Statistical approaches to detect epistasis in Genome Wide Association Studies

Virginie Stanislas
Supervisors: Cyril Dalmasso and Christophe Ambroise

Laboratoire de Mathématiques et Modélisation d’Évry

December 18th, 2017
Summary

1 General context
   - Complex diseases
   - GWAS
   - Epistasis

2 A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3 Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4 Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5 Conclusions
Summary

1. General context
   - Complex diseases
   - GWAS
   - Epistasis

2. A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3. Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4. Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5. Conclusions
Complex diseases

Monogenic disease

Complex disease

GWAS characteristics:

- **Objective:** find associations between genetic markers \((SNP_{i,j} \in \{0, 1, 2\})\) and a phenotypic trait \((Y_i \in \{0, 1\}\text{ or } Y_i \in \mathbb{R})\)
**GWAS characteristics:**

- **Objective:** find associations between genetic markers \((SNP_{i,j} \in \{0, 1, 2\})\) and a phenotypic trait \((Y_i \in \{0, 1\} \text{ or } Y_i \in \mathbb{R})\)
Genome-Wide Association Studies

**SNP analysis**
Differences between cases and controls at a specific SNP
SNP analysis
Differences between cases and controls at a specific SNP

GWAS limits:
- Reproductibility
- Genetic factors missing

Factors:
- High dimension \((p \gg n)\)
- Small effects
**Genome-Wide Association Studies**

**SNP analysis**
Differences between cases and controls at a specific SNP

**GWAS limits:**
- Reproductibility
- Genetic factors missing
- Missing heritability

**Missing heritability factors:**
- Non consideration of rare variants (MAF < 0.1%)
- Non consideration of structural variants (insertion, deletion, copy numbers...)
- Incorrect estimation measure of heritability
- Complex structure of genetic data
**Epistasis** - Definition

**Epistasis:** Interaction of alleles effects from different markers

<table>
<thead>
<tr>
<th>locus 1</th>
<th>locus 2</th>
<th>bb</th>
<th>bB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aA</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AA</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Epistasis - Definition

**Epistasis:** *Interaction of alleles effects from different markers*

<table>
<thead>
<tr>
<th>locus 1</th>
<th>bb</th>
<th>bB</th>
<th>BB</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>aA</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AA</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Different definitions according to disciplines with two major distinctions:

- **Biological epistasis**
- **Statistical epistasis**
Epistasis - Definition

**Biological epistasis:**
Physical interaction at the individual level

**Epistasis - Definition**

**Statistical epistasis:**
Deviation from additive effects of genetic variants at the population level

Epistasis - Definition

**Statistical epistasis:**
Deviation from additive effects of genetic variants at the population level

A possible model:

\[
\text{logit}[P(y = 1|x_1, x_2)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2
\]

with
- \(y\) a binary phenotype
- \(x_1, x_2\) the individual effect of both markers

## Epistasis - Definition

### Statistical epistasis:
Deviation from additive effects of genetic variants at the population level

A possible model:

$$\text{logit}[P(y = 1|x_1, x_2)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

with

- $y$ a binary phenotype
- $x_1, x_2$ the individual effect of both markers
Epistasis - Challenges to detect it

Methodological
Epistasis - Challenges to detect it

**Methodological**

→ $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
Epistasis - Challenges to detect it

Methodological

→ $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs

→ Curse of dimensionality
Epistasis - Challenges to detect it

Methodological

→ $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs

→ Curse of dimensionality

→ Correlation (linkage disequilibrium):
  - between observed markers
Epistasis - Challenges to detect it

**Methodological**

→ $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
→ Curse of dimensionality
→ Correlation (linkage disequilibrium):
  - between observed markers
  - between observed and causal markers
Epistasis - Challenges to detect it

Methodological
- $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
- Curse of dimensionality
- Correlation (linkage disequilibrium):
  - between observed markers
  - between observed and causal markers
- Distinction between marginal and interaction effects
Epistasis - Challenges to detect it

**Methodological**

- $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
- Curse of dimensionality
- Correlation (linkage disequilibrium):
  - between observed markers
  - between observed and causal markers
- Distinction between marginal and interaction effects

**Interpretation**

- Moving from statistical estimate of epistasis to biological epistasis
Epistasis - Challenges to detect it

**Methodological**

→ $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
→ Curse of dimensionality
→ Correlation (linkage disequilibrium):
  - between observed markers
  - between observed and causal markers
→ Distinction between marginal and interaction effects

**Interpretation**

→ Moving from statistical estimate of epistasis to biological epistasis

Epistasis is ubiquitous in human biology
Investigation indispensable to understand genetic data
Epistasis - Challenges to detect it

Methodological

- $5 \times 10^{11}$ pairwise interactions to investigate for a GWAS with $10^6$ SNPs
- Curse of dimensionality
- Correlation (linkage disequilibrium):
  - between observed markers
  - between observed and causal markers
- Distinction between marginal and interaction effects

Interpretation

- Moving from statistical estimate of epistasis to biological epistasis

Epistasis is ubiquitous in human biology
Investigation indispensable to understand genetic data

Large number of approaches proposed
Epistasis - A variety of methods
Epistasis - Scale of interactions

Existing methods: → mainly SNP × SNP
   → some at a group scale

Group definition:
→ genes
→ haplotypes
→ ...

