



HAL
open science

Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium

Charles-Elie Rabier, Simona Grusea

► **To cite this version:**

Charles-Elie Rabier, Simona Grusea. Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium. *Journal of the Royal Statistical Society: Series C Applied Statistics*, 2021, 70 (4), pp.1001-1026. 10.1111/rssc.12496 . hal-01987222v4

HAL Id: hal-01987222

<https://hal.science/hal-01987222v4>

Submitted on 20 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium

Charles-Elie Rabier

ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France.

IMAG, UMR 5149, CNRS, Université de Montpellier, France

LIRMM, UMR 5506, CNRS, Université de Montpellier, France

Simona Grusea

Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France.

Summary. 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. For this purpose, we introduce new proxies for the accuracy of the prediction. These proxies are suitable under sparse genetic map where it is likely to observe some imperfect linkage disequilibrium, i.e. the situation where the alleles at a gene location and at a marker located nearby vary. We show on rice real data that at least 1553 markers are required to implement GS.

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. These predictions should be accurate in order to allow us to select the best candidates for the breeding program. 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)) and coffee (Ferraio et al. (2018)). Note that, in medicine, the predictive ability of complex 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)).

Let us first recall the definition of a Quantitative Trait Locus (so-called QTL). A QTL is a section of the DNA that contains one or more genes influencing a quantitative trait which is able to be measured (see for instance Wu et al. (2007)). 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)), because of the high density of markers. In genetics, this correlation between a QTL and a marker is named Linkage Disequilibrium (LD): it refers to the non independence of alleles at 2 different loci (see Duret (2008) for more details). Contrary to Genome Wide Association Studies (GWAS) that look for QTLs, the goal in GS is to perform predictions using a large number of markers, without having to detect QTLs. The advantage of genomic predictions in GS over predictions based only on detected loci in GWAS, lies in the

fact that small-effects QTLs are very hard to detect and most of traits, characterized as complex trait, are governed by these small-effects QTLs.

Many factors are known to affect the prediction ability in GS. The size of the training set on which is learned the model, the architecture of the trait (cf. Section 1.3 on heritability), the density of markers, the LD and finally the relatedness between the training and the validation sets are key elements. Panel optimization (e.g. Rincent et al. (2012) and Mangin et al. (2019)), that investigates in detail this relatedness (see Wientjes et al. (2013)), is an important area of research in GS. It consists in choosing the best training individuals to predict breeding values for a given set of candidates.

In this manuscript, we propose a deep study on the LD factor, which is highly linked to the density of markers, and has a large influence on the prediction reliability. A usual estimator of LD is the statistic r^2 , which is the square of the correlation between the values of alleles at two loci in the same gamete (see for instance Weir (2008)). However, in presence of individuals from different genetic origins, this estimator is biased since it suggests for instance some LD between two independent loci with different allele frequencies. In the same way, when the statistic r^2 is computed using a sample of n related gametes, spurious LD is observed: the estimator is biased because the independence between gametes is a required assumption. As a consequence, Mangin et al. (2012) proposed new LD estimators (so called novel measures) that correct bias due to relatedness and population structure. For instance, in case of relatedness, the authors suggest to decorrelate the observations, by premultiplying the vector of observations by a matrix containing the correlations between all pairs of gametes. In GWAS, these LD measures play a role in the power of the association test (see Mangin et al. (2012) for more details). Last but not least, these measures are key elements in GS since they are also present in our general formula (Rabier et al. (2016)) on prediction in GS.

The aim of this paper is to study the accuracy of genomic prediction in GS under imperfect LD. Indeed, in our recent theoretical study (Rabier et al. (2018)), 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 after filtering), whereas 200 markers were sufficient for predicting a simple trait. In our study on GS in raygrass (Rabier et al. (2016)), we noticed that 24957 markers were unable to cover the entire genome (2.7 Gb). Furthermore, in a recent study on GS in coffee, Ferrao et al. (2018) showed that predictions relying on 4000 markers gave similar results as those based on 35000 markers. Last, in the study of Kriaridou et al. (2020) on aquaculture species, similar results were obtained when comparing 1000 markers and a high density of markers. In this context, we propose to tackle here the problem of imperfect LD in GS.

In presence of only a small density of markers, imputation methods (e.g. Beagle software from Browning et al. (2018)), that replace missing data with substituted values, are generally used to compensate this lack of information in GS (Habier (2009)). For instance, in the study on apple of Muranty et al. (2015), selection candidates were genotyped only for 364 markers, but imputation was performed thanks to a population genotyped for 7829 markers. However, most of research teams focusing on imputation, do not address the question of building a proxy suitable under imperfect LD for the prediction ability. Until now, in the literature, we denote only a few proxies appropriate under imperfect LD. de los Campos et al. (2013) derived an upper

bound for the prediction ability but computing this bound is unfeasible in practice on real data. Indeed, the bound relies on the coefficient of the linear regression of genomic relationships at markers on genomic relationships at QTLs. Recall that genomic relationships at QTLs between the training and the validation sets are unknown. In Lian et al. (2014), the authors introduced a proxy, relying on Daetwyler et al. (2008) seminal formula, but adjusted to handle imperfect LD. In particular, the authors assume the independence of each marker-QTL pair. A variant of this proxy was later proposed by our team in Rabier et al. (2016): the effective number of independent loci (Me) was replaced by a term that takes into account the fact that the number of markers is greater than the number of individuals. However, the independence of each marker-QTL pair can be viewed as a very strong assumption because of linkage and the fixed genome size. As a consequence, in this manuscript, we will relax this assumption and present new proxies suitable under imperfect LD. These proxies will help for determining the number of markers required for implementing GS, a topic close to research on the design of LD SNPs chips (e.g. Bolormaa et al. (2015); Corbin et al. (2014)). Indeed, as mentioned by Wu et al. (2016), despite their very good coverage, High Density (HD) SNPs chips may not optimize the ratio between the prediction accuracy and the chip price. In this context, our present work should be useful for designing LD SNP chips that reduce the genotyping costs by considering a moderate density of markers.

In what follows, we will investigate GS in rice with the help of the data of Spindel et al. (2015). We will concentrate on the rice flowering time (days to 50% flowering) collected in Los Banos, Philippines, during the dry season 2012. The data and programs used in our study are available at <https://github.com/rabier/GSImperfectLD>.

1.1. Causal linear model under imperfect LD

We start by introducing the causal statistical model. We are interested in a quantitative trait (i.e. the phenotype) which is observed on a sample of n training (TRN) individuals. Y_1, \dots, Y_n refers to the observations of the quantitative trait. We consider that we have m QTLs on the genome, which influence the quantitative trait.

For $1 \leq j \leq m$, we denote β_j^* the j -th QTL effect. We denote by X^* the matrix of size $n \times m$ containing the QTL alleles for the n individuals. The i -th row of X^* , written as $x_i' = (X_{i,1}^*, \dots, X_{i,p}^*)$, corresponds to the i -th individual, where $'$ denotes transposition.

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

$$Y = X^* \beta^* + \varepsilon, \quad (1)$$

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

We assume that X^* and ε are independent. Moreover, we suppose that genome information is available at p markers, with $p > n$ (high dimensional setting). Note that QTLs **do not necessarily** match markers (imperfect LD setting).

Let X be the matrix of size $n \times p$ containing the genome information at p markers for n individuals. Recall that in our previous work Rabier et al. (2018), we considered the perfect LD framework, where all QTLs are among markers locations. In that case, the considered model was

$$Y = X\beta + \varepsilon, \quad (2)$$

with $1 \leq j \leq p$, $\beta_j = 0$ if the marker is not a QTL, and $\beta_j \neq 0$ if the j th marker is a QTL.

Remark 1. Note that the perfect LD can be viewed as a particular case of imperfect LD. It is the case when the m QTLs are located on a few markers and β denotes the sparse vector of size p , containing the components of β^* .

In the present work, we are under the general case (imperfect LD) where QTLs do not necessarily match markers. r^* (resp. r) will denote the rank of the matrix X^* (resp. X), and $\mathcal{R}_{\text{rows}}(X^*)$ (resp. $\mathcal{R}_{\text{rows}}(X)$) will refer to the linear space generated by the rows of X^* (resp. X). In the same way, $\mathcal{R}_{\text{col}}(X^*)$ and $\mathcal{R}_{\text{col}}(X)$ will denote the corresponding linear spaces spanned by the columns. For the sake of readability, we drop the dependence on n in all the notations. In this work, we condition on $x_1^*, \dots, x_n^*, x_1, \dots, x_n$, to build proxies for GS under imperfect LD. Note however that, before conditioning, some correlation is present between the matrices X^* and X : for instance, due to the fixed genome size, x_i and x_i^* are necessarily correlated.

1.2. Test (TST) individual

We suppose that for a random supplementary individual, so-called test (TST) individual (denoted *new*), we dispose of its genotype but not of its phenotype. As before, let x_{new}^* denote the column vector containing the m QTLs information of the individual *new*.

The quantitative trait Y_{new} is such that:

$$Y_{new} = x_{new}^{*\prime} \beta^* + \varepsilon_{new},$$

where $\varepsilon_{new} \sim N(0, \sigma_e^2)$, and x_{new}^* , ε_{new} and ε are all independent.

Moreover, x_{new} refers to the random genome information at markers. Note that x_{new} and x_{new}^* are correlated because of the genetic linkage due to the fixed genome size.

In what follows, we will assume that Y , Y_{new} , x_{new} , x_{new}^* , the columns of X and the columns of X^* are centered.

1.3. Accuracy and prediction model

Based on the genome information available from the n individuals, we will construct an estimator \hat{Y}_{new} for the phenotypic value of the *new* individual.

Accuracy criteria

The prediction ability is quantified according to the so-called *phenotypic accuracy*, ρ_{ph} (e.g. Visscher et al. (2010)) or to the *genotypic accuracy*, ρ_g (e.g. Daetwyler et al. (2008, 2010)):

$$\rho_{ph} := \frac{\text{Cov}(\hat{Y}_{new}, Y_{new})}{\sqrt{\text{Var}(\hat{Y}_{new}) \text{Var}(Y_{new})}}, \quad \rho_g := \frac{\text{Cov}(\hat{Y}_{new}, x_{new}^{*\prime} \beta^*)}{\sqrt{\text{Var}(\hat{Y}_{new}) \text{Var}(x_{new}^{*\prime} \beta^*)}}. \quad (3)$$

Note that the accuracy criteria, very important in genetics, measures the correlation between predicted and true values. In particular, the accuracy is an essential component in the breeders equation (see for instance Lynch and Walsh (1998)).

Heritability

Since x_{new}^* , ε_{new} and ε are supposed to be all independent, we have the relationship $\rho_{ph}/\rho_g = h$, where h is the squared root of the heritability of the trait:

$$h^2 := \frac{\text{Var}(x_{new}^{*\prime} \beta^*)}{\text{Var}(Y_{new})}. \quad (4)$$

With the notation $\sigma_G^2 = \text{Var}(x_{new}^{*\prime} \beta^*)$, we have the relationship $h^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2)$.

Oracle situation

In the *oracle situation*, the QTLs locations and their effects are known. In this case, the natural predictor is $\hat{Y}_{new}^{oracle} = x_{new}' \beta^*$ and the the oracle accuracies are $\rho_g^{oracle} = 1$ and $\rho_{ph}^{oracle} = h$

L^2 prediction loss

In the main manuscript, we focused on the accuracy criteria. In Supplementary Material, we present a few results regarding the L^2 prediction loss, more familiar for statisticians.

Predictor based on Ridge regression

As in our previous study (Rabier et al, 2018), we will use as instrumental model, the Ridge regression which can be viewed as a Bayesian regression. Then, the Ridge regression estimator for the marker effects β (Tihonov (1963); Hoerl et al. (1970)) is:

$$\hat{\beta} = (X'X + \lambda I_p)^{-1} X'Y = X'V^{-1}Y, \quad (5)$$

where $V = XX' + \lambda I_n$, λ refers to a regularization (or tuning) parameter, and I_p denotes the identity matrix of size $p \times p$.

