^2\sum_{s\in\Omega_2}\frac{1}{(1+C_s)^2} + \xi(p,n)\sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} \\ &\sim \xi(p,n)\sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}. \end{aligned} \tag{4}$$

On the other hand, recall that

$$\hat{A}_1 = \frac{1}{n}\sum_{s=1}^{r}\beta^{\star\prime}\frac{d_s^2}{d_s^2+\lambda}\sum_{\ell=1}^{r^\star} Q^{\star(\ell)} d_\ell^\star P^{\star(\ell)\prime}P^{(s)}\sum_{j=1}^{r^\star} d_j^\star P^{(s)\prime}P^{\star(j)}Q^{\star(j)\prime}\beta^\star .$$

Since each "s" is allowed to tag only one ℓ, we have:

$$\begin{aligned} n\hat{A}_1 \sim & \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi(p,n)}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} \\ & + \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} \\ & + \sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star} + \sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star} \ . \end{aligned} \tag{5}$$

Let us study the first term of formula (5):

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi(p,n)}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} \sim \xi(p,n)\sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} \ .$$

Let us focus on the second term of formula (5):

$$\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} \sim \sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\xi_1^\ell\frac{n^{2\tau}}{r^\star} \ .$$

Besides, $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\xi_1^\ell\frac{n^{2\tau}}{r^\star} \leq \frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}$. Since $\#\Omega_2^\star = O(1)$ and using $(C4^\star)$, we have $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star} = o(1)$. Then, we have $\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}\xi_1^\ell = o(1)$.

Let us focus on the third term of formula (5):

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_1^\ell} d_\ell^{\star 2}\frac{\xi_1^\ell}{\#\Omega_1^\ell}\frac{n^{2\tau}}{r^\star} \sim \frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\xi_1^\ell$$

We have $\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\xi_1^\ell \leq \frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}$. Recall that we have already proved that $\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2} = o(1)$.

Let us handle the fourth term of formula (5):

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} \leq \frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_1^\star}\xi_2^\ell d_\ell^{\star 2}.$$

According to $(C7^\star)$, the right term is equal to $o(1)$.

Let us study the fifth term of formula (5):

$$\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_2^\ell} \frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star} \sim \sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\sum_{s\in\Omega_2^\ell}\frac{1}{1+C_s}.$$

We have :

$$\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}\sum_{s\in\Omega_2^\ell}\frac{1}{1+C_s}\le\sum_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}.$$

Since $\#\Omega_2^\star=O(1)$ and $\lambda\frac{n^{2\tau}}{r^\star}=o(1)$, we have $\sum\limits_{\ell\in\Omega_2^\star} d_\ell^{\star 2}\frac{n^{2\tau}}{r^\star}=o(1)$. As a consequence, $\sum\limits_{\ell\in\Omega_2^\star}\sum\limits_{s\in\Omega_2^\ell}\frac{d_\ell^{\star 2}}{1+C_s}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\frac{n^{2\tau}}{r^\star}=o(1)$.

Let us study the sixth term of formula (5). We have

$$\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\frac{\xi_2^\ell}{\#\Omega_2^\ell}\sum_{s\in\Omega_2^\ell}\frac{1}{1+C_s}\le\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\ .$$

Recall that we have already proved that $\frac{n^{2\tau}}{r^\star}\sum\limits_{\ell\in\Omega_3^\star} d_\ell^{\star 2}=o(1)$.

Let us consider the seventh term of formula (5), that is to say

$$\sum_{\ell\in\Omega_1^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star}\ .$$

We have

$$\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_1^\star} d_\ell^{\star 2}\sum_{s\in\Omega_3^\ell}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\le\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_1^\star}\xi_3^\ell d_\ell^{\star 2}\right)\left(\sum_{s\in\Omega_3}\frac{d_s^2}{\lambda}\right)\ .$$

According to $(C2)$ and $(C8^\star)$, the right term is equal to $o(1)$.