Advantages of group scale approaches:
→ genetic effects more detectable
→ reduce the number of variables
→ consideration of the correlation
→ results biologically interpretable

U.S. National Library of Medicine

V. Stanislas
Statistical approaches to detect epistasis in Genome Wide Association Studies
Gene level test outside a regression framework:

- Aggregating interaction tests
- Co-association tests
Epistasis - Gene scale methods

Gene level test outside a regression framework:
- Aggregating interaction tests
- Co-association tests

Gene level regression based approaches:

\[ \{ \text{PCA, PLS, Kernel} \} + \text{logistic regression} \]

Epistasis - Gene scale methods

**Gene level test outside a regression framework:**
- Aggregating interaction tests
- Co-association tests

**Gene level regression based approaches:**

\[
\begin{align*}
\text{PCA} & \quad \text{PLS} & \quad \text{Kernel} \\
\text{} & \quad + \quad \text{penalized regression} \quad (D'Angelo \ 2009, \ Wang \ X \ 2014)
\end{align*}
\]
Epistasis - Gene scale methods

**Gene level test outside a regression framework:**

- Aggregating interaction tests
- Co-association tests

**Gene level regression based approaches:**

\[
\begin{align*}
\text{PCA} & \quad \text{PLS} \\
\text{Kernel} & \quad + \quad \text{penalized regression} \quad (D'Angelo \ 2009, \ \text{Wang X} \ 2014)
\end{align*}
\]

**Objectives:** To develop a new gene scale method that:

- considers a more accurate definition of interaction variables,
- is applicable to numerous genes,
- resorts to a group penalty
Summary

1 General context
   - Complex diseases
   - GWAS
   - Epistasis

2 A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3 Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4 Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5 Conclusions
Group modeling approach

<table>
<thead>
<tr>
<th>SNP_{1,1}</th>
<th>\ldots</th>
<th>SNP_{1,p_1}</th>
<th>\ldots</th>
<th>SNP_{G,1}</th>
<th>\ldots</th>
<th>SNP_{G,p_G}</th>
<th>Pheno</th>
</tr>
</thead>
<tbody>
<tr>
<td>{\text{Ind}_1}</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>\text{y}_1</td>
<td></td>
</tr>
<tr>
<td>{\text{Ind}_2}</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>\text{y}_2</td>
<td></td>
</tr>
<tr>
<td>\ldots</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>\ldots</td>
<td></td>
</tr>
<tr>
<td>{\text{Ind}_i}</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>y_i</td>
<td></td>
</tr>
</tbody>
</table>

We note SNP_{1,1} = X_{1,1}

**model:**

\[ g(E[y|X]) = \sum_g \sum_{p_g} \beta_{g,p_g} X_{g,p_g} \]

\[ \beta = \begin{pmatrix} \beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,p_1}, \ldots, \beta_{G,1}, \ldots, \beta_{G,p_G} \end{pmatrix}^T \]
**Group modeling approach**

<table>
<thead>
<tr>
<th>SNP_{1,1}, .., SNP_{1,p_1}, .., SNP_{G,1}, .., SNP_{G,p_G}</th>
<th>Pheno</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Ind}_1 )</td>
<td>1</td>
</tr>
<tr>
<td>( \text{Ind}_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( . )</td>
<td>0</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>2</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>1</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>0</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>2</td>
</tr>
<tr>
<td>( \text{Pheno} )</td>
<td>( y_1 )</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>( y_2 )</td>
</tr>
<tr>
<td>( . )</td>
<td>( . )</td>
</tr>
<tr>
<td>( \text{Ind}_i )</td>
<td>( y_i )</td>
</tr>
</tbody>
</table>

We note \( SNP_{1,1} = X_{1,1} \)

\( r, s \) two genes

**model:**

\[
g(E[y|X]) = \sum_{g} \sum_{p_g} \beta_{g,p_g} X_{g,p_g} + \sum_{r,s} \gamma_{r,s} Z_{r,s}
\]

- **Main effects**
- **Interaction effects**

\[
\beta = \begin{pmatrix}
\beta_{1,1}, \beta_{1,2}, \cdots, \beta_{1,p_1}, & \cdots, & \beta_{G,1}, \cdots, \beta_{G,p_G}
\end{pmatrix}^T
\]

\[
\gamma = \begin{pmatrix}
\gamma_{12}, \cdots, \gamma_{1G}, \cdots, \gamma_{(G-1)G}
\end{pmatrix}
\]