The corresponding Ridge regression predictor for the phenotype of the test individual is:

$$\hat{Y}_{new} = x_{new}' \hat{\beta} = x_{new}' X' V^{-1} Y. \quad (6)$$

Note that this Ridge estimator is based on genome information at markers and is suitable in a high dimensional setting (i.e. $p > n$, see e.g. Shao and Deng (2012) and Bühlmann (2013)). In genetics, this estimator is also called random regression best linear unbiased predictor (RRBLUP) and it is known to be equivalent to genomic best linear unbiased predictor (GBLUP).

1.4. Our contributions and roadmap

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 causal matrix (i.e. at genes), and the one of the design matrix (i.e. at markers). Then, we state our Theorem 1 dealing with the *genotypic accuracy* under imperfect LD. This theorem is somewhat essential since the other results, appropriate under imperfect LD, are built on it.

In Section 3, Theorem 2 introduces an estimation $\hat{\rho}_g$ of ρ_g that does not require the genome information of TST individuals. TRN and TST are supposed to be sampled from the same probability distribution. According to this theorem, the projection of the regression function $X^* \beta^*$ on $\mathcal{R}_{col}(X)$ 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 $\mathcal{R}_{rows}(X)$ (as in Rabier et al. (2018)). Lemma 1 introduces, under imperfect LD, a lower bound for $\hat{\rho}_g$: it takes into account a global projection (same weights on each subspace) of $X^* \beta^*$ on the space spanned by the columns of X . Lemma 2 assumes that the signal β^* is spread out uniformly on each subspace of $\mathcal{R}_{rows}(X^*)$. The oracle accuracy is reached as soon as the limit of a loss factor (so called $1-\xi(n)$), due to imperfect LD, is equal to zero.

Section 4 of this manuscript introduces a modified predictor $\hat{\hat{\rho}}_g$ that improves the quality of the prediction. It relies on the projection of Y on a well chosen subspace of $\mathcal{R}_{col}(X)$. Lemma 3 proposes an estimation of that predictor's accuracy: as expected, under imperfect LD, it depends on the projection of the regression function $X^* \beta^*$ on the chosen subspace. After having introduced bounds for $\hat{\hat{\rho}}_g$ in Lemma 4, we will give a result that allows to compare $\hat{\hat{\rho}}_g$ and $\hat{\rho}_g$ under imperfect LD.

To conclude, in Section 5, we will present new imperfect LD proxies relying on our theoretical

results. Performances of these proxies will be illustrated on simulated data and on real data. Our proxies rely only on phenotypes and markers of TRN, but take into account the sparsity of the TST map. In order to build these imperfect LD proxies, we investigate a topic in GS that has not been studied before (as far as we know): the accuracy of the prediction of TST individuals 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^* and β^* required to compute our proxies. 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 β^* 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 analyze real data of Spindel et al. (2015) on GS in rice, considering different densities of markers. With the help of our imperfect LD proxies, we show that geneticists can evaluate the accuracy of their future predictions on TST individuals and figure out if they should redensify their TST genetic map to improve the reliability of their predictions. Last but not least, on these rice datasets, our imperfect LD proxies outperformed existing proxies based on Lian et al. (2014).

In the Supplementary Material, we present the mathematical proofs of our results, and show extra results regarding real data. In Section 10 of the Supplementary Material, we also present a few results regarding the L^2 prediction loss.

2. General expression for the genotypic accuracy

2.1. An existing formula suitable under imperfect LD

Recall that Y , Y_{new} , x_{new} , x_{new}^* , the columns of X and the columns of X^* are centered.

Conditioning on $x_1, \dots, x_n, x_1^*, \dots, x_n^*$, we have the following formula of Rabier et al. (2016) for the genotypic accuracy corresponding to the predictor \hat{Y}_{new} :

$$\rho_g = \frac{\beta^{*'} \mathbb{E}(x_{new}^* x_{new}') X' V^{-1} X^* \beta^*}{\left\{ \sigma_e^2 \mathbb{E} \left(\|x_{new}' X' V^{-1}\|^2 \right) + \beta^{*'} X^{*'} V^{-1} X \text{Var}(x_{new}) X' V^{-1} X^* \beta^* \right\}^{1/2} \sigma_G} \quad (7)$$

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

2.2. Alternative expression based on singular value decomposition (SVD)

We consider the singular value decomposition (SVD) of X^* (see Shao and Deng (2012) and Bühlmann (2013)):

$$X^* = P^* D^* Q^{*'}, \quad (8)$$

where P^* is a $n \times r^*$ matrix satisfying $P^{*'} P^* = I_{r^*}$, Q^* is a $m \times r^*$ matrix satisfying $Q^{*'} Q^* = I_{r^*}$, $D^* = \text{Diag}(d_1^*, \dots, d_{r^*}^*)$ with $d_1^* \geq \dots \geq d_{r^*}^* > 0$ and r^* is the rank of the matrix X^* .

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 $\mathcal{R}_{\text{rows}}(X^*) = \text{Span}\{Q^{*(1)}, \dots, Q^{*(r^*)}\}$.

In the same way, the SVD of X is

$$X = P D Q', \quad (9)$$

where P is a $n \times r$ matrix satisfying $P'P = I_r$, Q is a $p \times r$ matrix satisfying $Q'Q = I_r$, $D = \text{Diag}(d_1, \dots, d_r)$ with $d_1 \geq \dots \geq d_r > 0$, and r is the rank of the matrix X .

Using the SVD, the formula (7) gives the following alternative expression for the accuracy under imperfect LD

Theorem 1. *Let us assume that ε , x_{new} , x_{new}^* and ε_{new} are random. Then, conditionnally on X and X^* , the genotypic accuracy has the following expression*

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

where

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

The proof is given in Section 1 of the Supplementary Material. The phenotypic accuracy is obtained by replacing the term A_4 at the denominator by $A_4 + \sigma_e^2$. For a version of this result in the case of perfect LD, see Rabier et al. (2018).

Remark 2. *Note that we can express the L^2 prediction loss as follows:*

$$\mathbb{E} \left\{ (x_{new}' \hat{\beta} - x_{new}' \beta^*)^2 \right\} = A_2 + A_3 + A_4 - 2A_1 .$$

We will prove this formula in Section 10.1 of the Supplementary Material. Note that an alternative expression for A_1 is the following:

$$A_1 = \sum_{s=1}^r \frac{d_s}{d_s^2 + \lambda} \beta^{*'} \mathbb{E}(x_{new}^* x_{new}') Q^{(s)} Q^{(s)'} Q^{(s)} P^{(s)'} X^* \beta^* . \quad (10)$$

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 β^* . According to the above formula (10), we can notice that the term $d_s Q^{(s)} Q^{(s)'} \beta$ from Theorem 1 of Rabier et al. (2018) has been replaced here by the quantity $Q^{(s)} Q^{(s)'} Q^{(s)} P^{(s)'} X^* \beta^*$. In other words, under imperfect LD, we have to consider the projection of the vector $P^{(s)'} X^* \beta^* Q^{(s)}$ on $\text{Span}\{Q^{(s)}\}$, whereas under perfect LD, the projection of $d_s \beta$ on $\text{Span}\{Q^{(s)}\}$ is taken into account. Same remark holds for A_3 at the denominator.

Remark 3. *Since formulas obtained under imperfect LD are more general, we can easily retrieve 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. $X^* \beta^* = X \beta$), and $X^* \beta^* Q^{(s)}$ is obviously equal to $d_s \beta$.*

In what follows we are interested in estimating the genotypic accuracy ρ_g . The consistency of estimators of A_1 , A_2 , A_3 and A_4 guarantees the consistency of the estimator of ρ_g , thanks to Slutsky's lemma in the matrix case. However, as mentioned in our previous study, finding consistent estimators of A_1 , A_3 and A_4 is challenging in the high dimensional setting: the covariance matrix Σ needs to be estimated. As a consequence, we have chosen the empirical covariance estimator of Σ , as generally used by geneticists in practice.

3. Estimation of the genotypic accuracy

In the following, we consider that TRN and TST samples come from the same probability distribution. In this context, using the empirical covariances $X^{*'}X/n$, $X'X/n$ and $X^{*'}X^*/n$ as estimates for the covariances $\mathbb{E}(x_{new}^*x_{new}^*)$, $\mathbb{E}(x_{new}x_{new}')$ and $\mathbb{E}(x_{new}^*x_{new}^{*'})$ appearing in Theorem 1, we obtain the following theorem.

Theorem 2. *Let us assume that x_1, \dots, x_n and x_{new} are independent and identically distributed (i.i.d.). In the same way, let us assume that x_1^*, \dots, x_n^* and x_{new}^* are i.i.d. Then, conditionally on X and X^* , and assuming that ε , x_{new} and ε_{new} are random, 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

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

The proof is given in Section 2 of the Supplementary Material.

We can see that the term $d_s^2 \left\| Q^{(s)} Q^{(s)'} \beta \right\|^2$ from Theorem 2 of Rabier et al. (2018) obtained under perfect LD has been replaced by the quantity $\left\| P^{(s)} P^{(s)'} X^* \beta^* \right\|^2$ in the expressions of \hat{A}_1 and \hat{A}_3 . 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 $X^* \beta^* = X \beta$.

Remark 4. *This estimation $\hat{\rho}_g$ relies only on phenotypes and markers of TRN. As a consequence, this accuracy estimation can be used to evaluate GS accuracy before genotyping of the TST individuals.*

Remark 5. *In practice, the nuisance parameters X^* and β^* can be estimated with a penalized likelihood method, by considering a more dense map for TRN than for TST. We refer to the applications in Section 5 for more details.*

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*

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

The proof is given in Section 3 of the Supplementary Material.
 Note that \hat{A}_1 and \hat{A}_3 can be rewritten in the following way:

$$\begin{aligned}\hat{A}_1 &= \frac{1}{n} \sum_{s=1}^r \beta^{s'} \frac{d_s^2}{d_s^2 + \lambda} \sum_{\ell=1}^{r^*} Q^{*(\ell)} d_\ell^* P^{*(\ell)'} P^{(s)} \sum_{j=1}^{r^*} d_j^* P^{(s)'} P^{*(j)} Q^{*(j)'} \beta^*, \\ \hat{A}_3 &= \frac{1}{n} \sum_{s=1}^r \frac{d_s^4}{(d_s^2 + \lambda)^2} \left(\sum_{\ell=1}^{r^*} d_\ell^* P^{(s)'} P^{*(\ell)} Q^{*(\ell)'} \beta^* \right)^2.\end{aligned}$$

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

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

In order to study asymptotic properties of $\hat{\rho}_g$, we consider that

$$\begin{aligned}d_1^{*2} &\sim n^\psi \quad \text{with } 0 < \psi \leq 1, \\ d_{r^*}^{*2} &\sim n^\eta \quad \text{with } \eta \leq \psi \leq 1 \text{ and } \eta \text{ and } \psi \text{ not depending on } n.\end{aligned}$$

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

$$\|Q^* Q^{*'} \beta^*\|^2 \sim n^{2\tau}, \quad \text{with } \tau < \eta \text{ and } \tau \text{ not depending on } n.$$

Although r^* is bounded in our study, these conditions are somewhat inspired from Shao and Deng (2012) and Fan and Lv (2008).

Let us further consider a regularization parameter λ such as $\lambda \rightarrow \infty$ and $\lambda = o(d_1^{*2})$. Let us consider the following partition $\Omega_1^*, \Omega_2^*, \Omega_3^*$ of $\{1, \dots, r^*\}$:

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

Note that Ω_1^* contains at least the index 1. Moreover, let $\Omega_1, \Omega_2, \Omega_3$ be the following partition of $\{1, \dots, r\}$:

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

Recall that in our previous ‘‘perfect LD’’ study, we considered only these last 3 sets.