Let us consider the eighth term of formula (5). We have:

$$\sum_{\ell\in\Omega_2^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star}\sim\frac{n^{2\tau}}{r^\star}\sum_{\ell\in\Omega_2^\star}\frac{1}{C_\ell^\star}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell} d_s^2\ .$$

Besides, $\frac{n^{2\tau}}{r^\star}\sum\limits_{\ell\in\Omega_2^\star}\frac{1}{C_\ell^\star}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\sum_{s\in\Omega_3^\ell} d_s^2\le\frac{n^{2\tau}}{r^\star}\left(\sum\limits_{\ell\in\Omega_2^\star}\frac{1}{C_\ell^\star}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\right)\left(\sum\limits_{s\in\Omega_3} d_s^2\right)$

Using $(C4^\star)$, $(C2)$, and the fact that $\#\Omega_2^\star=O(1)$, the right term becomes equal to $o(1)$.

Let us study the ninth term of formula (5):

$$\sum_{\ell\in\Omega_3^\star}\sum_{s\in\Omega_3^\ell} d_\ell^{\star 2}\frac{d_s^2}{\lambda}\frac{\xi_3^\ell}{\#\Omega_3^\ell}\frac{n^{2\tau}}{r^\star}\le\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\right)\left(\sum_{s\in\Omega_3}\frac{d_s^2}{\lambda}\right)\ .$$

Since $\frac{n^{2\tau}}{r^\star}\left(\sum_{\ell\in\Omega_3^\star} d_\ell^{\star 2}\right)=o(1)$, the right term is equal to $o(1)$ using $(C2)$.

To conclude, we obtain:

$$n\hat{A}_1 \sim \sum_{\ell \in \Omega_1^\star} d_\ell^{\star 2} \xi(p,n) \frac{n^{2\tau}}{r^\star}. \tag{6}$$

Last,

$$n\hat{A}_4 \sim \sum_{\ell \in \Omega_1^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star} + \sum_{\ell \in \Omega_2^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star} + \sum_{\ell \in \Omega_3^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star}.$$

We have already shown that $\sum_{\ell \in \Omega_2^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star} = o(1)$ and $\sum_{\ell \in \Omega_3^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star} = o(1)$. Then

$$n\hat{A}_4 \sim \sum_{\ell \in \Omega_1^\star} d_\ell^{\star 2} \frac{n^{2\tau}}{r^\star}. \tag{7}$$

Finally, using formulae (4), (6) and (7), we have for large p and n, $\hat{\rho}_g \sim \sqrt{\xi(p,n)}$. It concludes the proof of the first two items of Lemma 2 of the manuscript.

Let us prove the third statement of Lemma 2 of the manuscript.

When $p \to +\infty$, the distance between markers and QTLs tend to zero. As a consequence, QTLs locations will match a few marker locations (i.e. perfect LD), and each column of $X^\star$ will be included in X. Then, we have $\mathcal{R}_{\text{col}}(X^\star) \subset \mathcal{R}_{\text{col}}(X)$. As a consequence, $\forall \ell \in \Omega_1^\star \cup \Omega_2^\star \cup \Omega_3^\star$, we have $PP'P^{\star\ell} = P^{\star\ell}$ and since $\left\|P^{\star(\ell)}\right\|^2 = 1$, we have the relationship $\xi_1^\ell + \xi_2^\ell + \xi_3^\ell = 1$. Let us recall condition $(C7^\star)$: $\frac{n^{2\tau}}{r^\star} \sum\limits_{\ell \in \Omega_1^\star} \xi_2^\ell d_\ell^{\star 2} = o(1)$. We have $\sum\limits_{\ell \in \Omega_1^\star} \xi_2^\ell d_\ell^{\star 2} \leq (\#\Omega_1^\star)\, d_1^{\star 2} \max_{\ell \in \Omega_1^\star} \xi_2^\ell$ and by definition, $d_1^{\star 2} \sim n^\psi$ with $0 < \psi \leq 1$. In this context, let us set $\forall \ell \in \Omega_1^\star$ $\xi_2^\ell = 1/n^{\theta_1}$ with $\theta_1 > \psi$. Since $d_1^{\star 2} \max_{\ell \in \Omega_1^\star} \xi_2^\ell \sim n^{\psi - \theta_1}$ and $\#\Omega_1^\star = O(1)$, it is clear that condition $(C7^\star)$ is fulfilled.

