Detecting Gene-Environment Interaction using Logistic Regression with Latent Exposure

F. Alarcon(1), Vittorio Perduca (1)G. Nuel(2).
(1) MAP5 UMR CNRS 8145, Université Paris Descartes, (2) INSMI CNRS Laboratory of probability (LPMA), Université Pierre et Marie Curie

GDR Statistique et Santé

Introduction

- Most complex diseases are affected in part by interactions between genes and genes (GG) and/or between genes and environmental (GE) factors .
- Accounting for GE interactions in GWAS can provide a better understanding of the disease and its risk factors.
- To date, only few loci that interact with environment have been discovery $\Rightarrow$ challenging problem.
- A possible cause :
  - The true exposure is seldom observed
  - Only proxy covariates are observed.
- Complex simulated dataset +PLINK analysis $\rightarrow$ presence of such latent exposure is a substantial obstacle to detect GE interactions in GWAS.
- We propose a model that introduce a binary latent exposure in the logistic regression model :
  - Validation on a simple simulated dataset.
  - Application to our complex simulated dataset

Plan

1. Presentation of the complex simulated dataset.
2. Loss of power in detection the GE interaction in presence of latent exposure with PLINK.
3. The Logistic Regression with Latent Exposure (LRLE) model.
4. Comparison of the LRLE model with two standards approaches :
   - The case-controls (CC) test
   - The case-only (CO) test

Simulated genotypes and covariates

- Genotypes $\rightarrow$ from HapMap phase III database of genetic variations.
- Covariates :
  - a latent exposure (`treatment`), correlated with :
  - body mass index (`bmi`) (itself depending on population belonging and `smoking`) ; `sex` and the population of belonging.
- Idea : the treatment is typically taken by women (and less often by men) trying to loose weight.

$$1/\mathbb{P}(\texttt{treatment}) = (1 + 2 \times 1_{\texttt{sex}=1}) \times [1 + \exp(-\texttt{bmi} + 25 + \gamma)]$$

where $\gamma \in \{-\inf, -0.1, 0, 0.15, -0.45, 0.35, 0.6, -0.4, 0.05, 0.1\}$ for population 1 to 11.

- observed covariates : `bmi` ; `sex` ; `smoking` ; $\texttt{pca}_i$, $i = 1, ...5$.
- non observed covariates : `treatment`.

Disease model and simulated phenotypes

- SNP chosen arbitrarily in a dense area of chromosome 6
- Binary susceptibility locus assuming a dominant effect :
  $\texttt{causalSNP} = 1$ in presence of at least one minor allele frequency
  $\texttt{causalSNP} = 0$ otherwise
- Disease model with strong GE interaction and baseline prevalence of 1% :

$$\mathbb{P}(\texttt{disease}) = 0.01 \times (1.0 + 50.0 \times 1_{\texttt{causalSNP}=1} \times 1_{\texttt{treatment}=1})$$

- Phenotypes simulated with the package `waffect` : 200 replicates under H0 of no association and 200 replicates of phenotypes under H1

Waffect (weighted affectation)

- Approach for simulating phenotypes under H1 that does not require generating new genotypes for each simulation.
- The disease model under H_1 is defined by the probability

$$\pi_i = \mathbb{P}(Y_i = 1 | X_i),$$

under the constraint $C = \{\sum_{i=1}^{n} Y_i = n_1\}$

- Under H_0, π_i have the same non-negative value for all individual i.
- Benefit : the number of case and controls is constant across replicates.
- One of the benefit of using Waffect to perform simulations under H1 is that it only requires a vector of probabilities as input.
- It implies that the choice of the disease model is unconstrained $\Rightarrow$ possibility to include G*E interactions