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

For every $\ell \in \{1, \dots, r^*\}$, we define the following sets Ω_k^ℓ , $k = 1, 2, 3$:

$$\Omega_k^\ell := \left\{ s \in \Omega_k \mid \left\| P^{(s)} P^{(s)'} P^{*(\ell)} \right\|^2 \neq 0 \right\}.$$

In other words, we assume that the projection of $P^{*(\ell)}$ on $\text{Span}\{P^{(1)}, \dots, P^{(r)}\}$ is spread out on the subspaces $\text{Span}\{P^{(s)}\}$, $\text{Span}\{P^{(s)}\}$, and $\text{Span}\{P^{(s)}\}$.

For every $k = 1, 2, 3$, we impose $\Omega_k^\ell \cap \Omega_k^{\ell'} = \emptyset, \forall \ell \neq \ell'$. In other words, a given “ s ” can not tag different “ ℓ ”.

Besides, $\forall \ell \in \Omega_1^*$, we will impose the corresponding set Ω_1^ℓ to be non empty: each “ ℓ ” associated to a large singular value of X^* is tagged by at least one “ s ” associated to large singular values of X . This implies that $\#\Omega_1^* \leq \#\Omega_1$, where $\#$ denotes the cardinality. 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 $\ell \in \Omega_1^*$ can also be tagged by some “ s ” that belong to Ω_2 or Ω_3 .

Moreover, for a general ℓ , with $1 \leq \ell \leq r^*$, we assume that within each subspace $\text{Span}_{s \in \Omega_k^\ell} \{P^{(s)}\}$, $k = 1, 2, 3$, the projection is spread out uniformly on each component $P^{(s)}$. As a consequence, taking into account the fact that $\|P^{*(\ell)}\|^2 = 1$, we define $\xi_k^{(\ell)} \in]0, 1]$, $k = 1, 2, 3$ by:

$$(C0^*) \text{ If } \#\Omega_k^\ell \neq 0, \left\| P^{(s)} P^{(s)'} P^{*(\ell)} \right\|^2 \sim \frac{\xi_k^{(\ell)}}{\#\Omega_k^\ell} \quad \forall s \in \Omega_k^\ell,$$

with $\sum_{k|\Omega_k^\ell \neq \emptyset} \xi_k^{(\ell)} \leq 1$.

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

$$\begin{aligned} \bullet (C1^*) \quad \frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} d_\ell^{*2} &\rightarrow +\infty & \bullet (C2) \quad \sum_{s \in \Omega_3} d_s^2 &= o(\lambda) \\ \bullet (C3) \quad \sum_{s \in \Omega_3} d_s^4 &= o(\lambda^2) & \bullet (C4^*) \quad \frac{n^{2\tau}}{r^*} &= o(1/\lambda) \\ \bullet (C5) \quad \#\Omega_1 &= O(1) & \bullet (C6) \quad \#\Omega_2 &= O(1) \\ \bullet (C7^*) \quad \frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} \xi_2^{(\ell)} d_\ell^{*2} &= o(1) & \bullet (C8^*) \quad \frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} \xi_3^{(\ell)} d_\ell^{*2} &= o(1). \end{aligned}$$

Because of conditions (C5) and (C6), since $p > n$, 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^* of the matrix X^* is bounded and Ω_1^* , Ω_2^* and Ω_3^* are finite sets. Some intuition and some explanations on these conditions are given in Section 4 of the Supplementary Material.

The following Lemma 2 assumes imperfect LD and that the signal is spread out uniformly on each subspace of $\mathcal{R}_{\text{rows}}(X^*)$.

Lemma 2 (Convergence to the oracle accuracy). *Let us consider same assumptions as in Theorem 2 and suppose that for every $k = 1, 2, 3$, we have $\Omega_k^\ell \cap \Omega_k^{\ell'} = \emptyset, \forall \ell \neq \ell'$. Besides, let us suppose that the projected signal is spread out uniformly on each subspace $\text{Span} \{Q^{*(\ell)}\}$, i.e.*

$$\left\| Q^{*(\ell)} Q^{*(\ell)'} \beta^* \right\|^2 \sim \frac{n^{2\tau}}{r^*}, \ell = 1, \dots, r^*. \quad (11)$$

Moreover, $\forall \ell \in \Omega_1^*$, let us assume that $\Omega_1^\ell \neq \emptyset$ and that $\xi_1^{(\ell)} = \xi(n)$ with $0 < b < \xi(n) \leq 1$. Then, assuming conditions (C0* – C1* – C2 – C3 – C4* – C5 – C6 – C7* – C8*):

- for large n , we have $\hat{\rho}_g \sim \sqrt{\xi(n)} \rho_g^{\text{oracle}}$

- if $\forall \ell \in \Omega_1^*$, $\xi_2^{(\ell)} = 1/n^{\theta_1}$ and $\xi_3^{(\ell)} = 1/n^{\theta_2}$ with $\theta_1 > \psi$ and $\theta_2 > \psi$, then we have $\hat{\rho}_g \rightarrow \rho_g^{\text{oracle}}$.

The proof is given in Section 5 of the Supplementary Material (see also Section 4 for some intuition). In Lemma 2 of Rabier et al. (2018), we obtain a simpler result in the case of perfect LD, in the case where the signal is spread out uniformly on the subspaces of $\mathcal{R}_{\text{rows}}(X)$.

Remark 6. For each $\ell \in \Omega_1^*$, $\xi(n)$ is the percentage of the L^2 norm of $P^{*(\ell)}$ represented on $\text{Span}\{P^{(s)}\}_{s \in \Omega_1^*}$. Note that under our conditions, we are only able to capture this percentage of the

L^2 norm of $P^{*(\ell)}$ (see Sections 4 and 5 of the Supplementary Material). $1 - \xi(n)$ can be viewed as a loss coefficient: it is the percentage of the L^2 norm of $P^{*(\ell)}$ that is unable to be captured (either from $\text{Span}\{P^{(s)}\}_{s \in \Omega_2^\ell}$, either $\text{Span}\{P^{(s)}\}_{s \in \Omega_3^\ell}$ or the complementary subspace).

Moreover, since $p \rightarrow +\infty$ when $n \rightarrow +\infty$, the distance between markers and QTLs tends to zero. As a consequence, QTLs locations match certain marker locations (i.e. perfect LD), and each column of X^* is included in X . Then, according to Lemma 2, the oracle accuracy is reached as soon as $\lim \sqrt{\xi(n)}$ is equal to one when $n \rightarrow +\infty$ (i.e. no loss). 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 in Section 6 of Supplementary Material the asymptotic behavior of our estimate $\hat{\rho}_g$ when the projected signal belongs only to one component (either $\text{Span}\{Q^{*(1)}\}$ or $\text{Span}\{Q^{*(r^*)}\}$).

4. An improved predictor

As before, we are interested in predicting the phenotype Y_{new} of a so-called test (TST) individual (denoted *new*), whose genome information is denoted x_{new} . 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 $1 \leq \tilde{r} \leq r$ and $\sigma(\cdot)$ a one-to-one map $\sigma : \{1, \dots, \tilde{r}\} \rightarrow \{1, \dots, r\}$. We thus have $\sigma(k) \neq \sigma(k')$ for $k \neq k'$. Let us consider the estimator

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

Besides, we set $\tilde{Q} := (Q^{\sigma(1)}, \dots, Q^{\sigma(\tilde{r})})$. Then, we propose the following predictor for the so-called *new* individual:

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

The corresponding genotypic accuracy is then :

$$\tilde{\rho}_g := \frac{\text{Cov}(\tilde{Y}_{\text{new}}, x'_{\text{new}}\beta)}{\sqrt{\text{Var}(\tilde{Y}_{\text{new}})\text{Var}(x'_{\text{new}}\beta)}}. \quad (12)$$

In Lemma 7.1 of Section 7 of the Supplementary Material, we give a more detailed formula for $\tilde{\rho}_g$.

Similarly to Theorem 2, we propose in the following theorem an estimation of the accuracy of the new predictor.

Lemma 3. *Let us consider same hypotheses as in Theorem 2. Then, an estimation of the quantity $\tilde{\rho}_g$ 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

$$\begin{aligned} \hat{A}_1 &:= \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \left\| P^{(\sigma(s))} P^{(\sigma(s))'} X^* \beta^* \right\|^2, \quad \hat{A}_2 := \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2 + \lambda)^2}, \\ \hat{A}_3 &:= \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2 + \lambda)^2} \left\| P^{(\sigma(s))} P^{(\sigma(s))'} X^* \beta^* \right\|^2, \quad \hat{A}_4 := \hat{A}_4. \end{aligned}$$

The proof is given in Section 8 of the Supplementary Material.

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

Lemma 4 (Bounds on $\hat{\rho}_g$). *Using same assumptions as in Theorem 2, we always have*

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

The proof relies heavily on the proof of Lemma 1, using the expressions of \hat{A}_1 , \hat{A}_2 and \hat{A}_3 given in Lemma 3. We can notice that at the denominator, the quantities \tilde{r} and $\left\| \tilde{P} \tilde{P}' X^* \beta^* \right\|^2$ replace now the quantities r and $\|PP' X^* \beta^*\|^2$ of Lemma 1. This decrease at the denominator will be profitable provided that the numerator does not decrease too much.

The following Lemma 5 is the analogue of Lemma 6 under imperfect LD of Rabier et al. (2018) and allows to compare the quantities $\hat{\rho}_g$ and $\hat{\rho}_g$ for fixed n .

Lemma 5. *Let us suppose that $\hat{A}_1 - \hat{A}_1 \neq 0$. Then, we have $\hat{\rho}_g \geq \hat{\rho}_g$ if and only if the following relation holds:*

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

Let us briefly recall the explanation given in Rabier et al. (2018). We have the decomposition $\hat{\beta} = \tilde{\beta} + \vec{\beta}$, with $\vec{\beta} := X'V^{-1}\tilde{P}\tilde{P}'Y$ where \tilde{P} denotes the matrix obtained from P by removing the column vectors $P^{\sigma(1)}, \dots, P^{\sigma(\tilde{r})}$. Similarly, we have $\hat{Y}_{\text{new}} = \tilde{Y}_{\text{new}} + \vec{Y}_{\text{new}}$, where $\vec{Y}_{\text{new}} := x'_{\text{new}} \vec{\beta}$ denotes the prediction.

Then, the different terms of the statement can be rewritten:

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

Last, in the same way as what has been done before, we tackle in Section 9 of Supplementary Material, a few extreme cases: the projected signal belongs either to $\text{Span}\{Q^{*(1)}\}$ or $\text{Span}\{Q^{*(r^*)}\}$.

5. Applications under imperfect LD

We propose to illustrate here our theoretical results, with the help of simulated and real data. To begin with, Subsection 5.1 describes the simulation framework. Next, Subsection 5.2 is devoted to a new topic in GS: the accuracy of the prediction on TST individuals when the genetic map of TRN differs from the one of TST. A dense TRN map will help for estimating nuisance parameters, required to compute our imperfect LD proxies. Subsection 5.3 studies the performances of the estimator $\hat{\beta}$, introduced in Section 4. Last, Section 5.4 is dedicated to GS in rice. We reanalyzed real data of Spindel et al. (2015) and compared performances of our proxies with existing proxies, using different densities of markers.

5.1. Simulation framework

The simulation framework is largely inspired of Rabier et al. (2016, 2018). The hypred R package (Technow (2014)) was used to produce genomic data. Populations were simulated by random mating between haploid individuals (i.e. with only one copy of each chromosome), during (a) 50, (b) 70 generations, or (c) 100 generations. In generation zero, eight haploid founder lines were crossed. The eight founder setup was supposed 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. Recall that by definition, according to the Haldane (1919) model, there are on average x crossovers on a genetic map of length x Morgan (see for instance Wu et al. (2007)). We considered 3 different densities of genetic markers equally spaced on the chromosome: (a) 500, (b) 1,000, or (c) 2,000 SNPs. These densities of markers were used for the TRN map. However, the TST map contains only half the number of markers of the TRN map.