\( q: \# \) of interaction variables for a couple
## Interaction variables construction:

### Based on literature proposal:

<table>
<thead>
<tr>
<th>methods</th>
<th>criteria</th>
<th>interaction term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Component analysis (PCA)</td>
<td>( \text{var}(X_r v) ) and ( \text{var}(X_s v) )</td>
<td>( \sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk}^r T_j^r T_k^s )</td>
</tr>
<tr>
<td>Partial Least Square (PLS)</td>
<td>( \text{cov}^2(YX_r c, X_s w) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^rs T_j^rs )</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCA)</td>
<td>( \text{cor}(X_r a, X_s b) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^rs U_j^r V_j^s )</td>
</tr>
</tbody>
</table>
Interaction variables construction:

Based on literature proposal:

<table>
<thead>
<tr>
<th>methods</th>
<th>criteria</th>
<th>interaction term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Component analysis (PCA)</td>
<td>(\text{var}(X_r \mathbf{v})) and (\text{var}(X_s \mathbf{v}))</td>
<td>(\sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk} T_j^r T_k^s)</td>
</tr>
<tr>
<td>Partial Least Square (PLS)</td>
<td>(\text{cov}^2(YX_r \mathbf{c}, X_s \mathbf{w}))</td>
<td>(\sum_{j=1}^{q} \gamma_{j}^rs T_j^rs)</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCA)</td>
<td>(\text{cor}(X_r \mathbf{a}, X_s \mathbf{b}))</td>
<td>(\sum_{j=1}^{q} \gamma_{j}^rs U_j^r V_j^s)</td>
</tr>
</tbody>
</table>

**Original proposition: Gene-Gene Eigen Epistasis (G-GEE)**

We consider \(f_u(X_r, X_s)\) to represent the interaction between genes \(r, s\).

We can choose \(f_u(X_r, X_s)\) following two conditions:
Interaction variables construction:

Based on literature proposal:

<table>
<thead>
<tr>
<th>methods</th>
<th>criteria</th>
<th>interaction term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Component analysis (PCA)</td>
<td>( \text{var}(X_r \mathbf{v}) ) and ( \text{var}(X_s \mathbf{v}) )</td>
<td>( \sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk} T_j^r T_k^s )</td>
</tr>
<tr>
<td>Partial Least Square (PLS)</td>
<td>( \text{cov}^2(YX_r c, X_s w) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^{rs} T_j^r )</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCA)</td>
<td>( \text{cor}(X_r a, X_s b) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^{rs} U_j^r V_j^s )</td>
</tr>
</tbody>
</table>

Original proposition: Gene-Gene Eigen Epistasis (G-GEE)
We consider \( f_u(X^r, X^s) \) to represent the interaction between genes \( r, s \).
We can choose \( f_u(X^r, X^s) \) following two conditions:

- \( \text{cov}^2((X_r, X_s), f_u(X^r, X^s)) \)
- \( \text{cov}^2(y, f_u(X^r, X^s)) \)
Interaction variables construction:

Based on literature proposal:

<table>
<thead>
<tr>
<th>methods</th>
<th>criteria</th>
<th>interaction term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal Component analysis (PCA)</td>
<td>( \text{var}(X_r \mathbf{v}) ) and ( \text{var}(X_s \mathbf{v}) )</td>
<td>( \sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk}^{rs} T_r^T T_s^T )</td>
</tr>
<tr>
<td>Partial Least Square (PLS)</td>
<td>( \text{cov}^2(Y X_r \mathbf{c}, X_s \mathbf{w}) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^{rs} T_j^{rs} )</td>
</tr>
<tr>
<td>Canonical Correlation Analysis (CCA)</td>
<td>( \text{cor}(X_r \mathbf{a}, X_s \mathbf{b}) )</td>
<td>( \sum_{j=1}^{q} \gamma_{j}^{rs} U_j^{r} V_j^{s} )</td>
</tr>
</tbody>
</table>

Original proposition: Gene-Gene Eigen Epistasis (G-GEE)

We consider \( f_u(X^r, X^s) \) to represent the interaction between genes \( r, s \).

We can choose \( f_u(X^r, X^s) \) following two conditions:

<table>
<thead>
<tr>
<th>criteria</th>
<th>methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{cov}^2((X_r, X_s), f_u(X^r, X^s)) )</td>
<td>G-GEE_{c1}</td>
</tr>
<tr>
<td>( \text{cov}^2(y, f_u(X^r, X^s)) )</td>
<td>G-GEE_{c2}</td>
</tr>
</tbody>
</table>
Interaction variables construction: G-GEE<sub>c1</sub>

