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

Charles-Elie Rabier, Brigitte Mangin, S Grusea

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

C.E. Rabier[a,b,c,d], B. Mangin[e], S. Grusea[a]

[a] *INSA de Toulouse, Institut de Mathématiques de Toulouse, Université de Toulouse, France*
[b] *MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France*
[c] *ISEM, Université de Montpellier, CNRS, France*
[d] *LIRMM, Université de Montpellier, CNRS, France*
[e] *LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France*

Abstract. Genomic selection, a hot topic in genetics, consists in predicting breeding values of selection candidates, using a large number of genetic markers, due to the recent progress in molecular biology. One of the most popular method chosen by geneticists is Ridge regression. In this context, we focus on some predictive aspects of Ridge regression and present theoretical results regarding the accuracy criteria, i.e., the correlation between predicted value and true value. We show the influence of the singular values, the regularization parameter, and the projection of the signal on the space spanned by the rows of the design matrix. Asymptotic results, in a high dimensional framework, are also given, and we prove that the convergence to an optimal accuracy highly depends on a weighted projection of the signal on each subspace. We discuss also on how to improve the prediction. Last, illustrations on simulated and real data are proposed.

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

1. Introduction and background

This year 2016, professor Michael Goddard and professor Theodorus Meuwissen are awarded The John J. Carty Award for the Advancement of Science by the National Academy of Science. They are considered as pioneers in the development of genomic selection (GS), because of their stimulating paper Meuwissen et al. (2001). In this context, our manuscript is devoted to methodological aspect of GS, a hot topic in genomics.

1.1. Preliminaries

For many years, geneticists focused on linkage analysis (LA), in order to detect Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome.

The most popular statistical method was Interval Mapping (Lander and Botstein (1989)), which consisted in scanning the genome, with the help of genetic markers, and in testing for the presence/absence of a QTL at every location in the genome.

The mathematical properties of Interval Mapping have been studied in detail by different research teams from all over the world (Cierco (1998); Chen and Chen (2005); Chang et al. (2009); Kim et al. (2009); Azaïs et al. (2014)). According to Wu et al. (2007), thousands of QTLs have been detected in plants, animals, and humans using the concept of Interval Mapping.

More recently, geneticists adopted genome-wide association studies (GWAS). In contrast to LA, GWAS are based on unrelated individuals and as a result, larger sample sizes can be considered. GWAS enabled the discovery of many SNP-trait associations in humans (e.g. Crohn's disease Barrett et al. (2008), human height Weedon et al. (2008)).

However, both approaches (LA and GWAS) suffered from the fact that they were unable to detect QTLs with very small effects. Recall that most traits of interest are called complex traits, since they are governed by a large number of small-effect QTLs (Goddard and Hayes (2008); Buckler et al. (2009)). It turns out that predictions based on selected SNPs could not be considered as reliable. This inability to capture all the genetic variation is known under the name of missing heritability.

Today, Genomic Selection, motivated by the seminal paper of Meuwissen et al. (2001), is an extremely popular technique in genetics. It consists in predicting breeding values of selection candidates, using a large number of genetic markers, due to the recent progress in molecular biology. The goal is not to detect QTL anymore, but to predict the future phenotype of young candidates as soon as their DNA has been collected. GS relies on the expectation that each QTL will be highly correlated with at least one marker (Schulz-Streeck et al. (2012)). In genetics, this large correlation is named strong Linkage Disequilibrium (LD). More precisely, LD refers to the non independence of alleles at 2 different loci (see Durett (2008) for more details), whereas Linkage Equilibrium (LE) denotes the independence of alleles at 2 different loci.

In contrast to LA and GWAS, where each marker is analyzed separately, GS considers all markers simultaneously (Whole genome regression analysis). GS was first applied to animal breeding (see Hayes et al. (2009)) and later to plant breeding (Jannink et al. (2010)): it was recently investigated on apple (Kumar et al. (2012)), sugar beet (Wurschum et al. (2013)), pea (Burstin et al. (2015)), and on inbred lines of rice (Spindel et al. (2015)).

1.2. A linear model

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

A fixed number of QTLs lie on the genome, having an effect on the quantitative trait. For $1 \leq j \leq p$, $\beta_j = 0$ means that the corresponding marker is not a QTL, whereas $\beta_j \neq 0$ refers to a QTL. In what follows, $\|\beta\|_0^0 := \sum_{j=1}^{p} |\beta|^0$ (with $0^0 = 0$) will denote the number of QTLs (i.e. non null marker effects).

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

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

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

In this manuscript, we will propose an analysis conditional on observed x_1, ..., x_n. However, before imposing this conditioning, we have to precise that the matrix X is independent of ε. Simulated data will be generated accordingly. In what follows, r will denote the rank of the matrix X, and $\mathcal{R}(X)$ will refer to the linear space generated by the rows of X.

1.3. Introducing a test individual

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

$$Y_{new} = x'_{new}\,\beta + \varepsilon_{new},$$

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

We suppose that x'_{new}, ε_{new} and ε are all independent.

1.4. Introducing the accuracy

In GS, we are interested in predicting either the genotypic value $x'_{new}\,\beta$, or the phenotypic value Y_{new}. In both cases, an estimator $\hat{Y}_{new}$ is constructed from a prediction model learned on n TRN individuals. $\hat{Y}_{new}$ is a function of the random variables x_{new} and ε. Then, the quality of the prediction is evaluated according to some accuracy criteria, i.e. the correlation between predicted and true values. This criteria, rarely studied in the statistical literature, is a key element in genetics: it plays a role in the rate of genetic gain. Indeed, the accuracy is one component present in the breeders equation (see for instance Lynch and Walsh (1998)). The *phenotypic accuracy*, ρ_{ph}, also called predictive ability, is defined in the following way

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

whereas the *genotypic accuracy*, ρ_g, is defined as

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

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

$$h^2 := \frac{\mathrm{Var}\left(x'_{new}\ \beta\right)}{\mathrm{Var}\left(Y_{new}\right)} = \frac{\mathrm{Var}\left(x'_{new}\ \beta\right)}{\mathrm{Var}\left(x'_{new}\ \beta\right) + \mathrm{Var}\left(\varepsilon_{new}\right)} = \frac{\beta'\mathrm{Var}\left(x_{new}\right)\beta}{\beta'\mathrm{Var}\left(x_{new}\right)\beta + \mathrm{Var}\left(\varepsilon_{new}\right)}\ . \tag{4}$$

In what follows, we set $\sigma^2_G = \mathrm{Var}\left(x'_{new}\ \beta\right)$, and as a consequence, we have the relationship $h^2 = \sigma^2_G/\left(\sigma^2_G + \sigma^2_e\right)$.

Depending on the authors, one focuses either on the phenotypic accuracy (e.g. Visscher et al. (2010)), or on the genotypic accuracy (e.g. Daetwyler et al. (2008, 2010)).

Penalized regression methods (see Li and Sillanpää (2012) for a review in GS), Bayesian methods (e.g. Kärkkäinen and Sillanpää (2012)), and reproducing kernel Hilbert spaces methods (de los Campos et al. (2010)) are essentially the methods used by geneticists to make prediction in GS. More recently, models combining Bayesian and deterministic approaches (i.e. including both random and fixed marker effects) are also investigated (e.g. Spindel et al. (2016)).

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

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

1.5. Some background on Ridge regression

In the present study, we propose to focus on Ridge regression. Indeed, it is one of the most popular methods for prediction of breeding values. Ridge regression (Tihonov (1963); Hoerl et al. (1970)) has been studied for many years. In genetics, this regression model, initially proposed by Meuwissen et al. (2001), is called random regression best linear unbiased predictor (RRBLUP) or genomic best linear unbiased predictor (GBLUP).

The Ridge estimator, suitable in a high dimensional setting (i.e. $p > n$), is the following:

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

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

Although Ridge regression is approximately 60 years old, statisticians keep studying this topic, and excellent papers have been published recently (e.g. Shao and Deng (2012); Bühlmann (2013); Dicker (2016)). Bühlmann (2013) focused on statistical inference in high dimension: he proposed to correct the bias of the Ridge estimator due to projection bias. In Dicker (2016), the author presented theoretical results when (Y, X) are jointly Gaussian and the columns of X are independent and identically distributed. However, in genomics, this hypothesis is too strong since the large number of markers (in the genome) can not be considered independent due to linkage and a fixed genome size.

Shao and Deng (2012) proposed a study where the design matrix X was treated as fixed. They focused on the estimation of θ, defined as the projection of β onto $\mathcal{R}(X)$ (i.e. linear space generated by the rows of X). Indeed, according to their Lemma 1, β is identifiable in model (1) if and only if $\beta \in \mathcal{R}(X)$, which is nonrealistic in practice. Recall that β is a vector of size p and that we have $\dim(\mathcal{R}(X)) \leq n$ when $p > n$.

Since the Ridge estimator $\hat{\beta}$ belongs to $\mathcal{R}(X)$ (cf. our Section 2.2), they studied convergence rates for the mean squared error $\mathbb{E}\left\{\left(\ell'\hat{\beta} - \ell'\theta\right)\right\}^2$, for any vector ℓ such as $\|\ell\| = 1$. They also obtained rates regarding the expected L^2 norm error, $\mathbb{E}\left(\left\|X\hat{\beta} - X\beta\right\|^2\right)$, for the Ridge regression estimator $X\hat{\beta}$ of $X\beta$ (cf. our Section 2.2).

1.6. Our contributions and roadmap

Our study starts, in Section 2, by recalling recent results on the accuracy. After a quick reminder on the singular value decomposition, we introduce our main theorem, Theorem 1, that presents a general formula for the genotypic accuracy, ρ_g. This is a key formula for the rest of the manuscript, since other theorems and lemmas are built on it. According to Theorem 1, ρ_g depends on the projection of the signal β on $\mathcal{R}(X)$. This projection can be named "weighted projection" since some weights depending on singular values and on the tuning parameter λ, act as multiplying factors.

Section 3 focuses on the case where TRN and TST samples come from the same probability distribution. In this context, Theorem 2 gives an estimation $\hat{\rho}_g$ of ρ_g and Lemma 1 introduces a lower bound for $\hat{\rho}_g$: as in ii) of Theorem 1 of Shao and Deng (2012), it only takes into account a global projection of the signal on $\mathcal{R}(X)$, with a global weight (i.e. same weights on each subspace).

Lemmas 2 and 3 propose a sharper analysis. In particular, Lemma 2 deals with the case where the projected signal is spread out uniformly on each vector of an orthonormal basis of $\mathcal{R}(X)$. It shows that under six given conditions, the estimation $\hat{\rho}_g$ tends to the oracle genotypic accuracy. These conditions are basically imposed on the singular values and on the ratio rank r of X to projected signal on $\mathcal{R}(X)$.

Lemma 3 investigates extreme cases: the projected signal belongs either to the subspace spanned by the vector of an orthonormal basis of $\mathcal{R}(X)$ associated to the largest singular value of X, or to the subspace spanned by the vector

associated to the smallest singular value. This setting is particularly interesting since Ridge regression imposes shrinkage, without taking into account of the signal.

In Section 4, we tackle the problem of TRN and TST samples not coming from the same probability distribution. Theorem 3 introduces an estimator of ρ_g, which relies on the scalar product between a random projection of the signal and the usual projection of the signal on $\mathcal{R}(X)$. Lemma 4 is the analogue of Lemma 1 under this new configuration.

Section 5 is devoted to Daetwyler et al. (2008)'s seminal formula for the accuracy in GS. Although the link between Ridge regression and Daetwyler et al. (2008) has already been addressed in one of our recent study (Rabier et al. (2016)), new results are given in Lemma 5. These results are potentially of interest for geneticists, since we give new substitutes for the effective number of independent loci M_e, a key quantity in the field (cf. Section 5 for more details).

Last, in Section 6, we propose another estimator for the genotypic accuracy; it is still derived from Ridge regression, and it may present better performances (cf. Theorem 6). We propose to project the vector Y on a well chosen subspace of the space spanned by the columns of X. We will give necessary conditions in Lemma 8, to observe an increase in terms of accuracy.

Our paper ends with an illustration on simulated data, mimicking the evolution of a population over time. We will show the impact of different probability distributions (between TRN and TST) on the quality of the estimated accuracy. Furthermore, we will highlight the fact that proxies built on our theoretical results outperform existing proxies in GS. Performances of the "modified" Ridge estimator will also be illustrated. Finally, a real data analysis is proposed; it relies on the recent paper of Spindel et al. (2015) dealing with GS in rice.

2. General expression for the accuracy

2.1. Introducing Ridge regression and the corresponding accuracy

Recall the expression of the Ridge estimator:

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

Since we have the well-known relationship

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

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

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

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

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

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

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

where $\|.\|$ is the L^2 norm, and $\operatorname{Var}(x_{new})$ is the covariance matrix of size $p \times p$. Note that this accuracy can be viewed as a conditional accuracy since this expression was obtained conditionally on the TRN design matrix X.

