Incorporating linkage disequilibrium blocks in Genome-Wide Association Studies

Alia Dehman, Christophe Ambroise, Pierre Neuvial

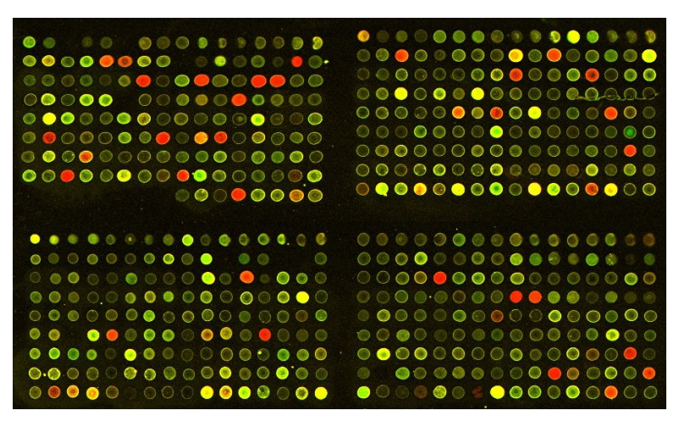
{alia.dehman, christophe.ambroise,pierre.neuvial}@genopole.cnrs.fr

Abstract

In genome-wide association studies, we are interested in finding genetic markers that are significantly associated with a phenotype of interest. Whole-genome single nucleotide polymorphism (SNP) data are collected for many thousands of SNP markers, leading to high-dimensional regression problems where the number of predictors greatly exceeds the number of observations. Moreover, these predictors are highly dependent, in particular due to linkage disequilibrium (LD).

We propose a two-step approach that explicitly takes advantage of the grouping structure induced by LD. In the first step, we infer LD blocks by performing a clustering of LD estimates with an adjacency constraint. In the second step, we perform Group Lasso regression on the inferred LD blocks.

GWA studies



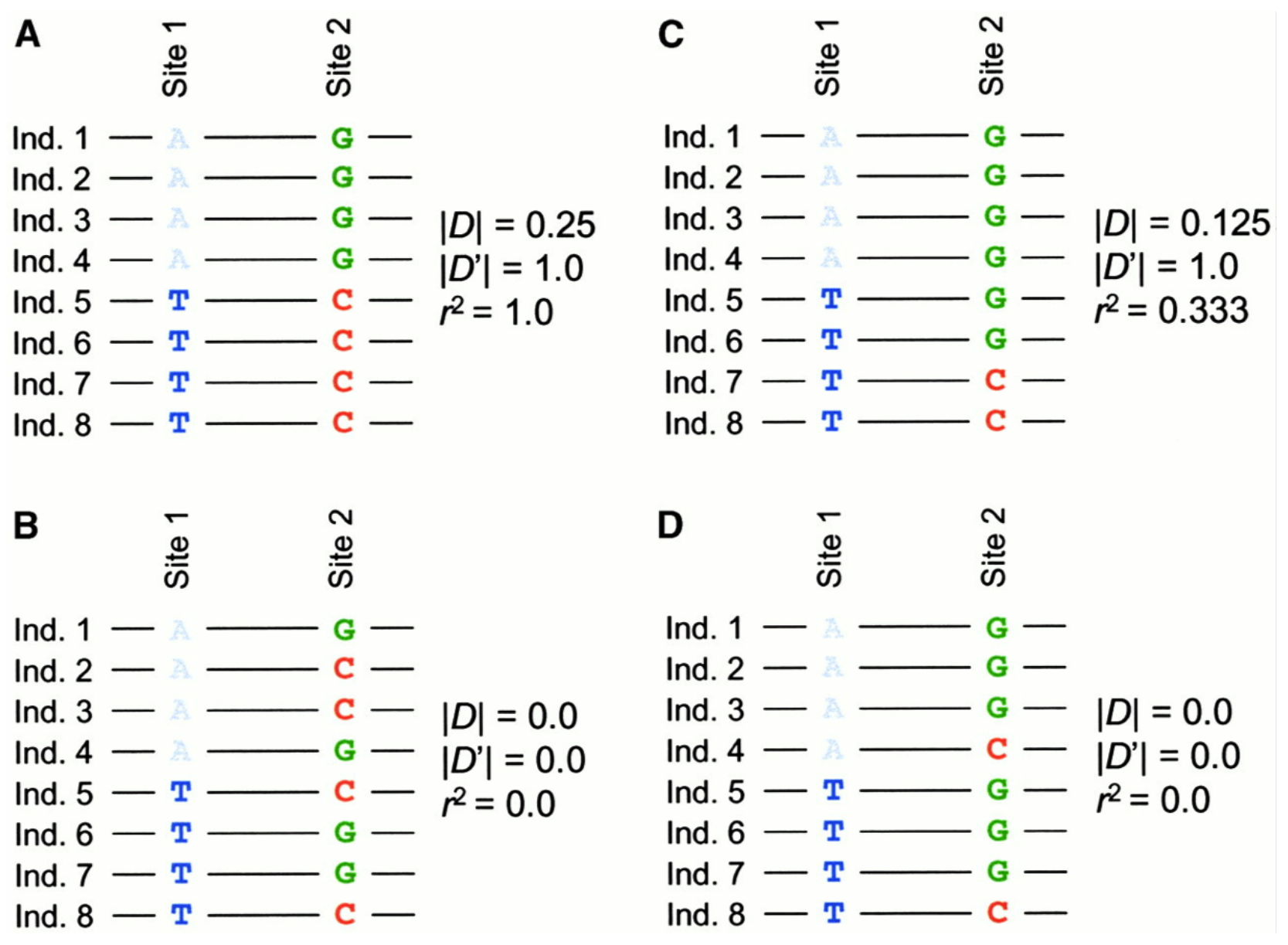
10^6 SNP can be genotyped in a single experiment

