Prediction in high dimensional linear models and application to genomic prediction with a sparse genetic map

Charles-Elie Rabier*, Simona Grusea♯
* IMAG, UMR 5149, CNRS, Université de Montpellier
♯ Institut de Mathématiques de Toulouse, Université de Toulouse, INSA

Summary

Genomic selection (GS) consists in selecting individuals on the basis of genomic predictions, using a large number of genetic markers. An important question in GS is to determine the number of markers required for a good prediction. In order to answer this question, we present here new statistical results regarding Ridge regression. We analyzed rice data from the Philippines and focused on the flowering time collected during the dry season 2012. Using different densities of markers, we show that at least 1553 markers are required to implement GS.

Causal model vs. Prediction model

Learning sample of size n

Causal model* (Q true regressors, with Q bounded)

$\theta^\star$ vector of effects, $M^\star$ matrix of measures, Y vector of phenotypes

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

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

Bayesian prediction model (K regressors, with $K >> n$)

θ vector of effects, M matrix of design

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

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

We assume that the prediction model
does not necessarily contain the true regressors
In other words, each column of $M^\star$ does not necessarily match a column of M

Validation sample + Accuracy criteria

— Let new denote an individual from the validation set

$$Y_{\text{new}} = m_{\text{new}}^{\star\prime}\theta^\star + e_{\text{new}} \qquad \text{where } e_{\text{new}} \sim N(0, \sigma_e^2)$$
and $m_{\text{new}}^\star$ vector of measures for the individual new

— Prediction of the continuous variable Y_{new}

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

⇒ Accuracy criteria

$$\rho = \frac{\text{Cov}\left(\hat{Y}_{\text{new}},\, Y_{\text{new}}\right)}{\sqrt{\mathbb{V}\left(\hat{Y}_{\text{new}}\right)\mathbb{V}\left(Y_{\text{new}}\right)}} \text{ with } m_{\text{new}} \text{ and } m_{\text{new}}^\star \text{ random, } M \text{ is known}$$

Component present in the breeder's equation (cf. Lynch and Walsh, 1998)

About the accuracy in the Ridge regression framework

The predictor is $\hat{Y}_{\text{new}} = m_{\text{new}}' \left(M'M + \lambda I_K\right)^{-1} M'Y$

Let us define :

$$A_1 := \theta^{\star\prime}\mathbb{E}\left(m_{\text{new}}^\star m_{\text{new}}'\right) M'V^{-1}M^\star\theta^\star \,,\; A_2 := \sigma_e^2\mathbb{E}\left(\left\|m_{\text{new}}'M'V^{-1}\right\|^2\right)$$
$$A_3 := \theta^{\star\prime}M^{\star\prime}V^{-1}M\mathbb{V}\left(m_{\text{new}}\right)M'V^{-1}M^\star\theta^\star \,,\; A_4 := \theta^{\star\prime}\mathbb{V}\left(m_{\text{new}}^\star\right)\theta^\star + \sigma_e^2$$

For the Ridge regression, we have

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

Under the oracle situation, we have $\hat{Y}_{\text{new}} = m_{\text{new}}^{\star\prime}\theta^\star$ and

$$\rho = \text{Cor}\left(m_{\text{new}}^{\star\prime}\theta^\star,\, Y_{\text{new}}\right) = \sqrt{\frac{\mathbb{V}\left(m_{\text{new}}^{\star\prime}\theta^\star\right)}{\mathbb{V}\left(Y_{\text{new}}\right)}} = h$$

In genetics, h^2 is called the heritability of the trait

Reference

— Rabier CE and Grusea S. **Prediction in high dimensional linear models and application to genomic selection** ..., *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, Vol 70(4), 2021

Results

Using the SVD decomposition of $M = UDW'$ and $M^\star = U^\star D^\star W^{\star\prime}$, an estimation of the accuracy for the Ridge regression is (under several conditions)