For the phenotypic model, we set the environmental variance σ_e^2 to 1. Besides, we considered either (a) 25 QTLs with effects 0.45, or (b) 100 QTLs with effects 0.30. The QTLs were equally spaced on the chromosome and were observed only in the TRN sample (i.e. not observed in the TST sample). For simulations regarding the improved predictor, a slightly different TST map and different QTLs locations were considered.

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. 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^* 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). To sum up, the genomes of the 500 TRN individuals do not vary across simulations, whereas the phenotypes vary. Besides, for the 100 TST individuals, the genomes and phenotypes are different for each simulation.

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 and the rrBLUP R package (Endelman (2011)).

In what follows, in order to make the reading easier, we will adopt the notation $\hat{\rho}_{ph}(X^*, \beta^*)$ for $\hat{\rho}_{ph}$. This will help for enumerating the nuisance parameters that have to be estimated.

Table 1. Comparison among different estimators of the phenotypic accuracy, when the QTLs are not observed in the TST sample. The TRN map contains 500 markers whereas the TST map contains only 250 markers. 25 QTLs are located along the genome. Emp. Acc. refers to the empirical phenotypic accuracy, whereas $\hat{\rho}_{ph}^{pLD}$ and $\tilde{\rho}_{ph}^{pLD}$ refer to complete LD proxies from Rabier et al. (2018). The Mean Squared Errors (MSE) with respect to the Empirical Accuracy are given in brackets, and their average over the 3 numbers of generations is denoted $\overline{\text{MSE}}$. For each genome length T, the proxy with the smallest $\overline{\text{MSE}}$ is highlighted in gray.

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

5.2. Imperfect LD proxies when TRN and TST do not share the same genetic map (Tables 1, 2, 3)

As previously said, we focus here on the prediction accuracy when the genetic map of TRN differs from the one of TST. The dense TRN map will help for estimating nuisance parameters, required to compute our imperfect LD proxies.

Recall that our imperfect LD proxies are built to estimate satisfactorily the accuracy of future predictions based on the TST map. Recall also that our proxies rely only on phenotypes and markers of TRN, but take into account the sparsity of the TST map. As a result, for a given TST map (i.e. a chosen number of markers for TST), we can first compute our proxies without having to genotype TST individuals, in order to evaluate the accuracy of future predictions based on the TST map. Next, once a satisfactory accuracy has been found, we will need to genotype TST individuals at the chosen density of markers, in order to perform predictions and to select the most promising candidates for GS.

In this context, let us consider a more dense map for TRN than for TST. Since the estimation $\hat{\rho}_{ph}(X^*, \beta^*)$ depends on nuisance parameters X^* and β^* , we propose to estimate these 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 $\hat{\rho}_{ph}(X^*, \beta^*)$, built on Theorem 2, relies on the design matrix X . In this context, using the same number of generations for TRN and TST, both samples (TRN and TST) share the same probability distribution, and it is reasonable to

Table 2. Same as Table 1 except that more markers are considered. The TRN map contains 1000 markers whereas the TST map contains only 500 markers.

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

consider the estimation $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}^*)$ as a proxy for the predictive ability. In order to estimate β^* 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). We refer to Hastie et al. (2009) for more details on these penalized likelihood methods. Tables 1, 2 and 3 compare the performances of our new proxies, that handle imperfect LD, with proxies suggested in Rabier et al. (2018) under perfect LD assumptions (using the Adaptive LASSO as a substitute for β). In what follows, $\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$ (resp. $\check{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$) will refer to the “perfect LD” proxies available before (resp. after) genotyping TST individuals.

Tables 1, 2 and 3 deal respectively with 250 markers, 500 markers and 1000 markers, equally spaced on a chromosome of length T. The dense TRN map contains twice the number of markers. According to Tables 1, 2 and 3, there is a clear advantage to handle explicitly imperfect LD for T=4 and T=6, whatever the density of markers: the proxies $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$ and $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$ gave always better performances than the quantities $\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$ and $\check{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$ relying on perfect LD. Recall that perfect LD proxies assume that QTLs are located on the TST map.

In contrast, when a chromosome of length 1M was studied, $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$ was the only proxy found to be more accurate than “perfect LD” proxies. Indeed, when p was set to 500 or 1000, $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$ was outperformed by “perfect LD” proxies. This result is not so surprising, since this genetic map is close to mimick 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 1 and 2 in Supplementary material).

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

Table 3. Same as Table 1 except that more markers are considered. The TRN map contains 2000 markers whereas the TST map contains only 1000 markers.

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

5.3. The quality of the prediction can be improved (Tables 4 and 5)

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 choose $\sigma(\cdot)$ such as $\frac{d_{\sigma(k)}^2}{d_{\sigma(k)}^2 + \lambda} \|P(\sigma(k))P(\sigma(k))'X^*\beta^*\|^2$ is the k -th largest term of the sequence $\frac{d_s^2}{d_s^2 + \lambda} \|P(s)P(s)'X^*\beta^*\|^2_{s=1, \dots, r}$.

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

During this procedure, β^* was estimated with the help of a penalized likelihood method. Table 4 compares the empirical correlations $\widehat{\text{Cor}}(\tilde{Y}_{\text{new}}, Y_{\text{new}})$ when subspaces were chosen according to the Adaptive LASSO or according to the LASSO. The table reports also the empirical accuracy, relying on the classical Ridge estimator.

In all the cases studied in Table 4, the empirical accuracy associated to the new estimator $\tilde{\beta}$ was always slightly greater than the classical empirical accuracy based on the Ridge estimator. Moreover, for the choice of the \tilde{r} subspaces, we could not establish the superiority of one penalized likelihood method over another.

Last, Table 4 investigates the case where the vector β^* belongs to $\mathcal{R}_{\text{rows}}(X^*)$. As expected, we observe a significant increase in terms of accuracy when the “modified predictor” is adopted.

5.4. 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

Table 4. Illustration of the predictions based on $\tilde{\beta}$. $\widehat{\text{Cor}}(\hat{Y}_{\text{new}}, Y_{\text{new}})$ refers to the empirical correlation between \hat{Y}_{new} and Y_{new} . $\widehat{\text{Cor}}(\tilde{Y}_{\text{new}}^{\text{ADLASSO}}, Y_{\text{new}})$ (resp. $\widehat{\text{Cor}}(\tilde{Y}_{\text{new}}^{\text{LASSO}}, Y_{\text{new}})$) refers to the empirical correlation between \tilde{Y}_{new} and Y_{new} , with the help of the Adaptive Lasso (resp. Lasso) for the choice of the subspace. The genome is of length T and 2 QTLs are located at 3cM and 80cM with effects +2 and -4 respectively ($\sigma_e^2 = 4$). 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 are not observed in the TST sample.

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

Table 5. Same as Table 4 except that the vector β^* belongs to $\mathcal{R}_{\text{rows}}(X^*)$. QTLs are located at marker locations of the TRN map on $[0, 1]$ ($\sigma_e^2 = 1$) and the vector β^* is such that $\beta^* = \omega Q^{*(1)} + \omega Q^{*(2)} + \omega Q^{*(3)}$.

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

Los Banos, Philippines, during the dry season 2012.

Among 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 TST 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. Each computed accuracy relies on 100 data sets: for each of the 10 sets, 100 draws were considered (with random individuals in TRN and TST sets). The nuisance parameters were estimated thanks to the most dense TRN map, i.e. the one containing the 73147 SNPs.

For each subset size, Table 6 reports the average performance of different GS proxies over the 10 random sets. In contrast, Tables 7 and 8 are dedicated to the configuration with 448 SNPs and 781 SNPs respectively, and provide results regarding each random set. Note that Tables 3 and 4 that handle 1553 and 3076 SNPs respectively, are included in Supplementary Material.

In order to compare our suggested proxies with existing proxies, two other proxies suitable under imperfect LD are also present in the different tables. Indeed, we denote in the literature only a few proxies appropriate under imperfect LD (cf. introduction). In particular, we considered the proxy of Lian et al. (2014) and a variant of this proxy proposed by our team in Rabier et al. (2016). The work of Lian et al. (2014) relies on Daetwyler et al. (2008) seminal formula, but the formula was adjusted by the authors to handle imperfect LD (cf. introduction). Our study in Rabier et al. (2016) is based on Lian et al. (2014), but we slightly modified the proxy in order to take into account the high dimensional setting.

To make the reading easier, let us recall the expression of the proxy of Lian et al. (2014):

$$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{\text{Me}}_{LJ}) = \hat{r}^2 \hat{h} \sqrt{\frac{\hat{h}^2/(1-\hat{h}^2)}{\widehat{\text{Me}}_{LJ}/n + \hat{r}^2 \hat{h}^2/(1-\hat{h}^2)}} \quad (13)$$

where \hat{h}^2 , \hat{r}^2 , $\widehat{\text{Me}}_{LJ}$ refer respectively to the estimated heritability, the estimated LD between each QTL and its associated marker, the effective number of independent loci estimated by the method of Li and Ji (2005). In our analyses, \hat{h}^2 was set to 0.4378 (see Table 1 of Spindel et al. (2015)). Besides, as mentioned in introduction, Lian et al. (2014) assumed a constant LD between each QTL and its associated marker. As a result, since our rice data can be viewed as a double haploid population, we used the relationship $\hat{r}^2 = \sqrt{r_{MM}^2}$ as advised by Lian et al. (2014), where r_{MM}^2 denotes the average LD between consecutive markers. On the other hand, the expression of our existing proxy is $\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{\text{F}}_{RM})$ where $\widehat{\text{F}}_{RM}$ stands for an estimation of the quantity $\mathbb{E}(\|x'_{new} X' V^{-1}\|^2)$.

Let us now describe the performances of the different proxies on these rice data sets. To begin with, we have to mention that in all cases studied, the two existing proxies were outperformed by our new proxies. Besides, according to Table 6, $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$ 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 $p = 448$, $\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$ and $\check{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$ were equal on average to 0.3168 and 0.3662 respectively, instead of 0.4789 (see also results from sets 3, 6, 8 and 10 in Table 7). In contrast, the imperfect LD proxy $\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$ was satisfactory for all densities 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 5.2), for such large genome size ($T = 13.101\text{M}$ in rice), it seems that we should choose the LASSO and not the Adaptive LASSO as a substitute for β^* 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.

6. Discussion and Conclusion

In this manuscript, we tackle the problem of imperfect LD in GS. Nowadays, GS is an important area of research in genetics that consists in selecting individuals on the basis of genomic predictions. Imperfect LD usually refers to the situation where QTLs do not match marker locations. In this case, the allele at a QTL usually differs from the alleles at its flanking markers. Imperfect LD can be observed for instance in some species where the number of markers is too small to cover the large genome size. More importantly, although genotyping costs have largely dropped during the last decade, breeders want to know the number of markers required to implement GS, in order to optimize their budget (see studies on eucalyptus, Tan et al. (2017), on coffee, Ferrao et al. (2018), and on aquaculture species Kriaridou et al. (2020)).

In most of these applied papers in GS, researchers start by generating random subsets with different numbers of markers. Then, for each density of markers, they evaluate the prediction ability by cross validation. As a result, they can figure out the optimal number of markers, i.e. the threshold from which no significant improvement is observed when adding extra markers. However, this kind of approach requires the availability of a high density of markers, in order to build subsets of different sizes.