We introduce the following notations

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

2.2. SVD decomposition

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

$$X = PDQ', \quad (8)$$

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

$$\theta := QQ'\beta$$

and, as mentioned in Shao and Deng (2012), we have the relationship

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

Then, the Ridge estimator $\hat{\beta}$ presents the particularity to belong to $\mathcal{R}(X)$. Note also that we have the relationship $X\theta = X\beta$. As a consequence, we have $\mathbb{E}\left(\left\|X\hat{\beta} - X\beta\right\|^2\right) = \mathbb{E}\left(\left\|X\hat{\theta} - X\theta\right\|^2\right)$, and rates presented in Theorem 1 of Shao and Deng (2012) for $X\theta$ are also suitable for $X\beta$.

2.3. Results

Our main result is the following.

Theorem 1. *Let $\Sigma = Var(x_{new})$ be the covariance matrix of size $p \times p$. Furthermore, let us assume that X is known, and that ε, x_{new} and ε_{new} are random. Then, the genotypic accuracy has the following expressions*

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

where

$$A_1 = \sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} \beta' \Sigma Q^{(s)} Q^{(s)\prime} \beta \ , \ A_2 = \sigma_e^2 \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \mathbb{E}\left(\left\| Q^{(s)} Q^{(s)\prime} x_{new} \right\|^2 \right)$$

$$A_3 = \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} Q^{(s)} Q^{(s)\prime} \beta \right)' \Sigma \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} Q^{(s)} Q^{(s)\prime} \beta \right) \ , \ A_4 = \beta' \Sigma \beta.$$

The proof is given in Section 8.1. The phenotypic accuracy is obtained by replacing the term A_4 at the denominator by $A_4 + \sigma_e^2$. Note that $Q^{(s)} Q^{(s)\prime} v$ is the projection of a column vector v of size p on the vector space spanned by $Q^{(s)}$. In view of Theorem 1, ρ_g depends on the projection $Q^{(s)} Q^{(s)\prime} \beta$ of the signal and also on the projection $Q^{(s)} Q^{(s)\prime} x_{new}$ of the genome information for the individual *new*.

Remark: In the backcross design (see for instance Azaïs et al. (2014)), we have the relationship $\Sigma_{kk'} = e^{-2|t_k - t_{k'}|}$, where t_k and $t_{k'}$ refer to locations of marker k and k' measured in Morgans.

In what follows we are interested in estimating the genotypic accuracy ρ_g. A consistent estimator of A_2 is easily derived from the Law of large numbers. Besides, by Slutsky's lemma in the matrix case, consistent estimators of A_1, A_3 and A_4 can be obtained provided that a consistent estimator of the covariance matrix Σ is used. This finally leads to a consistent estimator of ρ_g.

However, finding a consistent estimator for Σ is very challenging in the high dimensional setting; it is nowadays a hot topic in statistics. Some recent results (see e.g. Cai et al. (2010)) address this question, but the authors make quite restrictive assumptions on the covariance matrix Σ.

In our present work we have chosen the empirical covariance estimator, since it is the classical estimator used by geneticists in practice. We will show on simulated data that our estimators perform in a very satisfactory manner.

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

In this section, let us consider the case where the TRN and TST samples are from the same probability distribution. In this context, using the empirical

covariance matrix $X'X/n$ as an estimation of the covariance matrix Σ from Theorem 1, we obtain the following theorem.

Theorem 2. *Let us assume that $x_1, \ldots, x_n$ and x_{new} are independent and identically distributed (i.i.d.). Besides, let us consider that $x_1, \ldots, x_n$ have been observed (i.e. X is known), and that ε, x_{new} and ε_{new} are random. Then, an estimation of the genotypic accuracy is*

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

where

$$\hat{A}_1 = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^4}{d_s^2 + \lambda} \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \quad , \quad \hat{A}_2 = \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2 + \lambda)^2}$$

$$\hat{A}_3 = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2 + \lambda)^2} \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \quad , \quad \hat{A}_4 = \frac{1}{n} \sum_{s=1}^{r} d_s^2 \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 .$$

In contrast to Theorem 1, the projection $Q^{(s)} Q^{(s)\prime} x_{new}$ is not present in this new expression. Theoretical developments rely on the following estimation $\hat{A}_2$ of A_2:

$$\hat{A}_2 := \frac{\sigma_e^2}{n} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \mathrm{Tr} \left(X Q^{(s)} Q^{(s)\prime} Q^{(s)} Q^{(s)\prime} X' \right).$$

The proof is given in Section 8.2. Note that the unknown quantity β present in Theorem 2, can be estimated for instance by LASSO (Tibshirani (1996)), Adaptative LASSO (Zou (2006)) or Group LASSO (Yuan and Lin (2006)). We refer to our applications in Section 7.

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

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

$$\frac{\left\| QQ'\beta \right\|^2 \ \min_s \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2 \ r + \left\| QQ'\beta \right\|^2 \ \max_s \frac{d_s^6}{(d_s^2+\lambda)^2}} \ \sqrt{\left\| QQ'\beta \right\|^2 \ \max_s d_s^2}} \leq \hat{\rho}_g \leq \rho_g^{oracle}.$$

According to this lemma, the smaller the ratio $\frac{r}{\|QQ'\beta\|^2}$ is, the larger the lower bound is. Furthermore, the quantity $\min_s \frac{d_s^4}{d_s^2+\lambda}$ should be large enough, and the term $\max_s \frac{d_s^6}{(d_s^2+\lambda)^2}$ not too large. The proof is given in Section 8.3.

Although this lower bound can give a first indication on the quality of the prediction, a sharper analysis is needed (see below). Indeed, until now, as in ii) of Theorem 1 of Shao and Deng (2012), we have only taken into account a global projection $QQ'\beta$ of the signal on $\mathcal{R}(X)$, with a global weight.

3.1. Convergence of $\hat{\rho}_g$ to ρ_g^{oracle} when $n \to +\infty$ and $p \to +\infty$

Recall that $d_1 \geq d_2 \geq \ldots \geq d_r > 0$ are the singular values of X. To study asymptotic properties of $\hat{\rho}_g$, we consider that

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

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

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

These conditions are largely inspired from Shao and Deng (2012). However, we are mentioning the exact order of each term since our goal, in this section, is to study the behavior of the quantity $\hat{\rho}_g$, which is a ratio. For instance, Condition C2 of Shao and Deng (2012), which imposes $\left\|QQ'\beta\right\|^2 = O(n^{2\tau})$, is somewhat more general than ours.

On the other hand, in their Theorem 3, Shao and Deng (2012) suppose $d_1^2 = O(n)$, whereas Fan and Lv (2008) assume (in condition 4) $d_1^2 = O(n^{\upsilon})$ with $\upsilon \geq 0$. This way, our condition on d_1^2 can be viewed as a compromise between the conditions considered in these two papers. Note that all the results in the present paper are still valid even if $\psi > 1$.

Last, our condition on d_r^2 is inspired from condition C1 of Shao and Deng (2012).

To begin with, we propose to study the case where the signal is spread out uniformly on each subspace $Span\left\{Q^{(s)}\right\}$, i.e.

$$\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \sim \frac{n^{2\tau}}{r}, \; s = 1, \ldots, r. \tag{9}$$

Let us consider a regularization parameter λ such as :

- $\lambda \to \infty$
- $\lambda = o\left(d_1^2\right)$

The setting $\lambda \to \infty$ when $p \to \infty$ is somewhat classical in genomics. The heritability h^2 of a quantitative character is only approximately known by geneticists, and it is well known that the Ridge regression can be viewed in a Bayesian framework assuming same variance on each regressor. As a consequence, in order to obtain an estimated value of λ, the signal (linked to h^2) is generally spread out accross all the regressors. It leads to a tuning parameter which diverges to $+\infty$ and the $\hat{\beta}_k$'s are more and more shrinked when the number of regressors increases (see for instance our section on the regularization parameter in Rabier et al. (2016)).

Let us define three sets denoted Ω_1, Ω_2 and Ω_3:

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

Note that Ω_1 contains at least the index 1. On simulated data, after having chosen λ by Restricted Maximum Likelihood (Corbeil and Searle (1976)), so-called REML (cf. Section 7), these different sets were not empty. In what follows, we will call respectively "largest singular values" the ones whose index s belong to the set Ω_1. In the same way, "intermediate singular values" and "smallest singular values" refers to the sets Ω_2 and Ω_3 respectively.

Let us introduce a few extra conditions:

- (C1) $\frac{n^{2\tau}}{r}\sum_{s\in\Omega_1} d_s^2 \to +\infty$
- (C2) $\sum_{s\in\Omega_3} d_s^2 = o(\lambda)$
- (C3) $\sum_{s\in\Omega_3} d_s^4 = o(\lambda^2)$
- (C4) $n^{2\tau}/r = o(1/\lambda)$, i.e. $\lambda = o(r/n^{2\tau})$
- (C5) $\#\Omega_1 = O(1)$
- (C6) $\#\Omega_2 = O(1)$,

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

Before presenting our Lemma 2, let us give a few comments regarding the above conditions. Under (C2), the L^2 norm squared of the vector containing the largest singular values d_s for $s \in \Omega_1$ may diverge to $+\infty$ at a rate slower than λ. According to (C3), the L^2 norm squared of the vector whose components are the square of the smallest singular values may diverge to $+\infty$ at a rate slower than λ^2. Condition (C4) assumes that the ratio $r/n^{2\tau}$ diverges faster to $+\infty$ than the tuning parameter λ. Last, (C5) and (C6) impose that the number of large singular values and the number of intermediate singular values are bounded. In other words, when $p > n$, the rank r of the matrix X which is bounded by n, will diverge to $+\infty$ if and only if the number of small singular values tends to $+\infty$.

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

$$\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \sim \frac{n^{2\tau}}{r},\ s = 1,\ldots,r \tag{10}$$

and let us assume conditions (C1-C2-C3-C4-C5-C6). Then we have $\hat{\rho}_g \longrightarrow \rho_g^{oracle}$.

If we set $r = n^\gamma$ with $0 < \gamma \leq 1$, then the condition (C4) implies that $\tau < \gamma/2$. In other words, when trying to recover the oracle accuracy, the lower the rank r is, the weaker the signal can be.

Recall that the tuning parameter λ is such as $\lambda \to \infty$, $\lambda = o\left(d_1^2\right)$. Let us now introduce the following lemma, dealing with extreme cases.

Lemma 3 (Extreme cases). *Let us consider same hypotheses as in Theorem 2.*

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

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

   - *if* $2\tau + \psi > 1$*, then* $\hat{\rho}_g \to \rho_g^{oracle}$*.*
   - *if* $2\tau + \psi < 1$
     - *if* $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} = o\left(n^{2\tau+\psi}\right)$*, then* $\hat{\rho}_g \to \rho_g^{oracle}$
     - *if* $n^{2\tau+\psi} = o\left(\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}\right)$*, then* $\hat{\rho}_g \to 0$*.*

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

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

   moreover $\lambda \sim Cn^{\eta+\kappa}$ *with* $\kappa > \max(0, -\eta)$*,* $C > 0$*, then,*
   - *if* $\tau + \eta/2 - \kappa < 0$*, then* $\hat{\rho}_g \to 0$*.*
   - *if* $\tau + \eta/2 - \kappa > 0$
     - *if* $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} = o\left(n^{2\tau+\eta-2\kappa}\right)$*, then* $\hat{\rho}_g \to \rho_g^{oracle}$
     - *if* $n^{2\tau+\eta-2\kappa} = o\left(\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2}\right)$*, then* $\hat{\rho}_g \to 0$*.*

The proof is given in the Supplementary material.

Since $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} \leq r$, we have $r = o\left(n^{2\tau+\psi}\right)$ and thus the condition $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} = o\left(n^{2\tau+\psi}\right)$ can be replaced by $r = o\left(n^{2\tau+\psi}\right)$. In the same way, condition $\sum_{s=1}^{r} \frac{d_s^4}{(d_s^2+\lambda)^2} = o\left(n^{2\tau+\eta-2\kappa}\right)$ can be replaced by $r = o\left(n^{2\tau+\eta-2\kappa}\right)$.

According to this lemma, when the projected signal belongs only to *Span* $\left\{Q^{(r)}\right\}$, κ should be not too large in order to ensure $\tau+\eta/2-\kappa > 0$, and also to fulfill the condition $r = o\left(n^{2\tau+\eta-2\kappa}\right)$. As a consequence, the tuning parameter λ should be chosen appropriately.

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

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

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

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

Using $X'_{new}X_{new}/n_{new}$ as estimator of the covariance matrix Σ, we obtain the following Theorem 3, a random version of Theorem 2.

