Detecting Latent Exposure in Genome-Wide Association Studies using a Breakpoint Model for Logistic Regression

Flora Alarcon, Grégory Nuel

Detecting Latent Exposure in Genome-Wide Association Studies using a Breakpoint Model for Logistic Regression

Flora ALARCON[1], Gregory NUEL[2,3],

[1] Laboratoire MAP5, Université Paris Descartes and CNRS, Sorbonne Paris Cité, Paris France

[2] Institute of Mathematics (INSMI), National Center for French Research (CNRS)

[3] Stochastic and Biology Group, LPSM (CNRS 8001), Sorbonne Université, Paris, France

Abstract:

Detecting gene-environment (G ×E) interactions in the context of genome-wide association studies (GWAS) is a challenging problem since standard methods generally present a lack of power. An additional difficulty arises from the fact that the causal exposure is seldom observed and only a proxy of this exposure is observed. This leads to an additional drop in terms of power and it explains the failure of standard methods in detecting interactions, even very strong ones. In this article, we consider the latent exposure as a source of heterogeneity and we propose a new powerful method, named "Breakpoint Model for Logistic Regression" (BMLR), based on a breakpoint model, in order to detect G ×E interactions when causal exposure is unobserved. First, the BMLR method is compared to the ordered-subset analysis for case-control method, that has been developed for the same purpose, through simulations. This highlights the ability of BMLR to detect the heterogeneity, and therefore, to detect interaction with latent exposure. Finally, the BMLR method is compared to standard methods, such as Plink, to perform a GWAS on a published realistic benchmark.

Key words: Breakpoint Model, Logistic Regression, Gene-Environment Interaction, Con-

founding Factor, GWAS.

1 Introduction

In recent years, there has been a growing interest in detecting heterogeneity for complex human diseases (e.g., genetic heterogeneity, phenotypic heterogeneity, etc) in the context of genome-wide association studies (GWAS). Among others, a reason is that heterogeneity that is not considered (or detected) will involve a strong loss on the resulting power of the study.

On the other hand, detection of gene-environment interactions is of utmost interest in genetic epidemiology since it can help to identify high-risk subgroups in the population. This problem is well known to be challenging and several methods to detect gene-environment interactions (G $\times$ E) have been proposed (see e.g., [1, 2, 3, 4]). However, few loci that interact with environmental factors have been identified so far. On top of this low power, one of the reason that could explain this difficulty in detecting G $\times$ E interactions is that the causal exposure is seldom observed in practice. Indeed, it is frequent that only proxy covariates are observed instead of the causal exposure (*e.g.* body mass index (bmi) for appetite suppressant treatment). In such a situation, the latent exposure can be seen as a source of heterogeneity.

Moreover, the presence of an interaction between a genetic locus and a latent exposure can also be seen as an unexplained source of heterogeneity in GWAS. Indeed, the heterogeneity in term of exposure can be seen as a phenotypic heterogeneity (disease with or without exposure). However, classical methods for detecting G $\times$ E interactions have been shown to be inefficient when the causal exposure is unobserved [5]. Indeed, [5] has introduced a benchmark dataset simulated from data using a complex simulation framework. The goal of such benchmark dataset was to assess the power of detecting G $\times$ E interactions between a causal single-nucleotide polymorphism (SNP) and an unobserved environmental factor. Instead of this unobserved environmental factor, several proxy covariates were provided, all more or less correlated with the causal environmental factor. By comparing popular approaches, such as PLINK, Random Forest and FastLMM, this work has shown the lack of power in detecting G $\times$ E interactions when causal exposure are unobserved. Indeed, the disease model was simulated without marginal effect of the causal SNP and with a high relative risk of 50 for the interaction of the disease and the binary

latent exposure (which is an extreme scenario). Despite such a relative risk and the simulation of 595 cases and 596 controls, the power for detecting the G $\times$ E interaction on Chromosome 6 did not exceed 66% (best result obtained with the random forest) while this power was 100% when the causal exposure was artificially observed.

In the context of linkage analysis, Hauser *et al.* [6] (ordered-subset analysis) suggest identifying families subsets defined by the level of a trait-related covariate that provide maximal evidence for linkage and show that this approach might allow to uncover latent heterogeneity in the data. In [7] this idea was extended to case-control data. More precisely, to test whether the association between SNP genotypes and disease status is significantly stronger in a subset of individuals, individuals are sorted from small to high covariate values. Then, a 2×2 contingency table is created and an allelic association χ^2-statistic is computed for each subset of individuals. Finally, the subset with maximum association evidence is identified. This method, named OSACC for ordered-subset analysis for case-control), is simple but does not account for the complexity of the problem. Moreover, this method is computationally expensive and consequently, it cannot be applied to the whole genome.

In this paper, we follow these lines of works and propose a new method to detect the G $\times$ E interactions when causal exposure is unobserved by explicitly estimating the regression parameters. For that purpose, one deals with this problem as if it was a problem of detecting heterogeneity by considering the latent exposure as a source of heterogeneity. More precisely, the proposed approach allows to detect an interaction between a loci and a latent environmental exposure. This method is called BMLR for "Breakpoint Model for Logistic Regression". The main idea is to treat the latent exposure as longitudinal heterogeneity across individual in a proximity space (like the body mass index covariate or the age covariate) and to use a breakpoint model based on constrained Markov chain to detect the separation between the homogeneous groups. Then, we develop a statistic to test the presence of a genetic interaction with a latent exposure in the context of logistic regression. In comparison with OSACC, the proposed method is computationally efficient, more flexible and provides estimation for all parameters involved in the problem.

The paper is organized as follows: Section 2 introduces the BMLR model and presents validations on a simple simulated dataset. In section 3, we compare the proposed method to the

OSACC method, first on simple simulations and then on more realistic ones. Finally, our method is used to perform a GWAS on the published realistic benchmark and is compared to standard techniques (*e.g.* PLINK, random forests).