Design matrix with $n \ll p$

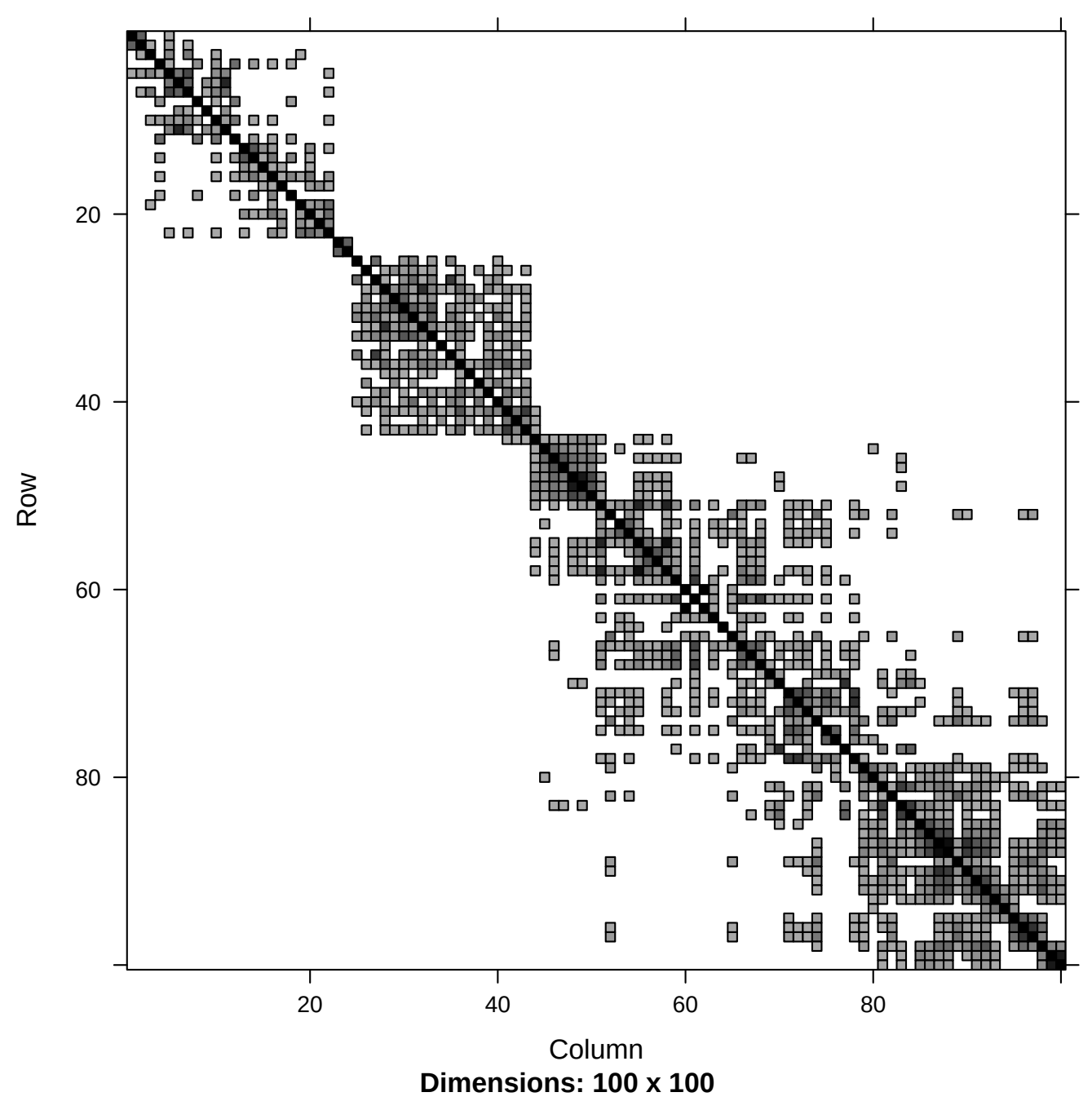
genotype data of p SNP are simultaneously collected for n patients