We set: \( f_u(X^r, X^s) = F^{rs} u \) with \( F^{rs} = \{X^r_{ij} X^s_{ik}\}^j_{i=1} \cdots , p_r ; ^k_{i=1} \cdots , p_s \)

\[ \hat{u} = \arg \min_{u, \|u\|=1} \text{cov}^2(X, F^{rs} u) \]

with \( X = (X^r, X^s) \)

\[ \min_{u, \|u\|=1} \|\text{cov}[F^{rs} u, X]\|^2 = \min_{u, \|u\|=1} u^T F^{rs T} X X^T F^{rs} u \]

\( u \): eigen vector associated to the smallest eigenvalue of \( F^{rs T} X X^T F^{rs} \)

We then obtain for each couple \((r, s) \rightarrow Z^{rs} = F^{rs} u\)
We set: \( f_u(X^r, X^s) = F^{rs} u \) with \( F^{rs} = \{ X^r_{ij}, X^s_{ik} \}_{i=1 \ldots n} \)

\[
\hat{u} = \arg \max_{\|u\|=1} \text{cov}^2(y, F^{rs} u)
\]

\[
\max_{\|u\|=1} \|\text{cov}(F^{rs} u, y)\|^2 = \max_{\|u\|=1} u^T F^{rs T} y y^T F^{rs} u
\]

\( u \): eigen vector associated to the largest eigenvalue of \( F^{rs T} y y^T F^{rs} \)

\[
\hat{u} = F^{rs T} y
\]

We then obtain for each couple \((r, s)\) \( Z^{rs} = F^{rs} u = F^{rs} F^{rs T} y \)
### Interaction variable modeling approaches comparison

<table>
<thead>
<tr>
<th>methods</th>
<th>criteria</th>
<th>interaction term</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G\text{-GEE}_{c1}$</td>
<td>$\text{cov}^2(X, f_u(X^r, X^s))$</td>
<td>$F^{rs} u \gamma^{rs}$</td>
</tr>
<tr>
<td>$G\text{-GEE}_{c2}$</td>
<td>$\text{cov}^2(Y, f_u(X^r, X^s))$</td>
<td>$F^{rs} u \gamma^{rs}$</td>
</tr>
<tr>
<td>PCA</td>
<td>$\text{var}(X_r v)$ and $\text{var}(X_s v)$</td>
<td>$\sum_{j=1}^{q} \sum_{k=1}^{q} \gamma_{jk} T_j^r T_k^s$</td>
</tr>
<tr>
<td>PLS</td>
<td>$\text{cov}^2(Y X_r c, X_s w)$</td>
<td>$\sum_{j=1}^{q} \gamma_{j}^{rs} T_j^{rs}$</td>
</tr>
<tr>
<td>CCA</td>
<td>$\text{cor}(X_r a, X_s b)$</td>
<td>$\sum_{j=1}^{q} \gamma_{j}^{rs} U_j^r V_j^s$</td>
</tr>
</tbody>
</table>
## Interaction variable modeling approaches comparison

<table>
<thead>
<tr>
<th>Methods</th>
<th>Criteria</th>
<th>Interaction Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-GEE(_c_1)</td>
<td>(\text{cov}^2(X, f_u(X^r, X^s)))</td>
<td>(F_{rs}^u \gamma_{rs})</td>
</tr>
<tr>
<td>G-GEE(_c_2)</td>
<td>(\text{cov}^2(Y, f_u(X^r, X^s)))</td>
<td>(F_{rs}^u \gamma_{rs})</td>
</tr>
<tr>
<td>PCA</td>
<td>(\text{var}(X_r \text{v})) and (\text{var}(X_s \text{v}))</td>
<td>(\sum_j \sum_k \gamma_{jk} T_j^r T_k^s)</td>
</tr>
<tr>
<td>PLS</td>
<td>(\text{cov}^2(YX_r \text{c}, X_s \text{w}))</td>
<td>(\sum_j \gamma_{j}^r T_j^s)</td>
</tr>
<tr>
<td>CCA</td>
<td>(\text{cor}(X_r \text{a}, X_s \text{b}))</td>
<td>(\sum_j \gamma_{j}^r U_j^r V_j^s)</td>
</tr>
</tbody>
</table>

\[
g(E[y|X]) = \sum_g \sum_{p_g} \beta_{g,p_g} X_{g,p_g} + \sum_{r,s} \gamma_{r,s} Z_{r,s}
\]
Coefficients estimation

**Group LASSO regression**

\[
(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta, \gamma} \left( \sum_i -\log L(y_i; X_i \beta + Z_i \gamma) + \lambda \left[ \sum_g \sqrt{p_g} \|\beta^g\|_2 + \sum_{rs} \sqrt{p_r p_s} \|\gamma^{rs}\|_2 \right] \right)
\]
Coefficients estimation

**Group LASSO regression**

\[
(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta, \gamma} \left( \sum_i -\log L(y_i; X_i\beta + Z_i\gamma) + \lambda \left[ \sum_g \sqrt{p_g} \|\beta^g\|_2 + \sum_{rs} \sqrt{p_r p_s} \|\gamma^{rs}\|_2 \right] \right)
\]