Power, ROC curves and AUC with PLINK

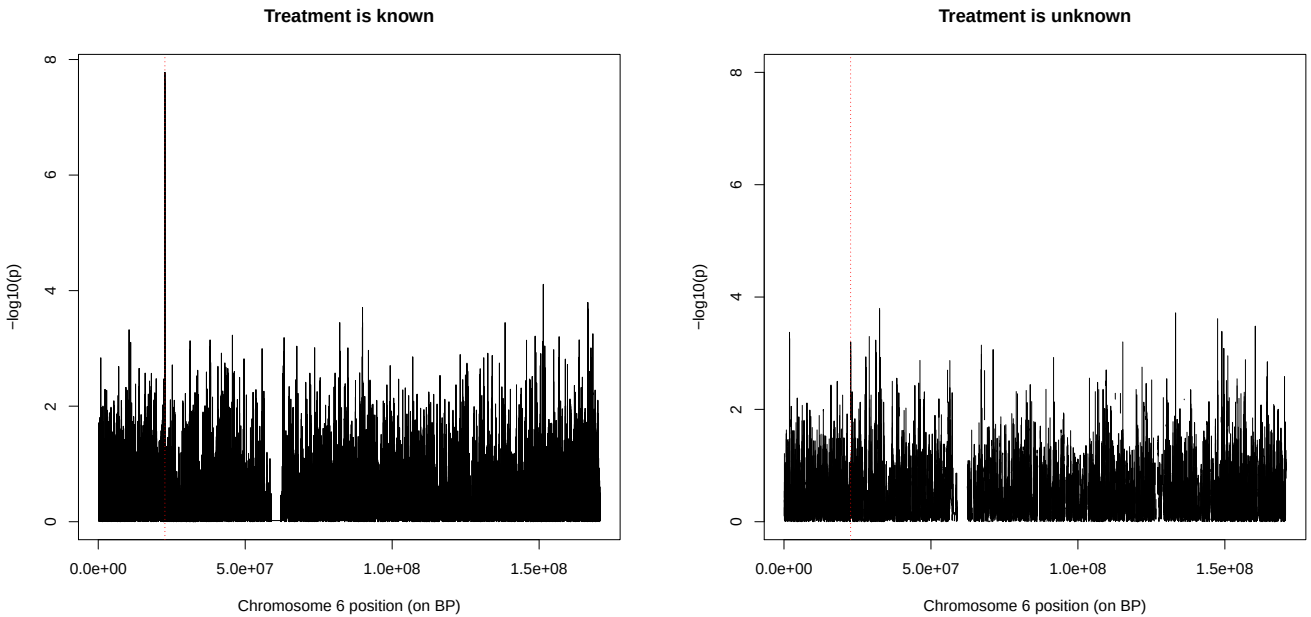
- $S = \min(p_{\text{value}}$ associated to the interaction between SNPs and `bmi` or `treatment`)
- $(S_{H0,1}, \ldots, S_{H0,200})$ and $(S_{H1,1}, \ldots, S_{H1,200})$

  $\rightarrow$ Area Under the Curve (AUC) corresponding to the receiver operating characteristic (ROC)

	PLINK SNP$\times$`treatment`	PLINK SNP$\times$`bmi`
AUC	**99.97**	**56.39**
95% Confidence Intervalle	*[99.93 - 100.0]*	*[50.69 - 62.08]*

Table: GWA performed on whole chromosome 6

Manhattan plots observing the covariate `treatment` or the coveriate `bmi`



(a) Observing the covariate `treatment`.

(b) Observing the covariate `bmi`.

Accounting for the latent exposure

The variables

* $y \in \{0,1\}^n$ is the disease status vector of n individuals.
* $E \in \{0,1\}^n$ is the latent exposures statute of the same n individuals.
* $H \in \mathbb{R}^{n \times m}$ is the proxy covariates matrix and $\eta \in \mathbb{R}^{m \times 1}$ the parameter corresponding.
* $X \in \mathbb{R}^{n \times p}$ is the nuisance covariates matrix and $\beta \in \mathbb{R}^{p \times 1}$ the parameter corresponding.
* $Z \in \mathbb{R}^{n \times q}$ is the exposure covariates matrix and $\delta \in \mathbb{R}^{q \times 1}$ the parameter corresponding.