2 Methods

The Breakpoint Model

The breakpoint model aims at finding the best (in terms of data separation) breakpoint that highlights the heterogeneity due to a latent exposure in the observed data. For that purpose, the idea is to treat the latent exposure as longitudinal heterogeneity across the ordered individuals in a proximity space (*e.g.* individuals ordered by increasing bmi).

Let us consider n observations, a binary response variable $y \in \{0,1\}^n$ and covariates $X \in \mathbb{R}^{n\times p}$. If we denote by bp $\in \{1,2,\ldots,n-1\}$ the breakpoint, one decomposes $I = \{1,\ldots,n\}$ into the partition $I_1 \cup I_2$ with $I_1 = \{1,\ldots,\text{bp}\}$ (n_1 observations) and $I_2 = \{\text{bp}+1,\ldots,n\}$ (n_2 observations), and the vector y can therefore be written as $(y_1\ y_2)^T$ where $y_1 \in \mathbb{R}^{n_1\times 1}$ corresponds to the ordered observation between 1 and bp and $y_2 \in \mathbb{R}^{n_2\times 1}$ corresponds to the ordered observation between bp$+1$ and n. Similarly, we have $X = (X_1\ X_2)^T$, with $X_1 \in \mathbb{R}^{n_1\times p}$, and $X_2 \in \mathbb{R}^{n_2\times p}$.

For all breakpoint bp, we want to test the following hypothesis: $H0 : \{\text{logit } y = X\beta\}$ versus $H1 : \{\text{logit } y_1 = X_1\beta_1 \text{ and logit } y_2 = X_2\beta_2\}$ where $\beta, \beta_1, \beta_2 \in \mathbb{R}^{p\times 1}$.

For that purpose, our objective is then to find the most probable segmentation defined as the breakpoint configuration achieving the highest likelihood when fitting our regression model on each segment. Thus, the breakpoint is such that it maximizes the criterion

$$\texttt{crit}(\text{bp}) = \left\{ \max_{\beta_1} \text{loglik}(\beta_1|y_1, \underbrace{X[1:\text{bp},\ ]}_{X_1}) + \max_{\beta_2} \text{loglik}(\beta_2|y_2, \underbrace{X[\text{bp}+1:n,]}_{X_2}) \right\},$$

where loglik denote the log-likelihood function. Finally, the breakpoint is such as:

$$\text{bp}^* = \arg\max_{\text{bp}} \texttt{crit}(\text{bp})$$

Figure 1 shows likelihood values according to the breakpoint position on data simulated under a model M0 for individuals ordered from 1 to 2000 and under a model M0' (see below for a description of the various simulation models) for individuals ordered from 2001 to 3000. We note that the criterion is maximal very closed to the true breakpoint location which is in 2000.

One drawback of this strategy is that the computation of the criterion for all possible segmentations is time-consuming. This is why an efficient approach using the constrained Hidden Markov Model (HMM) introduced in Luong *et al.*(2013) [8] will be preferred here. Let us note that this method also has been used recently for detecting heterogeneity in survival responses [9].

Computational speed-up

In this section, we present a faster way to find the breakpoint that maximizes the criterion. Our approach is directly inspired from [8, 9] where the segmentation of n points into K segments ($K = 2$ in our case) is modelled through an hidden constrained Markov model.

Constrained Markov chain for breakpoint models

Let us denote by $y \in \{0,1\}^n$ the vector of (ordered) binary response variables, and by $S \in \{1,2\}^n$ the (hidden) segmentation process. Without any constraint on S, we define our model as follows:

$$\mathbb{P}(y, S; \beta) \propto \underbrace{\mathbb{P}(S_1) \prod_{i=2}^{n} \mathbb{P}(S_i|S_{i-1})}_{\text{Markov part}} \underbrace{\prod_{i=1}^{n} \mathbb{P}(y_i|S_i;\beta)}_{\text{logit part}}$$

with $\log \mathbb{P}(y_i|S_i = k;\beta) = \mathbf{1}_{y_i=1} X_i\beta_k - \log\left(1 + e^{X_i\beta_k}\right)$ and $\mathbb{P}(S_k) = \mathbf{1}_{k=1}$, $\mathbb{P}(S_i = k|S_{i-1} = j) = \mathbf{1}_{(k-j)\in\{0,1\}}$. Note that the Markov transition are improper since they do not sum to 1 which is not a problem since we only define this distribution up to a normalizing constant. Indeed, in order for this model to correspond to our breakpoint framework, we need to work conditionally to the constraint $\mathcal{C} = \{S_n = 2\}$. Using this constraint, with $n = 10$ and bp $= 6$ we get for example $S = 1111112222$.

Max forward / max backward

Now, following [8, 9] we need to introduce the so-called max-forward (denoted $F^{\max}$) and max-backward (denoted $B^{\max}$) quantities for all $i \in \{1, \ldots, n\}$ and $k \in \{1, 2\}$:

$$\left\{ \begin{array}{l} F_i^{\max}(k) = \max_{S_1,\ldots,S_{i-1}} \mathbb{P}(S_1, \ldots, S_{i-1}, y_1, \ldots y_i, S_i = k) \\ B_i^{\max}(k) = \max_{S_{i+1},\ldots,S_n} \mathbb{P}(S_{i+1}, \ldots, S_n, y_{i+1}, \ldots y_n, \mathcal{C} | S_i = k \ ) \end{array} \right.$$

with the convention that $B_n^{\max}(\cdot) \equiv 1$.