$$\begin{pmatrix} x_1^1 & x_1^2 & \dots & x_1^p \\ \vdots & & & \\ x_n^1 & x_n^2 & \dots & x_n^p \end{pmatrix}$$

Linkage Disequilibrium



Two SNP sites typed from eight individuals (Gaut et. al. 2003).



r^2 coefficients among the **first 100 SNP** of Chromosome 6 in Dalmasso et al (2008).

LD-level inference

The problem of SNP selection is ill-posed

- **biologically**: associated SNP may not be genotyped
- **statistically**: strong dependence between SNP (due to LD) raises an identifiability problem.

State of the art: use of tag SNPs

Our proposal: selecting LD blocks associated with the phenotype.

A two-step approach

Inference of blocks (from X only)

- A $p \times p$ matrix of LD pairwise measures is calculated.
- Ward's Hierarchical Clustering with an adjacency constraint (R package `rioja`)

Selection of associated blocks

- The Group Lasso: well-adapted to group-structured variables:

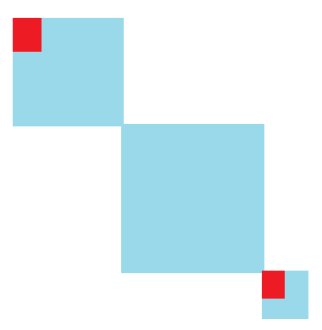
$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} (||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}_{\mathbf{g}}||_2).$$

Simulation study

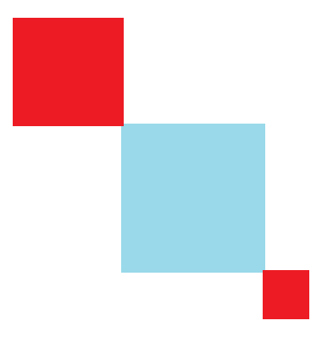
Parameters

- $n = 200$, $p = 512$, $K = 9$ groups of sizes $(2, 2, 4, 8, 16, 32, 64, 128, 256)$.
- The first 2 SNPs of groups of sizes $2, 2, 4, 8$ are associated with the phenotype.
- $cov(X_{.j}, X_{.j'}) = \rho \ \mathbf{1}_{j=j'}$.
- Coefficient of determination: $R^2 = 0.2$.

