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

Charles-Elie Rabier[1,2], Brigitte Mangin[3] and Simona Grusea[4]

[1] Institut des Sciences de l'Evolution (ISEM), Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France

[2] Institut Alexander Grothendieck Montpellier Institute (IMAG), Université de Montpellier, CNRS, France

[3] Laboratoire des Intéractions Plantes Microorganismes (LIPM), Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

[4] Institut de Mathématiques de Toulouse (IMT), Université de Toulouse, INSA de Toulouse, France

Corresponding author: charles-elie.rabier@umontpellier.fr

Reference paper:* Rabier *et al.* (2019). On the accuracy in high dimensional linear models and its application to genomic selection. *Scandinavian Journal of Statistics, 46(1).
https://onlinelibrary.wiley.com/doi/full/10.1111/sjos.12352

Keywords: Genomic Selection, Prediction Accuracy, High Dimension, Ridge Regression, Rice data

For many years, geneticists focused on linkage analysis (LA) in order to detect on a given chromosome a Quantitative Trait Locus, so-called QTL: a QTL is a section of the DNA that contains one or more genes influencing a quantitative trait which is able to be measured. In this context, the most popular statistical method was Interval Mapping (Lander and Botstein, 1989). It consists in performing statistical tests along the genome. Using the information brought by genetic markers, the presence of a QTL is tested at every location in the genome. Later, geneticists moved on to 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. age-related macular degeneration, Fritsche et al., 2016, autisum spectrum disorder, Connolly et al., 2017). 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 governed by a large number of small-effect QTLs (Goddard and Hayes, 2009, Buckler et al., 2009). It turns out that predictions based on selected SNPs could not be considered as reliable.

Today, Genomic Selection (GS), motivated by the seminal paper of Hayes 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, thanks to the recent progress in molecular biology. The goal is not to detect QTLs 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). GS was first applied to animal breeding (see Hayes et al, 2009) and GS is nowadays extensively investigated in plants. For instance, we can mention studies on apple (Muranty et al. (2015)), eucalyptus (Tan et al. (2017)), japanese pears (Minamikawa et al. (2018)), strawberry (Gezan et al. (2017)), banana (Nyine et al. (2018)) and coffea (Ferrao et al. (2018)).

In GS, the quality of the prediction is evaluated according to some accuracy criteria, i.e. the correlation between predicted and true values. This criteria 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). One of the most popular methods, for prediction of breeding values, is Ridge regression. In genetics, this regression model, initially proposed by Hayes et al. (2001) and Whittaker et al. (2000), is called random regression best linear unbiased predictor (RRBLUP) or genomic best linear unbiased predictor (GBLUP). We focus here on some predictive aspects of Ridge regression and present theoretical results regarding the accuracy criteria. 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. On simulated data, proxies built on our theoretical results outperformed existing proxies in GS, built on Daetwyler et al. (2008)'s seminal formula. Next, we will discuss on how to improve the prediction, using a "modified" predictor derived from Ridge regression. Finally, a real data analysis is proposed; it relies on the paper of Spindel et al. (2015) dealing with GS in rice.