Limits of the group LASSO regression:

- \( P(S^* \subset \hat{S}) \xrightarrow{n \to +\infty} 1 \) but \( |\hat{S}| \gg |S^*| \)

- Difficult to compute p-value or confidence interval
Group LASSO regression

\[
(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta, \gamma} \left( \sum_i -\log L(y_i; X_i\beta + Z_i\gamma) + \lambda \left[ \sum_g \sqrt{p_g} \|\beta_g\|_2 + \sum_{rs} \sqrt{p_r p_s} \|\gamma_{rs}\|_2 \right] \right)
\]

Limits of the group LASSO regression:
- \( P(S^* \subset \hat{S}) \xrightarrow{n \to +\infty} 1 \) but \( |\hat{S}| \gg |S^*| \)
- Difficult to compute p-value or confidence interval

Adaptive-Ridge Cleaning  Becu JM et al., 2017
Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $H\theta = X\beta + Z\gamma$
Split randomly $H$ in two subsets $H_1$ and $H_2$ of size $n/2$
Coefficients estimation: Adaptive-Ridge Cleaning

Setting: \( H\theta = X\beta + Z\gamma \)

Split randomly \( H \) in two subsets \( H_1 \) and \( H_2 \) of size \( n/2 \)

**First stage:** Screening on \( H_1 \)

\[
\hat{\theta} = \arg\min_{\theta} \left( \sum_i -\log L(y_i; H_1i\theta) + \lambda \left[ \sum_g \sqrt{p_g} \|\theta^g\|_2 \right] \right)
\]
Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $H\theta = X\beta + Z\gamma$
Split randomly $H$ in two subsets $H_1$ and $H_2$ of size $n/2$

First stage: Screening on $H_1$

$$\hat{\theta} = \arg\min_{\theta} \left( \sum_i -\log L(y_i; H_1i\theta) + \lambda \left[ \sum_g \sqrt{p_g} \left\| \theta^g \right\|_2 \right] \right)$$

$\hat{S}$ : support of the selected groups
Coefficients estimation: Adaptive-Ridge Cleaning

Setting: \( H\theta = X\beta + Z\gamma \)
Split randomly \( H \) in two subsets \( H_1 \) and \( H_2 \) of size \( n/2 \)

**First stage:** Screening on \( H_1 \)

\[
\hat{\theta} = \arg\min_{\theta} \left( \sum_i -\log L(y_i; H_1i\theta) + \lambda \left[ \sum_g \sqrt{p_g} \|\theta^g\|_2 \right] \right)
\]

\( \rightarrow \hat{S} : \) support of the selected groups

**Second stage:** Cleaning on \( H_2 \)

\[
\tilde{\theta} = \arg\min_{\theta \ ; \ \theta_j = 0 \ if \ j \notin \hat{S}} \left( \sum_i -\log L(y_i; H_2i\theta) + \mu \left[ \sum_g \sum_{j \in g} \frac{\lambda\sqrt{p_g}}{\|\hat{\theta}^g\|_2} \theta_j^2 \right] \right)
\]
Coefficients estimation: Adaptive-Ridge Cleaning

**Significance of \( \tilde{\theta} \):** Permutation test based on a Fisher test approach
Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

$$F_g = \frac{\sum_i (y_i - \hat{y}_i^\omega)^2 - \sum_i (y_i - \hat{y}_i^\Omega)^2}{\sum_i (y_i - \hat{y}_i^\Omega)^2}$$

With:
- $\hat{y}_i^\omega$: predicted values obtained without the group $g$
- $\hat{y}_i^\Omega$: predicted values using all groups $g \in \hat{S}$
Coefficients estimation: Adaptive-Ridge Cleaning

**Significance of $\tilde{\theta}$:** Permutation test based on a Fisher test approach

\[
F_g = \frac{\sum_i (y_i - \hat{y}_i^\omega)^2 - \sum_i (y_i - \hat{y}_i^\Omega)^2}{\sum_i (y_i - \hat{y}_i^\Omega)^2}
\]

\[
F_g^* = \frac{\sum_i (y_i - \hat{y}_i^\omega)^2 - \sum_i (y_i - \hat{y}_i^{\Omega*})^2}{\sum_i (y_i - \hat{y}_i^{\Omega*})^2}
\]

With:
- $\hat{y}_i^\omega$: predicted values obtained without the group $g$
- $\hat{y}_i^\Omega$: predicted values using all groups $g \in \hat{S}$
- $\hat{y}_i^{\Omega*}$: predicted values using all groups $g \in \hat{S}$ on the matrix $H^*$ of permuted elements for columns corresponding to group $g$
Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

\[ F_g = \frac{\sum_i (y_i - \hat{y}_i^\omega)^2 - \sum_i (y_i - \hat{y}_i^\Omega)^2}{\sum_i (y_i - \hat{y}_i^\Omega)^2} \]

\[ F_g^* = \frac{\sum_i (y_i - \hat{y}_i^\omega)^2 - \sum_i (y_i - \hat{y}_i^{\Omega*})^2}{\sum_i (y_i - \hat{y}_i^{\Omega*})^2} \]