In the same way, if we set $\forall \ell \in \Omega_1^\star$ $\xi_3^\ell = 1/n^{\theta_2}$ with $\theta_2 > \psi$, condition $(C8^\star)$ is fulfilled. Then, using the new expressions of ξ_2^ℓ and ξ_3^ℓ, we have $\xi_1^\ell = 1 - \xi_2^\ell - \xi_3^\ell = 1 - 1/n^{\theta_1} - 1/n^{\theta_2}$. Moreover, since $\xi_2^\ell \to 0$ and $\xi_3^\ell \to 0$, we can deduce that $\xi_1^\ell \to 1$. As a result, using the notation $\xi(p,n)$ for ξ_1^ℓ, we obtain the desired limits $\xi(p,n) \longrightarrow 1$ and $\hat{\rho}_g \longrightarrow \rho_g^{oracle}$. It concludes the proof.

6. Proof of Lemma 3 of the main manuscript

6.1. *The projected signal belongs only to Span* $\left\{Q^{\star(1)}\right\}$*, and is tagged by one* $s \in \Omega_1$

If the projected signal belongs only to *Span* $\left\{Q^{\star(1)}\right\}$, that is to say

$$\left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|^2 \sim n^{2\tau}, \quad \left\|Q^{\star(\ell)}Q^{\star(\ell)\prime}\beta^\star\right\|^2 = 0, \text{ for } 1 < \ell \leq r^\star \ .$$

Let us consider that $\ell = 1$ is tagged by only one "s" that belongs to Ω_1, i.e. $\left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \sim \xi(p,n)$ only for that s.

Using Theorem 2, we have:

$$\hat{\rho}_g = \frac{\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|}{\left(\sigma_e^2 \sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} + \frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{(1)\prime}\beta^\star\right\|^2\right)^{1/2}} \tag{8}$$

Using further the fact that $d_1^{\star 2} \sim n^\psi$ and $\lambda = o(d_1^2)$ (since $s \in \Omega_1$), we obtain

$$\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\| \sim \xi(p,n)\; n^{\tau+\psi/2}$$

$$\frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|^2 \sim \xi(p,n)\; n^{2\tau+\psi}$$

If $2\tau+\psi > 1$, then $n = o(n^{2\tau+\psi})$. As a consequence, since $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} \leq r \leq n$, we have $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$.

Let us now consider the case $2\tau+\psi < 1$. Then, it is obvious from expression (8), that we need to impose $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{2\tau+\psi}\right)$ in order to obtain $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$.

6.2. The projected signal belongs only to Span$\left\{Q^{\star(1)}\right\}$, and is tagged by one $s \in \Omega_2$

Recall that

$$\hat{\rho}_g = \frac{\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|}{\left(\sigma_e^2 \sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} + \frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{(1)\prime}\beta^\star\right\|^2\right)^{1/2}} \tag{9}$$

Using further the fact that $d_1^{\star 2} \sim n^\psi$, we obtain

$$\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\| \sim \frac{\xi(p,n)\; n^{\tau+\psi/2}}{1+C_s}$$

Besides,

$$\frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|^2 \sim \frac{\xi(p,n)\; n^{2\tau+\psi}}{(1+C_s)^2}$$

If $2\tau+\psi > 1$, then $n = o(n^{2\tau+\psi})$. As a consequence, since $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} \leq r \leq n$, we have $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$.

Let us now consider the case $2\tau+\psi < 1$. Then, it is obvious from expression (10), that we need to impose $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{2\tau+\psi}\right)$ in order to obtain $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$.

6.3. The projected signal belongs only to Span$\left\{Q^{\star(1)}\right\}$, and is tagged by one $s \in \Omega_3$

Recall that

$$\hat{\rho}_g = \frac{\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\|}{\left(\sigma_e^2 \sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} + \frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{(1)\prime}\beta^\star\right\|^2\right)^{1/2}} \tag{10}$$

Let us suppose that $\lambda \sim Cn^{\kappa+\eta}$ with $\kappa > \max(0, -\eta)$. Besides, we set $d_s \sim n^\gamma$, with $\gamma < (\kappa+\eta)/2$. Using further the fact that $d_1^{\star 2} \sim n^\psi$, we obtain

$$\frac{d_s^2 d_1^\star}{d_s^2+\lambda} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{\star(1)\prime}\beta^\star\right\| \sim \frac{\xi(p,n)}{C}\, n^{2\gamma+\tau+\psi/2-\kappa-\eta}.$$