Theorem 3. *Let us assume that X is given and that X_{new} is random, with its rows being i.i.d. Then, an estimator of the genotypic accuracy is*

$$\check{\rho}_g = \frac{\check{A}_1}{\left(\check{A}_2 + \check{A}_3\right)^{1/2} \left(\check{A}_4\right)^{1/2}}, \tag{11}$$

where

$$\check{A}_1 = \frac{1}{n_{new}} \sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} \left(\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \ < Z^{(\alpha)} Z^{(\alpha)\prime} \beta, Q^{(s)} Q^{(s)\prime} \beta > \right),$$

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

$$\check{A}_3 = \frac{1}{n_{new}} \sum_{s=1}^{r} \frac{d_s^2}{d_s^2 + \lambda} Q^{(s)\prime} \beta \sum_{\ell=1}^{r} \frac{d_\ell^2}{d_\ell^2 + \lambda} Q^{(\ell)\prime} \beta \left(\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \ < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)}, Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} > \right),$$

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

Note that the expression in Equation (11) was obtained with the help of the estimator $\check{A}_2$, defined in the following way

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

The proof is given in the Supplementary material.

Let $< ., . >$ denote the usual scalar product. We now introduce Lemma 4, the analogue of Lemma 1 in this new framework.

Lemma 4 (Bounds on $\check{\rho}_g$). *Under the same hypotheses as in Theorem 3, we always have*

$$\frac{B_1}{\left(B_2 + B_3\right)^{1/2} B_4^{1/2}} \leq \check{\rho}_g \leq \rho_g^{oracle},$$

where

$$B_1 = \min_{1 \leq s \leq r} \frac{d_s^2}{d_s^2 + \lambda} \min_{1 \leq \alpha \leq r_{new}} f_\alpha^2 \ < ZZ'\beta \ , \ QQ'\beta >,$$

$$B_2 = \sigma_e^2 \ r \ r_{new} \max_{1 \leq s \leq r} \frac{d_s^2}{(d_s^2 + \lambda)^2} \max_{1 \leq \alpha \leq r_{new}} f_\alpha^2 \max_{s,\alpha} \left\| Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2,$$

$$B_3 = \max_{1 \leq s \leq r} \frac{d_s^4}{(d_s^2 + \lambda)^2} \left\| QQ'\beta \right\|^2 \max_{1 \leq \alpha \leq r_{new}} f_\alpha^2 \ r^2,$$

$$B_4 = \left\| ZZ'\beta \right\|^2 \max_{1 \leq \alpha \leq r_{new}} f_\alpha^2.$$

Note that it is possible to replace B_2 by the quantity

$$\sigma_e^2 \, r_{new} \, \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} \left(f_\alpha^2 < Q^{(s)}, Z^{(\alpha)} >^2 \right).$$

entailing another lower bound for $\hat{\rho}_g$. The proof is given in the Supplementary material.

Contrary to the lower bound introduced in Lemma 1, this lower bound can take negative values, since the scalar product $< ZZ'\beta \, , \, QQ'\beta >$ is present in the numerator. This may happen when the rows of X are not i.i.d., or when the probability distributions of TRN and TST are very different.

In Section 7, we will illustrate performances of $\check{\rho}_{ph}$ and $\hat{\rho}_{ph}$, on simulated and real data.

5. Link with a seminal formula in GS

A large number of formulas for accuracy are now available in the literature. One of the most popular was proposed in Daetwyler et al. (2008). In their study, the authors assumed that the gene locations are known (i.e. indices of the non null coefficients of β are perfectly known). Furthermore, they focused on an orthogonal design. According to Rabier et al. (2016), a general version of Daetwyler et al. (2008) formula regarding the genotypic accuracy, is

$$\sqrt{\frac{h^2/(1-h^2)}{\frac{\|\beta\|_0^0}{n} \, + \, \frac{h^2}{1-h^2}}} \, . \tag{12}$$

Recall that $\|\beta\|_0^0 = \sum_{j=1}^{p} |\beta|^0$ with $(0^0 = 0)$. Note that in Daetwyler et al. (2008), the authors analyzed the case $\sigma_G^2 + \sigma_e^2 = 1$. In that sense, formula (12) is somewhat general since it does not rely on such assumptions.

Later, in Daetwyler et al. (2010), the authors extended their previous work, in order not to deal with a more general design (not only orthogonal). Indeed, they allowed for the presence of a large number of loci (in the genome) that can not be considered independent due to linkage and a fixed genome size. They proposed, in particular, to substitute the effective number of independent loci M_e for $\|\beta\|_0^0$, into their original formula. Subsequently, a large number of research groups built on this concept and proposed different ways of estimating M_e. Those methods are either based on the effective population size (e.g., Goddard et al. (2011)), or on the number of independent tests in association mapping (Li and Ji (2005)).

Let us come back to our present study. Our theoretical results allow us to introduce now the following lemma.

Lemma 5. *Instead of substituting the effective number of independent loci M_e for $\|\beta\|_0^0$, we should substitute the quantity nA_2 from Theorem 1 into Daetwyler et al. (2008) formula. It can be estimated by*

- $n\hat{A}_2$ *from Theorem 2, when* $x_1, \ldots, x_n$ *and* x_{new} *are i.i.d.*
- $n_{new}\check{A}_2$ *from Theorem 3, when* $x_1, \ldots, x_n$ *are not i.i.d., provided that the rows of* X_{new} *are i.i.d.*

This way, after having replaced $\|\beta\|_0^0$ by $n\hat{A}_2$ or $n_{new}\check{A}_2$, we obtain new accuracy proxies, that should be of interest for geneticists. However, as other suggested proxies, these proxies are not optimal since with such simple expressions, there is a loss of information. Recall that the true expression of ρ_g is presented in our Theorem 1.

6. How to improve the quality of the prediction

In this section, we propose to introduce another estimator, derived from Ridge regression, and that may present, in some cases, better performances than previously studied estimators. We propose to project the vector Y on a well chosen subspace of the space spanned by the columns of X. Let $1 \leq \tilde{r} \leq r$ and $\sigma(.)$ a one-to-one map $\sigma : \{1, \ldots, \tilde{r}\} \to \{1, \ldots, r\}$. We thus have $\sigma(k) \neq \sigma(k')$ for $k \neq k'$.

Let us consider the estimator

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

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

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

We refer to Section 7.2.4, where we describe a procedure for choosing $\sigma(.)$ and $\tilde{r}$. Let $\tilde{\rho}_g$ be the analogue of ρ_g, with $\hat{Y}_{new}$ replaced by $\tilde{Y}_{new}$ (cf. formula (3)):

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

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

Lemma 6. *Let us consider same hypotheses as in Theorem 2, then an estimation of the quantity* $\tilde{\rho}_g$ *is*

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

where

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

$$\hat{\tilde{A}}_3 = \frac{1}{n}\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^6}{(d_{\sigma(s)}^2+\lambda)^2} \left\| Q^{(\sigma(s))} Q^{(\sigma(s))\prime}\beta \right\|^2 \ , \ \hat{\tilde{A}}_4 = \hat{A}_4.$$

The proof is given in the Supplementary material. Note that the quantities $\tilde{A}_1, \ldots, \tilde{A}_4$, are the analogues of $A_1, \ldots, A_4$ in this new setting.

Let us introduce our Lemma 7, which is the analogue of Lemma 1 regarding bounds for the genotypic accuracy.

Lemma 7 (Bounds on $\hat{\tilde{\rho}}_g$). *Let us consider same hypotheses as in Theorem 2, then we always have*

$$\frac{\left\|\tilde{Q}\tilde{Q}'\beta\right\|^2 \min\limits_{1\le s\le \tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2+\lambda}}{\sqrt{\sigma_e^2\ \tilde{r} + \left\|\tilde{Q}\tilde{Q}'\beta\right\|^2 \max\limits_{1\le s\le \tilde{r}} \frac{d_{\sigma(s)}^6}{(d_{\sigma(s)}^2+\lambda)^2}}\ \sqrt{\|QQ'\beta\|^2 \max\limits_{1\le s\le r} d_s^2}} \le \hat{\tilde{\rho}}_g \le \rho_g^{oracle}.$$

The proof relies heavily on the proof of Lemma 1, using the expressions of $\hat{\tilde{A}}_1$, $\hat{\tilde{A}}_2$, $\hat{\tilde{A}}_3$ given in Lemma 6. We can notice that at the denominator, the quantities $\tilde{r}$ and $\left\|\tilde{Q}\tilde{Q}'\beta\right\|$ replace now the quantities r and $\|QQ'\beta\|$ of Lemma 1. This way, this decrease at the denominator will be profitable as soon as the numerator does not decrease too much.

For fixed n, we can obtain the following comparison between $\hat{\tilde{\rho}}_g$ and $\hat{\rho}_g$.

Lemma 8. *We have $\hat{\tilde{\rho}}_g \ge \hat{\rho}_g$ if and only if the following relation holds:*

$$\left|\hat{\tilde{A}}_1 - \frac{(\hat{A}_1-\hat{\tilde{A}}_1)(\hat{\tilde{A}}_2+\hat{\tilde{A}}_3)}{\hat{A}_2+\hat{A}_3-(\hat{\tilde{A}}_2+\hat{\tilde{A}}_3)}\right| \ge \frac{(\hat{A}_1-\hat{\tilde{A}}_1)\sqrt{(\hat{A}_2+\hat{A}_3)(\hat{\tilde{A}}_2+\hat{\tilde{A}}_3)}}{\hat{A}_2+\hat{A}_3-(\hat{\tilde{A}}_2+\hat{\tilde{A}}_3)}.$$

The proof is given in Section 8.5. In other words, under this condition, the accuracy has been improved and as a consequence, we should choose the estimator $\tilde{Y}_{new}$ instead of the classical estimator $\hat{Y}_{new}$.

Remark: If $\left\|Q^{(\ell)}Q^{(\ell)\prime}\beta\right\|^2 = 0$, for all $\ell \notin \{\sigma(1), \ldots, \sigma(\tilde{r})\}$ then $\hat{\tilde{\rho}}_g \ge \hat{\rho}_g$. Indeed, in this case we have $\hat{\tilde{A}}_1 = \hat{A}_1$ and the condition in Lemma 8 is obviously fulfilled.

We introduce the following notations : for $i = 1, \ldots, 3$,

$$\tilde{\Omega}_i := \Omega_i \cap \{\sigma(1), \ldots, \sigma(\tilde{r})\}.$$

We then have the following analogue of Lemma 2 which treats the case when the signal is spread out uniformly among the different subspaces.

Lemma 9. *Let us consider the same hypotheses as in Lemma 2. Moreover, we suppose that we have the relation*

$$\sum_{s\in\tilde{\Omega}_1} d_s^2 \sim \sum_{s\in\Omega_1} d_s^2.$$

Then we have $\hat{\tilde{\rho}}_g \longrightarrow \rho_g^{oracle}$ *and* $\hat{\rho}_g \longrightarrow \rho_g^{oracle}$.

The proof is given in the Supplementary material. In other words, we have to impose that the L^2 norm squared of the singular values that belong to $\tilde{\Omega}_1$, and the L^2 norm squared of the singular values that belong to Ω_1, are equivalent. In the same way as for the classical Ridge estimator, let us focus on a few extreme cases.

Lemma 10 (Extreme cases). *Let us consider same hypotheses as in Theorem 2.*

1. *If* $1 \in \{\sigma(1),\dots,\sigma(\tilde{r})\}$ *and the projected signal belongs only to Span* $\left\{Q^{(1)}\right\}$, *that is to say*

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

   and moreover $2\tau+\psi<1$ *and the following two conditions hold*

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

   then $\hat{\tilde{\rho}}_g \to \rho_g^{oracle}$ *whereas* $\hat{\rho}_g \to 0$.
2. *If* $r \in \{\sigma(1),\dots,\sigma(\tilde{r})\}$ *and the projected signal belongs only to Span* $\left\{Q^{(r)}\right\}$, *that is to say*

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

   and moreover $\lambda \sim Cn^{\eta+\kappa}$ *with* $\kappa > \max(0,-\eta)$, $C>0$, $\tau+\eta/2-\kappa>0$ *and the following two conditions hold*

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

   then $\hat{\tilde{\rho}}_g \to \rho_g^{oracle}$ *whereas* $\hat{\rho}_g \to 0$.

The proof is largely inspired from the proof of Lemma 3. According to this lemma, there are a few cases where at the same time, the new accuracy $\hat{\tilde{\rho}}_g$ is optimal and the classical accuracy $\hat{\rho}_g$ is null.

Note that the condition $\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{\left(d_{\sigma(s)}^2+\lambda\right)^2} = o\left(n^{2\tau+\psi}\right)$ can be replaced by $\tilde{r} = o\left(n^{2\tau+\psi}\right)$. In the same way, the condition $\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{\left(d_{\sigma(s)}^2+\lambda\right)^2} = o\left(n^{2\tau+\eta-2\kappa}\right)$ can be replaced by $\tilde{r} = o\left(n^{2\tau+\eta-2\kappa}\right)$.

In Supplementary material we also investigate the same setting as in Theorem 3, when X_{new} is random. Lemma 2 and 3 (in Supplementary material) are the analogues of Theorem 3 and Lemma 4, respectively.

7. Applications

In this section, we propose to illustrate our theoretical results, with the help of simulated data.