With:
$\hat{y}^\omega$: predicted values obtained without the group $g$
$\hat{y}^\Omega$: predicted values using all groups $g \in \hat{S}$
$\hat{y}^{\Omega*}$: predicted values using all groups $g \in \hat{S}$ on the matrix $H^*$ of permuted elements for columns corresponding to group $g$

Empirical p-values:

\[ p_g = \frac{1}{B} \sum_{b=1}^{B} 1\{F_g \leq F_g^*\} \]

with $B$ the number of permutations
Summary

1. General context
   - Complex diseases
   - GWAS
   - Epistasis

2. A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3. Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4. Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5. Conclusions
Simulations design: Genotype

**Completely simulated genotype:**

\[ X_i \sim \mathcal{N}_p(0, \Sigma) \text{ with } \Sigma \text{ a block diagonal correlation matrix} \]

\( (\rho \text{ correlation level for two SNPs in the same gene}) \)

\[ MAF_j \sim \mathcal{U}[0.05, 0.5] \text{ with fixed } MAF_j \text{ if } j \text{ causal SNP} \]

**Genotype from real data:**

From a real data set composed of 763 individuals and 63,340 SNPs structured in 7216 genes.
Simulations design: Phenotype

from Wang X et al., 2014:

\[
g(E[y_i|(X_i, Z_i)]) = \beta_0 + \sum_g \beta_g \left( \sum_{k \in C} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left( \sum_{(j,k) \in C^2} X_{ij}^r X_{ik}^s \right)
\]

PCA model:

\[
g(E[y_i|(X_i, Z_i)]) = \beta_0 + \sum_g \beta_g \left( \sum_{k \in C} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} C_{i1}^r C_{i1}^s.
\]
Simulations design

Adjustment of the strength of association for continuous outcomes

- \( \epsilon_i \) generated from \( \mathcal{N}(0, \sigma^2) \)
- \( \sigma^2 \) determined from \( R^2 \) coefficient

We note \( H \theta = [X, Z] \begin{bmatrix} \beta^T \\ \gamma \end{bmatrix} \), and

\[
R^2 = \frac{\sum (H_i \theta - \bar{y})^2}{\sum (H_i \theta + \epsilon_i - \bar{y})^2}
\]

We can determined an expression for \( \sigma^2 \)

\[
\sigma^2 = \frac{(1 - R^2) \sum (H_i \theta - \bar{y})^2}{R^2 (n - 2)}
\]
Simulations studies

**First comparison:** PCA, PLS and CCA  
Choosing the parameters

**Second comparison:** with G-GEE_{c1} and G-GEE_{c2}  
Using completely simulated genotype  
Using genotype from a real data set

**Third comparison:** Case-control methods

**Fourth comparison:** Investigation of new interaction variable definitions
First comparison: methods issued of the literature

**Design:** Completely simulated genotype
Continuous phenotype from Wang X et al., 2014

**Parameters:**

- Correlation among SNPs $\rho$
- MAF values of causal SNPs
- Values of $\beta$ and $\gamma$
- Number of components
- $R^2$
- Number of genes

- Number of SNPs by genes
- Number of causal SNPs by causal genes
- Number of subjects
- Marginal or/and interaction effects
First comparison: methods issued of the literature

**Design:** Completely simulated genotype
Continous phenotype from Wang X et al., 2014

**Parameters:**
- Correlation among SNPs $\rho$
- MAF values of causal SNPs
- Values of $\beta$ and $\gamma$
- Number of components
- $R^2$
- Number of genes
- Number of SNPs by genes
- Number of causal SNPs by causal genes
- Number of subjects
- Marginal or/and interaction effects
First comparison: methods issued of the literature
First comparison: methods issued of the literature

Parameters:
- $\rho = 0.8$
- MAF = 0.2
- $\beta = \gamma = 2$
- Number of components = 2
- $R^2$
- Number of genes = 6
- Number of SNPs by genes = 6
- Number of causal SNPs by causal genes = 2
- Number of subjects = 600
- Marginal or/and interaction effects
Simulations studies

**First comparison:** PCA, PLS and CCA
Choosing the parameters

**Second comparison:** with G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

**Third comparison:** Case-control methods

**Fourth comparison:** Investigation of new interaction variable definitions
Second comparison: G-GEE and simulated genotypes

- **Wang model**
- **PCA model**

→ Main effects:
- gene 1
- gene 2

→ Interaction effects:
- gene 1 x gene 2
- gene 3 x gene 4
Discoveries matrix - an example
Second comparison: G-GEE and simulated Genotypes - $R^2 = 0.2$