The model

$$\text{logit}\, \mathbb{P}(\boldsymbol{E} = 1) = \boldsymbol{H}^T \boldsymbol{\eta} \tag{1}$$

$$\text{logit}\, \mathbb{P}(\boldsymbol{y} = 1 | \boldsymbol{E}) = \boldsymbol{X}^T \boldsymbol{\beta} + \mathbb{1}_{\boldsymbol{E}=1} \boldsymbol{Z}^T \boldsymbol{\delta} \tag{2}$$

The likelihood

By integrating the likelihood of the model over the unobserved exposure $\boldsymbol{E}$ we obtain after some simplification the following log-likelihood:

$$\begin{aligned}\text{loglik}(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\eta}) = \log \mathbb{P}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{H}; \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\eta}) = \sum_{i, y_i=0} \log\left[1 + \exp(b_i + d_i) + \exp(b_i + h_i) + \exp(h_i)\right] \\ + \sum_{i, y_i=1} \left\{b_i + \log\left[1 + \exp(b_i + d_i) + \exp(d_i + h_i) + \exp(b_i + d_i + h_i)\right]\right\} \\ - \sum_{i=1}^{n} \log[1 + \exp(h_i)] - \sum_{i=1}^{n} \log[1 + \exp(b_i)] - \sum_{i=1}^{n} \log[1 + \exp(b_i + d_i)] \quad (3)\end{aligned}$$

where $\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{\beta}$, $\boldsymbol{d} = \boldsymbol{Z}^T\boldsymbol{\delta}$, and $\boldsymbol{h} = \boldsymbol{H}^T\boldsymbol{\eta}$.

Optimization problem

- Explicit expressions of the first two derivates of $\text{loglik}(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\eta})$.
- Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with exact first derivates

Statistic of test

The statistic of test S_λ :

$$S_\lambda = 2(L_1 - L_0),$$

with

$$L_1 = \max_{\beta_1, \delta_1, \eta_1} \left\{ \log \mathbb{P}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{H}; \beta_1, \delta_1, \eta_1) - \frac{\lambda}{2} \left(\beta_1^T \beta_1 + \delta_1^T \delta_1 + \eta_1^T \eta_1 \right) \right\}$$

$$L_0 = \max_{\beta_0, \delta_0, \eta_0} \left\{ \log \mathbb{P}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}_0, \boldsymbol{H}; \beta_0, \delta_0, \eta_0) - \frac{\lambda}{2} \left(\beta_0^T \beta_0 + \delta_0^T \delta_0 + \eta_0^T \eta_0 \right) \right\}$$

Notations

* $\lambda \geqslant 0$ = The regularization parameter.
* L_1 = The log-likelihood under the alternative hypothesis H1
* L_0 = The log-likelihood under the null hypothesis H0 of no association.
* $\boldsymbol{X} = 1 + \text{nuisance} + \text{gene}$;
* $\boldsymbol{Z} = 1 + \text{gene}$
* $\boldsymbol{Z}_0 = 1$
* $\boldsymbol{H} = 1 + \text{proxy}$

Validation : simple simulated dataset

$$\boldsymbol{X} = \boldsymbol{Z} = 1 + \text{snp} \quad \boldsymbol{Z}_0 = 1 \quad \boldsymbol{H} = 1 + \text{proxy},$$

* snp $\in \{0, 1, 2\}$ drawn independently with probability $\mathbb{P}(\text{snp}) = (0.80\ 0.15\ 0.05)$.
* proxy $\in \mathbb{R}$ drawn independently according to standard gaussian distribution.
* $\boldsymbol{\beta} = (\beta_1 = -0.3, \beta_2 = 0.1)$
* $\boldsymbol{\delta} = (\delta_1 = 0.5, \delta_2 = \text{GE})$, GE $\in \mathbb{R}$ is the gene-environment interaction effect
* $\boldsymbol{\eta} = (\delta_1 = 0.3, \eta_2 = \text{PXY})$, PXY $\in \mathbb{R}$ is the proxy effect.
*

The larger |GE|, the easier to detect the interaction signal, the larger |PXY|, the more informative the proxy variables.