At the denominator, we have:

$$\frac{d_s^4 d_1^{\star 2}}{(d_s^2+\lambda)^2} \left\|P^{(s)}P^{(s)\prime}P^{\star(1)}\right\|^2 \left\|Q^{\star(1)}Q^{(1)\prime}\beta^\star\right\|^2 \sim \frac{\xi(p,n)}{C^2} n^{4\gamma-2\kappa-2\eta+2\tau+\Psi}$$

If $4\gamma - 2\kappa - 2\eta + 2\tau + \Psi > 1$, then $n = o(n^{4\gamma-2\kappa-2\eta+2\tau+\Psi})$. As a consequence, since $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} \leq r \leq n$, we have $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$. When $4\gamma - 2\kappa - 2\eta + 2\tau + \Psi < 1$, we need to impose $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{4\gamma-2\kappa-2\eta+2\tau+\Psi}\right)$ in order to obtain $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}$.

6.4. The projected signal belongs only to Span$\left\{Q^{\star(r^\star)}\right\}$

Lemma 2. *Let us consider same assumptions as in Theorem 2 of the main manuscript. Besides, let us suppose that the projected signal belongs only to Span$\left\{Q^{\star(r^\star)}\right\}$ that is to say*

$$\left\|Q^{\star(r^\star)}Q^{\star(r^\star)\prime}\beta^\star\right\|^2 \sim n^{2\tau}, \quad \left\|Q^{\star(s)}Q^{\star(s)\prime}\beta^\star\right\|^2 = 0, \text{ for } 1 \leq s < r^\star.$$

Moreover, let us assume that $\ell = r^\star$ is tagged only by one s such as $\left\|P^{(s)}P^{(s)\prime}P^{\star(r^\star)}\right\|^2 \sim \xi(p,n)$ with $0 < \xi(p,n) \leq 1$, and $\left\|P^{(u)}P^{(u)\prime}P^{\star(r^\star)}\right\|^2 = 0$ $\forall u \neq s$. Then

- *If $s \in \Omega_1 \cup \Omega_2$, and*
  - *if $2\tau + \eta > 1$, then $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}\, \rho_g^{oracle}$.*
  - *if $2\tau + \eta < 1$, then*
    - *if $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{2\tau+\eta}\right)$, then $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}\, \rho_g^{oracle}$*
    - *if $n^{2\tau+\eta} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right)$, then $\hat{\rho}_g \to 0$.*
- *If $s \in \Omega_3$, $\lambda \sim Cn^{\kappa+\eta}$, $d_s \sim n^\gamma$ with $\kappa > \max(0, -\eta)$, $\gamma < (\kappa+\eta)/2$*

– *if* $4\gamma - 2\kappa - 2\eta + 2\tau + \eta > 1$, *then* $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}\ \rho_g^{oracle}$

– *if* $4\gamma - 2\kappa - 2\eta + 2\tau + \eta < 1$, *then*

  * *if* $\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2} = o\left(n^{4\gamma-2\kappa-2\eta+2\tau+\eta}\right)$, *then* $\hat{\rho}_g \to \lim \sqrt{\xi(p,n)}\ \rho_g^{oracle}$
  * *if* $n^{4\gamma-2\kappa-2\eta+2\tau+\eta} = o\left(\sum_{u=1}^{r} \frac{d_u^4}{(d_u^2+\lambda)^2}\right)$, *then* $\hat{\rho}_g \to 0$.

The proof is largely inspired of the one of Lemma 2 of the main manuscript, as soon as we replace ψ by η.

Table 1: Same as Table 11 of the main manuscript, except that 781 SNPs are used for the TST sample.

