Incorporating linkage disequilibrium blocks in Genome-Wide Association Studies

Alia Dehman
Christophe Ambroise
Pierre Neuvial

Laboratoire Statistique et Génome

July 2^{nd}, 2013

1. Genome-Wide Association Studies
    - The regression model
    - Sparsity and high-dimension contexts
    - Biological context : LD

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods
3. Results
   - True number of clusters
   - Misspecified number of clusters

1. Genome-Wide Association Studies
    - The regression model
    - Sparsity and high-dimension contexts
    - Biological context : LD
2. Taking the group structure into account
    - Classical approach
    - A Two-Step Approach
    - Competing methods
3. Results
    - True number of clusters
    - Misspecified number of clusters
4. Current works

Outline

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods
3. Results
   - True number of clusters
   - Misspecified number of clusters
4. Current works

The regression model

- To identify genetic markers that are significantly associated with a phenotype of interest.

- **Phenotypic trait :** qualitative or quantitative
  Genetic markers : Single Nucleotide Polymorphisms (SNP)

- **The regression model**

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i \ \ , i = 1, \dots, n$$

  - n : number of individuals
  - p : number of covariates
  - Y_i : response for the individual i
  - $X_{.j}$: observations for covariate j (coded in 0, 1 or 2)

Sparsity and high-dimension contexts

Sparsity : Only a subset of SNPs is significantly associated with the phenotype.

$$Card\{j, \beta_j \neq 0\} \ll p$$

High-dimension : Many thousands of markers vs a few hundred observations.

$$p \gg n$$

The LD measures

Linkage Disequilibrium (or Gametic Disequilibrium) : Is the non-random association of alleles at two or more loci. Its amount depends on the difference between observed allelic frequencies and those expected from a homogenous, randomly distributed model.

- Z_j the indicator of the presence of minor allele for SNP j.
- $Z_j \sim \mathcal{B}(p_j)$

$$D(j,k) = cov(Z_j, Z_k)$$
$$r^2(j,k) = corr(Z_j, Z_k)$$

How to estimate it ?

snp	vv	vV	VV
uu	a	b	c
uU	d	e	f
UU	g	h	i

snp	v	V
u	α	β
U	γ	δ

Only the genotype data table is observed

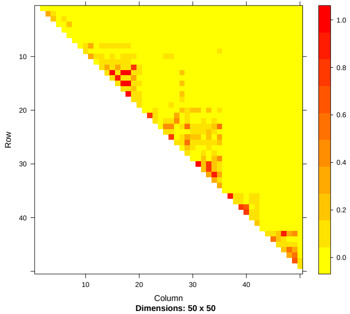
- α, β, γ, δ are estimated
- a system of equations. e.g : $\alpha = 2a + b + d + pe$

with p the « probability » of the haplotype (uv, UV).

$\Rightarrow$ estimating p, then (α, β, γ, δ) and finally $D = p_{UV} - p_U p_V$.

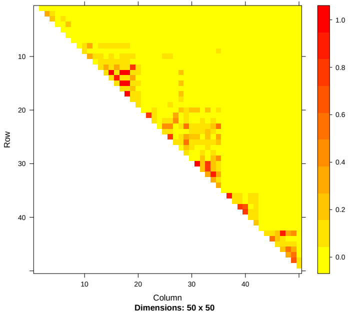
The LD-block structure

- the r^2 coefficients among the **50 first SNP** of the Chromosome 22 (Dalmasso et al. 2008)



The LD-block structure

- the r^2 coefficients among the **50 first SNP** of the Chromosome 22 (Dalmasso et al. 2008)
- LD structured in blocks



Outline

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods
3. Results
   - True number of clusters
   - Misspecified number of clusters
4. Current works

Classical approach : tag-SNP

To deal with high-dimensional problems and dependence among SNP :

- based on LD
- selection of « representative » SNP of each LD-block : *tagging*

⚠ **Loss of information**

Loss of power : tag-SNP not necessarly the causal SNP.