We now introduce the log-evidence as:

$$\log e_i(k) = \mathbf{1}_{y_i=1} X_i \beta_k - \log\left(1 + e^{X_i \beta_k}\right)$$

and using the results from [8] (by simply replacing the sum operator by max), we easily establish the following recursions:

$$\left\{ \begin{array}{l} \log F_1^{\max}(k) = \log\left(\mathbf{1}_{k=1} e_1(k)\right) \\ \log F_i^{\max}(k) = \log e_i(k) + \max(\log F_{i-1}^{\max}(k-1), \log F_{i-1}^{\max}(k)) \text{ for } i = 2, \ldots, n \end{array} \right.$$

and

$$\left\{ \begin{array}{l} \log B_n^{\max}(k) = 0 \text{ and } \log B_{n-1}^{\max}(k) = \log\left(\mathbf{1}_{k=2} e_n(k)\right) \\ \log B_{i-1}^{\max}(k) = \max(\log e_i(k) + \log B_i^{\max}(k), \log e_i(k+1) + \log B_i^{\max}(k+1)) \text{ for } i = n-1, \ldots, 1 \end{array} \right.$$

Once the max-forward and max-backward quantities have been computed, one can easily derive from them our quantities of interest:

$$\texttt{crit}(\theta) = \left\{ \max_{\beta_1} \text{loglik}(\beta_1 | y_1, X_1) + \max_{\beta_2} \text{loglik}(\beta_2 | y_2, X_2) \right\} = \log F_n^{\max}(2)$$

where $\theta = (\beta_1, \beta_2)$, and if we denote by S^* the segmentation maximizing this criterion, we obtain:

$$S_i^* = \arg\max_k \left\{ \log F_i^{\max}(k) + \log B_i^{\max}(k) \right\}$$

method	bp*	β_{11}	β_{12}	β_{21}	β_{22}	time
brute force	2005	-0.63077	0.30422	0.19324	-0.37095	62.7 s
max-forward	2005	-0.63078	0.30424	0.19321	-0.37090	24.1 ms

Table 1: Comparison of the performance on the brute force algorithm and max-forward on a simple example. Computations performed on a standard laptop computer.

and bp* $= i$ such that $S^*_{i+1} - S^*_i = 1$.

As a consequence, it is therefore possible to compute for any θ the objective function $\texttt{crit}(\theta)$ and the corresponding bp* with a complexity $\mathcal{O}(n)$. The optimization on θ can then be obtained through any multidimensional optimizer (*e.g.* Newton-Raphson). As an example, let us consider the following design: $n = 3000$, $\theta = (-0.7, 0.3, 0.2, -0.5)$ and bp* $= 2005$. We can see in Table 1 that both the brute force approach and the max-forward one are giving very similar results but with a noticeable difference: max-forward is roughly 3000 time faster than the brute force approach.

Likelihood Ratio (LR) test

Now, if we want to test for heterogeneity, the idea is to perform a likelihood ratio test to compare the best segmentation with $K = 2$ segments to the best (and only one) segmentation with $K = 1$ segment. In the same way, the test allows to test the interaction between a genetic covariate and a latent environmental exposure provided that the observations are ordered such as they can display an heterogeneity. For example, the data can be collected over time, or we may use a PCA over the covariate space and project the data on the first component. Then, one can fit a model with one breakpoint against the heterogeneous model. P-values can be produced using the likelihood ratio statistics. We consider the logistic regression model : logit $y = \beta_0 + \beta_1 X_P + \beta_2 G$, where X_P is considered as a nuisance covariate, correlated with a latent exposure (i.e. X_P is a proxy of the causal exposure). The binary hypothesis testing formulation is:

- $H0$: homogeneous model ($\theta = (\beta_0, \beta_1, \beta_2)$)
- $H1$: breakpoint model with one breakpoint ($\theta_1 = (\beta_{01}, \beta_{11}, \beta_{21})$ et $\theta_2 = (\beta_{02}, \beta_{12}, \beta_{22})$)

And the test statistic is defined by:

$$\text{LR} = 2 \times \left(\max_{\text{bp}} \texttt{crit}(\text{bp}) - \max_{\theta} \text{loglik}(\theta) \right).$$

Obviously, we can easily adjust on confounding factors in the previous model, increasing the dimension of the parameters vector θ.

Since the two models are clearly nested, it might seem a good idea to assume that the distribution of the LR statistic under H0 should follow a χ^2-distribution with 2 degrees of freedom (df). A simple simulated dataset was performed (as described in the following section where $n = 3000$ and bp $= 2000$). Figure 2 shows the histogram of the distribution of the LR statistic under H0 and the distribution of a $\chi^2(\text{df} = 2)$ (blue curve). We observe that it does not work since this LR test is indeed the maximum of this statistic over all possible segmentations. If all tests were independent (which is *not* the case), the resulting statistics would be distributed as the maximum of n $\chi^2(\text{df} = 2)$ which is not distributed as a χ^2.

However, the empirical distribution of the LR statistic can be easily obtained by permuting the response variable y and the observed covariates X.

Constrained breakpoint model

The method can be easily extended if parameter coordinates are fixed. For exemple, if we consider the previous logistic regression model logit $y = \beta_0 + \beta_1 X_P + \beta_2 G$ and if we are interested to test only the interaction between the genetic covariate G and the latent exposure E, we can constrained parameters β_{01} and β_{02} to be the same under $H1$ as well as β_{11} and β_{12} . Thus, the hypotheses become:

- $H0$: homogeneous model ($\theta = (\beta_0, \beta_1, \beta_2)$)
- $H1$: breakpoint model with one breakpoint ($\theta_1 = (\beta_0, \beta_1, \beta_{21})$ et $\theta_2 = (\beta_0, \beta_1, \beta_{22})$)

Simulations

In this section, we compare our BMLR method with unconstrained or constrained parameters (respectively denoted as BMLR and cBMLR) with the OSACC method on several simulated datasets.