7.1. Simulation framework

Genomic data were generated by means of the hypred R package Technow (2014), and according to the same process as in Rabier et al. (2016). In particular, populations were simulated by random mating between haploid individuals (i.e. with only one copy of each chromosome), during (a) 30, (b) 50, or (c) 70 generations. Recombination was modeled according to Haldane (1919). Recall that Haldane modeling assumes that the number of recombination follow a standard Poisson process.

In generation zero, two haploid founder lines were crossed. These two lines were completely different genetically: the first (resp. the other) line with allele $+1$ (resp. -1) at each marker. Generation 1 consisted of (a) 400 or (b) 500 haploid offsprings of these two founders. After that, the population kept evolving by random mating with a constant size at each generation. This type of simulation mimics recombinant inbred line (RIL) evolving populations. In the final generation, under the 400 offsprings scenario, 2 individuals were randomly selected, and 100 full sibs were generated in order to get some closely related individuals (as in classical genomic studies). Then, it allows to deal with two kinds of TRN populations, both based on 500 individuals: one contains 100 full sibs, whereas the other does not contain any full sib. The prediction model was evaluated on 100 TST (in all cases), that were produced in the last generation.

The focus was on one chromosome of length 1 Morgan. We considered 3 different densities of genetic markers equally spaced on the chromosome: (a) 100, (b) 1,000, or (c) 2,000 SNPs. We studied two configurations for the phenotypic model: (a) 2 QTLs located at 3cM and 80cM with effects $+1$ and -2, respectively, or (b) 100 QTLs located every centimorgan, with the same effect $+0.15$. The environmental variance σ_e^2 was set to 1.

In what follows, we will focus on the phenotypic accuracy criteria: $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$ will denote the analogue of the quantities $\hat{\rho}_g$ and $\check{\rho}_g$ for the phenotypic accuracy. As in Rabier et al. (2016), we set the value of σ_e^2 to 1, and we will consider this true value in the expressions of $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$. Recall that ρ_{ph} is

obtained by replacing the term A_4 by $A_4 + \sigma_e^2$, in our Theorem 1. Indeed, in what follows, since we consider h unknown, we can not use the expression $\rho_{ph} = h\rho_g$.

The empirical accuracy was computed in the R software, with the empirical correlation between the predicted values and the true values. Note also that all the quantities presented in the different tables, are averages based on 100 simulations. Since we analyze the case where X does not vary across replicates, one simulation consists (a) in regenerating 100 TST individuals, by random mating between individuals from the penultimate generation, and (b) in regenerating new phenotypes (TRN+TST).

The regularization parameter λ was estimated by REML. The rrBLUP R package and in particular its function kin.blup were used to compute the variance components.

7.2. Illustrations on simulated data

7.2.1. Different probability distributions

To begin with, we propose to investigate the long-term behavior of GS, i.e. the reliability of the predicted model as a function of time (Habier et al. (2007); Goddard et al. (2009)). For instance, in plants, since a large number of generations can be obtained easily, the fitted model is usually not readjusted at each generation, in order to save time or costs due to genotyping.

In this context, Table 1 compares different estimators of the phenotypic accuracy as a function of the number of generations during which the TST sample evolved. The TRN sample was always based on 30 generations. We can notice that $\check{\rho}_{ph}(\beta)$ matches the empirical accuracy whatever the number of generations for TST. In contrast, $\hat{\rho}_{ph}(\beta)$ deteriorates overtime. It was expected since $\check{\rho}_{ph}(\beta)$ handles explicitly the TST matrix X_{new}, which is not the case of $\hat{\rho}_{ph}(\beta)$ that relies on the TRN matrix X.

Table 2 considers the same number of generations for TRN and TST, and focuses on the case where a few siblings (100 or none) are included in the TRN sample. Recall that when full sibs are incorporated, the TRN and TST samples do not come from the same probability distribution. According to the table, even in presence of 100 full sibs, we observe a good agreement between the empirical accuracy and estimations based on $\hat{\rho}_{ph}(\beta)$. In view of Tables 1 and 2, it seems that not readjusting the model overtime has more impact on prediction than the presence of full sibs in the TRN set. To sum up, $\check{\rho}_{ph}(\beta)$ appears to be a reliable estimator whatever the simulation framework.

7.2.2. Behavior of the accuracy when β is estimated

In fact, the vector β containing the marker effects, is an unknown quantity. Then, we propose to consider here different estimators of β, suitable in a high-dimensional setting. We will concentrate on the LASSO (Tibshirani (1996)), the Adaptative LASSO (Zou (2006)) and on the Group LASSO (Yuan and Lin (2006)). Note that other estimators could have been chosen. Recall that the LASSO is a L^1 penalization method, and that the Adaptative LASSO replaces

Table 1: Comparison among different estimators of the phenotypic accuracy as a function of the number of generations during which the TST sample evolved (TRN sample is always based on 30 generations). The chromosome is of length 1M and 2 QTLs are located at 3cM and 80cM with effects +1 and −2, respectively ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$). Emp. Acc. refers to the empirical *phenotypic accuracy*.

Nb Markers	Nb TST generations	Emp. Acc.	$\hat{\rho}_{ph}(\beta)$	$\check{\rho}_{ph}(\beta)$
100	30	0.6901	0.6959	0.6827
	50	0.6587	0.6845	0.6523
	70	0.6419	0.6800	0.6406
1,000	30	0.686	0.6941	0.6773
	50	0.6511	0.7104	0.6438
	70	0.6224	0.7078	0.6143
2,000	30	0.6900	0.6876	0.6791
	50	0.6076	0.6872	0.5973
	70	0.5652	0.6829	0.5613

Table 2: Comparison among different estimators of the phenotypic accuracy as a function of the number of siblings in the TRN sample (TRN and TST samples based on the same number of generations). The chromosome is of length 1M and 2 QTLs are located at 3cM and 80cM with effects +1 and −2, respectively ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$). Emp. Acc. refers to the empirical *phenotypic accuracy*.

Nb Markers	Nb generations	Nb Siblings	Emp. Acc.	$\hat{\rho}_{ph}(\beta)$	$\check{\rho}_{ph}(\beta)$
100	30	0	0.6933	0.6908	0.6834
		100	0.6941	0.6890	0.6772
	50	0	0.6819	0.6765	0.6695
		100	0.6871	0.6571	0.6822
	70	0	0.6708	0.6717	0.6660
		100	0.6937	0.6869	0.6768
1,000	30	0	0.6843	0.6910	0.6841
		100	0.6735	0.6739	0.6594
	50	0	0.6602	0.6597	0.6570
		100	0.6431	0.6058	0.6214
	70	0	0.6728	0.6852	0.6663
		100	0.6042	0.6116	0.5917
2,000	30	0	0.6744	0.6719	0.6660
		100	0.6858	0.7053	0.6857
	50	0	0.6327	0.6255	0.6202
		100	0.6758	0.6913	0.6598
	70	0	0.6813	0.6845	0.6673
		100	0.6711	0.7033	0.6656

the L^1 penalty by a weighted penalty. Zou (2006) proved that Adaptative LASSO enjoyed oracle properties. Last, the Group LASSO differs from his cousins, since it allows to handle a group structure for β. We used the glmnet, parcor and gglasso R packages to compute the LASSO, the Adaptative LASSO

and the Group LASSO, respectively.

Tables 3 and 4 focus on the scenario with 2 large QTLs and 100 small QTLs, respectively. According to Table 3, the Adaptative LASSO presents better performances than his cousins, whatever the density of markers and the number of generations. As expected, the best estimators are the ones assuming known β. Note that since the TRN and TST are based on the same number of generations, we did not observe significative differences between $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$.

Table 4 shows that the accuracy based on LASSO and cousins, deteriorates slightly with a high density of markers (1,000 or 2,000). It also decreases when the number of generations increases. In view of the two tables, the Adaptative LASSO is closer to the empirical accuracy under the 2 QTLs scenario. Indeed, when 2 large QTLs well separated lied on the genome, the Adaptative LASSO was able to recover perfectly those genes, whereas the 100 QTLs scenario makes the signal recovery less trivial.

To complete our simulation study, it is worth to consider the case of a mixture between major genes and multiple small QTLs which mimics probably better the common architecture for a lot of traits. So, we generated two large QTLs located at 3cM and 80cM, and 98 small QTLs located every centimorgan (except at 3cM and 80cM). We considered three scenarios: (a) large QTLs with effects $+0.5$ and -0.6, small QTLs with the same effect $+0.07$, (b) large QTLs with effects $+1$ and -0.7, small QTLs with the same effect $+0.1$, (c) large QTLs with effects $+2$ and -2, small QTLs with the same effect $+0.1$. According to Table 5, under these new configurations, the performances are still fair even if it deteriorates slightly in presence of a high density of markers.

To conclude, in view of all our results presented in this section, the Adaptative LASSO seems to be the most appropriate method for substituting $\hat{\beta}$ into the expressions of $\hat{\rho}_{ph}$ and $\check{\rho}_{ph}$.

7.2.3. Comparison with existing methods

Table 6 shows a comparison of performance of seven different proxies in terms of the phenotypic accuracy. Three of these proxies, the ones based on M_{e1}, M_{e2}, and M_{e3}, rely on the effective population size (e.g., Goddard et al. (2011)), whereas the M_{LJ}-based proxy, comes from association studies Li and Ji (2005).

Recall that these proxies consist in substituting the effective number of independent loci M_e for $\|\beta\|_0^0$, into Daetwyler et al. (2008) original formula (cf. Section 5). The expressions of M_{e1}, M_{e2}, and M_{e3} are the following:

$$M_{e1} = \frac{2N_eL}{\log(4N_el)}\ ,\ M_{e2} = \frac{2N_eL}{\log(2N_el)}\ ,\ M_{e3} = \frac{2N_eL}{\log(N_el)}$$

where L, l, and N_e denote the genome length, average chromosome length, and effective population size respectively. M_{e1} was proposed by Goddard et al. (2009), whereas M_{e2} and M_{e3} are from Goddard et al. (2011). We refer to Rabier et al. (2016) for more details on the estimation of N_e, based on Hill and Weir (1998). The fifth proxy is the one introduced in Rabier et al. (2016). Note that

Table 3: Comparison among different estimators of the phenotypic accuracy, in presence of a few major genes ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$). The chromosome is of length 1M and the 2 QTLs are located at 3cM and 80cM with effects $+1$ and -2, respectively. $\hat{\beta}_{LASSO}$, $\hat{\beta}_{ADLASSO}$, $\hat{\beta}_{GPLASSO}$ refer to the LASSO, Adaptative LASSO and Group LASSO estimators of β, respectively. Emp. Acc. refers to the empirical phenotypic accuracy.

Nb markers	Method	30 generations	50 generations	70 generations
100	Emp. Acc.	0.6967	0.6804	0.6708
	$\hat{\rho}_{ph}(\beta)$	0.6969	0.6765	0.6717
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.5962	0.5767	0.5735
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.6927	0.6675	0.6676
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.5934	0.5595	0.5484
	$\check{\rho}_{ph}(\beta)$	0.6915	0.6731	0.6654
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.5907	0.5742	0.5677
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.6872	0.6712	0.6614
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.5857	0.5580	0.5411
1,000	Emp. Acc.	0.7015	0.6683	0.6713
	$\hat{\rho}_{ph}(\beta)$	0.7155	0.6597	0.685
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.6197	0.5354	0.5720
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.7066	0.6488	0.675
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.6244	0.5471	0.586
	$\check{\rho}_{ph}(\beta)$	0.6889	0.6576	0.6642
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.5965	0.5347	0.5544
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.6812	0.6454	0.6548
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.6022	0.5495	0.5708
2,000	Emp. Acc.	0.6977	0.6316	0.4174
	$\hat{\rho}_{ph}(\beta)$	0.6933	0.6254	0.4600
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.5872	0.4794	0.2790
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.6783	0.6134	0.4399
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.5904	0.4814	0.2801
	$\check{\rho}_{ph}(\beta)$	0.6881	0.6264	0.4095
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.5842	0.4831	0.2522
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.6830	0.6138	0.3890
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.5902	0.4878	0.2601

the heritability h^2 was estimated with the help of variance components obtained by the rrBLUP R package. Last, the remaining proxies are those suggested in this paper: $\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$ and $\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$.

Table 6 reports the Mean Squared Error (MSE) associated to each method, and based on 15 architectures. An architecture refers to a fixed number of: (a) SNPs; (b) QTL numbers, effects, and locations. The number 15 comes from the fact that we considered (a) either 100, either 1,000 or 2,000 SNPs, and (b) either 2 large QTLs, either 100 small QTLs, or the 3 scenarios of Table 5.

According to Table 6, $\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$ and $\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$ are the most

Table 4: Comparison among different estimators of the phenotypic accuracy, in presence of multiple small QTLs ($n = 500$, $n_{new} = 100$, $\sigma_e^2 = 1$). The chromosome is of length 1M and 100 QTLs with the same effect +0.15, are located every centimorgan. Same notations as in Table 3.