		β_1	β_2	δ_1	GE	η_1	PXY
reference		−0.3000	0.1000	0.5000	1.0000	0.3000	5.0000
$n = 60,000$	mean	−0.3000	0.0986	0.4998	1.0041	0.3040	5.0434
	sd	0.0149	0.0274	0.0225	0.0487	0.1060	0.4656
$n = 6,000$	mean	−0.3005	0.0965	0.5014	1.0153	0.3241	6.3590
	sd	0.0486	0.0814	0.0752	0.1784	0.4985	7.2436
$n = 600$	mean	−0.3793	−0.1150	0.6065	1.6978	0.7581	31.9937
	sd	0.9076	1.1949	0.9438	2.4341	10.7990	49.1062

Table: Empirical behavior of the maximum likelihood estimate for the validation model from sample size 200 ($\lambda = 0$).

Penalty calibration

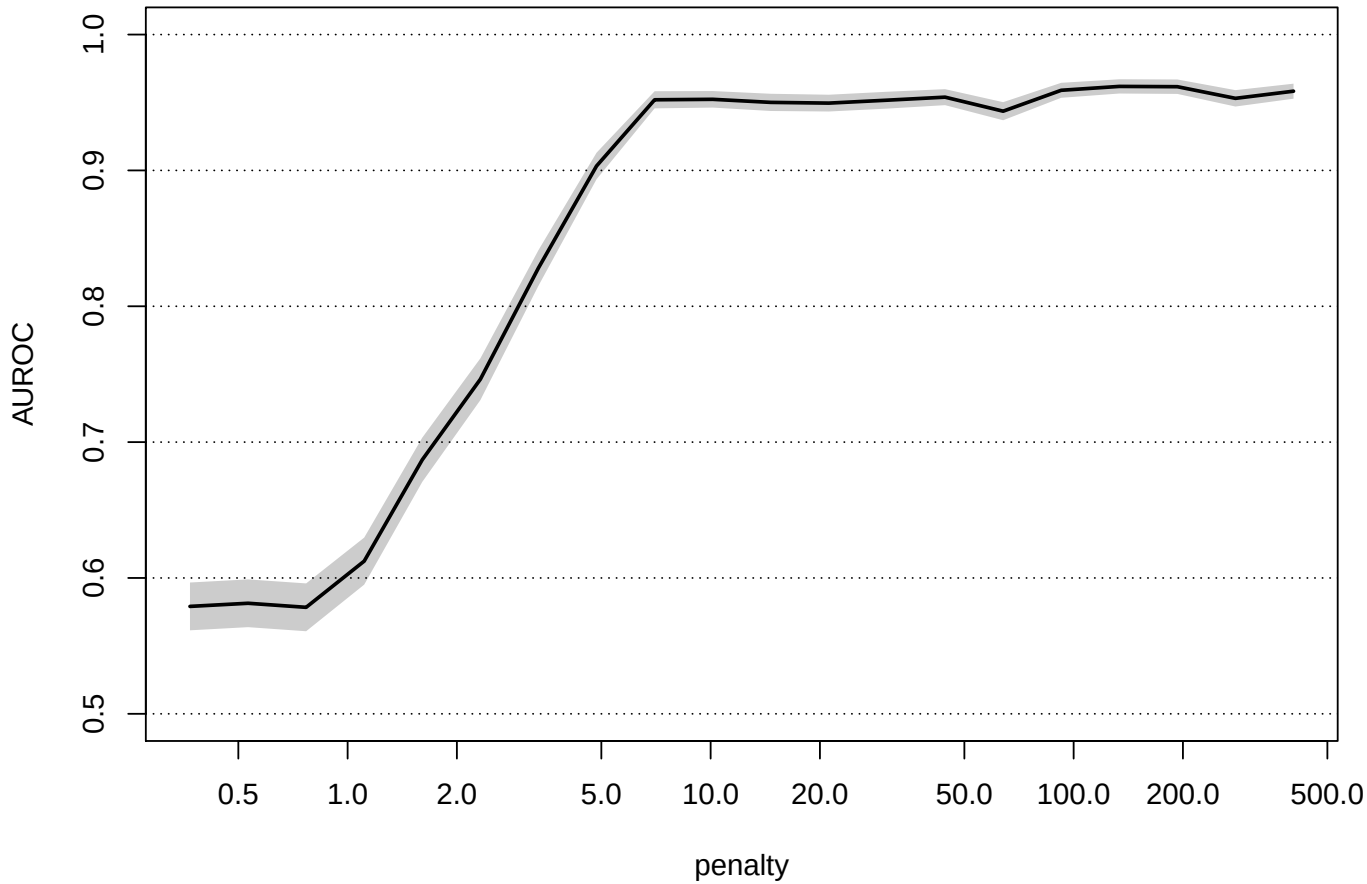


Figure: AUROC estimate according to the value of the penalty λ with $\beta = (-0.3, 0.1)$, $\delta = (0.5, \text{GE} = 0.8)$ and $\eta = (0.3, \text{PXY} = -0.5)$. AUROC and 95% confidence interval estimated using a 2, 000 samples.