The performance of the two approaches is assessed through the area under the receiver operating characteristic (AUROC). Let us recall that ROC curves provide a graphical representation of the specificities and sensitivities that can be obtained for all possible values of the threshold of significance [10]. All AUROC presented in this paper (including 95% confidence intervals) have been empirically obtained using finite sample size under $H0$ and $H1$ and the R package pROC [11]. An AUROC from 0.5 to 0.7, 0.7 to 0.8, 0.8 to 0.9, or above 0.9 can be respectively interpreted as a classifier of 'weak', 'good', 'excellent', or 'perfect' statistical power.

In all simulations of case-control dataset, we used the package waffect [12] to generate the phenotypic status (case or control) for a chosen causal disease model. The package waffect are freely available on the CRAN website of R package [13]. Moreover, simulations under $H0$ are simply obtained by performing a permutation on the individual sample.

A simple simulated dataset

First, the two approaches (BMLR and OSACC) are evaluated and compared on a simple scenario. We consider the following logistic regression model:

$$\text{logit } y = \beta_0 + \beta_1 G,$$

where G represents the genotypic observed exposure taking the values $\{0, 1, 2\}$ with respective probabilities $\{0.8, 0.15, 0.05\}$. $y \in \{0, 1\}$ is the binary phenotypic variable (generated with waffect package). $n = 1200$ ordered individuals are simulated, among which 600 are cases and 600 are controls. The breakpoint is arbitrarily chosen at position bp $= 500$ (so, after the 499th individual) and the number of replicates is set to 500 for the AUROC estimation with the corresponding 95% confidence intervals. Thus, the $n_1 =$ bp first observations are simulated with parameter $\theta_1 = (\beta_{01}, \beta_{11})$ and the $n_2 = n -$ bp following observations with parameters $\theta_2 = (\beta_{02}, \beta_{12})$.

In the first scenario (S1), that refers to the constrained model, β_{01} and β_{02}, the intercept for the two models before and after the breakpoint, are fixed to the same value equal to -5. β_{11} are fixed to 3.0 and β_{12} varies from 0 (i.e. an important heterogeneity before and after the breakpoint) to 3 (i.e. no heterogeneity). In the second scenario (S2), that is the unconstrained model, parameters are fixed as follow : $\beta_{11} = 3.0$ and $\beta_{12} = 2.5$. The intercept β_{02} is fixed to

-5.0 while β_{01} varies from -6.0 to -4.0. So, when β_{01} is equal to -5.0, the model is weakly heterogeneous.

A more realistic simulated dataset

This section presents a more realistic simulated dataset in order to compare the two methods. Here, we distinguish our approach between two sub-approaches: BMLR and BMLRc where coefficient of the model parameter could be fixed. First, a covariate named `bmi` (representing the body mass index) is simulated with respect to a uniform distribution between 18 and 35. The latent exposure $E \in \{0, 1\}$ is simulated according to the logit model such as the probability to be exposed for individual with $\texttt{bmi} = 18$ is 0.01 and the probability to be exposed for individual with $\texttt{bmi} = 35$ is 0.99 (i.e. $\mathbb{P}(E = 1|\texttt{bmi} = 18) = 0.01$ and $\mathbb{P}(E = 1|\texttt{bmi} = 35) = 0.99$). A genotypic exposure G is simulated as previously with values $\{0, 1, 2\}$ with respective probabilities $\{0.8, 0.15, 0.05\}$. The response covariate y is simulated with waffect according to the following model:

$$\text{logit } y = \beta_0 + \beta_1 \texttt{bmi} + \beta_2 G + \beta_3 E + \beta_4(\texttt{bmi} \times E) + \beta_5(G \times E)$$

Again 1200 ordered individuals are partitioned between 600 cases and 600 controls. We consider 100 replicates and we denote by θ the parameter vector $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

Five different models are considered in these simulations. In each model, except for model M5, we consider a marginal effect of the genotypic exposure G (corresponding to $\beta_2 = 0.7$) and a marginal effect of the observed covariate `bmi` (corresponding to $\beta_1 = 0.05$). Models are detailed here below:

- Model M0 with $\theta = (-8.3; 0.05; 0.7; 0.0; 0.0; 0.0)$ represents the null model, without any effect of latent exposure.
- Model M0' with $\theta = (-8.3, 0.05, 0.7, 1.0, 0.0, 0.0)$ represents a model with only a marginal effect of latent exposure.
- Model M3 where $\theta = (-8.3, 0.05, 0.7, 0.0, 0.0, 2.0)$ is one with an effect of interaction between the genotypic exposure and the latent exposure (no marginal effect of latent exposure neither interaction between `bmi` and latent exposure).

- Model M4 where $\theta = (-8.3, 0.05, 0.7, 1.0, 0.0, 2.0)$ is a model with a marginal effect of latent exposure and an effect of the interaction between the genotypic exposure and the latent exposure (no interaction between `bmi` and latent exposure).
- Model M5 where $\theta = (-8.3, 0.0, 0.0, 0.0, 0.0, 5.0)$ is the model that gives an advantage to the OSACC method as it does not contain any effect except a strong interaction between the genotypic exposure and the latent exposure.

Figure 3 represents boxplots of the `bmi` distribution. It should be noticed that this enlightens the link between the observed covariate `bmi` and the latent exposure E. Indeed, figure 3(a) shows that exposed individuals have a higher `bmi` than the unexposed ones. Moreover, figure 3(b) shows the `bmi` boxplots from a disease response variable Y simulated with model M3. In this case, the `bmi` distributions illustrated by the boxplots are not as different as from the exposure case.