Nb markers	Method	30 generations	50 generations	70 generations
100	Emp. Acc.	0.8504	0.8055	0.7056
	$\hat{\rho}_{ph}(\beta)$	0.8346	0.8007	0.6938
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.7990	0.7010	0.6043
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8366	0.8036	0.6998
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.7813	0.7370	0.5611
	$\check{\rho}_{ph}(\beta)$	0.8434	0.7941	0.6981
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.8020	0.7471	0.60
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8426	0.7959	0.7029
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.7889	0.7250	0.5611
1,000	Emp. Acc.	0.8700	0.8143	0.7233
	$\hat{\rho}_{ph}(\beta)$	0.8781	0.8086	0.7308
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.8558	0.7635	0.6532
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8495	0.7627	0.6718
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.8508	0.7581	0.6466
	$\check{\rho}_{ph}(\beta)$	0.8604	0.8045	0.7162
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.8299	0.7502	0.6233
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8226	0.7489	0.6452
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.8273	0.7479	0.6224
2,000	Emp. Acc.	0.8590	0.8045	0.7387
	$\hat{\rho}_{ph}(\beta)$	0.8464	0.8113	0.7319
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.8116	0.7662	0.6503
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8102	0.7641	0.6697
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.8062	0.7607	0.6495
	$\check{\rho}_{ph}(\beta)$	0.8510	0.7936	0.7300
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.8096	0.7339	0.6317
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.8093	0.7358	0.6542
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.8074	0.7322	0.6364

competitive proxies. They outperformed our recent proxy Rabier et al. (2016), and classical proxies used by geneticists. As expected, $\check{\rho}_{ph}(\beta)$ yielded the best performances. Recall that it cannot be computed in practice because it depends on the unknown β.

7.2.4. The quality of the prediction can be improved

In this section, we propose to illustrate the quality of predictions based on $\tilde{\beta}$. Recall that this estimator is built after having projected the vector Y on a well chosen subspace of the space spanned by the columns of X.

In order to find an appropriate subspace, we used the following procedure.

Table 5: Comparison among different estimators of the phenotypic accuracy, in presence of a mixture of major genes and small QTLs (50 generations, $n = 500$, $n_{new} = 100$, p =1,000, $\sigma_e^2 = 1$). Three scenarios considered (a) 2 large QTLs with effects +0.5 and −0.6, 98 small QTLs with the same effect +0.07, (b) 2 large QTLs with effects +1 and −0.7, 98 small QTLs with the same effect +0.1, (c) 2 large QTLs with effects +2 and −2, 98 small QTLs with the same effect +0.1. The chromosome is of length 1M and the large QTLs are located at 3cM and 80cM, whereas the small QTLs are located every centimorgan (except at 3cM and 80cM). Same notations as in Table 3.

Nb markers	Method	Scenario (a)	Scenario (b)	Scenario (c)
100	Emp. Acc.	0.5479	0.7012	0.8074
	$\hat{\rho}_{ph}(\beta)$	0.5362	0.6900	0.8013
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.3792	0.6096	0.7614
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.5400	0.6678	0.8049
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.3500	0.5909	0.7419
	$\check{\rho}_{ph}(\beta)$	0.5296	0.6868	0.7962
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.3628	0.6016	0.7550
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.5313	0.6942	0.7999
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.3370	0.5720	0.7324
1,000	Emp. Acc.	0.5867	0.7374	0.8307
	$\hat{\rho}_{ph}(\beta)$	0.5738	0.7316	0.8276
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.4187	0.6575	0.7935
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.5077	0.6639	0.7918
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.4127	0.6526	0.7843
	$\check{\rho}_{ph}(\beta)$	0.5768	0.7274	0.8209
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.4055	0.6411	0.7833
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.4973	0.6478	0.7811
	$\check{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.4036	0.6401	0.7773
2,000	Emp. Acc.	0.5446	0.7063	0.8038
	$\hat{\rho}_{ph}(\beta)$	0.5502	0.7132	0.8029
	$\hat{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.3710	0.6297	0.7633
	$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.4867	0.6445	0.7594
	$\hat{\rho}_{ph}(\hat{\beta}_{GPLASSO})$	0.3578	0.6197	0.7488
	$\check{\rho}_{ph}(\beta)$	0.5378	0.6972	0.7937
	$\check{\rho}_{ph}(\hat{\beta}_{LASSO})$	0.3407	0.5958	0.7502
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.4627	0.6190	0.7525
	$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	0.3317	0.5886	0.7379

We decided that $\frac{d^4_{\sigma(k)}}{d^2_{\sigma(k)}+\lambda}\left\|Q^{(\sigma(k))}Q^{(\sigma(k))\prime}\beta\right\|^2$ was the k-th largest term of the sequence $\left(\frac{d^4_s}{d^2_s+\lambda}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2\right)_{s=1,\ldots,r}$. $\tilde{r}$ was chosen as the largest value satisfying the condition $\hat{\tilde{A}}_1/A_1 \leq \upsilon$, where υ denotes a tuning parameter. The corresponding accuracy was then computed for a given value of υ.

Since υ was unknown, we performed an optimization over the grid

Table 6: Mean squared error (with respect to the Empirical accuracy) corresponding to 7 proxies. The MSE corresponding to $\check{\rho}_{ph}(\beta)$ is also shown . MSE= $\sum_{a=1}^{15}\left(\mathrm{AccP}_a - \mathrm{AccE}_a\right)^2/15$ where 15 is the number of studied architectures. AccE_a and AccP_a are averages on 100 replicates, and denote respectively, for architecture a, the Empirical Accuracy and the Accuracy based on the chosen proxy (30 generations for TRN).

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

$\{0.7, 0.8, 0.9, 0.925, 0.95, 0.975, 0.99\}$, and kept the highest accuracy.

Tables 7 and 8 focus on the cases $n = 500$ and $n = 800$, respectively. According to the tables, in all studied cases, the quantity $\hat{\tilde{\rho}}_{ph}(\beta)$ was greater than $\hat{\rho}_{ph}(\beta)$. In the same way, the empirical accuracy associated to the new estimator (i.e. $cor\left(\tilde{Y}_{new}, Y_{new}\right)$), was always greater than the classical empirical accuracy based on the Ridge estimator (i.e. $cor\left(\hat{Y}_{new}, Y_{new}\right)$).

Last, Table 9 focuses on the case where the vector β belongs to $\mathcal{R}(X)$. In particular, we considered $\beta = 0.3Q^{(1)} + 0.3Q^{(2)} + 0.3Q^{(3)}$. As expected (cf. remark below Lemma 8), $\hat{\tilde{\rho}}_{ph}(\beta)$ took greater values than $\hat{\rho}_{ph}(\beta)$.

Table 7: Illustration of the predictions based on $\tilde{\beta}$. $cor\left(\hat{Y}_{new}, Y_{new}\right)$ (resp. $cor\left(\tilde{Y}_{new}, Y_{new}\right)$) refers to the empirical correlation between $\hat{Y}_{new}$ (resp. $\tilde{Y}_{new}$) and Y_{new}. The chromosome is of length 4M and 100 QTLs with the same effect +0.15, are located every centimorgan on $[0, 1\mathrm{M}]$. 4,000 markers (p=4,000) are equally spaced on $[0, 4\mathrm{M}]$ ($n = 500$, $n_{new} = 100$).

σ_e^2	Method	50 generations	100 generations
1	$cor\left(\hat{Y}_{new}, Y_{new}\right)$	0.7478	0.5959
	$cor\left(\tilde{Y}_{new}, Y_{new}\right)$	0.7682	0.6132
	$\hat{\rho}_{ph}(\beta)$	0.7399	0.6352
	$\hat{\tilde{\rho}}_{ph}(\beta)$	0.7570	0.6541
9	$cor\left(\hat{Y}_{new}, Y_{new}\right)$	0.2874	0.1949
	$cor\left(\tilde{Y}_{new}, Y_{new}\right)$	0.3152	0.2163
	$\hat{\rho}_{ph}(\beta)$	0.3023	0.2320
	$\hat{\tilde{\rho}}_{ph}(\beta)$	0.3306	0.2604

Table 8: Same as Table 7 except that $n = 800$.

σ_e^2	Method	50 generations	100 generations
1	$cor\left(\hat{Y}_{new}, Y_{new}\right)$	0.7911	0.6127
	$cor\left(\tilde{Y}_{new}, Y_{new}\right)$	0.8087	0.6301
	$\hat{\rho}_{ph}(\beta)$	0.7824	0.6509
	$\hat{\tilde{\rho}}_{ph}(\beta)$	0.7965	0.6663
9	$cor\left(\hat{Y}_{new}, Y_{new}\right)$	0.3725	0.1981
	$cor\left(\tilde{Y}_{new}, Y_{new}\right)$	0.4044	0.2302
	$\hat{\rho}_{ph}(\beta)$	0.3766	0.2248
	$\hat{\tilde{\rho}}_{ph}(\beta)$	0.4041	0.2494

Table 9: Comparison among the quantities $\hat{\rho}_{ph}(\beta)$ and $\hat{\tilde{\rho}}_{ph}(\beta)$, when the vector β belongs to $\mathcal{R}(X)$. The chromosome is of length 1M, $\beta = 0.3Q^{(1)} + 0.3Q^{(2)} + 0.3Q^{(3)}$ and $n_{new} = 100$.

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

7.3. Real data: GS in rice

To conclude this article, we propose to analyze some data from the recent paper of Spindel et al. (2015) dealing with GS in rice.

We considered the dataset of 13,101 SNPs, randomly chosen by the authors from their 73,147 collected SNPs. We decided to focus on two rice traits: flowering and yield. Besides, our analysis relies on the dry season 2012. 80% of the observations 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 $n = 252$ for flowering, and $n = 248$ for yield. In both cases, we considered $n_{new} = 63$. Table 10 shows a comparison of performance of seven different proxies in terms of the phenotypic accuracy. The computed MSE rely on 100 data sets (with random individuals in TRN and TST sets).

In order to compute proxies based on M_{e1}, M_{e2}, M_{e3}, we used the value $L = 13.101$ for the genome length (from Section "GS using marker subsets" of Spindel et al. (2015)), and $l = 1.09175$ for the average chromosome length. Recall that the rice presents 12 chromosomes. In order to make calculation easier, the effective population size N_e was obtained by using only 1,007 SNPs spread out on the genome (a SNP every 0.012 Morgan). Furthermore, we used

the Adaptative LASSO to compute our suggested proxies, $\hat{\rho}_{ph}(\beta)$ and $\check{\rho}_{ph}(\beta)$. Note that since σ_e^2 was unknown, we considered the estimation of σ_e^2 given by REML.

According to Table 10, $\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$ is the most interesting proxy. Indeed, for flowering and yield, the associated MSE was the smallest among all proxies, and the associated mean accuracies were pretty close to the empirical accuracies (0.5485 vs. 0.5576 for flowering, and 0.2650 vs. 0.3361 for yield). As a consequence, the results presented in this manuscript should be of interest for geneticists.

Table 10: Mean squared error (with respect to the Empirical accuracy) corresponding to 7 proxies, and based on rice data from Spindel et al. (2015) (dry season 2012). The computed MSE rely on 100 data sets (with random individuals in TRN and TST sets). The average, for each proxy, is given in brackets. The Empirical accuracy was 0.5576 for flowering, and 0.3361 for yield ($n = 252$ for flowering, $n = 248$ for yield, $n_{new} = 63$ in both cases).

MSE based on	Flowering	Yield
$\check{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	1.6248×10^{-2} (0.5485)	2.807×10^{-2} (0.2650)
$\hat{\rho}_{ph}(\hat{\beta}_{ADLASSO})$	2.41×10^{-2} (0.6201)	4.85×10^{-2} (0.4571)
Plos One (2016)	7.08×10^{-2} (0.7903)	1.25×10^{-1} (0.6647)
M_{e1}	4.49×10^{-2} (0.7055)	5.70×10^{-2} (0.5234)
M_{e2}	4.18×10^{-2} (0.6917)	5.10×10^{-2} (0.5064)
M_{e3}	3.83×10^{-2} (0.6741)	4.43×10^{-2} (0.4854)
M_{LJ}	4.71×10^{-2} (0.7142)	6.27×10^{-2} (0.5383)

8. Proofs

8.1. Proof of Theorem 1

By definition,

$$A_1 = \beta' \, \mathrm{Var}\,(x_{new}) \, X'V^{-1}X\beta.$$

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

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

Using formula (8), we easily have

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

As a consequence, since $\Sigma = \mathbb{E}\,(x_{new}\; x'_{new})$, we have the relationship

$$A_1 = \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} \beta' \, \Sigma \, Q^{(s)}Q^{(s)\prime}\beta. \tag{16}$$

By definition,

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

According to formula (14) and , we have

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

Furthermore, we have

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

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

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

Finally,

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

By definition,

$$A_3 = \beta' X' V^{-1} X \mathrm{Var}\left(x_{new}\right) X' V^{-1} X \beta.$$

Then, according to formula (15),

$$A_3 = \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, Q^{(s)} Q^{(s)\prime} \beta\right)' \; \Sigma \; \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, Q^{(s)} Q^{(s)\prime} \beta\right).$$

8.2. Proof of Theorem 2

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

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

where $\hat{\Sigma} := X'X/n$ is the empirical covariance matrix.

