

# Incorporating linkage disequilibrium blocks in Genome-Wide Association Studies

Alia Dehman

Director: Christophe Ambroise

Co-director: Pierre Neuvial

Laboratoire "Statistique et Génome"

"Des Génomes Aux Organismes"

June 7<sup>th</sup>, 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
  - Under- and over-estimated clustering

- 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
  - Under- and over-estimated clustering
- 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
  - Under- and over-estimated clustering
- 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.

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

**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.

- $Z_j^m$  the indicator of the presence of minor allele for SNP  $j$  on maternal copy.
- $Z_j^p$  the indicator of the presence of minor allele for SNP  $j$  on paternal copy.
- $Z_j \sim \mathcal{B}(p_j)$

$$D(j, k) = \text{cov}(Z_j, Z_k)$$

$$r^2(j, k) = \text{corr}(Z_j, Z_k)$$

## How to estimate it?

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

snp	v	V
u	$\alpha$	$\beta$
U	$\gamma$	$\delta$



Only the genotype data table is observed :  $\mathbf{X}_{.j} = Z_j^m + Z_j^p$ .

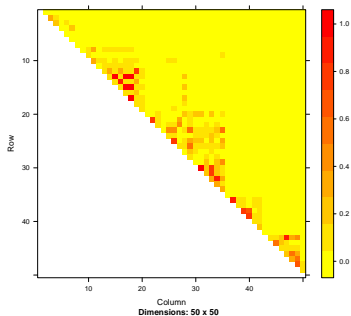
- $\alpha, \beta, \gamma, \delta$  are estimated
- a system of equations. exmpl :  $\alpha = 2a + b + d + pe$

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

$\Rightarrow$  estimating  $p$ , then  $(\alpha, \beta, \gamma, \delta)$  and finally  $D$ .

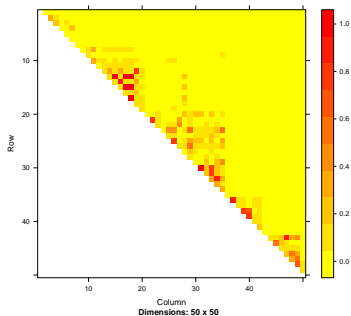
# The LD-block structure

- 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
  - Under- and over-estimated clustering
- 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* (Chloé Friguet presentation)



**Loss of information**

**Loss of power** : tag-SNP not necessarily 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 Hierarchical Clustering (*R* package rioja)

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

## Selection of blocks associated with phenotype :

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

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

# Competing methods

**Lasso :**

$$\hat{\beta}^{l1} = \arg \min_{\beta} \sum_i (y_i - \mathbf{X}_i \cdot \beta)^2 + \lambda \|\beta\|_1,$$

**Elastic-Net :**

$$\hat{\beta}^{EN} = \arg \min_{\beta} \sum_i (y_i - \mathbf{X}_i \cdot \beta)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2,$$

with  $\lambda$ ,  $\lambda_1$  and  $\lambda_2$  three regularization parameters.



SNP-selection & block-selection :

- **SNP-level evaluation** : causal SNP.
- **block-level evaluation** : SNP in the same block that a causal SNP.

# Simulation study

- $n = 200$ ,  $p = 512$ .
- 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.
- If  $j \neq j'$  are in the same group,  $cov(X_{.j}, X_{.j'}) = \rho$  else  $cov(X_{.j}, X_{.j'}) = 0$ .
- $R^2 = 0.2$  the coefficient of determination.

# 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
  - Under- and over-estimated clustering
- 4 Current works

# True number of clusters

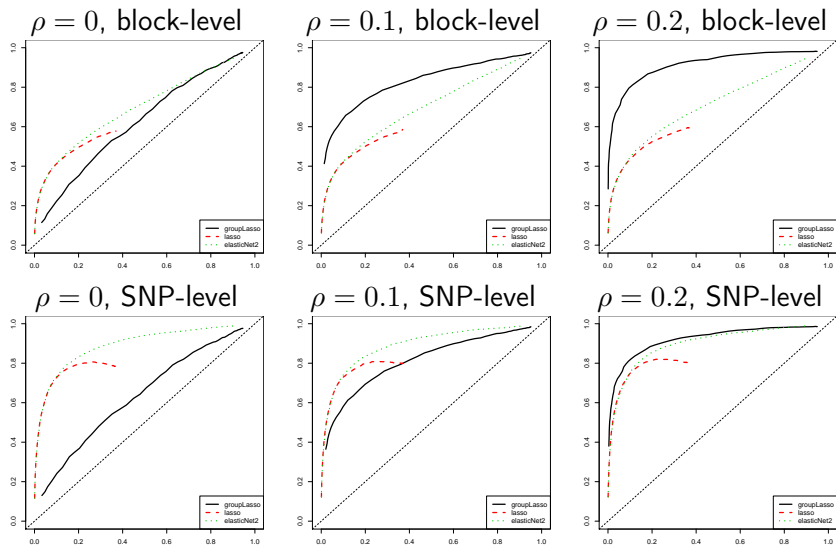


Figure: The number of clusters is set to 9.