3 Results

Comparison between BMLR and OSACC on a simple simulated dataset

In the simple simulated dataset, the goal is to detect the underlying heterogeneity. To that end, one compares our BMLR method with the OSACC one according to both parameters β_{12} and β_{01}. Figure 4 shows the AUROC with the 95% confidence intervals for the two methods in the scenario S1 (i.e. according to β_{12}). One can notice that both methods give the same perfect statistical power when the model before the breakpoint (i.e. with $\theta_1 = (\beta_{01} = -5.0\ ;\ \beta_{11} = 3.0)$) is very different from the model after (i.e. with $\theta_2 = (\beta_{02} = -5.0\ ;\ \beta_{12} = 0.0)$). Then, as expected when the two models become closer (i.e. when β_{12} get closer to $\beta_{11} = 3.0$), the statistical power decreases for both methods and converges to a weak statistical power when $\beta_{12} = \beta_{11} = 3.0$. However, the BMLR method performs always better than the OSACC one in all these scenarios, demonstrating the interest of the method. Indeed, when $\beta_{12} = 1.0$, the AUROC obtained with OSACC is around 0.5 while the AUROC obtained with BMLR is around 1.00. A blue vertical lines is drawn at $\beta_{12} = 2.5$ where the two methods have a weak statistical power. Now, we take this point ($\beta_{11} = 3.0$ and $\beta_{12} = 2.5$) and plot the AUC with the 95% confidence intervals for the two

methods , according to the β_{01} parameter (see Figure 5). This scenario do not penalized OSACC method as much.

When β_{01} diverts from β_{02}, BMLR performance rapidly increases while OSACC performance remains constant, and consequently poor. As expected, as β_{12} is fixed to 2.5, the OSACC method does not have any power as well as BMLR method when $\beta_{01} = \beta_{02} = -5.0$. However, the power obtained with BMLR is very good as soon as β_{01} deviates from β_{02}, reaching an AUC of 1 when $\beta_{01} > -4.5$ or $\beta_{01} < -5.5$.

In conclusion, BMLR performance is very promising for detecting heterogeneities while OSACC method poorly performs, even in these toy-examples scenarios.

Comparison between BMLR and OSACC on a more realistic simulated dataset

This section provides comparisons between the different methods in a more realistic scenario. First let us recall that the exposure E is not observed (i.e. it is a latent exposure). Thus, to perform the power studies, we consider the following logistic regression model:

$$\text{logit } Y = \beta_0 + \beta_1 \texttt{bmi} + \beta_2 G$$

In addition to the OSACC method, three different extensions of the BMLR method are considered:

1) The BMLR method in which the breakpoint is defined as the maximum of the quantity

$$\left\{ \max_{\theta_1} \text{loglik}(\theta_1 | y_1, X_1) + \max_{\theta_2} \text{loglik}(\theta_2 | y_2, X_2) \right\},$$

where $\theta_1 = (\beta_{01}, \beta_{11}, \beta_{21})$ is different to $\theta_2 = (\beta_{02}, \beta_{12}, \beta_{22})$

2) The BMLRc2, where both intercept and `bmi` coefficients are supposed to be fixed: $\theta_1 = (\beta_0, \beta_1, \beta_{21})$ and $\theta_2 = (\beta_0, \beta_1, \beta_{22})$. Thus we are testing for an interaction between the genotypic and a latent exposure.

3) The BMLRc1 where only the `bmi` coefficient is fixed: $\theta_1 = (\beta_{01}, \beta_1, \beta_{21})$ and $\theta_2 = (\beta_{02}, \beta_1, \beta_{22})$. Thus we are again testing for an interaction between the genotypic and a latent exposure,

Model	OSACC	BMLR	BMLRc2	BMLRc1
M0	$0.50\ [0.42-0.58]$	$0.50\ [0.42-0.58]$	$0.53\ [0.45-0.61]$	$0.58\ [0.50-0.66]$
M0'	$0.51\ [0.43-0.59]$	$0.58\ [0.50-0.66]$	$0.58\ [0.50-0.66]$	$0.60\ [0.52-0.68]$
M3	$0.53\ [0.45-0.61]$	$0.86\ [0.80-0.91]$	$1.00\ [0.99-1.00]$	$0.99\ [0.98-1.00]$
M4	$0.52\ [0.44-0.60]$	$0.87\ [0.82-0.93]$	$0.99\ [0.98-1.00]$	$0.98\ [0.96-0.99]$
M5	$0.64\ [0.46-0.62]$	$0.94\ [0.91-0.98]$	$1.00\ [1.00-1.00]$	$1.00\ [1.00-1.00]$

Table 2: AUROC for the methods according to the five simulated models.

but without fixing the intercept coefficient.

Table 2 shows the power of the methods according to the five models M0, M0', ..., M5. The M0 model is the null model, without any heterogeneity. Thus, as expected, each method has a weak power.

For models M0' to M5, the OSACC method has the weakest AUROC, demonstrating the limits of this method. Otherwise, BMLR is a powerful method to detect the interaction between the genotypic exposure and the latent exposure (models M3 to M5) with a power always greater than 0.86. Moreover, constrained BMLR methods perform better than reaching AUROC greater than 0.98 in models M3 to M5. On top of that, notice that even for model M0', that is not particularly in favor of BMLR, this method performs better than OSACC for both constrained and unconstrained cases. For completeness, we also applied the classical linear regressions of PLINK to the same models, and obtained for all of them an AUROC of 0.50 (with $[0.42-0.58]$ confidence interval), hence showing that PLINK has absolutely no power for detecting the interaction in our simulations.

In conclusion, BMLR generally achieves a very good power to detect interaction with a latent exposure while the OSACC performance shows that this method fails to detect interaction.