Then, using the SVD decomposition $X = PDQ'$, we obtain

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\beta' \; X'X \; Q^{(s)}Q^{(s)\prime}\beta \\
&= \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\beta' \; QD^2Q' \; Q^{(s)}Q^{(s)\prime}\beta \\
&= \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\beta' \; \left(\sum_{u=1}^{r} d_u^2 Q^{(u)}Q^{(u)\prime}\right) \; Q^{(s)}Q^{(s)\prime}\beta.
\end{aligned}$$

Since $Q'Q = I_r$, we further deduce

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\beta' \; d_s^2 Q^{(s)}Q^{(s)\prime} \; Q^{(s)}Q^{(s)\prime}\beta \\
&= \frac{1}{n}\sum_{s=1}^{r} \frac{d_s^4}{d_s^2+\lambda} \; \left\| Q^{(s)}Q^{(s)\prime}\beta \right\|^2.
\end{aligned}$$

A natural estimation of A_2 is

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

Note that

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

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

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

Let us consider the following estimation of A_3:

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

Note that

$$XQ^{(s)}Q^{(s)\prime}\beta = PDQ'Q^{(s)}Q^{(s)\prime}\beta = d_s P e_s Q^{(s)\prime}\beta = d_s P^{(s)}Q^{(s)\prime}\beta.$$

As a consequence,

$$\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, XQ^{(s)}Q^{(s)\prime}\beta = \sum_{s=1}^{r} \frac{d_s^3}{d_s^2+\lambda} P^{(s)}Q^{(s)\prime}\beta.$$

Last, we obtain

$$\begin{aligned}\hat{A}_3 &= \frac{1}{n}\left(\sum_{\ell=1}^{r} \frac{d_\ell^3}{d_\ell^2+\lambda}\beta' Q^{(\ell)}P^{(\ell)\prime}\right)\left(\sum_{s=1}^{r} \frac{d_s^3}{d_s^2+\lambda} P^{(s)}Q^{(s)\prime}\beta\right) \\ &= \frac{1}{n}\sum_{\ell=1}^{r} \frac{d_\ell^3}{d_\ell^2+\lambda}\sum_{s=1}^{r} \frac{d_s^3}{d_s^2+\lambda}\beta' Q^{(\ell)}P^{(\ell)\prime}P^{(s)}Q^{(s)\prime}\beta \\ &= \frac{1}{n}\sum_{\ell=1}^{r} \frac{d_\ell^6}{(d_\ell^2+\lambda)^2}\beta' Q^{(\ell)}Q^{(\ell)\prime}\beta \\ &= \frac{1}{n}\sum_{\ell=1}^{r} \frac{d_\ell^6}{(d_\ell^2+\lambda)^2}\left\|Q^{(\ell)}Q^{(\ell)\prime}\beta\right\|^2.\end{aligned}$$

Finally, let us consider the following estimation of A_4:

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

We have

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

8.3. Proof of Lemma 1

$$\begin{aligned}
\hat{A}_1 &= \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{d_s^2+\lambda}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \\
&= \frac{1}{n}\sum_{s=1}^{r}\left(\frac{d_s^3}{d_s^2+\lambda}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|\right)\left(d_s\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|\right) \\
&\leq \frac{1}{n}\left(\sum_{s=1}^{r}\frac{d_s^6}{(d_s^2+\lambda)^2}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2\right)^{1/2}\left(\sum_{s=1}^{r}d_s^2\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2\right)^{1/2} \\
&= \hat{A}_3^{1/2}\hat{A}_4^{1/2},
\end{aligned}$$

using the Cauchy-Schwartz inequality. Since $\hat{A}_2 \geq 0$, we obtain

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

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

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

Then,

$$\begin{aligned}
n\hat{A}_1 &= \sum_{s=1}^{r}\frac{d_s^4}{d_s^2+\lambda}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \geq \|QQ'\beta\|^2 \ \min_s \frac{d_s^4}{d_s^2+\lambda} \\
n\hat{A}_3 &= \sum_{s=1}^{r}\frac{d_s^6}{(d_s^2+\lambda)^2}\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \leq \|QQ'\beta\|^2 \ \max_s \frac{d_s^6}{(d_s^2+\lambda)^2} \\
n\hat{A}_4 &= \sum_{s=1}^{r}d_s^2\left\|Q^{(s)}Q^{(s)\prime}\beta\right\|^2 \leq \|QQ'\beta\|^2 \max_s d_s^2.
\end{aligned}$$

Since $\frac{d_s^4}{(d_s^2+\lambda)^2}$ is bounded by one, we 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,

$$\hat{\rho}_g \geq \frac{\|QQ'\beta\|^2 \ \min_s \frac{d_s^4}{d_s^2+\lambda}}{\sqrt{\sigma_e^2\, r + \|QQ'\beta\|^2 \ \max_s \frac{d_s^6}{(d_s^2+\lambda)^2}}\ \sqrt{\|QQ'\beta\|^2 \ \max_s d_s^2}}.$$

8.4. Proof of Lemma 2

Using Theorem 2, we have:

$$n\hat{A}_1 \sim \sum_{s\in\Omega_1} d_s^2\frac{n^{2\tau}}{r} \ + \ \sum_{s\in\Omega_2}\frac{d_s^4}{d_s^2+C_s d_s^2}\frac{n^{2\tau}}{r} \ + \ \sum_{s\in\Omega_3}\frac{d_s^4}{\lambda}\frac{n^{2\tau}}{r}.$$

According to our conditions (C3) and (C4),

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

Then,

$$n\hat{A}_1 \sim \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r} \; + \; \sum_{s\in\Omega_2} \frac{d_s^2}{1+C_s}\frac{n^{2\tau}}{r}. \tag{17}$$

We have

$$\sum_{s\in\Omega_2} \frac{d_s^2}{1+C_s}\frac{n^{2\tau}}{r} \leq \sum_{s\in\Omega_2} d_s^2 \frac{n^{2\tau}}{r} \leq \frac{n^{2\tau}}{r}\#\Omega_2 \tilde{C}\lambda,$$

with $\tilde{C} > 0$.

Since $\#\Omega_2 = O(1)$ by (C6) and $\lambda\frac{n^{2\tau}}{r} = o(1)$, we have

$$\sum_{s\in\Omega_2} d_s^2 \frac{n^{2\tau}}{r} = o(1) \tag{18}$$

and thus $\sum_{s\in\Omega_2} \frac{d_s^2}{1+C_s}\frac{n^{2\tau}}{r} = o(1)$. Therefore

$$n\hat{A}_1 \sim \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r}. \tag{19}$$

In the same way, using condition (C3), we have

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

Let us now focus on the quantity $\hat{A}_3$.

$$n\hat{A}_3 \sim \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r} \; + \; \sum_{s\in\Omega_2} \frac{d_s^2}{(1+C_s)^2}\frac{n^{2\tau}}{r} \; + \; \sum_{s\in\Omega_3} \frac{d_s^6}{\lambda^2}\frac{n^{2\tau}}{r}.$$

Since $\sum_{s\in\Omega_3} d_s^6 \leq \sum_{s\in\Omega_3} d_s^2 \sum_{s\in\Omega_3} d_s^4$, we have $\sum_{s\in\Omega_3} d_s^6 = o(\lambda^3)$ (cf. (C2) and (C3)). Then, according to (C4), $\sum_{s\in\Omega_3} \frac{d_s^6}{\lambda^2}\frac{n^{2\tau}}{r} = o(1)$. This yields,

$$n\hat{A}_2 + n\hat{A}_3 \sim \sigma_e^2 \#\Omega_1 \; + \; \sigma_e^2 \sum_{s\in\Omega_2} \frac{1}{(1+C_s)^2} \; + \; \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r} \; + \; \sum_{s\in\Omega_2} \frac{d_s^2}{(1+C_s)^2}\frac{n^{2\tau}}{r}.$$

We further have

$$\sum_{s\in\Omega_2} \frac{d_s^2}{(1+C_s)^2}\frac{n^{2\tau}}{r} \leq \sum_{s\in\Omega_2} d_s^2 \frac{n^{2\tau}}{r}.$$

Using the previous relation (18), we have $\sum_{s\in\Omega_2} \frac{d_s^2}{(1+C_s)^2}\frac{n^{2\tau}}{r} = o(1)$. As a result,

$$n\hat{A}_2 + n\hat{A}_3 \sim \sigma_e^2 \#\Omega_1 \ + \ \sigma_e^2 \sum_{s\in\Omega_2} \frac{1}{(1+C_s)^2} \ + \ \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r}.$$

Then, conditions (C1), (C5) and (C6) ensure that

$$n\hat{A}_2 + n\hat{A}_3 \sim \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r}. \tag{20}$$

Last,

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

According to conditions (C4) and (C2), $\sum_{s\in\Omega_3} d_s^2 \frac{n^{2\tau}}{r} = o(1)$. Using again the relation (18) we deduce

$$n\hat{A}_4 \sim \sum_{s\in\Omega_1} d_s^2 \frac{n^{2\tau}}{r}. \tag{21}$$

To conclude, using formulae (19), (20) and (21), we have $\hat{\rho}_g \to 1$.

8.5. Proof of Lemma 8

To simplify notations, let us put

$$u := \hat{\tilde{A}}_1, \ \delta_1 := \hat{A}_1 - \hat{\tilde{A}}_1,$$
$$v := \hat{\tilde{A}}_2 + \hat{\tilde{A}}_3, \ \delta_2 := \hat{A}_2 + \hat{A}_3 - (\hat{\tilde{A}}_2 + \hat{\tilde{A}}_3).$$

With these notations the condition $\hat{\tilde{\rho}}_g \geq \hat{\rho}_g$ reads

$$\frac{u+\delta_1}{\sqrt{v+\delta_2}} \leq \frac{u}{\sqrt{v}},$$

which is further equivalent to

$$\delta_2 u^2 - 2u\delta_1 v - \delta_1^2 v \geq 0.$$

The discriminant in the u variable equals $\Delta = 4\delta_1^2 v(v+\delta_2)$ and is positive. The above second order inequation is thus satisfied for

$$\left| u - \frac{\delta_1}{\delta_2} v \right| \geq \frac{\delta_1}{\delta_2} \sqrt{v(v+\delta_2)},$$

which gives, after few simplifications, the desired statement.

Supporting information. Additional information for this article is available below

Text S1 : Supplementary material containing a few proofs.

References

Azaïs, J.M., Delmas, C. & Rabier, C.E. (2014). Likelihood ratio test process for Quantitative Trait Locus detection. *Statistics*. **48**, (4), 787-801.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., et al (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*. **40**, (8), 955-962.

Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., et al (2009). The genetic architecture of maize flowering time. *Science*. **325**, (5941), 714-718.

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

Burstin, J., Salloignon, P., Martinello, M., Magnin-Robert, J.B., Siol, M., Jacquin, F., et al (2015). Genetic diversity and trait genomic prediction in a pea diversity panel. *BMC genomics*. **16**, (1), 105.

Cai, T.T., Zhang, C., & Zhou, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38**, (4), 2118-2144.

Chang, M. N., Wu, R., Wu, S. S., & Casella, G. (2009). Score statistics for mapping quantitative trait loci. *Statistical applications in genetics and molecular biology*. **8**, (1), 1-35.

Chen, Z. & Chen H. (2005). On some statistical aspects of the interval mapping for QTL detection. *Statistica Sinica*. **15**, (4), 909-925.

Cierco, C. (1998). Asymptotic distribution of the maximum likelihood ratio test for gene detection. *Statistics*. **31**, (3), 261-285.

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

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

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

de los Campos, G., Gianola, D., Rosa, G.J., Weigel, K.A. & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research*. **92**, (04), 295-308.

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

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

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

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

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

Goddard, M.E. & Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics.* **10**, (6), 381-391.

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

Goddard, M.E., Hayes, B.J., & Meuwissen, T.H.E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics.* **128**, (6), 409-421.

Goldenshluger, A., & Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis.* **84**, (1), 40-60.

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

Jannink, J.L., Lorenz, A.J. & Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics.* **9**, (2), 166-177.

Kärkkäinen, H.P. & Sillanpää, M.J. (2012). Back to basics for Bayesian model building in genomic selection. *Genetics.* **191**, (3), 969-987.

Kim, D.Y., Cui, Y., & Zhao, O. (2009). Asymptotic test of mixture model and its applications to QTL interval mapping. *Journal of Statistical Planning and Inference.* **143**, (8), 1320-1329.

Kumar, S., Chagné, D., Bink, M.C., Volz, R.K., Whitworth, C. & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (Malus× domestica Borkh.). *PloS One.* **7**, (5), e36674.

Habier, D., Fernando, R. & Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics.* **177**, (4), 2389-2397.

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.

Hill, W. & Weir, B. (1998). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical population biology*. **33**, (1), 54-78.

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

Lander, E.S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. **121**, (1), 185-199.