Wang model

PCA model

→ Main effects:
  - gene 1
  - gene 2

→ Interaction effects:
  - gene 1 x gene 2

→ Interaction effects:
  - gene 3 x gene 4
Second comparison: G-GEE and real genotypes

**Settings**

- **Main effects:** gene 1, gene 2
- **Interaction effects:** gene 1 x gene 2

**Wang simulation model**

- **Main effects:** gene 1, gene 2
- **Interaction effects:** gene 3 x gene 4

**PCA simulation model**

- **Main effects:** gene 1, gene 2
- **Interaction effects:** gene 1 x gene 2
Second comparison: G-GEE and real genotypes - $R^2 = 0.2$

**Settings**

**Main effects:**
- gene 1
- gene 2

**Interaction effects:**
- gene 1 x gene 2

**Wang simulation model**

**PCA simulation model**

**Evaluation and comparison**

**Main effects:**
- gene 1
- gene 2

**Interaction effects:**
- gene 3 x gene 4

**Main effects:**
- gene 1
- gene 2

**Interaction effects:**
- gene 1 x gene 2

**Conclusions**
Simulations studies

**First comparison:** PCA, PLS and CCA
Choosing the parameters

**Second comparison:** G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

**Third comparison:** Case-control methods

**Fourth comparison:** Investigation of new interaction variable definitions
Third comparison: Case-control methods

Methods defined outside a regression framework

- **Aggregating tests**
  - minP (Emily et al., 2016)
  - GATES (Li et al., 2011)

- **Co-association test**
  - PLSPM (Zhang et al., 2013)

- **LD based test**
  - CLD (Rajapakse et al., 2012)

- **Entropy based method**
  - GBIBM (Li et al., 2015)

**Package R:** GeneGeneInteR (Emily et al., 2017)
Third comparison: Case-control methods

Design:
Real Genotypes
Continuous phenotype simulation from *Wang X et al., 2014*:

Main effects:
gene 1
gene 2

Interaction effects:
gene 1 × gene 2

Main effects:
gene 1
gene 2

Interaction effects:
gene 3 × gene 4

V. Stanislas
Statistical approaches to detect epistasis in Genome Wide Association Studies
Third comparison: Case-control methods

Design:
Completely simulated Genotypes
Continuous phenotype simulation from Wang X et al., 2014:

Main effects:
gene 1
gene 2

Interaction effects:
gene 1 × gene 2
-
Simulations studies

First comparison: PCA, PLS and CCA
Choosing the parameters

Second comparison: with $G\text{-GEE}_{c1}$ and $G\text{-GEE}_{c2}$
Using completely simulated genotype
Using genotype from a real data set

Third comparison: Case-control methods

Fourth comparison: Investigation of new interaction variable definitions
Fourth comparison: Machine Learning based approaches

With G-GEE, we looked for:

\[
\hat{u} = \arg \max_{u, \|u\|=1} \text{cov}^2(y, f_u(X^r, X^s))
\]

with \( f_u(X^r, X^s) = F^{rs}u \) and \( F^{rs} = \{X^r_{ij}X^s_{ik} \}_{j=1,\ldots,p_r; k=1,\ldots,p_s} \)

We now find new functions \( f_u(X^r, X^s) \) that maximized the criteria:

\[
E_{X^r, X^s, Y}[(y - f_u(X^r, X^s))^2]
\]

With the following non parametric approaches:

- Random Forests
- Boosting
- SVM
- Neural Network
Fourth comparison: Machine Learning based approaches - $R^2 = 0.4$

**Design:**

Real Genotypes

Continuous phenotype simulation from *Wang X et al., 2014*:

**Main effects:**
- gene 1
- gene 2

**Interaction effects:**
- gene 1 x gene 2
- gene 3 x gene 4

---

V. Stanislas

Statistical approaches to detect epistasis in Genome Wide Association Studies 42 / 54
Fourth comparison: Machine Learning based approaches - $R^2 = 0.4$

Design:
Real Genotypes
Continuous phenotype simulation from Wang X et al., 2014:

Main effects: gene 1, gene 2
Interaction effects: gene 1 × gene 2

Main effects: gene 1, gene 2
Interaction effects: gene 3 × gene 4
Summary

1. General context
   - Complex diseases
   - GWAS
   - Epistasis

2. A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3. Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4. Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5. Conclusions
Ankylosing Spondylitis

Chronic inflammatory disease of the axial skeleton

**Epidemiology:**
- Age at first symptoms: 20 - 30 years
- Sexe: predominance for men (sex ratio 2M:1W)
- Prevalence: depend of populations (0.1% - 1.4%)

**Risk factors:**
- Strong genetic component (heritability >90%)
- Importance of HLA complex

**HLA complex:**
- Localized on chromosome 6
- Regroup about 200 genes
- Coding the immunity system
- Antigen HLA-B27: associated to SPA
Crohn’s Disease