Application on a published realistic benchmark

Finally, these methods are applied and compared with standard methods (Plink, random forest, mixed model) on a published realistic benchmark described in [5]. Briefly, the dataset has been simulated in order to mimic a situation in which the causal exposure is unobserved but some covariates correlating with this hidden exposure are observed, such as `bmi`, `smoking`, `sex`. This dataset is based on the publicly available HapMap project datasets [14, 15] for real genotypes

with population structures (i.e. simulation of Single Nucleotide Polymorphism, SNP). Analyses were restricted to Chromosome 6 that contains more than 10 000 SNP. As in our previous simulations, waffect was used to generate phenotypes for a chosen causal disease model. Thus, we adjust for all observed covariates in the model as well as the five first PCA components. Moreover, we constrain parameters to be the same before and after the breakpoint except for the `snp` parameter. Finally, the model is written as follows:

$$\text{logit } y = \alpha + \sum_{i=1}^{5} \beta_i \texttt{pca}_i + \gamma \texttt{ sex} + \delta \texttt{ smoking} + \varepsilon \texttt{ bmi} + \eta \texttt{ snp}$$

The aim is to perform a GWAS to estimate the power of detecting the SNP that interacts with the latent exposure. Several genotypic regions are considered on Chromosome 6, centered on the causal SNP. BMLR and OSACC are compared with the power obtained in the different regions with popular methods (PLINK, random forest, linear mixed models) [5].

The OSACC method is only applied in cases where the genotypic region analyzed on Chromosome 6 was small (i.e. on causal SNP and on the region of 200 SNP centered on causal SNP). Indeed, the OSACC method is very slow and it takes several weeks to analyze the whole Chromosome 6 for 100 replicates.

Table 3 shows the AUROC estimated with the two methods appropriate for the detection of the causal SNP in the context where the causal exposure is unknown : OSACC and BMLR, according to the region on the Chromosome 6, centered on the causal SNP. As previously, two cases have been considered for BMLR : BMLR where no parameter was fixed after and before the breakpoint, and BMLRc, that fixes all parameters except the SNP coefficient (i.e. η).Concerning the OSACC method, table 3 shows that the estimated statistical power is weak, even when the region is restricted to the causal SNP.

The BMLR method reaches an excellent power to detect the SNP that interacts with the latent exposure. In particular, the BMLRc method that focuses on the G$\times$E interactions (where E denote the latent exposure) gives excellent statistical power and outperforms clearly the other methods. Indeed, the AUROC obtained for the whole Chromosome 6 is $0.96\ [0.94 - 0.99]$ and it is $0.99\ [0.99 - 1.00]$ when the region is restricted to the causal SNP.

Table 4 shows the AUROC obtained with popular methods such as PLINK, random forests (RF)

AUROC (%)	OSACC	BMLR	BMLRc
whole Chromosome 6 (100 replicats)	NA	$0.77\ [0.70 - 0.84]$	$0.96\ [0.94 - 0.99]$
8 000 SNPs region	NA	$0.80\ [0.74 - 0.87]$	$0.93\ [0.89 - 0.97]$
2 000 SNPs region	NA	$0.79\ [0.72 - 0.86]$	$0.97\ [0.95 - 0.99]$
800 SNPs regions	NA	$0.81\ [0.75 - 0.88]$	$0.96\ [0.93 - 0.98]$
200 SNPs region	$0.60\ [0.52 - 0.68]$	$0.81\ [0.76 - 0.87]$	$0.94\ [0.92 - 0.97]$
causal SNP	$0.56\ [0.48 - 0.64]$	$0.86\ [0.81 - 0.91]$	$0.99\ [0.99 - 1.00]$

Table 3: AUROC performed on the benchmark according to the region on Chromosome 6 with OSACC and BMLR approaches. Restricted regions are centered on causal SNP.

and linear mixed models (Fast-LMM) to detect G×E interaction when the causal exposure (E) is unknown, adapted from [5]. The approach referred to as "PLINK SNP" consisted in performing analysis regardless of G×E interactions by looking at the p-value associated to the significant coefficient for the SNPs, while the approach referred to as "PLINK SNP×`bmi`" accounted for interactions between the analyzed SNPs and `bmi` through the p-value associated to the significance coefficients of such interactions. For all popular methods considered, power is low, particularly when estimation is done on whole Chromosome 6.

The RF method gives results comparable to those obtained with the BMLR method, even if the BMLR method is more powerful for larger regions, which is a more realistic situation. For example, on the whole Chromosome 6, the AUROC estimated with BMLR is $0.77\ [0.70 - 0.84]$ while it is $0.66\ [0.62 - 0.73]$ with the RF method. The best estimated powers are obtained with BMLRc method, with AUROC estimations between 0.93 and 1.00.

However, the low power estimated with popular approaches shows that this methods does not suited to this context.

Moreover, the statistical power increases when the genotypic region length decreases both with BMLR estimation and with RF, while the power does not vary much with the BMLRc estimation. BMLR methods is thus able to detect interactions between a genetic locus and a latent environmental exposure in GWAS, that can be seen as a source of heterogeneity.

AUROC (%)	PLINK SNP	PLINK SNP $\times$ bmi	Fast-LMM	RF
whole Chromosome 6 (100 replicats)	$0.65\ [0.59 - 0.70]$	$0.56\ [0.51 - 0.62]$	$0.62\ [0.56 - 0.67]$	$0.66\ [0.62 - 0.73]$
8 000 SNPs region	$0.72\ [0.67 - 0.77]$	$0.55\ [0.50 - 0.61]$	$0.69\ [0.64 - 0.74]$	$0.72\ [0.67 - 0.77]$
2 000 SNPs region	$0.74\ [0.69 - 0.79]$	$0.58\ [0.52 - 0.64]$	$0.71\ [0.66 - 0.76]$	$0.76\ [0.71 - 0.81]$
800 SNPs regions	$0.82\ [0.77 - 0.86]$	$0.60\ [0.55 - 0.66]$	$0.81\ [0.76 - 0.85]$	$0.79\ [0.75 - 0.84]$
200 SNPs region	$0.85\ [0.85 - 0.92]$	$0.69\ [0.63 - 0.74]$	$0.87\ [0.83 - 0.90]$	$0.85\ [0.81 - 0.89]$
causal SNP	$0.99\ [0.98 - 1.00]$	$0.89\ [0.85 - 0.92]$	$0.99\ [0.98 - 1.00]$	$0.89\ [0.86 - 0.92]$

Table 4: AUROC performed on the benchmark according to the region on Chromosome 6 with classical approaches (PLINK, Fast-LMM, Random Forest). Restricted regions are centered on causal SNP (source: [5]).