Li, J. & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*. **95**, (3), 221-227.

Li, Z. & Sillanpää, M.J. (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*. **125**, (3), 419-435.

Lynch, M. & Walsh, B. (1998). *Genetics and analysis of quantitative traits*, Sinauer Sunderland, MA.

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

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

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

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

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

Spindel, J.E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*. **116**, 395-408.

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

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

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

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

Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., et al (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*. **40**, (5), 575-583.

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

Würschum, T., Reif, J.C., Kraft, T., Janssen, G. & Zhao, Y. (2013). Genomic selection in sugar beet breeding populations. *BMC genetics*. **14**, (1), 85.

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

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

Charles-Elie Rabier (charles-elie.rabier@umontpellier.fr)
ISEM, Université de Montpellier, CNRS, France.
Brigitte Mangin (brigitte.mangin@inra.fr)
LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France
Simona Grusea (grusea@insa-toulouse.fr)
INSA de Toulouse, Institut de Mathématiques de Toulouse, Université de Toulouse, France.

Text S1: Supplementary material of "On the accuracy in high dimensional linear models and its application to genomic selection"

C.E. Rabier[a,b,c,d], B. Mangin[e], S. Grusea[a]

[a] *INSA de Toulouse, Institut de Mathématiques de Toulouse, Université de Toulouse, France*
[b] *MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France*
[c] *ISEM, Université de Montpellier, CNRS, France*
[d] *LIRMM, Université de Montpellier, CNRS, France*
[e] *LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France*

---

1. Introduction

Lemma 1. *Let us consider same hypotheses as in Theorem 1 of the main manuscript. Then, the quantity $\tilde{\rho}_g$ defined in Section 6 of the main manuscript has the following expression*

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

where

$$\tilde{A}_1 = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda}\, \beta' \,\Sigma\, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \;\;,\;\; \tilde{A}_2 = \sigma_e^2 \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{(d_{\sigma(s)}^2 + \lambda)^2}\, \mathbb{E}\left(\left\| Q^{(\sigma(s))} Q^{(\sigma(s))\prime} x_{new} \right\|^2 \right)$$

$$\tilde{A}_3 = \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda}\, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right)' \,\Sigma\, \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda}\, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right),\;\; \tilde{A}_4 = A_4.$$

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

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

As a result, let us define

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

Using the fact that $X'V^{-1} = Q\overline{D}P'$ and the fact that $\Sigma = \mathbb{E}\left(x_{new}\, x'_{new}\right)$, we have

$$\begin{aligned}\tilde{A}_1 &= \beta'\, \Sigma X'V^{-1}\tilde{P}\tilde{P}'X\beta \\ &= \beta'\, \Sigma Q\overline{D}P'\tilde{P}\tilde{P}'X\beta.\end{aligned}$$

After some simple algebra, we obtain

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

Then,

$$\begin{aligned}\tilde{A}_1 &= \beta'\, \Sigma \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}Q^{(\sigma(s))}P^{(\sigma(s))\prime}\right)\left(\sum_{s=1}^{r} d_s P^{(s)}Q^{(s)\prime}\right)\beta \\ &= \beta'\, \Sigma \left(\sum_{s=1}^{\tilde{r}} \frac{d^2_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}Q^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta\right) \\ &= \sum_{s=1}^{\tilde{r}} \frac{d^2_{\sigma(s)}}{d^2_{\sigma(s)}+\lambda}\beta'\, \Sigma\, Q^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta.\end{aligned}$$

Let us now consider $\tilde{A}_2$. We have

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

According to formula (1), we obtain

$$Q\overline{D}P'\tilde{P}\tilde{P}'P\overline{D}Q' = \sum_{s=1}^{\tilde{r}} \frac{d^2_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}Q^{(\sigma(s))}Q^{(\sigma(s))\prime}$$

and

$$\begin{aligned}x'_{new}Q\overline{D}P'\tilde{P}\tilde{P}'P\overline{D}Q'x_{new} &= \sum_{s=1}^{\tilde{r}} \frac{d^2_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}x'_{new}Q^{(\sigma(s))}Q^{(\sigma(s))\prime}x_{new} \\ &= \sum_{s=1}^{\tilde{r}} \frac{d^2_{\sigma(s)}}{(d^2_{\sigma(s)}+\lambda)^2}\left\|Q^{(\sigma(s))}Q^{(\sigma(s))\prime}x_{new}\right\|^2.\end{aligned}$$

The last equality comes from the fact that $Q^{(\sigma(s))}Q^{(\sigma(s))\prime}$ is an idempotent matrix. To conclude, we have

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

Furthermore, recall that

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

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

$$\tilde{A}_3 = \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right)' \Sigma \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right).$$

□

Lemma 2. *Let us consider same hypotheses as in Theorem 3 of the main manuscript. Then, a natural estimator of the quantity $\tilde{\rho}_g$ is the following:*

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

where

$$\check{\tilde{A}}_1 = \frac{1}{n_{new}} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \left(\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \ < Z^{(\alpha)} Z^{(\alpha)\prime} \beta, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta > \right),$$

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

$$\check{\tilde{A}}_3 = \frac{1}{n_{new}} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} Q^{(\sigma(s))\prime} \beta \sum_{\ell=1}^{\tilde{r}} \frac{d_{\sigma(\ell)}^2}{d_{\sigma(\ell)}^2 + \lambda} Q^{(\sigma(\ell))\prime} \beta \left(\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \ < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\sigma(s))}, Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\sigma(\ell))} > \right),$$

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

Proof. In the same way as before, we consider the estimators $\check{\tilde{A}}_1$, $\check{\tilde{A}}_2$ and $\check{\tilde{A}}_3$, of $\tilde{A}_1$, $\tilde{A}_2$ and $\tilde{A}_3$, respectively:

$$\check{\tilde{A}}_1 := \frac{1}{n_{new}} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \, \beta' \, X_{new}' X_{new} \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta,$$

$$\check{\tilde{A}}_2 := \frac{\sigma_e^2}{n_{new}} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{(d_{\sigma(s)}^2 + \lambda)^2} \mathrm{Tr}\left(X_{new} Q^{(\sigma(s))} Q^{(\sigma(s))\prime} Q^{(\sigma(s))} Q^{(\sigma(s))\prime} X_{new}' \right),$$

$$\check{\tilde{A}}_3 = \frac{1}{n_{new}} \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right)' X_{new}' X_{new} \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right).$$

After some easy computations we can deduce the stated formulas. □

Lemma 3. *Let us consider same hypotheses as in Theorem 3 of the manuscript. Then we always have*

$$\frac{\tilde{B}_1}{\left(\tilde{B}_2+\tilde{B}_3\right)^{1/2}\tilde{B}_4^{1/2}} \leq \overset{\circ}{\rho}_g \leq \rho_g^{oracle},$$

where

$$\begin{aligned}
\tilde{B}_1 &= \min_{1\leq s\leq \tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2+\lambda} \min f_\alpha^2 \quad < ZZ'\beta\ ,\ \tilde{Q}\tilde{Q}'\beta >,\\
\tilde{B}_2 &= \sigma_e^2\ \tilde{r}\ r_{new}\ \max_{1\leq s\leq \tilde{r}} \frac{d_{\sigma(s)}^2}{(d_{\sigma(s)}^2+\lambda)^2} \max f_\alpha^2 \max_{1\leq s\leq \tilde{r},\alpha} \left\| Q^{(\sigma(s))\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2,\\
\tilde{B}_3 &= \max_{1\leq s\leq \tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2+\lambda)^2} \left\| \tilde{Q}\tilde{Q}'\beta \right\|^2 \max f_\alpha^2\ \ \tilde{r}^2,\\
\tilde{B}_4 &= B_4.
\end{aligned}$$

The proof relies heavily on the proof of Lemma 4 of the main manuscript, provided that we consider the expressions of $\overset{\circ}{\tilde{A}}_1$, $\overset{\circ}{\tilde{A}}_2$, $\overset{\circ}{\tilde{A}}_3$ given in Lemma 2 above.

2. Proof of Lemma 3 of the main manuscript

2.1. The projected signal belongs only to Span $\left\{Q^{(1)}\right\}$

Using Theorem 2, we have:

$$\hat{\rho}_g = \frac{\frac{d_1^3}{d_1^2+\lambda}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|}{\left(\sigma_e^2\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}+\frac{d_1^6}{(d_1^2+\lambda)^2}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|^2\right)^{1/2}}. \tag{2}$$

From Lemma 1 and the fact that $\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2} \leq r \leq n$, we deduce that

$$1 \geq \hat{\rho}_g \geq \frac{\frac{d_1^3}{d_1^2+\lambda}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|}{\left(\sigma_e^2 n+\frac{d_1^6}{(d_1^2+\lambda)^2}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|^2\right)^{1/2}}. \tag{3}$$

Using further the fact that $d_1^2 \sim n^\psi$ and $\lambda = o(d_1^2)$, we obtain

$$\frac{d_1^6}{(d_1^2+\lambda)^2}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|^2 \sim n^{2\tau+\psi}\ \ ,\ \ \frac{d_1^3}{d_1^2+\lambda}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\| \sim n^{\tau+\psi/2}.$$

If $2\tau+\psi>1$, then

$$\frac{\frac{d_1^3}{d_1^2+\lambda}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|}{\left(\sigma_e^2 n+\frac{d_1^6}{(d_1^2+\lambda)^2}\left\|Q^{(1)}Q^{(1)\prime}\beta\right\|^2\right)^{1/2}}\longrightarrow 1.$$

Finally, according to formula (3), $\hat{\rho}_g\to 1$.

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

In contrast, if $n^{2\tau+\psi}=o\left(\sum_{s=1}^r \frac{d_s^4}{(d_s^2+\lambda)^2}\right)$ then $\hat{\rho}_g\to 0$.

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

Using again Theorem 2, we have:

$$\hat{\rho}_g=\frac{\frac{d_r^3}{d_r^2+\lambda}\left\|Q^{(r)}Q^{(r)\prime}\beta\right\|}{\left(\sigma_e^2\sum_{s=1}^r\frac{d_s^4}{(d_s^2+\lambda)^2}+\frac{d_r^6}{(d_r^2+\lambda)^2}\left\|Q^{(r)}Q^{(r)\prime}\beta\right\|^2\right)^{1/2}}. \tag{4}$$

Recall that $d_r^2\sim n^\eta$ with $\eta<\psi\leq 1$. If we suppose moreover that $\lambda\sim Cn^{\kappa+\eta}$ with $\kappa>\max(0,-\eta)$ and $C>0$, then we have

$$\frac{d_r^3}{d_r^2+\lambda}=\frac{d_r}{1+\lambda/d_r^2}\sim\frac{1}{C}n^{\eta/2-\kappa}$$
$$\frac{d_r^3}{d_r^2+\lambda}\left\|Q^{(r)}Q^{(r)\prime}\beta\right\|\sim\frac{1}{C}n^{\tau+\eta/2-\kappa}.$$

It is obvious that $\hat{\rho}_g\to 0$ when $\tau+\eta/2-\kappa<0$. Indeed, at the denominator, since $d_1^2=o(n)$, we have $\sigma_e^2\frac{d_1^4}{(d_1^2+\lambda)^2}\sim\sigma_e^2$ which is bounded away from 0.

If $\tau+\eta/2-\kappa>0$, then we have to separate two different cases. If $\sum_{s=1}^r\frac{d_s^4}{(d_s^2+\lambda)^2}=o\left(n^{2\tau+\eta-2\kappa}\right)$, then $\hat{\rho}_g\to 1$.

In contrast, if $n^{2\tau+\eta-2\kappa}=o\left(\sum_{s=1}^r\frac{d_s^4}{(d_s^2+\lambda)^2}\right)$, then $\hat{\rho}_g\to 0$.