In our present study, we propose to guide the experimenter prior to proceeding to high density genotyping. In particular, we introduce new imperfect LD proxies relying on a theoretical analysis, which provide good estimates for accuracy of predictions on TST individuals. These imperfect LD proxies will also help for determining the number of markers required for implementing GS. For this purpose, we suggest to consider two different genetic maps: a dense map for TRN and a sparse map for TST. The dense map only helps to estimate nuisance parameters present in the expression of the proxies, such as the QTL effects β^* and the QTL alleles X^* . Note that our imperfect LD proxies rely only on phenotypes and on markers of TRN. In that sense, high density genotyping has to be performed only for the TRN set. Although the proxies do not require genotyping and phenotyping of TST, they take into account the sparsity of the TST map. So, in order to exploit our proxies in an efficient way, the breeders should first compute the proxies with a very sparse TST map. Next, if the accuracy is too low, the breeders should redensify the TST map and recompute the proxies. For instance, in our rice data analysis, we considered successively 448, 781, 1553 and 3076 markers, and showed a good agreement between our proxies and the empirical accuracy. Last, once a satisfactory accuracy has been found, we have to genotype TST individuals at the chosen density of markers (e.g. 1553 markers), in order to perform predictions and to select the most promising candidates for GS.

In this paper, we have estimated the nuisance parameters β^* by penalized regression because of the high dimensional setting. Using extensive simulations, we have compared three estimators for β^* : the LASSO (Tibshirani (1996)), the Adaptive LASSO (Zou (2006)) and the Group LASSO (Yuan and Lin (2006)). The LASSO is found to be the best substitute for β^* when the TST map is sparse, whereas the Adaptive LASSO is more appropriate when the TST map is dense. As previously said, the superiority of the Adaptive LASSO was expected in presence of a large number of markers for TST. Indeed, we are close to mimick a perfect LD situation and we have already shown in a previous study (Rabier et al. (2018)) that we should use the Adaptive LASSO under perfect LD. More importantly, in our theoretical analysis, we did not investigate the loss of accuracy due to the estimation of β^* . It would have made the mathematical treatment even more complex. In the future, this aspect should be addressed in order to have a deep understanding of the behavior of the LASSO and of the Adaptive LASSO, as a function of the sparsity of the

TST map.

On the other hand, results on penalized regressions could have been exploited in a different way. In Mangin et al. (2019), the authors assume perfect LD and only use the penalized regression to locate QTLs. It is only during a second step that QTL effects are estimated thanks to the putative causal genome. This extra step is supposed to avoid the bias of the estimator due to the high dimensional setting. Under imperfect LD, this approach appears to be feasible and could be interesting to study. On the other hand, The MultiLocus Mixed Model (MLMM) forward selection of Segura et al. (2012) should also be investigated in the future under imperfect LD. Indeed, it gave very satisfactory results on real data in Mangin et al. (2019).

A deep treatment of the imperfect LD scenario was necessary, since this topic is more complex than the perfect LD, and there are only a few methodological papers in the literature. Our mathematical study helps to understand this concept in detail, and to provide relevant proxies to geneticists. From a theoretical point of view, we have shown that under imperfect LD, the projection of the regression function $X^*\beta^*$ on $\mathcal{R}_{\text{col}}(X)$ plays a key role for the accuracy. Our analytical treatment also highlights the presence of a loss factor (see Lemma 2), due to imperfect LD, that does not allow to reach the oracle (i.e. optimal) accuracy. From a practical point of view, we have shown on simulated data that the performances of perfect LD proxies deteriorate heavily when the LD between QTLs and markers decreases. In contrast, our imperfect LD proxies that model explicitly the fact that QTLs do not match exactly marker locations do not suffer from this drawback. Besides, on real data, our suggested proxies outperformed existing proxies based on the work of Lian et al. (2014) and devised for imperfect LD. This is not surprising, since our proxies rely on more complex formulas than in Lian et al. (2014), whose proxies are extended versions of Daetwyler et al. (2008) seminal formula that was derived assuming drastic conditions.

To conclude, in this study we have introduced new imperfect LD proxies that give satisfactory results on simulated and real data. These proxies should help breeders to figure out the number of markers required for implementing GS. Although many proxies are already present in the literature on perfect LD (see Morota (2017) for a review), only a few proxies have been devised until now for imperfect LD. In this context, our work is intended to fill this gap and to provide useful tools that were missing for the GS community. These tools could be interesting for designing cost-effective SNP chips. Indeed, as mentioned in Wu et al. (2016), LD SNP chips can be viewed as a good compromise between chip prices and increased prediction accuracies.

Supplementary Material: The online version of this article offers Supplementary Material that gives the mathematical proofs of the results of the main manuscript.

Acknowledgements

We would like to thank the two referees for their helpful suggestions which improved the paper.

References

- Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, and Inouye M (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS genetics*; 10(2) e1004137.
- Bolormaa S, Gore K, Van Der Werf JHJ, Hayes BJ, and Daetwyler HD (2015). Design of a low density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Animal Genetics*; 46(5) 544-556.

Table 6. 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 densities of markers for the TST samples are studied. The Mean Squared Error (MSE) with respect to the Empirical Accuracy is given in brackets. For each density of markers, the proxy with the tiniest MSE is highlighted in gray. MSE refers to the average over the 4 densities of markers.

Method	448 SNPs	781 SNPs	1553 SNPs	3076 SNPs	MSE
Emp. Acc.	0.4789	0.4919	0.5275	0.5242	
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4269 (0.0355)	0.4379 (0.0376)	0.4520 (0.0419)	0.4461 (0.0430)	0.0395
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4621 (0.0244)	0.4653 (0.0226)	0.4737 (0.0254)	0.4728 (0.0263)	0.0247
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3168 (0.0529)	0.3571 (0.0364)	0.4233 (0.0264)	0.4115 (0.0290)	0.0362
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3662 (0.0454)	0.4202 (0.0281)	0.4919 (0.0215)	0.4952 (0.0342)	0.0323
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.2147 (0.2790)	0.1941 (0.3065)	0.1669 (0.3685)	0.1409 (0.3915)	0.3364
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3771 (0.1370)	0.3877 (0.1356)	0.4006 (0.1485)	0.4187 (0.1342)	0.1388

Table 7. Same as Table 6 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.

Dataset ID	Set 1	Set 2	Set 3	Set 4
Emp. Acc.	0.5993	0.5445	0.4117	0.5054
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4764 (0.0429)	0.4441 (0.0441)	0.4053 (0.0322)	0.4358 (0.0356)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.5125 (0.0271)	0.4847 (0.02486)	0.4380 (0.0236)	0.4808 (0.0207)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.5065 (0.0171)	0.4712 (0.0154)	0.1580 (0.0959)	0.4222 (0.0176)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.5404 (0.0124)	0.5128 (0.0128)	0.2059 (0.0867)	0.4663 (0.0153)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.2160 (0.3758)	0.2208 (0.3357)	0.2113 (0.2318)	0.2012 (0.3117)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3658 (0.2297)	0.3708 (0.1914)	0.3806 (0.0961)	0.3538 (0.1658)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.4676	0.4081	0.4878	0.4455
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4309 (0.0353)	0.4070 (0.0348)	0.4362 (0.0373)	0.4214 (0.0348)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4653 (0.0233)	0.4207 (0.0232)	0.4676 (0.0227)	0.4508 (0.0244)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3251 (0.0398)	0.1774 (0.0907)	0.3732 (0.0286)	0.2726 (0.0668)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3953 (0.0298)	0.2179 (0.0823)	0.4343 (0.0211)	0.3274 (0.0586)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.1964 (0.2928)	0.2195 (0.2256)	0.2182 (0.2772)	0.2239 (0.2463)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3552 (0.1461)	0.3955 (0.0964)	0.3810 (0.1297)	0.3919 (0.1064)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4427	0.4696		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4117 (0.0218)	0.4130 (0.0366)		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4622 (0.0316)	0.4382 (0.0229)		
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.2789 (0.0404)	0.1829 (0.1179)		
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3255 (0.0314)	0.2366 (0.1036)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.2194 (0.2322)	0.2206 (0.2609)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3864 (0.0976)	0.3897 (0.1110)		

Table 8. Same as Table 7, 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
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4398 (0.0318)	0.4289 (0.0334)	0.4285 (0.0383)	0.4462 (0.0463)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4360 (0.0226)	0.4537 (0.0211)	0.4622 (0.0216)	0.4869 (0.0263)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.2349 (0.05634)	0.2664 (0.0619)	0.3380 (0.0329)	0.5296 (0.0105)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3008 (0.0415)	0.3378 (0.0441)	0.4027 (0.0221)	0.6032 (0.0126)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.1983 (0.2586)	0.1907 (0.2898)	0.2020 (0.2837)	0.1892 (0.3717)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.4085 (0.0929)	0.3931 (0.1116)	0.3962 (0.1121)	0.3608 (0.2047)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5449	0.5161	0.4121	0.5078
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4458 (0.0414)	0.4447 (0.0382)	0.4184 (0.0331)	0.4451 (0.0397)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4737 (0.0247)	0.4771 (0.0220)	0.4324 (0.0230)	0.4811 (0.0234)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.4045 (0.0313)	0.3893 (0.0284)	0.1965 (0.0743)	0.4053 (0.0244)
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.4691 (0.0201)	0.4502 (0.0192)	0.2298 (0.0684)	0.4629 (0.0187)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.1953 (0.3537)	0.1867 (0.3386)	0.1876 (0.2474)	0.2020 (0.2831)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3738 (0.1828)	0.3734 (0.1609)	0.4012 (0.0973)	0.3962 (0.1121)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4881	0.5119		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4412 (0.0360)	0.4419 (0.0374)		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4749 (0.0189)	0.4763 (0.0216)		
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.3574 (0.0278)	0.4493 (0.0158)		
$\hat{\rho}_{ph}^{pLD}(\hat{\beta}_{ADLASSO})$	0.4176 (0.0208)	0.5277 (0.0137)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{MeLJ})$	0.1944 (0.3090)	0.1946 (0.3290)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \widehat{F}_{RM})$	0.3931 (0.1292)	0.3809 (0.1525)		

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

Browning BL, Zhou Y, and Browning SR (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*; 103(3) 338-348.

Corbeil RR, and Searle SR (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*; 18(1) 31-38.

Corbin LJ, Kranis A, Blott SC, Swinburne JE, Vaudin M, Bishop SC, and Woolliams JA (2014). The utility of low-density genotyping for imputation in the Thoroughbred horse. *Genetics Selection Evolution*; 46(1) 9.

Daetwyler HD, Villanueva B, and Woolliams JA (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*; 3(10) e3395.

Daetwyler HD, Villanueva B, and Woolliams JA (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics*; 185(3) 1021-1031.

de los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013). Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet*; 9(7) e1003608.

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

- Endelman JB (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*; 4(3) 250-255.
- Fan J and Lv J (2008). Sure independence screening for ultrahigh dimensional feature space. *J.R. Statist. Soc. B*; 70(5) 849-911.
- Ferrao LFV, Ferrao RG, Ferrao MAG, Fonseca A, Carbonetto P, Stephens M, and Garcia AAF (2018). Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity*.
- Hastie T, Tibshirani R, and Friedman J (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science Business Media.
- Gezan SA, Osorio LF, Verma S, and Whitaker VM (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*; 4 16070.
- Hayes B, Bowman P, Chamberlain A, and 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.
- Habier D, Fernando RL, and Dekkers JC (2009). Genomic selection using low-density marker panels. *Genetics*; 182(1) 343-353.
- Hoerl AE, and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*; 12(1) 55-67.
- Kriaridou C, Tsairidou S, Houston RD, and Robledo D (2020). Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. *Frontiers in Genetics*; 11 124.
- Lee SH, Weerasinghe WSP, Wray NR, Goddard ME, and Van Der Werf JH (2017). Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific reports*; 7 42091.
- Lian L, Jacobson A, Zhong S, and Bernardo R (2014). Genomewide prediction accuracy within 969 maize biparental populations. *Crop Science*; 54(4) 1514-1522.
- Li J and Ji L (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*; 95(3) 221-227.
- Lynch M and Walsh B (1998). *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA.
- Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, and Cierco-Ayrolles C (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*; 108(3) 285.
- Mangin, B., Rincent, R., Rabier, C. E., Moreau, L., Goudemand-Dugue, E. (2019). Training set optimization of genomic prediction by means of EthAcc. *PloS one*; 14(2) e0205629.
- Meuwissen TH, Hayes B, and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*; 157(4) 1819-1829.