4 Conclusion

In this article, we have proposed an original and powerful method, the BMLR, based on a breakpoint model for logistic regression, that is able to detect interaction with a latent exposure. This method is also very useful to detect SNPs that interact with a non observed exposure in GWAS. Moreover, the method allows to distinguish confounding factors from causal factors. Of course, it can also detect an effect of a latent exposure in the absence of any interaction. In addition to the latter, an important advantage of the method is its speed in maximizing the likelihood on all possible breakpoint thanks to efficiently execute max-forward/backward recursions.

The proposed method is used to perform a GWAS on a dataset previously described in [5] where a GWAS had been performed with standard methods (Plink, Random Forest and Fast-LMM), all providing a weak statistical power to detect the interaction between the causal SNP and the latent exposure. In this case, the method we proposed is shown to be able to reach a perfect statistical power. Moreover, statistical power estimations obtained with Plink were better when only the marginal effect of the SNP was tested (i.e. $0.65\ [0.59 - 0.70]$). However, if we compare our results with the results on marginal effect, again, BMLR method performs better (data not shown). The proposed method is also compared with OSACC that fails to perform well. One reason could be that OSACC maximizes the likelihood only on one side (before or after the breakpoint).

Simulations were focused on a proximity space of dimension one, since there was only one proxy of the latent exposure, easy to order (i.e. the bmi). When the dimension of proximity space is higher, principal components analysis can be performed in order to sort the data.

The purpose of the article was to build a statistical test and, as a consequence, only one

breakpoint is sufficient. However, in case of multiple breakpoints, future works can include the extension of the proposed method. Following the same way, the proposed method could be extended to deal with problems of detecting phenotypic heterogeneity. Although tis problem is more difficult, it seems to feasible to develop the appropriate methodology based on BMLR.

Finally, since we deal with test statistic, an important perspective is to derive theoretical performance, i.e., to derive the test statistic distribution under $H0$, in order to better adjust the proposed method. Furthermore, one relies on a chi-square test under $H0$ and it should be interesting to use more complex tests, such as Wald test or Rao test.

Conflict of interest. The authors declare no conflict of interest.

Aknowledgements. This work was supported by ANR SAMOGWAS.

References

[1] Peter Kraft, Y-C Yen, Daniel O Stram, John Morrison, and W James Gauderman. Exploiting gene-environment interaction to detect genetic associations. *Human heredity*, 63(2):111–119, 2007.

[2] Cassandra E Murcray, Juan Pablo Lewinger, and W James Gauderman. Gene-environment interaction in genome-wide association studies. *American journal of epidemiology*, 169(2):219–226, 2009.

[3] James Y Dai, Benjamin A Logsdon, Ying Huang, Li Hsu, Alexander P Reiner, Ross L Prentice, and Charles Kooperberg. Simultaneously testing for marginal genetic association and gene-environment interaction. *American journal of epidemiology*, 176(2):164–173, 2012.

[4] James Y Dai, Charles Kooperberg, Michael Leblanc, and Ross L Prentice. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, page ass044, 2012.

[5] Flora Alarcon, Vittorio Perduca, and Gregory Nuel. Is it possible to detect g$\times$ e interactions in gwas when causal exposure is unobserved? *Journal of Epidemiological Research*, 2(1):p109, 2015.

[6] Elizabeth R Hauser, Richard M Watanabe, William L Duren, Meredyth P Bass, Carl D Langefeld, and Michael Boehnke. Ordered subset analysis in genetic linkage mapping of complex traits. *Genetic epidemiology*, 27(1):53–63, 2004.

[7] Xuejun Qin, Elizabeth R Hauser, and Silke Schmidt. Ordered subset analysis for case-control studies. *Genetic epidemiology*, 34(5):407–417, 2010.

[8] The Minh Luong, Yves Rozenholc, and Gregory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden markov model. *Computational Statistics & Data Analysis*, 68:129–140, 2013.

[9] Olivier Bouaziz and Grégory Nuel. A change-point model for detecting heterogeneity in ordered survival responses. *Statistical Methods in Medical Research*, page 0962280217707231, 2016.

[10] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[11] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[12] Vittorio Perduca, Christine Sinoquet, Raphaël Mourad, and Gregory Nuel. Alternative Methods for H1 Simulations in Genome-Wide Association Studies. *Human Heredity*, 73(2):95–104, 2012.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[14] Gudmundur A Thorisson, Albert V Smith, Lalitha Krishnan, and Lincoln D Stein. The international hapmap project web site. *Genome research*, 15(11):1592–1593, 2005.

