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

Charles-Elie Rabier*†[1] and Simona Grusea[2]

[1]IMAG – Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier – France
[2]IMT – Institut Mathématiques de Toulouse, Institut National des Sciences Appliquées - Toulouse – France

Abstract

This poster is a summary of the article by Rabier and Grusea (JRSS C, 2021) on L2 regularization in genomics.
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. We present here new statistical results regarding Ridge regression. We show that the projection of the regression function on the space spanned by column space of the design matrix is a key element for the accuracy of the prediction. Besides, the "oracle accuracy" is reached as soon as the limit of a loss factor, due to the number of markers and their locations, is equal to zero. We also introduce a modified predictor to improve performances of the Ridge estimator. Last, we analyze rice data from Los Banos, Philippines and focus 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.

Keywords: Ridge regression, Prediction, Genomics, Accuracy

*Speaker
†Corresponding author: charles-elie.rabier@umontpellier.fr