Definition of associated SNPs



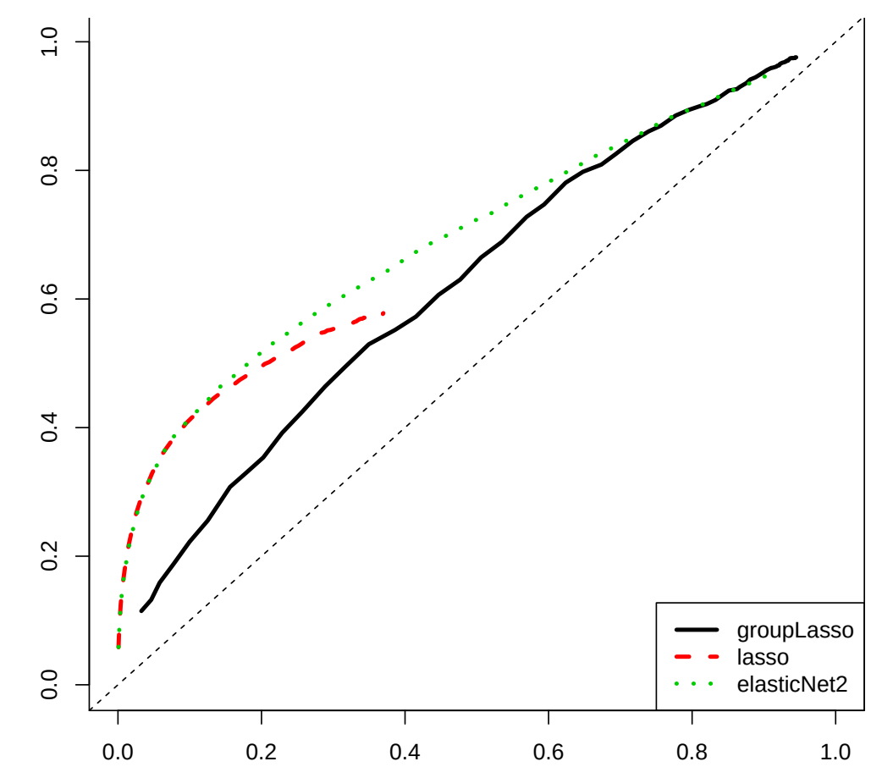
SNP-level



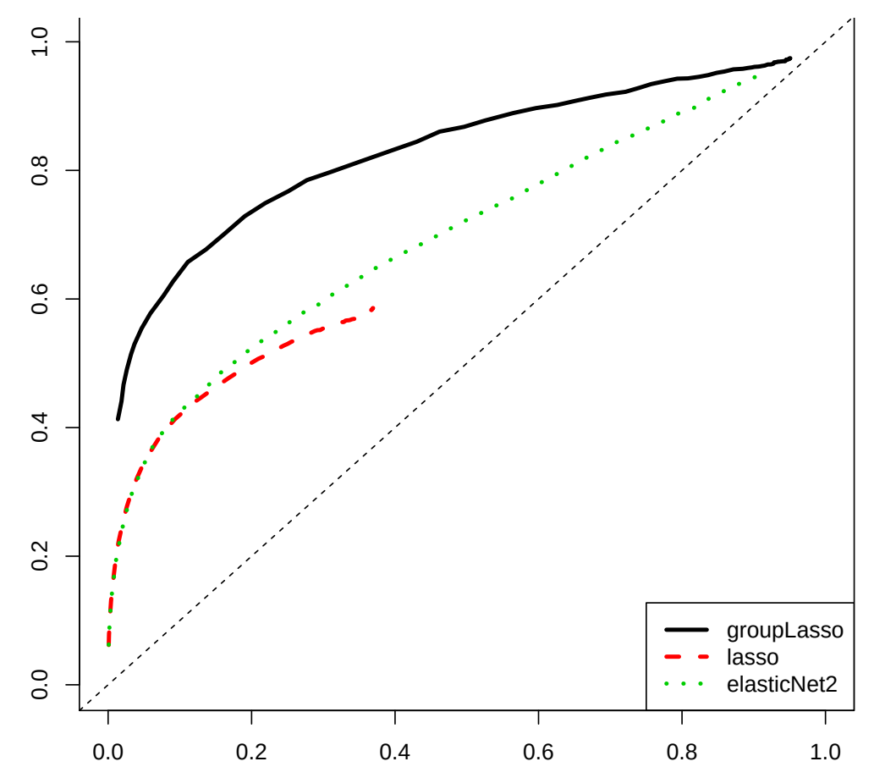
Block-level

Known true number of clusters

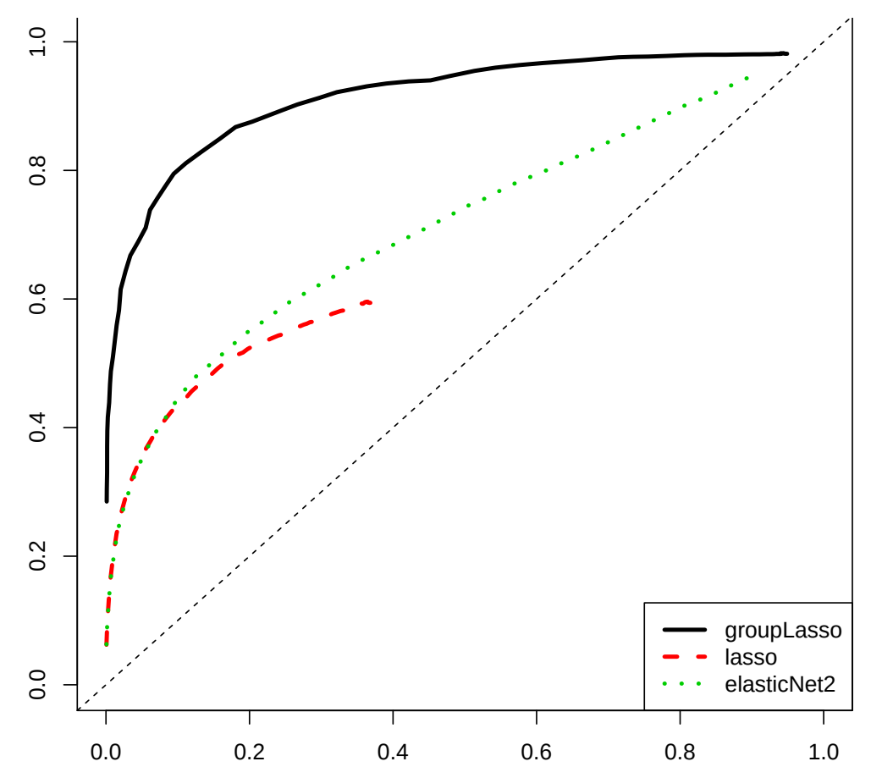
The proposed method is well-adapted to LD-structured data