We set $\lambda = 200$

The two main approaches

The Cases-controls (CC) approach

Simple logistic regression based on case-control data :

$$\text{logit}\,\mathbb{P}(y = 1|G, E, S) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E + \beta^T S$$

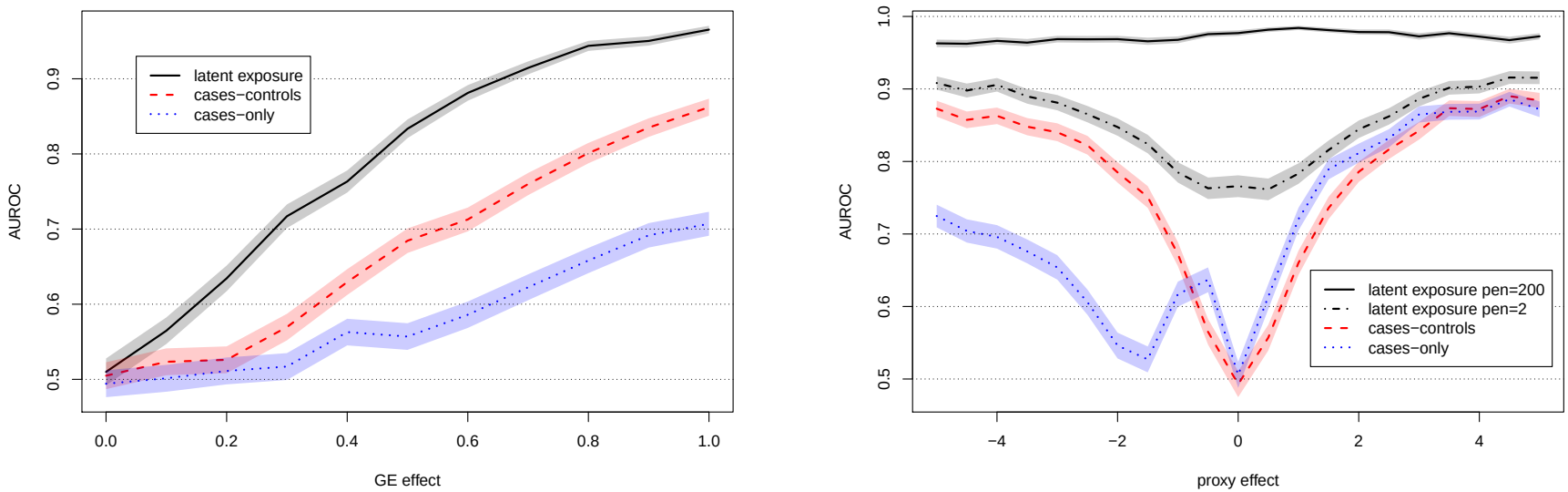
Wald test for $H0 : \beta_{EG} = 0$, based on maximum likelihood estimation, or the corresponding likelihood ratio chi-squared test.

Cases-only (CO) approach

$$\text{logit}\,\mathbb{P}(G = g|E, S, y = 1) = \gamma_0 + \gamma_{gE} E + \gamma_s^T S, \quad g = 1, 2$$

Under the assumption of G-E independence conditional on S, the likelihood ratio test for $H0 : \gamma_{1E} = \gamma_{2E} = 0$ among cases is a valid test for interaction effects.