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

where

$$\widehat{A_1} = \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^2}{d_s^2+\lambda}\left\|U^{(s)}U^{(s)\prime}M^\star\theta^\star\right\|^2 \,,\; \widehat{A_2} = \frac{\sigma_e^2}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}$$
$$\widehat{A_3} = \frac{1}{n}\sum_{s=1}^{r}\frac{d_s^4}{(d_s^2+\lambda)^2}\left\|U^{(s)}U^{(s)\prime}M^\star\theta^\star\right\|^2 \,,\; \widehat{A_4} = \frac{1}{n}\sum_{s=1}^{r^\star} d_s^{\star 2}\left\|W^{\star(s)}W^{\star(s)\prime}\theta^\star\right\|^2 + \sigma_e^2$$

We discuss the link between
— the singular values $d_1^\star \geq d_2^\star \geq \ldots \geq d_{r^\star}^\star > 0$ and $d_1 \geq d_2 \geq \ldots \geq d_r > 0$
— the regularization parameter λ
— the projection of the signal $M^\star\theta^\star$ on the different subspaces

For that, we consider $\lambda \to +\infty$ with $\lambda = o\left(d_1^{\star 2}\right)$, and also the partitions $\Omega_1^\star, \Omega_2^\star, \Omega_3^\star$ of $\{1, \ldots, r^\star\}$ and $\Omega_1, \Omega_2, \Omega_3$ of $\{1, \ldots, r\}$, such as

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

and several technical conditions

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

to show the convergence to the oracle accuracy. For large n, $\hat{\rho} \sim \sqrt{\xi(n)}\,h$. $1-\xi(n)$ can be viewed as a loss coefficient : it is the percentage of the L^2 norm of $U^{\star(\ell)}$ that is unable to be captured

Applications on simulated data and on rice real data

Is $\hat{\rho}$ a good proxy for the accuracy ρ ?

We need to estimate the nuisance parameters $\theta^\star$ and $M^\star$
⇒ we consider more markers for TRN than for TST

Comparison between :
— new proxy with more markers for TRN than for TST (imperfect linkage disequilibrium)
— old proxy with the same number of markers for TRN and TST (complete linkage)

Simulated data (panmictic population)
A small example with 1,000 markers for TRN and 500 markers for TST

Method	50 generations	70 generations	100 generations	$\overline{\text{MSE}}$
Emp. Acc.	0.3909	0.3772	0.3217	
$\hat{\rho}(\hat{M}^\star, \hat{\theta}^\star_{LASSO})$	0.3397 (0.0112)	0.3436 (0.0132)	0.2629 (0.0146)	0.0130
$\hat{\rho}(\hat{M}^\star, \hat{\theta}^\star_{GPLASSO})$	0.2413 (0.0334)	0.3059 (0.0179)	0.2178 (0.0228)	0.0247
$\hat{\rho}(\hat{M}^\star, \hat{\theta}^\star_{ADLASSO})$	0.4677 (0.01293)	0.4821 (0.0222)	0.4093 (0.0164)	0.0172
$\hat{\rho}^{pLD}(\hat{\theta}_{ADLASSO})$	0.2970 (0.0336)	0.3182 (0.0306)	0.0986 (0.0693)	0.0445

Rice flowering time (data from Spindel et al, Plos Genetics 2015)
— K = 73,147 for TRN
— 4 densities of markers for TST (448, 781, 1553 and 3076)
— 252 TRN, 63 TST (i.e. 80% and 20%) + 100 samplings

Number of markers (SNPs) required for predicting the TST individuals

Method	448 SNPs	781 SNPs	1553 SNPs	3076 SNPs	$\overline{\text{MSE}}$
Emp. Acc.	0.4789	0.4919	0.5275	0.5242	
$\hat{\rho}(\hat{M}^\star, \hat{\theta}^\star_{LASSO})$	0.4621 (0.0244)	0.4653 (0.0226)	0.4737 (0.0254)	0.4728 (0.0263)	0.0247
$\hat{\rho}(\hat{M}^\star, \hat{\theta}^\star_{ADLASSO})$	0.4269 (0.0355)	0.4379 (0.0376)	0.4520 (0.0419)	0.4461 (0.0430)	0.0395
$\hat{\rho}^{pLD}(\hat{\theta}_{ADLASSO})$	0.3662 (0.0454)	0.4202 (0.0281)	0.4919 (0.0215)	0.4952 (0.0342)	0.0323