$\rho = 0$, block-level

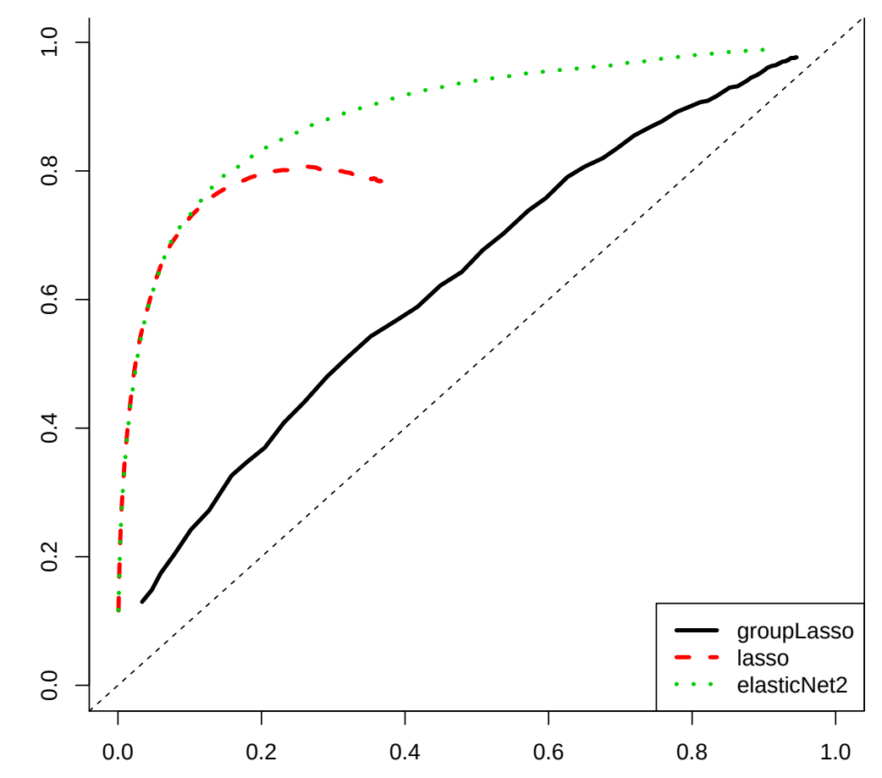


$\rho = 0.1$, block-level

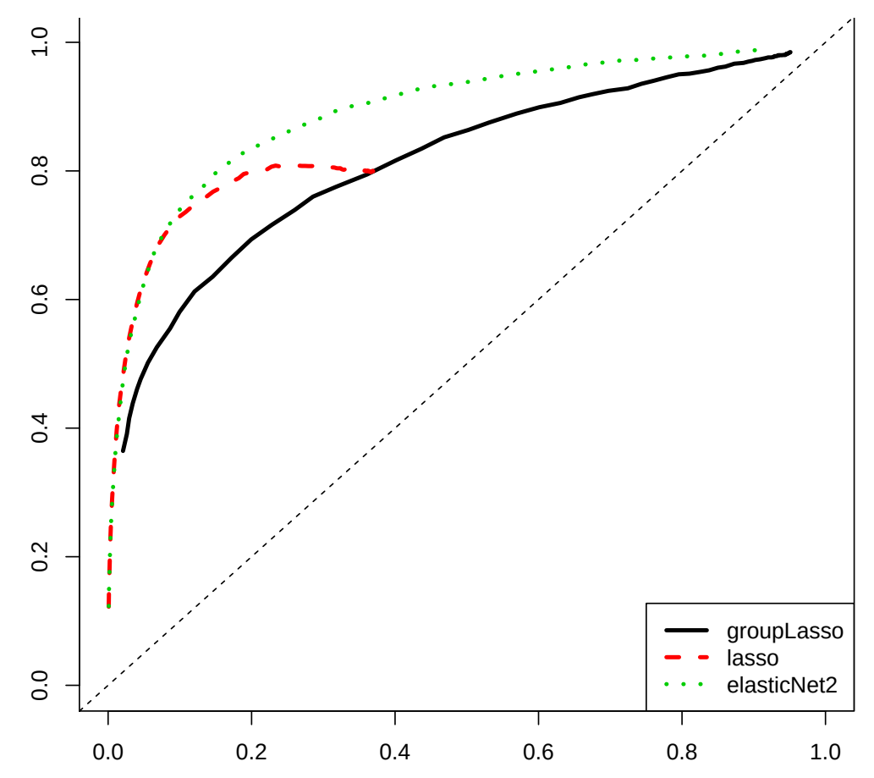