Comparison of LRLE method, CC approach and CO approach according to GE effect and proxy effect



(a) (b)

Figure: Comparison of respectively the LRLE, the CC and the CO approach according to: (a) GE $\in [0, 1]$ and PXY $= -5$; (b) GE $= 1.0$ and PXY $\in [-5, 5]$. Simulations have been performed with the following set of parameters : $\beta = (-0.3, 0.1)$, $\delta = (0.5, \text{GE})$ and $\eta = (0.3, \text{PXY})$. AUROC and 95% confidence interval estimated using a $2,000$ samples.

Application to our complex simulated dataset

Comparison of LRLE method, CC approach and CO approach according to the number of SNP centered on causal SNP

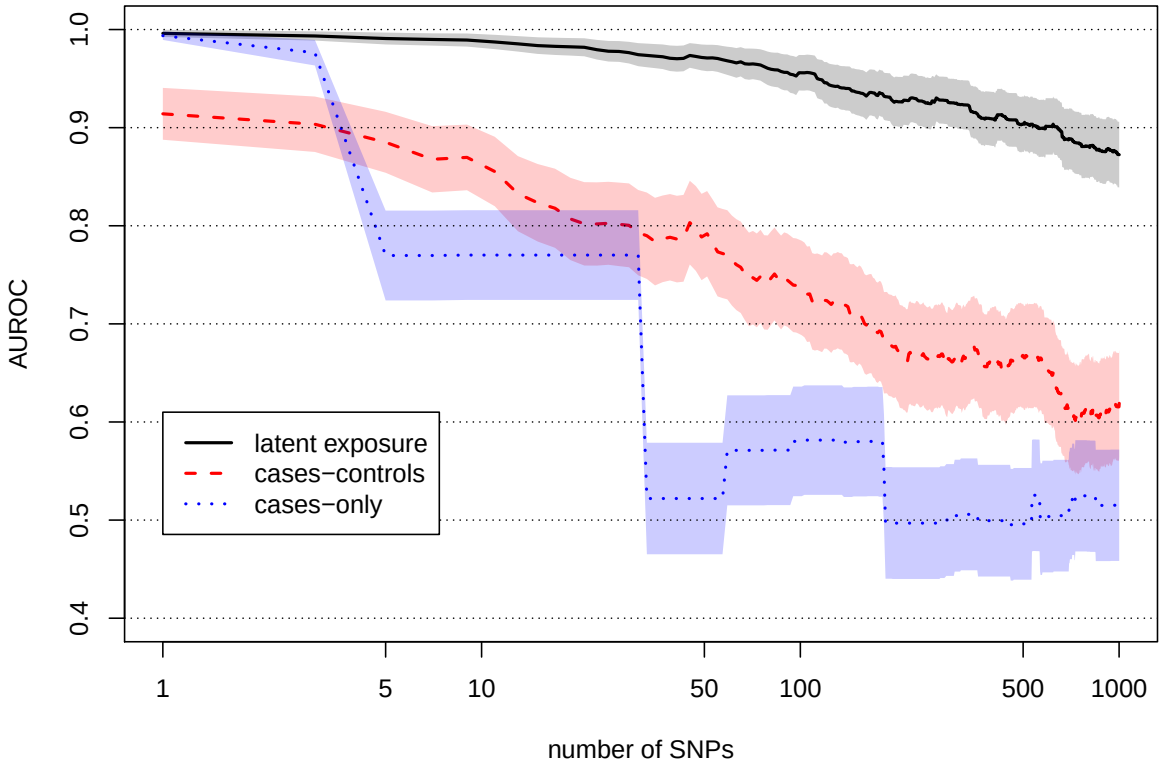


Figure: AUROC obtained with the three approaches (LRLE, CC and CO), according to the number of SNPs centered on the causal SNP. AUROC and 95% confidence interval estimated using 2, 000 samples.

Conclusion

- Presence of a latent causal exposure is an obstacle to detect GE interactions in GWAS.
- Presentation of an useful simulated dataset with realistic genotypes and covariates.
- Presentation of a new model of Logistic Regression with Latent binary Exposure (LRLE) and with a Ridge penalization that is much more powerful than the standard approaches.
- Weakness of the LRLE approach : the high regularized penalty involves a loss of information caught from the proxy.

Perspectives

- Study possible solutions to gain from the proxy information.
- To generalize the model to large set of SNPs at the same time.