3. Proof of Theorem 3 of the main manuscript

Let us consider the following natural estimator of A_1:

$$\check{A}_1=\frac{1}{n_{new}}\sum_{s=1}^r\frac{d_s^2}{d_s^2+\lambda}\,\beta'\,X_{new}'X_{new}\,Q^{(s)}Q^{(s)\prime}\beta.$$

We have

$$\check{A}_1=\frac{1}{n_{new}}\sum_{s=1}^r\frac{d_s^2}{d_s^2+\lambda}\,\beta'\,ZF^2Z'\,Q^{(s)}Q^{(s)\prime}\beta$$
$$=\frac{1}{n_{new}}\sum_{s=1}^r\frac{d_s^2}{d_s^2+\lambda}\beta'\sum_{\alpha=1}^{r_{new}}f_\alpha^2Z^{(\alpha)}Z^{(\alpha)\prime}Q^{(s)}Q^{(s)\prime}\beta.$$

Further, a natural estimator of A_2 is

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

We can easily see that

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

which gives

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

A natural estimator of A_3 is:

$$\begin{aligned}\check{A}_3 &= \frac{1}{n_{new}} \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, Q^{(s)} Q^{(s)\prime} \beta\right)' X_{new}' X_{new} \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, Q^{(s)} Q^{(s)\prime} \beta\right) \\ &= \frac{1}{n_{new}} \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, X_{new} Q^{(s)} Q^{(s)\prime} \beta\right)' \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda}\, X_{new} Q^{(s)} Q^{(s)\prime} \beta\right).\end{aligned}$$

Using the fact that

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

we deduce

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

Consequently,

$$\begin{aligned}\check{A}_3 &= \frac{1}{n_{new}} \sum_{s=1}^{r} \sum_{\ell=1}^{r} \frac{d_s^2\, d_\ell^2}{(d_s^2+\lambda)(d_\ell^2+\lambda)} \sum_{\alpha=1}^{r_{new}} f_\alpha Q^{(s)\prime} Z^{(\alpha)} Q^{(s)\prime} \beta W^{(\alpha)\prime} \sum_{\vartheta=1}^{r_{new}} f_\vartheta Q^{(\ell)\prime} Z^{(\vartheta)} Q^{(\ell)\prime} \beta W^{(\vartheta)} \\ &= \frac{1}{n_{new}} \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} Q^{(s)\prime} \beta \sum_{\ell=1}^{r} \frac{d_\ell^2}{d_\ell^2+\lambda} Q^{(\ell)\prime} \beta \sum_{\alpha=1}^{r_{new}} f_\alpha^2 < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)}, Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} > .\end{aligned}$$

4. Proof of Lemma 4 of the main manuscript

To begin with, let us focus on the upper bound. First, we have to notice that we have the relationship

$$\check{A}_1 = \frac{1}{n_{new}} \sum_{\alpha=1}^{r_{new}} < f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} \beta \ , \ \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta > .$$

Then, applying two times the Cauchy-Schwartz inequality, we obtain

$$\begin{aligned}
\check{A}_1 &\leq \frac{1}{n_{new}} \sum_{\alpha=1}^{r_{new}} \left(\left\| f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} \beta \right\| \left\| \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta \right\| \right) \\
&\leq \frac{1}{n_{new}} \left(\sum_{\alpha=1}^{r_{new}} \left\| f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} \beta \right\|^2 \right)^{1/2} \left(\sum_{\alpha=1}^{r_{new}} \left\| \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} f_\alpha Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \right)^{1/2} \\
&= \check{A}_4^{1/2} \frac{1}{\sqrt{n_{new}}} \left(\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \left\| \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \right)^{1/2} .
\end{aligned}$$

We have

$$\begin{aligned}
&\sum_{\alpha=1}^{r_{new}} f_\alpha^2 \left\| \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \\
&= \sum_{\alpha=1}^{r_{new}} f_\alpha^2 \left(\sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} \beta' Q^{(s)} Q^{(s)\prime} Z^{(\alpha)} Z^{(\alpha)\prime} \right) \left(\sum_{\ell=1}^{r} \frac{d_\ell^2}{d_\ell^2+\lambda} Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} Q^{(\ell)\prime} \beta \right) \\
&= \sum_{s=1}^{r} \frac{d_s^2}{d_s^2+\lambda} \beta' Q^{(s)} \sum_{\ell=1}^{r} \frac{d_\ell^2}{d_\ell^2+\lambda} \beta' Q^{(\ell)} \sum_{\alpha=1}^{r_{new}} f_\alpha^2 < Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(s)} \ , \ Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} > \\
&= n_{new} \check{A}_3.
\end{aligned}$$

Thus

$$\check{A}_1 \leq \check{A}_4^{1/2} \check{A}_3^{1/2}.$$

Since $\check{A}_2 \geq 0$, we finally obtain

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

Let us now move on to the lower bound. We have the relationship:

$$\begin{aligned}
\check{A}_2 &\leq \frac{\sigma_e^2}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \quad \left(\sum_{s=1}^{r} \left\| \sum_{\alpha=1}^{r_{new}} Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2 \right) \\
&\leq \frac{r\ \sigma_e^2}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \max_{1\leq s\leq r} \left\| \sum_{\alpha=1}^{r_{new}} Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2 \\
&\leq \frac{r\ \sigma_e^2}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \max_{1\leq s\leq r} \sum_{\alpha=1}^{r_{new}} \left\| Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2 \\
&\leq \frac{r\ r_{new}\ \sigma_e^2}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \max_{s,\alpha} \left\| Q^{(s)\prime} Z^{(\alpha)} W^{(\alpha)} \right\|^2 .
\end{aligned}$$

Coming back to the expression of $\check{A}_2$, we also have:

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

We can notice that $\sum_{i=1}^{n_{new}} \left(\sum_{\omega=1}^{r_{new}} W_i^{(\omega)} \right)^2 = Tr(WW') = Tr(W'W) = r_{new}$.

As a consequence, another bound is the following

$$\check{A}_2 \leq \frac{\sigma_e^2\ r_{new}}{n_{new}} \sum_{s=1}^{r} \frac{d_s^2}{(d_s^2+\lambda)^2} \max_{1\leq\alpha\leq r_{new}} \left(f_\alpha^2 < Q^{(s)}, Z^{(\alpha)} >^2 \right).$$

On the other hand, we have

$$\begin{aligned}
\check{A}_1 &\geq \min_{1\leq s\leq r} \frac{d_s^2}{d_s^2+\lambda} \min_{1\leq\alpha\leq r_{new}} f_\alpha^2 \quad \sum_{s=1}^{r} \left(\sum_{\alpha=1}^{r_{new}} \beta'\ Z^{(\alpha)}\ Z^{(\alpha)\prime} Q^{(s)} Q^{(s)\prime} \beta \right) \\
&= \frac{1}{n_{new}} \min_{1\leq s\leq r} \frac{d_s^2}{d_s^2+\lambda} \min_{1\leq\alpha\leq r_{new}} f_\alpha^2 \quad \beta'\ Z\ Z'QQ'\beta \\
&= \frac{1}{n_{new}} \min_{1\leq s\leq r} \frac{d_s^2}{d_s^2+\lambda} \min_{1\leq\alpha\leq r_{new}} f_\alpha^2 \quad < Z\ Z'\beta\ ,\ QQ'\beta > .
\end{aligned}$$

Last,

$$\begin{aligned}
\check{A}_3 &\leq \frac{1}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^4}{(d_s^2+\lambda)^2} \max_{1\leq s\leq r} \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \sum_{s=1}^{r} \sum_{\ell=1}^{r} \sum_{\alpha=1}^{r_{new}} Q^{(s)\prime} Z^{(\alpha)} Z^{(\alpha)\prime} Q^{(\ell)} \\
&= \frac{1}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^4}{(d_s^2+\lambda)^2} \max_{1\leq s\leq r} \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \sum_{s=1}^{r} \sum_{\ell=1}^{r} Q^{(s)\prime} ZZ' Q^{(\ell)} \\
&\leq \frac{1}{n_{new}} \max_{1\leq s\leq r} \frac{d_s^4}{(d_s^2+\lambda)^2} \max_{1\leq s\leq r} \left\| Q^{(s)} Q^{(s)\prime} \beta \right\|^2 \max_{1\leq\alpha\leq r_{new}} f_\alpha^2 \\
&\quad \times \left\{ r \max_{1\leq s\leq r} \left\| ZZ' Q^{(s)} \right\|^2 \ +\ r(r-1) \max_{s\neq\ell} < ZZ' Q^{(s)}, ZZ' Q^{(\ell)} > \right\}.
\end{aligned}$$

Since ZZ' is an idempotent matrix and $Q^{(s)\prime}Q^{(s)} = 1$ for all $1 \leq s \leq r$, we have

$$\left\| ZZ'Q^{(s)} \right\|^2 \leq 1.$$

Besides, according to Cauchy-Schwartz inequality,

$$|< ZZ'Q^{(s)}, ZZ'Q^{(\ell)} >| \leq \left\| ZZ'Q^{(s)} \right\| \left\| ZZ'Q^{(\ell)} \right\| \leq 1.$$

Finally, since QQ' is an idempotent matrix, and putting together all the above considerings, we obtain

$$\check{A}_3 \leq \frac{r^2}{n_{new}} \max_{1 \leq s \leq r} \frac{d_s^4}{(d_s^2 + \lambda)^2} \left\| QQ'\beta \right\|^2 \max_{1 \leq \alpha \leq r_{new}} f_\alpha^2,$$

which finishes the proof.

5. Proof of Lemma 6 of the main manuscript

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

$$\tilde{A}_1 = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \beta' \, \Sigma \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta.$$

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

$$\hat{\tilde{A}}_1 := \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \beta' \, \hat{\Sigma} \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta,$$

where $\hat{\Sigma} = X'X/n$ is the empirical covariance matrix.

We have

$$\begin{aligned} \hat{\tilde{A}}_1 &= \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \beta' \, \hat{\Sigma} \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \\ &= \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2 + \lambda} \beta' Q D^2 Q' \, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta. \end{aligned}$$

It is easy to see that

$$QD^2Q' \, Q^{(\sigma(s))} = d_{\sigma(s)}^2 Q^{(\sigma(s))}.$$

Therefore,

$$\hat{\tilde{A}}_1 = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2 + \lambda} \beta' Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta = \frac{1}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{d_{\sigma(s)}^2 + \lambda} \left\| Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right\|^2.$$

Let us recall the expression $\tilde{A}_2$ given previously:

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

A natural estimation of $\tilde{A}_2$ is

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

Note that

$$D Q' Q^{(\sigma(s))} = d_{\sigma(s)} e_{\sigma(s)},$$

where $e_{\sigma(s)}$ denotes the $\sigma(s)$-th vector of the canonical basis of $\mathbb{R}^r$. As a result,

$$\begin{aligned}
\hat{\tilde{A}}_2 &= \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2+\lambda)^2} \mathrm{Tr}\left(P e_{\sigma(s)} e'_{\sigma(s)} P' \right) \\
&= \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2+\lambda)^2} \mathrm{Tr}\left(P' P e_{\sigma(s)} e'_{\sigma(s)} \right) \\
&= \frac{\sigma_e^2}{n} \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^4}{(d_{\sigma(s)}^2+\lambda)^2}.
\end{aligned}$$

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

$$\hat{\tilde{A}}_3 := \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2+\lambda}\, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right)' \hat{\Sigma} \left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2+\lambda}\, Q^{(\sigma(s))} Q^{(\sigma(s))\prime} \beta \right).$$

We have the following relations

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

As $X = PDQ'$, we have

$$XQ^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta = d_{\sigma(s)}Pe_{\sigma(s)}Q^{(\sigma(s))\prime}\beta = d_{\sigma(s)}P^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta$$

and thus

$$\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^2}{d_{\sigma(s)}^2+\lambda}\, XQ^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta = \sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^3}{d_{\sigma(s)}^2+\lambda} P^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta.$$

Last, we obtain

$$\begin{aligned}
\hat{\tilde{A}}_3 &= \frac{1}{n}\left(\sum_{\ell=1}^{\tilde{r}} \frac{d_{\sigma(\ell)}^3}{d_{\sigma(\ell)}^2+\lambda}\beta' Q^{(\sigma(\ell))}P^{(\sigma(\ell))\prime}\right)\left(\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^3}{d_{\sigma(s)}^2+\lambda} P^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta\right) \\
&= \frac{1}{n}\sum_{\ell=1}^{\tilde{r}} \frac{d_{\sigma(\ell)}^3}{d_{\sigma(\ell)}^2+\lambda}\sum_{s=1}^{\tilde{r}} \frac{d_{\sigma(s)}^3}{d_{\sigma(s)}^2+\lambda}\beta' Q^{(\sigma(\ell))}P^{(\sigma(\ell))\prime}P^{(\sigma(s))}Q^{(\sigma(s))\prime}\beta \\
&= \frac{1}{n}\sum_{\ell=1}^{\tilde{r}} \frac{d_{\sigma(\ell)}^6}{(d_{\sigma(\ell)}^2+\lambda)^2}\beta' Q^{(\sigma(\ell))}Q^{(\sigma(\ell))\prime}\beta \\
&= \frac{1}{n}\sum_{\ell=1}^{\tilde{r}} \frac{d_{\sigma(\ell)}^6}{(d_{\sigma(\ell)}^2+\lambda)^2}\left\|Q^{(\sigma(\ell))}Q^{(\sigma(\ell))\prime}\beta\right\|^2.
\end{aligned}$$

6. Proof of Lemma 9 of the main manuscript

Using Lemma 6 of the main manuscript and proceeding in the same way as in the proof of Lemma 2 of the main manuscript, we obtain

$$\begin{aligned}
n\hat{\tilde{A}}_1 &\sim \sum_{s\in\tilde{\Omega}_1} d_s^2\frac{n^{2\tau}}{r}, \\
n\hat{\tilde{A}}_2 + n\tilde{\hat{A}}_3 &\sim \sum_{s\in\tilde{\Omega}_1} d_s^2\frac{n^{2\tau}}{r}, \\
n\hat{\tilde{A}}_4 = n\hat{A}_4 &\sim \sum_{s\in\Omega_1} d_s^2\frac{n^{2\tau}}{r},
\end{aligned}$$

and the stated result follows.

References

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