$\rho = 0.2$, block-level

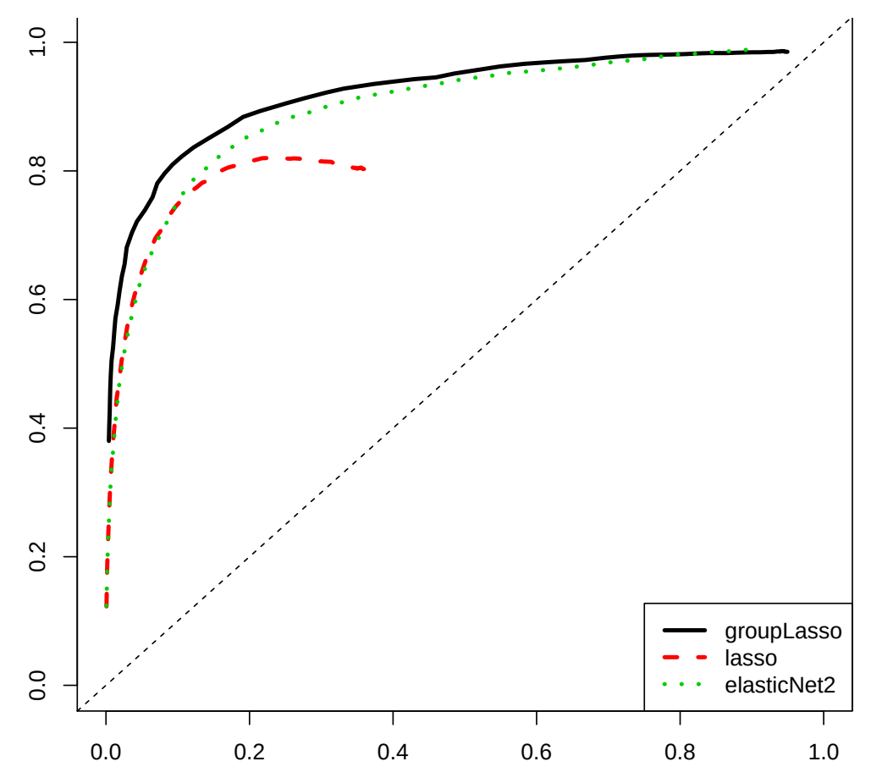
Lasso and Elastic-Net outperformed at the SNP-level for $\rho \geq 0.2$