Dataset ID	Set 1	Set 2	Set 3	Set 4
Emp. Acc.	0.4289	0.4709	0.4753	0.5638
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4398 (0.0318)	0.4289 (0.0334)	0.4285 (0.0383)	0.4462 (0.0463)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4360 (0.0226)	0.4537 (0.0211)	0.4622 (0.0216)	0.4869 (0.0263)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.2349 (0.05634)	0.2664 (0.0619)	0.3380 (0.0329)	0.5296 (0.0105)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3008 (0.0415)	0.3378 (0.0441)	0.4027 (0.0221)	0.6032 (0.0126)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5449	0.5161	0.4121	0.5078
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4458 (0.0414)	0.4447 (0.0382)	0.4184 (0.0331)	0.4451 (0.0397)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4737 (0.0247)	0.4771 (0.0220)	0.4324 (0.0230)	0.4811 (0.0234)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4045 (0.0313)	0.3893 (0.0284)	0.1965 (0.0743)	0.4053 (0.0244)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4691 (0.0201)	0.4502 (0.0192)	0.2298 (0.0684)	0.4629 (0.0187)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4881	0.5119		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4412 (0.0360)	0.4419 (0.0374)		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4749 (0.0189)	0.4763 (0.0216)		
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3574 (0.0278)	0.4493 (0.0158)		
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4176 (0.0208)	0.5277 (0.0137)		

References

References

Rabier, C.E., Barre, P., Asp, T., Charmet, G. & Mangin, B. (2016). On the Accuracy of Genomic Selection. *PloS One*. **11**, (6), e0156086. doi:10.1371/journal.pone.0156086.

Rabier, C. E., Mangin, B., Grusea, S. (2018). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics*.

Table 2: Same as Table 11 of the main manuscript, except that 1553 SNPs are used for the TST sample.

Dataset ID	Set 1	Set 2	Set 3	Set 4
Emp. Acc.	0.5668	0.5151	0.4889	0.5089
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4535 (0.0403)	0.4489 (0.0422)	0.4438 (0.0379)	0.4379 (0.0394)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4823 (0.0273)	0.4778 (0.0258)	0.4722 (0.0235)	0.4594 (0.0241)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5072 (0.0143)	0.4267 (0.0205)	0.3497 (0.0322)	0.2822 (0.0814)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5526 (0.0121)	0.5081 (0.0156)	0.4205 (0.0203)	0.3587 (0.0625)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5730	0.5091	0.5142	0.5242
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4909 (0.0511)	0.4456 (0.0391)	0.4497 (0.0369)	0.4520 (0.0429)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4579 (0.0288)	0.4686 (0.0244)	0.4825 (0.0222)	0.4805 (0.0259)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4816 (0.0209)	0.4134 (0.0227)	0.4830 (0.0099)	0.4293 (0.0233)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5314 (0.0166)	0.4977 (0.0164)	0.5714 (0.0149)	0.4922 (0.0179)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.5590	0.5156		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4496 (0.0483)	0.4409 (0.0407)		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4831 (0.0276)	0.4723 (0.0245)		
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4936 (0.0176)	0.3664 (0.0357)		
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5403 (0.01615)	0.4461 (0.0227)		

Table 3: Same as Table 11 of the main manuscript, except that 3076 SNPs are used for the TST sample.

Dataset ID	Set 1	Set 2	Set 3	Set 4
Emp. Acc.	?	0.5639	0.4662	0.4851
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	?	0.4494 (0.0478)	0.4351 (0.0364)	0.4456 (0.0377)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	?	0.4813 (0.0288)	0.4587 (0.0241)	0.4684 (0.0237)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	?	0.5304 (0.0111)	0.2552 (0.0758)	0.3152 (0.0415)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	?	0.6094 (0.1449)	0.3328 (0.0607)	0.4079 (0.0210)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5581	0.5096	0.5349	0.5717
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4526 (0.0482)	0.4411 (0.0403)	0.4481 (0.0449)	0.4521 (0.0499)
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4806 (0.0293)	0.4648 (0.0253)	0.4762 (0.0263)	0.4856 (0.0288)
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4818 (0.0191)	0.4002 (0.0249)	0.4237 (0.0269)	0.5277 (0.0148)
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.5469 (0.0145)	0.4784 (0.0167)	0.4832 (0.0206)	0.6113 (0.0148)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4969	0.5266		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{ADLASSO})$	0.4421 (0.0389)	0.4533 (0.0410)		
$\hat{\rho}_{ph}(\hat{X}^\star, \hat{\beta}^\star_{LASSO})$	0.4637 (0.0242)	0.4798 (0.0244)		
$\check{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.3419 (0.0354)	0.4439 (0.0201)		
$\hat{\rho}^{scand}_{ph}(\hat{\beta}_{ADLASSO})$	0.4312 (0.0354)	0.5138 (0.0148)		