A different approach :

- a block-selection

A Two-Step Approach

Inference of blocks

- only the genotype data $\mathbf{X}$ are used.
- a $p \times p$ matrix LD pairwise measures is calculated.
- Ward Constrained Hierarchichal Clustering (R package `rioja`)

Selection of blocks associated with phenotype

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

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

Competing methods

Lasso

$$\hat{\boldsymbol{\beta}}^{l1} = \arg\min_{\beta} \sum_{i} (y_i - \mathbf{X}_{i.}\boldsymbol{\beta})^2 + \lambda ||\boldsymbol{\beta}||_1,$$

Elastic-Net

$$\hat{\boldsymbol{\beta}}^{EN} = \arg\min_{\beta} \sum_{i} (y_i - \mathbf{X}_{i.}\boldsymbol{\beta})^2 + \lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2^2,$$

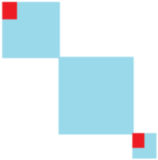
with λ, λ_1 and λ_2 three regularization parameters.
(R package `quadrupen`)

Evaluation

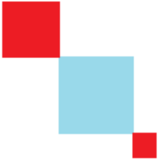
Parameters

- $n = 200$, $p = 512$, $K = 9$ 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

Outline

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods
3. Results
   - True number of clusters
   - Misspecified number of clusters
4. Current works

True number of clusters

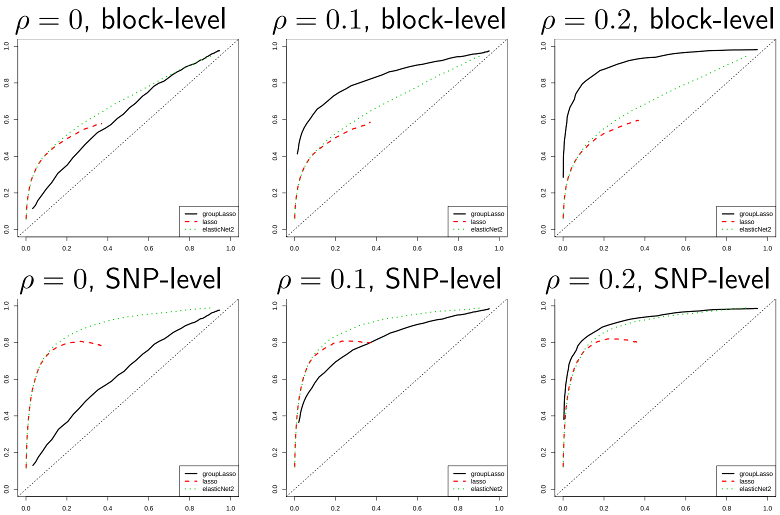


Figure: The number of clusters is set to 9.

Misspecified number of clusters : too few

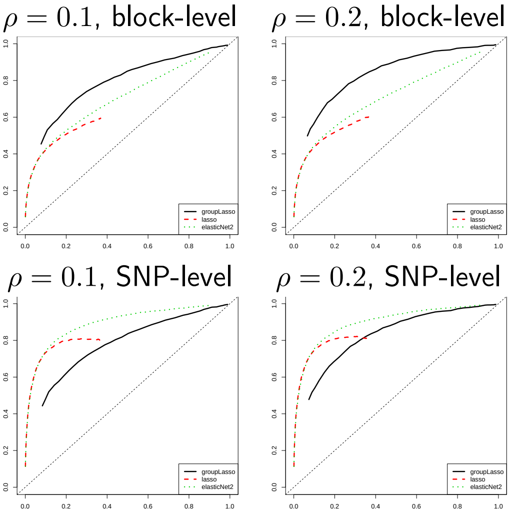
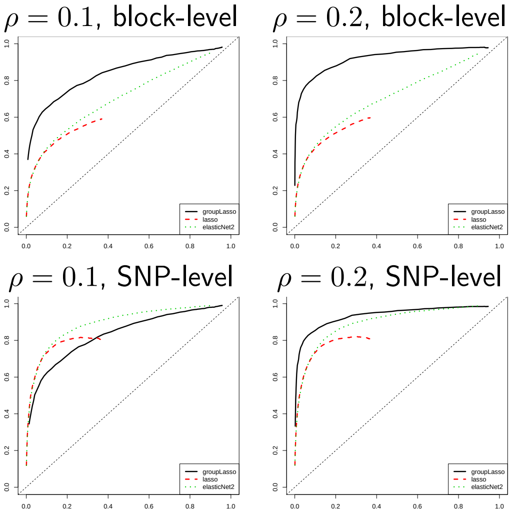


Figure: The number of clusters is set to 5.