- Minamikawa MF, Takada N, Terakami S, Saito T, Onogi A, Kajiya-Kanegae H ..., and 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 AA, Sheikhi A, Kranis A, Tusell L, Morota G, ..., and Gianola D (2018). Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific reports*; 8.
- Morota G (2017). ShinyGPAS: interactive genomic prediction accuracy simulator based on deterministic formulas *Genetics Selection Evolution*; 49(1) 91.
- Muranty H, Troggio M, Sadok IB, ..., and 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, ..., and Dolezel 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 CE, Barre P, Asp T, Charmet G, and Mangin B (2016). On the Accuracy of Genomic Selection. *PloS One*; 11(6) e0156086. doi:10.1371/ journal.pone.0156086.
- Rabier CE, Mangin B, and Grusea S (2018). On the accuracy in high dimensional linear models and its application to genomic selection. *Scand. J. Statist.*; 46(1) 289-313.
- Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, ... and Schön CC (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*; 192(2) 715-728.
- Schulz-Streeck T, Ogutu J, Karaman Z, Knaak C, and Piepho H (2012). Genomic selection using multiple populations. *Crop Science*; 52(6) 2453-2461.
- Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*; 44(7) 825.
- Shao J and Deng X (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.*; 40(2) 812-831.
- Spindel J, Begum H, Akdemir D, Virk P, Collard B, and 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 GS, Ferreira KZ, Sundberg B, and Ingvarsson PK (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*.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*; 267-288.
- Tikhonov AN (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk. SSSR* 151 501-504.

- Visser PM, Yang J, and Goddard ME (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.
- Weir, B. S. (2008). Linkage disequilibrium and association mapping. *Annu. Rev. Genomics Hum. Genet.*; 9 129-142.
- Wientjes YC, Veerkamp RF, and Calus MP (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*; 193(2) 621-631.
- Wu R, Ma C, and Casella G (2007). *Statistical genetics of quantitative traits: linkage, maps and QTL*. Springer Science & Business Media.
- Wu XL, Xu J, Feng G, Wiggans GR, Taylor JF, He J, ... , and Bauck S (2016). Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications. *PloS one*; 11(9) e0161719.
- Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*; 68(1) 49-67.
- Zhang X, Perez-Rodriguez P, Semagn K, Beyene Y, Babu R, Lopez-Cruz MA, ..., and Prasanna BM (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. *J. Am. Statist. Ass*; 101(476) 1418-1429.

Supplementary Material of “Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium”

Charles-Elie Rabier and Simona Grusea

ISE-M, UMR 5554, CNRS, IRD, Université de Montpellier, France
IMAG, UMR 5149, CNRS, Université de Montpellier, France
LIRMM, UMR 5506, CNRS, Université de Montpellier, France
Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France
e-mail: ce.rabier@gmail.com; grusea@insa-toulouse.fr

1. Proof of Theorem 1 of the main manuscript

By definition,

$$A_1 = \beta^{*'} \mathbb{E}(x_{new}^* x_{new}') X' V^{-1} X^* \beta^*.$$

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

$$X' V^{-1} = Q \bar{D} P'. \quad (1)$$

Recall that $X^* = P^* D^* Q^{*'}$. After easy calculations, we obtain

$$X' V^{-1} P^* D^* Q^{*'} \beta^* = \sum_{s=1}^r \frac{d_s}{d_s^2 + \lambda} Q^{(s)} P^{(s)'} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)'} \beta^*. \quad (2)$$

Then,

$$A_1 = \sum_{s=1}^r \frac{d_s}{d_s^2 + \lambda} \beta^{*'} \mathbb{E}(x_{new}^* x_{new}') Q^{(s)} P^{(s)'} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)'} \beta^*.$$

By definition,

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

According to Theorem 1 of [2], we also have

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

By definition,

$$A_3 = \beta^{*\prime} X^{*\prime} V^{-1} X \text{Var}(x_{new}) X' V^{-1} X^* \beta^*.$$

According to formula (2), we obtain the desired result

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

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

$$A_4 = \beta^{*\prime} \mathbb{E}(x_{new} x_{new}') \beta^*.$$

2. Proof of Theorem 2 of the main manuscript

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

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

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

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

$$X^{*\prime} X Q^{(s)} = d_s \sum_{\ell=1}^{r^*} Q^{*(\ell)} d_\ell^* P^{*(\ell)\prime} P^{(s)}.$$

We deduce

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

According to Theorem 2 of [2], a natural estimation of A_2 is

$$\hat{A}_2 = \frac{\sigma_e^2}{n} \sum_{s=1}^r \frac{d_s^2}{(d_s^2 + \lambda)^2} \sum_{i=1}^n \left\| Q^{(s)} Q^{(s)\prime} x_i \right\|^2,$$

and it leads to the following expression

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

Let us consider the following estimation of A_3

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

We have

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

Note that

$$X Q^{(s)} Q^{*(j)'} \beta^* = P D Q' Q^{(s)} Q^{*(j)'} \beta^* = d_s P e_s Q^{*(j)'} \beta^* = d_s P^{(s)} Q^{*(j)'} \beta^*$$

where e_s denotes the s -th vector of the canonical basis of \mathbb{R}^r . As a consequence,

$$\sum_{s=1}^r \frac{d_s}{d_s^2 + \lambda} X Q^{(s)} \sum_{j=1}^{r^*} d_j^* Q^{*(j)'} \beta^* P^{(s)'} P^{*(j)} = \sum_{s=1}^r \frac{d_s^2}{d_s^2 + \lambda} P^{(s)} \sum_{j=1}^{r^*} d_j^* P^{(s)'} P^{*(j)} Q^{*(j)'} \beta^*.$$

Last, we obtain

$$\hat{A}_3 = \frac{1}{n} \sum_{s=1}^r \frac{d_s^4}{(d_s^2 + \lambda)^2} \left(\sum_{j=1}^{r^*} d_j^* P^{(s)'} P^{*(j)} Q^{*(j)'} \beta^* \right)^2.$$

Finally, let us consider the following estimation of A_4 :

$$\hat{A}_4 = \frac{1}{n} \beta^{*'} X^{*'} X^* \beta^*.$$

We have

$$\begin{aligned} \hat{A}_4 &= \frac{1}{n} \beta^{*'} Q^* D^{*2} Q^{*'} \beta^* = \frac{1}{n} \sum_{s=1}^r d_s^{*2} \beta^{*'} Q^{*(s)} Q^{*(s)'} \beta^* \\ &= \frac{1}{n} \sum_{s=1}^r d_s^{*2} \beta^{*'} Q^{*(s)} Q^{*(s)'} Q^{*(s)} Q^{*(s)'} \beta^* = \frac{1}{n} \sum_{s=1}^r d_s^{*2} \left\| Q^{*(s)} Q^{*(s)'} \beta^* \right\|^2. \end{aligned}$$

3. Proof of Lemma 1 of the main manuscript

To begin with, we have to notice that

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

Then, using the Cauchy-Schwartz inequality and the fact that $X^*\beta^*$ belongs to $\text{Span}(P^{*(1)}, \dots, P^{*(r^*)})$, we have

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

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

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

Since $\frac{d_s^4}{(d_s^2+\lambda)^2} \leq 1$, 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:

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

4. Some intuition on the different conditions and on the proof of Lemma 2 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 (see e.g. the terms $d_s^2/(d_s^2 + \lambda)$ in \hat{A}_1). In contrast, the singular values d_ℓ^* of X^* are not directly affected by the shrinkage. Recall that the shrinkage parameter λ is necessary in order to handle the high dimensional setting $p \gg n$.

Let us consider a “ ℓ ” that belongs to Ω_1^* . 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*) (resp. (C8*)) will ensure that the projection $\xi_2^{(\ell)}$ (resp. $\xi_3^{(\ell)}$) of $P^{*(\ell)}$ on $\text{Span}_{s \in \Omega_2} \{P^{(s)}\}$ (resp. $\text{Span}_{s \in \Omega_3} \{P^{(s)}\}$) is small enough. In that sense, the loss due to the shrinkage will have no impact. In contrast, the projection $\xi_1^{(\ell)}$ of $P^{*(\ell)}$ on $\text{Span}_{s \in \Omega_1} \{P^{(s)}\}$ 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 Ω_2^* and Ω_3^* . However, the related terms will be negligible because of conditions (C4*) and because of the order of d_ℓ^* compared to λ . We refer to the proof of Lemma 2 for more details (see below).

5. Proof of Lemma 2 of the main manuscript

According to the proof of Lemma 2 in Rabier et al. [2] (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 $\hat{A}_3 = \frac{1}{n} \sum_{s=1}^r \frac{d_s^4}{(d_s^2+\lambda)^2} \left(\sum_{\ell=1}^{r^*} d_\ell^* P^{(s)\prime} P^{*(\ell)} Q^{*(\ell)\prime} \beta^* \right)^2$.

Then,

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

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

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

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

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

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

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

Besides, $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \xi_1^{(\ell)} \leq \sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*}$. Since by definition the cardinality of Ω_2^* is bounded, and since $\lambda \frac{n^{2\tau}}{r^*} = o(1)$ (Condition (C4*)), we have $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} = o(1)$, that implies $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \xi_1^{(\ell)} = o(1)$.

Let us further consider the third term of formula (3):

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

We have $\sum_{\ell \in \Omega_3^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \xi_1^{(\ell)} \leq \sum_{\ell \in \Omega_3^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*}$. Since Ω_3^* is bounded, $\sum_{\ell \in \Omega_3^*} d_\ell^{*2} = o(\lambda)$. Then, according to (C4*), $\sum_{\ell \in \Omega_3^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} = o(1)$. As a consequence, $\sum_{\ell \in \Omega_3^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \xi_1^{(\ell)} = o(1)$.

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

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

We have:

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

According to Condition (C7*), $\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} \xi_2^{(\ell)} d_\ell^{*2} = o(1)$, that implies

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

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

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

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

As a consequence, $\sum_{\ell \in \Omega_2^*} \sum_{s \in \Omega_2^\ell} \frac{d_\ell^{*2}}{(1+C_s)^2} \frac{\xi_2^{(\ell)}}{\#\Omega_2^\ell} \frac{n^{2\tau}}{r^*} = o(1)$.

Let us consider the sixth term of formula (3):

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

We have

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

Since Ω_3^* is bounded, $\sum_{\ell \in \Omega_3^*} d_\ell^{*2} = o(\lambda)$. Then, according to (C4^{*}), we have

$$\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_3^*} d_\ell^{*2} = o(1). \text{ It implies } \sum_{\ell \in \Omega_3^*} \sum_{s \in \Omega_2^\ell} \frac{d_\ell^{*2}}{(1+C_s)^2} \frac{\xi_2^{(\ell)}}{\#\Omega_2^\ell} \frac{n^{2\tau}}{r^*} = o(1).$$

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

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

We have,

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

According to (C3) and (C8^{*}), 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):

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

We have

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

Using (C4*), (C3) and the fact that $\#\Omega_2^*$ 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):

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

We have:

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

Besides,

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

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

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

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

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

Conditions (C5), (C6), and (C1*) and the fact that $\xi(n)$ is bounded away from zero, ensure that

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

On the other hand, recall that

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

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

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

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

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

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

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

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

have $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} = o(1)$. Then, we have $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \xi_1^{(\ell)} = o(1)$.