$\rho = 0$, SNP-level



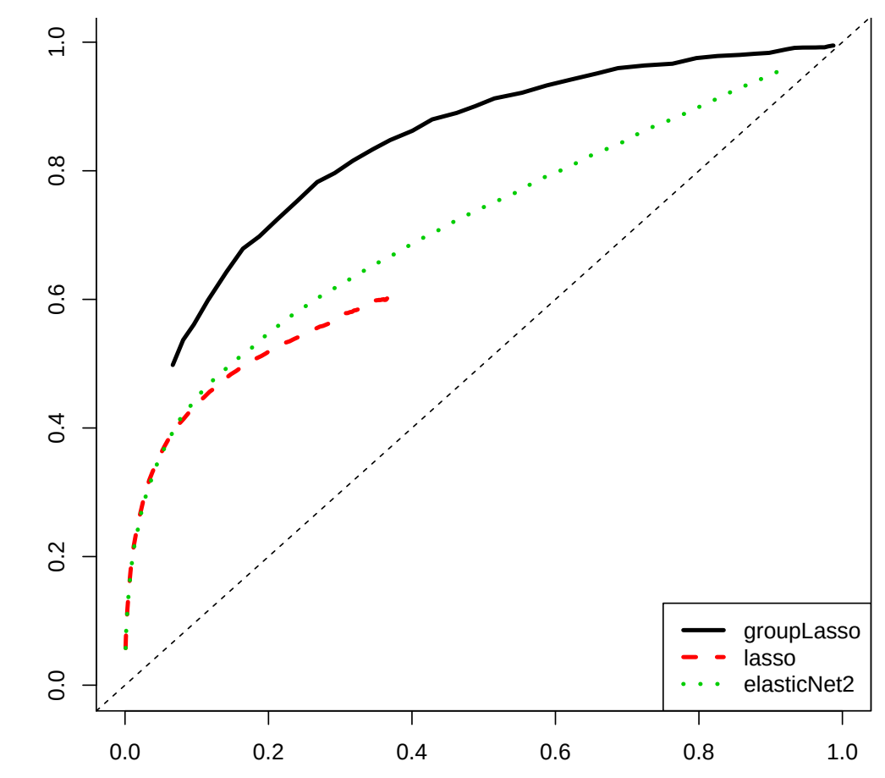
$\rho = 0.1$, SNP-level



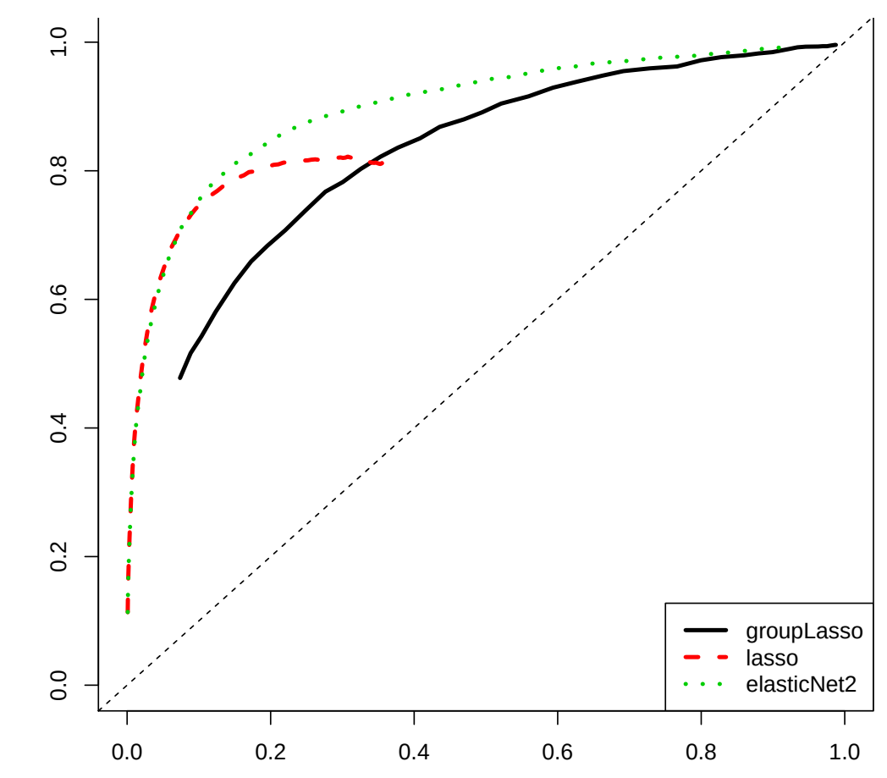
$\rho = 0.2$, SNP-level

Misspecified number of clusters

Forcing 5 groups to be selected when K=9 ($\rho = 0.2$)