Form of chronic inflammation bowel disease

Epidemiology:

- Prevalence: 10-30 per 100,000 (Europe and North America)
- More common in the industrialized world
- Median onset of disease: 30 years

Multiple risk factors:

- Environmental
- Microbiota
- Genetic

Genetic factors:

→ NOD2, first identified mutation

Potential interactions:

- NOD2 and TLR proteins
- NOD2 and CTLA4
- IL23R and CTLA4
- NOD2 and IBD5
- IBD5, ATGL16L1 and IL23R

Quality controls and filtering

Markers filtering:
- SNP call rate $\leq 95\%$
- MAF $\leq 5\%$
- Deviation from Hardy Weinberg Equilibrium in controls ($p < 1 \times 10^{-5}$)
- Duplicates
- SNPs not belonging to one unique gene

Subject filtering:
- Sample call rate $\leq 93\%$
- Duplicates
Ankylosing Spondylitis

**Data set:** International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6,611 genes
51,287 SNPs
Ankylosing Spondylitis

**Data set:** International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6,611 genes
51,287 SNPs
Ankylosing Spondylitis

Data set: International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6,611 genes
51,287 SNPs

→ 29 known genes
→ 62 genes from an univariate analysis
→ 91 genes to investigate
Ankylosing Spondylitis

**Data set:** International Genetics of Ankylosing Spondylitis study

- 401 cases
- 357 controls
- 6,611 genes
- 51,287 SNPs

→ 29 known genes
→ 62 genes from an univariate analysis
→ **91 genes to investigate**

<table>
<thead>
<tr>
<th>Statistical approaches</th>
<th>Significant results</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-GEE</td>
<td>NKX2-3 × HCG27</td>
</tr>
<tr>
<td>PLS</td>
<td>HLA-B</td>
</tr>
<tr>
<td></td>
<td>HCP5</td>
</tr>
<tr>
<td></td>
<td>HLAB × HCG27</td>
</tr>
<tr>
<td>PCA</td>
<td>HLAB-B</td>
</tr>
<tr>
<td></td>
<td>EOMES × HCP5</td>
</tr>
<tr>
<td></td>
<td>IL1R2 × MICB</td>
</tr>
<tr>
<td></td>
<td>ZFP57 × LOC101929772</td>
</tr>
<tr>
<td></td>
<td>TRIM31 × HCG26</td>
</tr>
</tbody>
</table>
Crohn’s Disease

Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17 304 genes
140 487 SNPs
Crohn’s Disease

Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17,304 genes
140,487 SNPs
Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17,304 genes
140,487 SNPs

→ 72 known genes
→ 60 genes from an univariate analysis
   (22 known)
→ 110 genes to investigate
**Crohn’s Disease**

**Data set:** Wellcome Trust Case-Control Consortium

- 1938 cases
- 1500 controls
- 17,304 genes
- 140,487 SNPs

→ 72 known genes
→ 60 genes from an univariate analysis (22 known)
→ **110 genes to investigate**

<table>
<thead>
<tr>
<th>Statistical approach</th>
<th>Significant results</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-GEE</td>
<td>LOC105369715 × STAT1</td>
</tr>
<tr>
<td></td>
<td>STAT1 × CD6</td>
</tr>
<tr>
<td>PLS</td>
<td>IFNGR1 × SBNO2</td>
</tr>
<tr>
<td></td>
<td>IRGM × NOD2</td>
</tr>
<tr>
<td>PCA</td>
<td>IRGM</td>
</tr>
<tr>
<td></td>
<td>LOC101929544 × TLR4</td>
</tr>
<tr>
<td></td>
<td>BATF × IL10</td>
</tr>
</tbody>
</table>
Summary

1. General context
   - Complex diseases
   - GWAS
   - Epistasis

2. A new method
   - General modeling approach
   - Interactions construction
   - Coefficients estimation

3. Evaluation and comparison
   - Simulation designs and scenarios
   - Setting parameters
   - Comparison with G-GEE
   - Case-control methods comparisons
   - Non parametric interaction modeling approach

4. Application
   - Ankylosing Spondylitis
   - Crohn’s Disease
   - Analysis and results

5. Conclusions
Conclusions and perspectives

Contributions:

→ Proposition of a new Group LASSO framework
→ Proposition of an original interaction modeling

Publication, software and presentations:

→ Package G-GEE available on Github


→ 4 talks and 3 posters in international conferences
Conclusions and perspectives

Limitations:

- Number of SNPs by genes to analyze
- Computation costs for estimation coefficients
- Choice of the genes to consider
- Confusion phenomenon
- Sensitive to group definition
Perspectives:

→ Explore new $f_u(X^r, X^s)$ definitions
→ Optimization of the computational cost of $F^{rs}$
→ New selection of the parameter $\lambda$
→ Using another penalization regression framework
→ Gene selection using biological knowledge
→ Investigate other grouping definitions
Thank you for your attention!