Misspecified number of clusters : too many



$\rho = 0.1$, block-level

$\rho = 0.2$, block-level

$\rho = 0.1$, SNP-level

$\rho = 0.2$, SNP-level

Figure: The number of clusters is set to 13.

Outline

1. Genome-Wide Association Studies
   - The regression model
   - Sparsity and high-dimension contexts
   - Biological context : LD
2. Taking the group structure into account
   - Classical approach
   - A Two-Step Approach
   - Competing methods
3. Results
   - True number of clusters
   - Misspecified number of clusters
4. Current works

Memory requirement of the clustering

Ward Constrained Hierarchical Clustering

$$d(A,B) = \frac{n_A n_B}{n_A + n_B}\left(\frac{1}{n_A^2}S_{A,A} + \frac{1}{n_B^2}S_{B,B} - \frac{2}{n_A n_B}S_{A,B}\right)$$

	rioja	**cWard**
Type of entry	$p \times p$ dissimilarity matrix	the $n \times p$ design matrix
Time complexity	$\mathcal{O}(np^2)$	$\mathcal{O}(np^2)$
Memory complexity	$\mathcal{O}(p^2)$	$\mathcal{O}(np)$

Automatic model selection

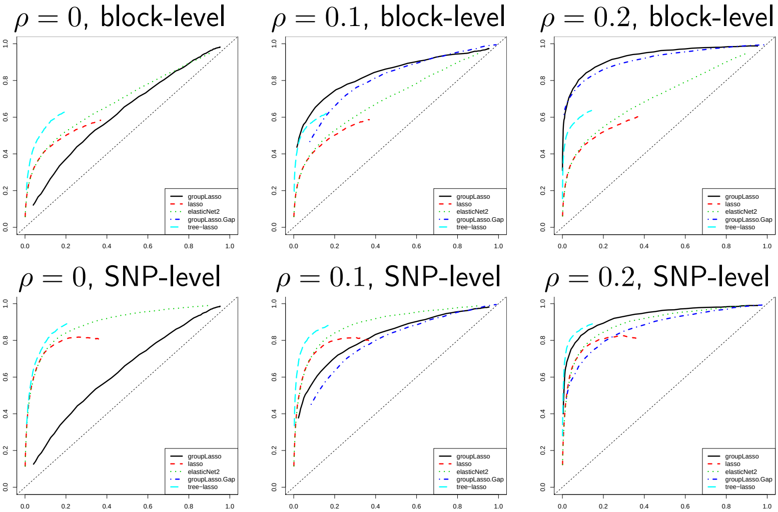
Inferring the number of clusters :

- maximal gap (Bühlmann et. al., 2012, arXiv :1209.5908v1)
- BIC criterion
- Gap Statistic (Tibshirani et. al., 2001, JRSSB)

Tree-Group Lasso

$$\hat{\boldsymbol{\beta}}^{Tree} = \arg\min_{\beta} \sum_i (y_i - \mathbf{X}_{i.}\boldsymbol{\beta})^2 + \lambda \sum_{h=0}^{d} \sum_{g=1}^{G_h} \omega_g^h ||\boldsymbol{\beta}_{\boldsymbol{g}}^{\boldsymbol{h}}||_2.$$

Automatic model selection


$\rho = 0$, block-level

$\rho = 0.1$, block-level

$\rho = 0.2$, block-level

$\rho = 0$, SNP-level

$\rho = 0.1$, SNP-level

$\rho = 0.2$, SNP-level

Figure: $\rho = 0 : 1$ cluster, $\rho = 0.1 : 5$ clusters, $\rho = 0.2 : 6$ clusters

Thank you for your attention !