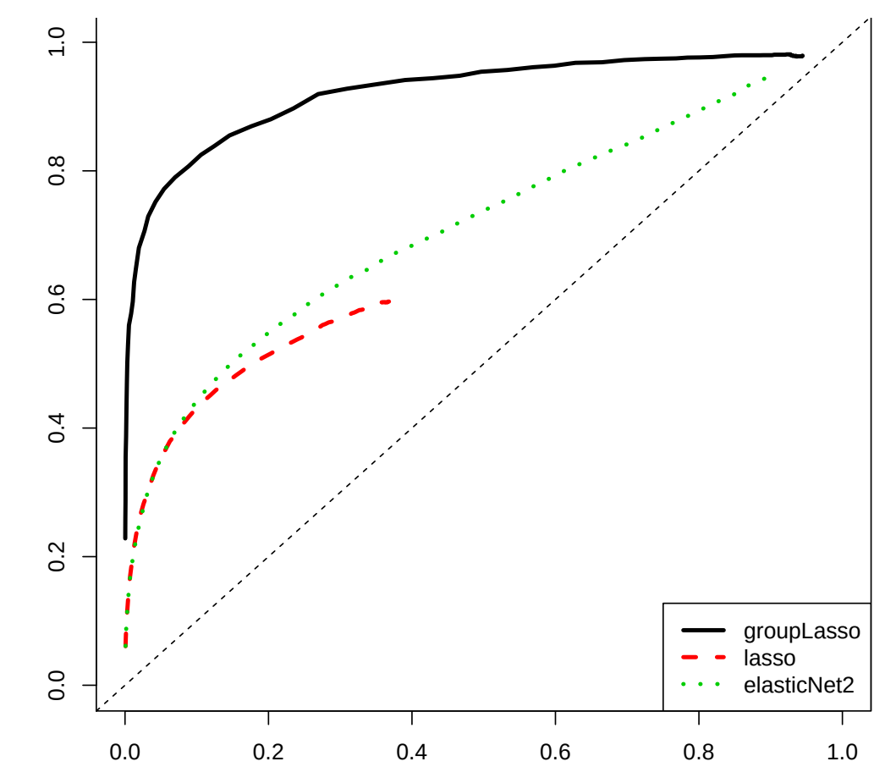
block-level



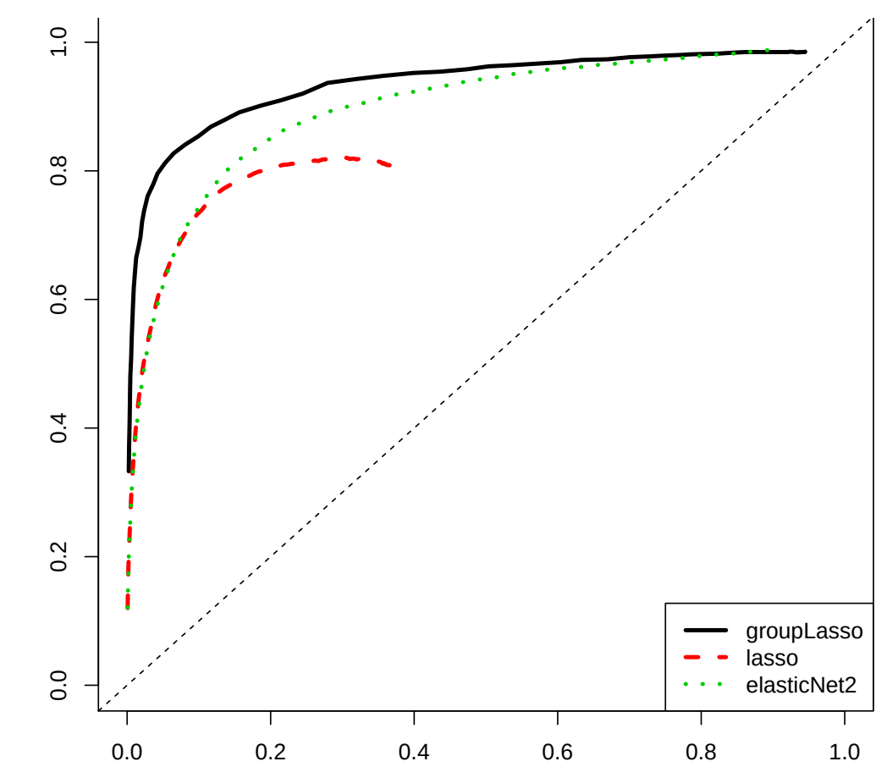
SNP-level

The Group Lasso makes errors by canceling or activating too large groups

Forcing 13 groups to be selected when K=9 ($\rho = 0.2$)



block-level



SNP-level

The Group Lasso can activate the right small blocks among the ones clustered