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

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

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

$$\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_3^*} d_\ell^{*2} = o(1).$$

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

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

According to (C7*), the right term is equal to $o(1)$.

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

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

We have :

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

Since $\#\Omega_2^* = O(1)$ and $\lambda \frac{n^{2\tau}}{r^*} = o(1)$, we have $\sum_{\ell \in \Omega_2^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} = o(1)$. As a conse-

quence, $\sum_{\ell \in \Omega_2^*} \sum_{s \in \Omega_2^\ell} \frac{d_\ell^{*2}}{1+C_s} \frac{\xi_2^{(\ell)}}{\#\Omega_2^\ell} \frac{n^{2\tau}}{r^*} = o(1)$.

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

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

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

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

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

We have

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

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

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

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

Besides,

$$\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_2^*} \frac{1}{C_\ell^*} \frac{\xi_3^{(\ell)}}{\#\Omega_3^\ell} \sum_{s \in \Omega_3^\ell} d_s^2 \leq \frac{n^{2\tau}}{r^*} \left(\sum_{\ell \in \Omega_2^*} \frac{1}{C_\ell^*} \frac{\xi_3^{(\ell)}}{\#\Omega_3^\ell} \right) \left(\sum_{s \in \Omega_3} d_s^2 \right).$$

Using (C4*), (C2), and the fact that $\#\Omega_2^* = O(1)$, the right term equals $o(1)$.

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

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

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

To conclude, we obtain:

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

Last,

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

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

$$n\hat{A}_4 \sim \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*}. \quad (7)$$

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

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

Since $p \rightarrow +\infty$ when $n \rightarrow +\infty$, the distance between markers and QTLs tends to zero. As a consequence, QTLs locations will match a few marker locations (i.e. perfect LD), and each column of X^* will be included in X . Then, we have $\mathcal{R}_{\text{col}}(X^*) \subset \mathcal{R}_{\text{col}}(X)$. As a consequence, $\forall \ell \in \Omega_1^* \cup \Omega_2^* \cup \Omega_3^*$, we have $PP'P^{*\ell} = P^{*\ell}$ and since $\|P^{*(\ell)}\|^2 = 1$, we have the relationship $\xi_1^{(\ell)} + \xi_2^{(\ell)} + \xi_3^{(\ell)} = 1$.

Let us recall condition (C7*): $\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} \xi_2^{(\ell)} d_\ell^{*2} = o(1)$. We have $\sum_{\ell \in \Omega_1^*} \xi_2^{(\ell)} d_\ell^{*2} \leq (\#\Omega_1^*) d_1^{*2} \max_{\ell \in \Omega_1^*} \xi_2^{(\ell)}$ and by definition, $d_1^{*2} \sim n^\psi$ with $0 < \psi \leq 1$. In this context, let us set $\forall \ell \in \Omega_1^* \xi_2^{(\ell)} = 1/n^{\theta_1}$ with $\theta_1 > \psi$. Since $d_1^{*2} \max_{\ell \in \Omega_1^*} \xi_2^{(\ell)} \sim n^{\psi-\theta_1}$ and $\#\Omega_1^* = O(1)$, it is clear that condition (C7*) is fulfilled.

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

6. Some extreme cases

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

6.1. The projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$

Lemma 6.1. *Let us consider same assumptions as in Theorem 2. Besides, let us suppose that the projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$ that is to say*

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

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

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

Proof. The proof is divided in three parts, called a), b) and c).

a) The projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$, and is tagged by one $s \in \Omega_1$

Let us suppose that the projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$, that is to say

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

Let us consider that $\ell = 1$ is tagged by only one “ s ” that belongs to Ω_1 , i.e. $\|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \sim \xi(n)$ only for that s , with $0 < b < \xi(n) \leq 1$.

Using Theorem 2, we have:

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

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

$$\frac{d_s^2 d_1^*}{d_s^2 + \lambda} \|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \|Q^{*(1)}Q^{*(1)'}\beta^*\| \sim \xi(n) n^{\tau + \psi/2},$$

$$\frac{d_s^4 d_1^{*2}}{(d_s^2 + \lambda)^2} \|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \|Q^{*(1)}Q^{*(1)'}\beta^*\|^2 \sim \xi(n) n^{2\tau + \psi}.$$

If $2\tau + \psi > 1$, then $n = o(n^{2\tau + \psi})$. As a consequence, since $\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2} \leq r \leq n$ and $0 < b < \xi(n)$, we have $\hat{\rho}_g \sim \sqrt{\xi(n)}$.

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

b) The projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$, and is tagged by one $s \in \Omega_2$

Recall that

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

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

$$\frac{d_s^2 d_1^*}{d_s^2 + \lambda} \|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \|Q^{*(1)}Q^{*(1)'}\beta^*\| \sim \frac{\xi(n) n^{\tau + \psi/2}}{1 + C_s}.$$

Besides,

$$\frac{d_s^4 d_1^{*2}}{(d_s^2 + \lambda)^2} \|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \|Q^{*(1)}Q^{*(1)'}\beta^*\|^2 \sim \frac{\xi(n) n^{2\tau + \psi}}{(1 + C_s)^2}.$$

If $2\tau + \psi > 1$, then $n = o(n^{2\tau + \psi})$. As a consequence, since $\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2} \leq r \leq n$ and $0 < b < \xi(n)$, we have $\hat{\rho}_g \sim \sqrt{\xi(n)}$.

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

c) *The projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$, and is tagged by one $s \in \Omega_3$*

Recall that

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

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

$$\frac{d_s^2 d_1^*}{d_s^2 + \lambda} \|P^{(s)} P^{(s)'} P^{*(1)}\|^2 \|Q^{*(1)} Q^{*(1)'} \beta^*\| \sim \frac{\xi(n)}{C} n^{2\gamma + \tau + \psi/2 - \kappa - \eta}.$$

At the denominator in formula (10), we have:

$$\frac{d_s^4 d_1^{*2}}{(d_s^2 + \lambda)^2} \|P^{(s)} P^{(s)'} P^{*(1)}\|^2 \|Q^{*(1)} Q^{*(1)'} \beta^*\|^2 \sim \frac{\xi(n)}{C^2} n^{4\gamma - 2\kappa - 2\eta + 2\tau + \Psi}.$$

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 $\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2} \leq r \leq n$ and $0 < b < \xi(n)$, we have $\hat{\rho}_g \sim \sqrt{\xi(n)}$. When $4\gamma - 2\kappa - 2\eta + 2\tau + \psi < 1$, we need to impose $\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2} = o(n^{4\gamma - 2\kappa - 2\eta + 2\tau + \psi})$ in order to obtain $\hat{\rho}_g \sim \sqrt{\xi(n)}$. This concludes the proof.

6.2. The projected signal belongs only to $\text{Span}\{Q^{*(r^*)}\}$

Lemma 6.2. *Let us consider same assumptions as in Theorem 2 of the main manuscript. Besides, let us suppose that the projected signal belongs only to $\text{Span}\{Q^{*(r^*)}\}$, that is to say*

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

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

- If $s \in \Omega_1 \cup \Omega_2$:
 - if $2\tau + \eta > 1$, then $\hat{\rho}_g \sim \sqrt{\xi(n)} \rho_g^{oracle}$.
 - if $2\tau + \eta < 1$, then
 - * if $\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2} = o(n^{2\tau + \eta})$, then $\hat{\rho}_g \sim \sqrt{\xi(n)} \rho_g^{oracle}$
 - * if $n^{2\tau + \eta} = o\left(\sum_{u=1}^r \frac{d_u^4}{(d_u^2 + \lambda)^2}\right)$, then $\hat{\rho}_g \rightarrow 0$.

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

The proof is largely inspired from the one of Lemma 6.1 above, as soon as we replace ψ by η .

7. Explicit formula for the accuracy $\tilde{\rho}_g$ of the improved predictor

Lemma 7.1. *Let us consider same hypotheses as in Theorem 1 of the main manuscript. Then, the quantity $\tilde{\rho}_g$ defined in Section 4 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

$$\begin{aligned} \tilde{A}_1 &= \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} \beta^{*\prime} \mathbb{E}(x_{new}^* x'_{new}) Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^*, \\ \tilde{A}_2 &= \sigma_e^2 \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{(d_{\sigma(s)}^2 + \lambda)^2} \mathbb{E} \left(\left\| Q^{(\sigma(s))} Q^{(\sigma(s))\prime} x_{new} \right\|^2 \right), \\ \tilde{A}_3 &= \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^* \right)' \mathbb{E}(x_{new} x'_{new}) \\ &\quad \times \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^* \right), \\ \tilde{A}_4 &= A_4. \end{aligned}$$

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

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

As a result, let us define

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

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

$$\tilde{A}_1 = \beta^{*\prime} \mathbb{E}(x_{new}^* x'_{new}) Q\bar{D}P' \tilde{P} \tilde{P}' X^* \beta^*.$$

After some simple algebra, we obtain

$$Q\bar{D}P' \tilde{P} = \left(\frac{d_{\sigma(1)}}{d_{\sigma(1)}^2 + \lambda} Q^{(\sigma(1))}, \dots, \frac{d_{\sigma(\tilde{r})}}{d_{\sigma(\tilde{r})}^2 + \lambda} Q^{(\sigma(\tilde{r}))} \right). \quad (11)$$

Then,

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

Let us now consider \tilde{A}_2 . According to Rabier et al. [2], we have

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

Furthermore, recall that

$$\tilde{A}_3 = \beta^{*\prime} X^{*\prime} \tilde{P} \tilde{P}' V^{-1} X \text{Var}(x_{new}) X' V^{-1} \tilde{P} \tilde{P}' X^* \beta^*.$$

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

$$\begin{aligned} \tilde{A}_3 &= \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^* \right)' \mathbb{E}(x_{new} x'_{new}) \\ &\quad \times \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^* \right). \end{aligned}$$

8. Proof of Lemma 3 of the main manuscript

To begin with, let us recall the expression \tilde{A}_1 given in Lemma 7.1 above:

$$\tilde{A}_1 = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} \beta^{*\prime} \mathbb{E}(x_{new}^* x'_{new}) Q^{(\sigma(s))} P^{(\sigma(s))\prime} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)\prime} \beta^*. \quad (12)$$

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

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

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

$$X^{*'}XQ^{(\sigma(s))} = d_{\sigma(s)} \sum_{\ell=1}^{r^*} d_{\ell}^* Q^{*(\ell)} P^{*(\ell)'} P^{(\sigma(s))}.$$

Then,

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

According to [2],

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

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

$$\begin{aligned} \hat{A}_3 &= \frac{1}{n} \left(\sum_{s=1}^{\tilde{r}} X \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))'} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)'} \beta^* \right)' \\ &\quad \times \left(\sum_{s=1}^{\tilde{r}} X \frac{d_{\sigma(s)}}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} P^{(\sigma(s))'} \sum_{j=1}^{r^*} d_j^* P^{*(j)} Q^{*(j)'} \beta^* \right). \end{aligned}$$

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

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

9. Some extreme cases using the improved predictor

Let us now introduce a new result dealing with an extreme case:

Lemma 9.1. *Let us consider same assumptions as in Theorem 2 of the main manuscript. Besides, let us suppose that the projected signal belongs only to $\text{Span}\{Q^{*(1)}\}$, that is to say*

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

Moreover, let us assume that $\ell = 1$ is tagged only by one $s \in \{\sigma(1), \dots, \sigma(\tilde{r})\}$ such as $\|P^{(s)}P^{(s)'}P^{*(1)}\|^2 \sim \xi(n)$ with $0 < b < \xi(n) \leq 1$, and $\|P^{(u)}P^{(u)'}P^{*(1)}\|^2 = 0 \forall u \neq s$. Then

1. If $s \in \Omega_1 \cup \Omega_2$, $2\tau + \psi < 1$ and the following two conditions hold

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

we have $\hat{\rho}_g \sim \sqrt{\xi(n)}\rho_g^{oracle}$, whereas $\hat{\rho}_g \rightarrow 0$.