# Under-estimated clustering

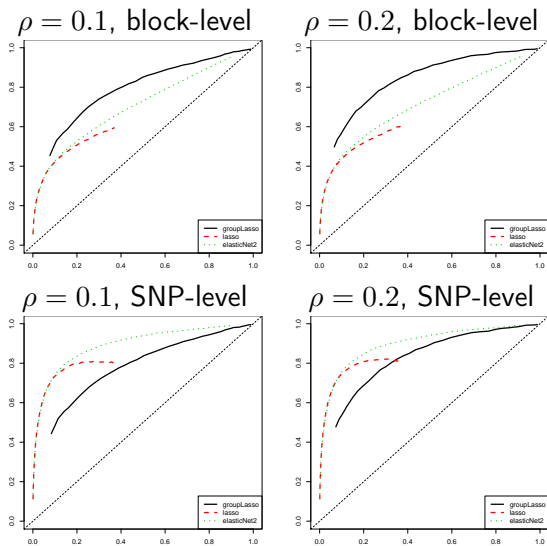


Figure: The number of clusters is set to 5.

# Over-estimated clustering

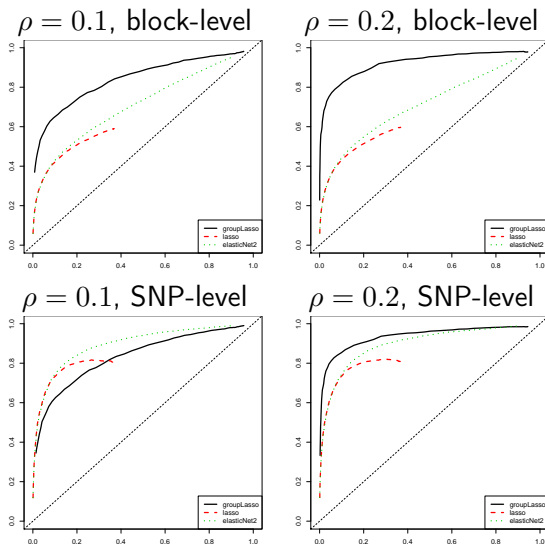


Figure: The number of clusters is set to 13.

# 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
  - Under- and over-estimated clustering
- 4 Current works

## Memory requirement of the clustering

	<b>rioja</b>	<b>cWard</b>
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	$\mathcal{O}(p^2)$	$\mathcal{O}(np)$



## Inferring the number of clusters :

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

## Tree-Group Lasso

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

# Automatic model selection

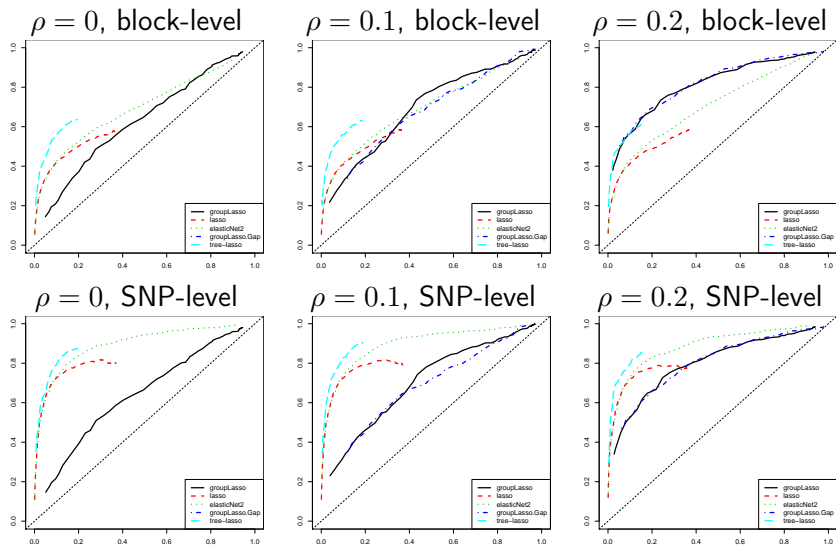


Figure: 1 cluster is inferred for  $\rho = 0$ , 5 clusters for  $\rho = 0.1$  and 6 clusters for  $\rho = 0.2$

# Current works

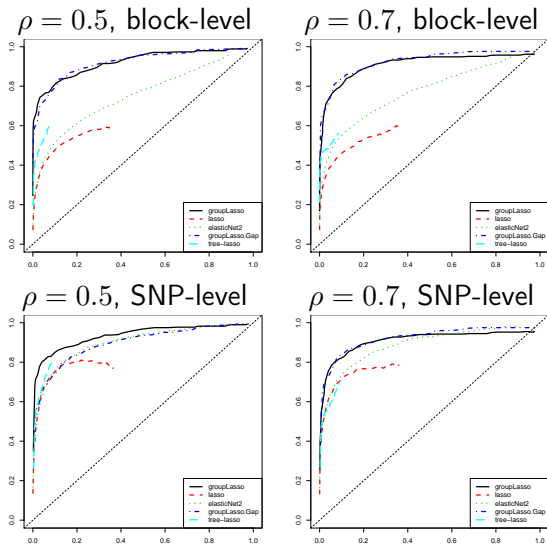


Figure: 7 clusters are inferred for  $\rho = 0.5$  and 8 clusters for  $\rho = 0.7$ .