2. If $s \in \Omega_3$, $\lambda \sim Cn^{\kappa + \eta}$, $d_s \sim n^\gamma$ with $C > 0$, $\kappa > \max(0, -\eta)$, $\gamma < (\kappa + \eta)/2$, $4\gamma - 2\kappa - 2\eta + 2\tau + \psi < 1$, and the following two conditions hold

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

we have $\hat{\rho}_g \sim \sqrt{\xi(n)}\rho_g^{oracle}$, whereas $\hat{\rho}_g \rightarrow 0$.

The proof is largely inspired from the proof of Lemma 6.1 of this Supplementary Material. According to this lemma, there are some cases where, at the same time, the new accuracy $\hat{\rho}_g$ is not negligible (asymptotically equivalent to $\sqrt{\xi(n)}\rho_g^{oracle}$) and the classical accuracy $\hat{\rho}_g$ is null.

Note that the analogue of this lemma, for a projected signal belonging only to $\text{Span}\{Q^{*(r^*)}\}$ can be easily deduced.

10. Some results regarding the L^2 prediction loss

We first prove the Remark 2 of the main manuscript in which we give an expression for the L^2 prediction loss.

10.1. Proof of Remark 2 of the main manuscript

We have

$$\begin{aligned}
& \mathbb{E} \left\{ (x'_{new}\hat{\beta} - x'^*_{new}\beta^*)^2 \mid x_{new}, x^*_{new} \right\} \\
&= \mathbb{E} \left\{ (x'_{new}X'V^{-1}Y - x'_{new}X'V^{-1}X^*\beta^* + x'_{new}X'V^{-1}X^*\beta^* - x'^*_{new}\beta^*)^2 \mid x_{new}, x^*_{new} \right\} \\
&= \mathbb{E} \left[\{x'_{new}X'V^{-1}(Y - X^*\beta^*)\}^2 \mid x_{new} \right] \\
&\quad + \mathbb{E} \left[\{x'_{new}X'V^{-1}X^*\beta^* - x'^*_{new}\beta^*\}^2 \mid x_{new}, x^*_{new} \right] \\
&\quad + 2(x'_{new}X'V^{-1}X^*\beta^* - x'^*_{new}\beta^*) \mathbb{E} [x'_{new}X'V^{-1}(Y - X^*\beta^*) \mid x_{new}] \\
&= \sigma_e^2 \|x'_{new}X'V^{-1}\|^2 + \beta^{*\prime}X^*V^{-1}Xx_{new}x'_{new}X'V^{-1}X^*\beta^* \\
&\quad + \beta^{*\prime}x^*_{new}x^*_{new}\beta^* - 2\beta^{*\prime}x^*_{new}x'_{new}X'V^{-1}X^*\beta^*.
\end{aligned}$$

As a result,

$$\begin{aligned}
& \mathbb{E} \left\{ (x'_{new} \hat{\beta} - x'^*_{new} \beta^*)^2 \right\} \\
&= \sigma_e^2 \mathbb{E} \left\{ \|x'_{new} X' V^{-1}\|^2 \right\} + \beta^{*'} X^{*'} V^{-1} X \text{Var}(x_{new}) X' V^{-1} X^* \beta^* + \sigma_G^2 \\
&\quad - 2\beta^{*'} \mathbb{E}(x^*_{new} x'_{new}) X' V^{-1} X^* \beta^* \\
&= A_2 + A_3 + A_4 - 2A_1 .
\end{aligned}$$

This gives the expression of the L^2 prediction loss.

10.2. Estimation of the L^2 prediction loss, when TRN and TST samples come from the same probability distribution

A natural estimation is the following

$$\hat{\mathbb{E}} \left\{ (x'_{new} \hat{\beta} - x'^*_{new} \beta^*)^2 \right\} = \hat{A}_2 + \hat{A}_3 + \hat{A}_4 - 2\hat{A}_1 .$$

According to formulae (4), (6) and (7),

$$\begin{aligned}
n\hat{A}_4 &\sim \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \\
n\hat{A}_2 + n\hat{A}_3 &\sim \xi(n) \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} \\
n\hat{A}_1 &\sim \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \xi(n) \frac{n^{2\tau}}{r^*} .
\end{aligned}$$

As a result, we have:

$$\hat{A}_2 + \hat{A}_3 + \hat{A}_4 - 2\hat{A}_1 \sim \frac{1 - \xi(n)}{n} \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} .$$

By definition, the loss coefficient $1 - \xi(n)$ is bounded by 0 and 1. In order to ensure that the quantity $\hat{\mathbb{E}} \left\{ (x'_{new} \hat{\beta} - x'^*_{new} \beta^*)^2 \right\}$ tends to 0, it suffices to have

$$\sum_{\ell \in \Omega_1^*} d_\ell^{*2} \frac{n^{2\tau}}{r^*} = o(n) .$$

Indeed, recall that under condition $(C1^*)$, we have $\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} d_\ell^{*2} \rightarrow +\infty$.

As a result, it is sufficient that $\frac{n^{2\tau}}{r^*} \sum_{\ell \in \Omega_1^*} d_\ell^{*2}$ diverges to $+\infty$ at a rate slower than n .

10.3. How to improve the quality of the prediction

The L^2 prediction loss, associated to the new estimator $\tilde{\beta}$ is

$$\mathbb{E} \left\{ (x'_{new} \tilde{\beta} - x'^*_{new} \beta^*)^2 \right\} = \tilde{A}_2 + \tilde{A}_3 + \tilde{A}_4 - 2\tilde{A}_1 .$$

Assuming that TRN and TST samples come from the same probability distribution, an estimation of this quantity is the following

$$\hat{\mathbb{E}} \left\{ (x'_{new} \tilde{\beta} - x'^*_{new} \beta^*)^2 \right\} = \hat{A}_2 + \hat{A}_3 + \hat{A}_4 - 2\hat{A}_1$$

where \hat{A}_2 , \hat{A}_3 , \hat{A}_4 and \hat{A}_1 are given in Lemma 3 of the main manuscript.

Remark 10.1. Note that the prediction is improved if

$$\hat{\mathbb{E}} \left\{ (x'_{new} \tilde{\beta} - x'^*_{new} \beta^*)^2 \right\} < \hat{\mathbb{E}} \left\{ (x'_{new} \hat{\beta} - x'^*_{new} \beta^*)^2 \right\} ,$$

i.e.

$$\hat{A}_2 - \hat{A}_2 + \hat{A}_3 - \hat{A}_3 + 2(\hat{A}_1 - \hat{A}_1) > 0 .$$

According to the main text (below Lemma 5),

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

As a result, this condition can be rewritten

$$\widehat{Var}(\vec{Y}_{new}) > 2\widehat{Cov}(\vec{Y}_{new}, Y_{new}) .$$

References

- [1] RABIER, C.E., BARRE, P., ASP, T., CHARMET, G., and MANGIN, B. (2016). On the Accuracy of Genomic Selection. *PloS One*, **11(6)** e0156086. doi:10.1371/ journal.pone.0156086.
- [2] RABIER, C. E., MANGIN, B., and GRUSEA, S. (2018). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics*, **46(1)** 289-313.

TABLE 1

Same as Table 1 of the main manuscript except that 100 QTLs are now considered on $[0, T]$. Recall that the TRN map contains 1000 markers whereas the TST map contains only 500 markers.

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

TABLE 2

Same as Table 1 except that more markers are considered. The TRN map contains 2000 markers whereas the TST map contains only 1000 markers.

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

TABLE 3
Same as Table 7 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
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4535 (0.0403)	0.4489 (0.0422)	0.4438 (0.0379)	0.4379 (0.0394)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4823 (0.0273)	0.4778 (0.0258)	0.4722 (0.0235)	0.4594 (0.0241)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5072 (0.0143)	0.4267 (0.0205)	0.3497 (0.0322)	0.2822 (0.0814)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5526 (0.0121)	0.5081 (0.0156)	0.4205 (0.0203)	0.3587 (0.0625)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1702 (0.3974)	0.1620 (0.3680)	0.1700 (0.3374)	0.1732 (0.3496)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.3977 (0.1772)	0.3970 (0.1462)	0.4141 (0.1197)	0.4109 (0.1279)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5730	0.5091	0.5142	0.5242
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4909 (0.0511)	0.4456 (0.0391)	0.4497 (0.0369)	0.4520 (0.0429)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4579 (0.0288)	0.4686 (0.0244)	0.4825 (0.0222)	0.4805 (0.0259)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.4816 (0.0209)	0.4134 (0.0227)	0.4830 (0.0099)	0.4293 (0.0233)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5314 (0.0166)	0.4977 (0.0164)	0.5714 (0.0149)	0.4922 (0.0179)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1650 (0.4110)	0.1634 (0.3525)	0.1652 (0.3561)	0.1660 (0.3588)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.3951 (0.1880)	0.3955 (0.1354)	0.4022 (0.1347)	0.4021 (0.1457)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.5590	0.5156		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4496 (0.0483)	0.4409 (0.0407)		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4831 (0.0276)	0.4723 (0.0245)		
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.4936 (0.0176)	0.3664 (0.0357)		
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5403 (0.01615)	0.4461 (0.0227)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1664 (0.3899)	0.1675 (0.3544)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.3921 (0.1723)	0.3999 (0.1379)		

TABLE 4
Same as Table 7 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.	0.5288	0.5639	0.4662	0.4851
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4417 (0.0449)	0.4494 (0.0478)	0.4351 (0.0364)	0.4456 (0.0377)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4692 (0.0281)	0.4813 (0.0288)	0.4587 (0.0241)	0.4684 (0.0237)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.4387 (0.0213)	0.5304 (0.0111)	0.2552 (0.0758)	0.3152 (0.0415)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5372 (0.0151)	0.6094 (0.1449)	0.3328 (0.0607)	0.4079 (0.0210)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1419 (0.3964)	0.1384 (0.4270)	0.1406 (0.2416)	0.1412 (0.3531)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.4152 (0.1380)	0.4078 (0.1676)	0.4253 (0.0950)	0.4266 (0.1026)
Dataset ID	Set 5	Set 6	Set 7	Set 8
Emp. Acc.	0.5581	0.5096	0.5349	0.5717
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4526 (0.04282)	0.4411 (0.0403)	0.4481 (0.0449)	0.4521 (0.0499)
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4806 (0.0293)	0.4648 (0.0253)	0.4762 (0.0263)	0.4856 (0.0288)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.4818 (0.0191)	0.4002 (0.0249)	0.4237 (0.0269)	0.5277 (0.0148)
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.5469 (0.0145)	0.4784 (0.0167)	0.4832 (0.0206)	0.6113 (0.0148)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1392 (0.4270)	0.1414 (0.3803)	0.1426 (0.3989)	0.1393 (0.4319)
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.4150 (0.1616)	0.4185 (0.1247)	0.4208 (0.1369)	0.4093 (0.1707)
Dataset ID	Set 9	Set 10		
Emp. Acc.	0.4969	0.5266		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{ADLASSO}^*)$	0.4421 (0.0389)	0.4533 (0.0410)		
$\hat{\rho}_{ph}(\hat{X}^*, \hat{\beta}_{LASSO}^*)$	0.4637 (0.0242)	0.4798 (0.0244)		
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.3419 (0.0354)	0.4439 (0.0201)		
$\hat{\rho}_{ph}^{LD}(\hat{\beta}_{ADLASSO})$	0.4312 (0.0354)	0.5138 (0.0148)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \text{MeLJ})$	0.1433 (0.3633)	0.1414 (0.3952)		
$\hat{\rho}_{ph}^{Lian}(\hat{h}^2, \hat{r}^2, \hat{F}_{RM})$	0.4281 (0.1106)	0.4205 (0.1341)		