[15] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

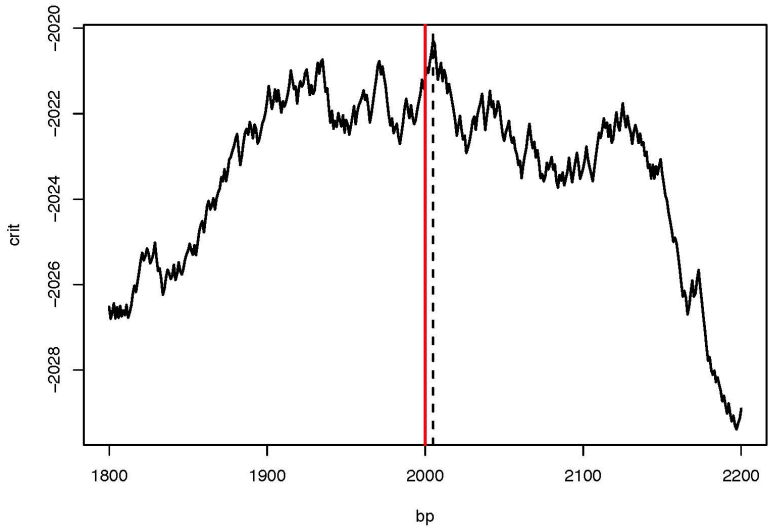


Figure 1: Criterion Values according to bp position.

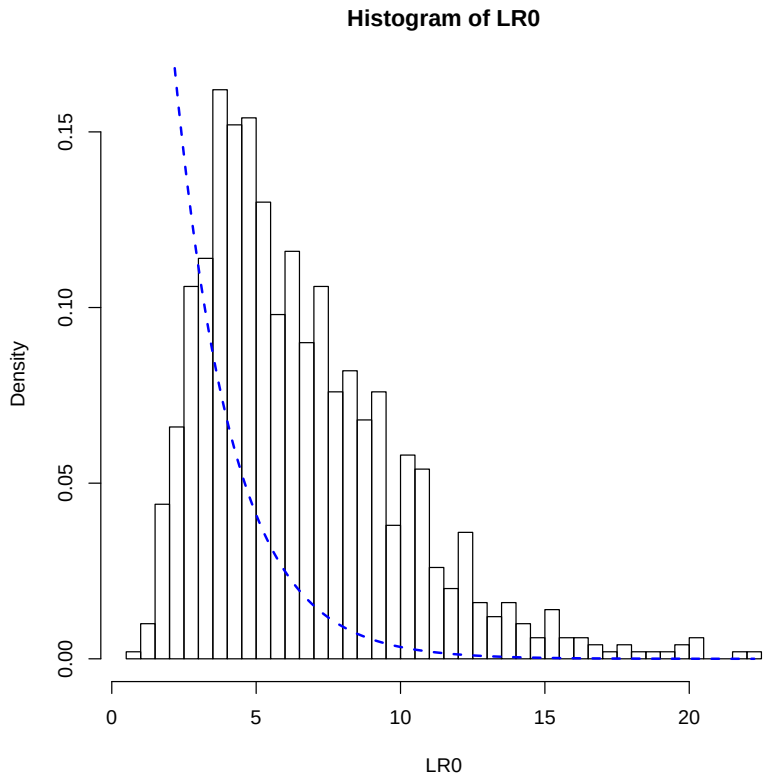
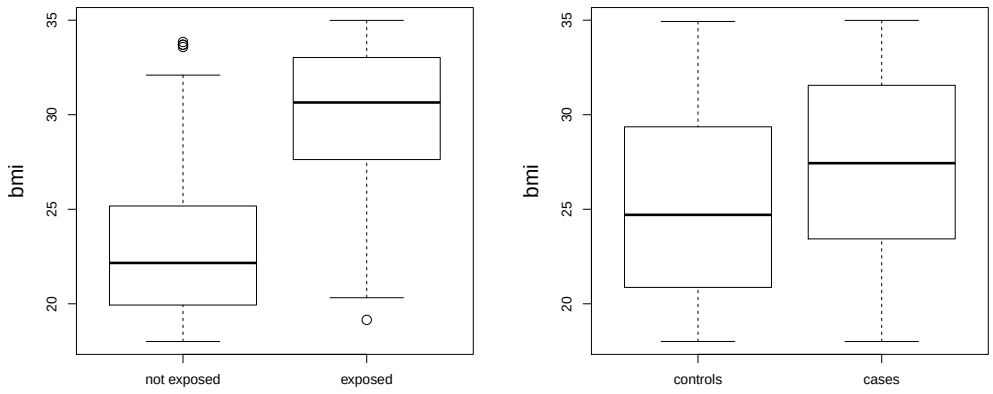


Figure 2: Histogram of LR under $H0$.



(a) Boxplot of `bmi` according to latent exposure E

(b) Boxplot of `bmi` according to case/control status in M3 simulation

Figure 3: Distribution of the observed covariate `bmi`

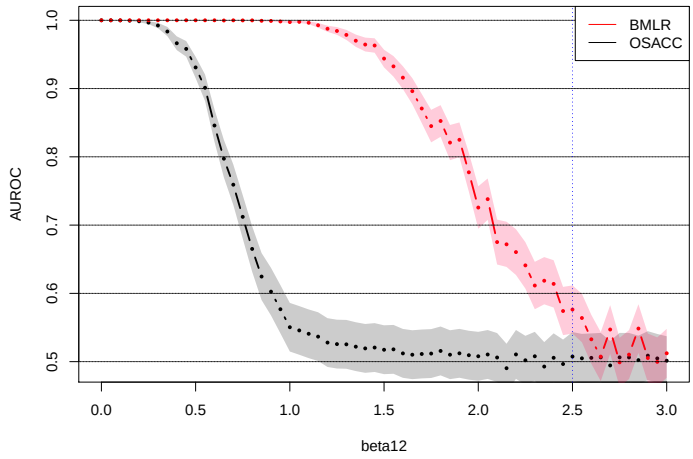


Figure 4: Comparison between BMLR and OSACC according to β_{12}, where $\beta_{01} = \beta_{02} = -5.0$ and $\beta_{11} = 3$.

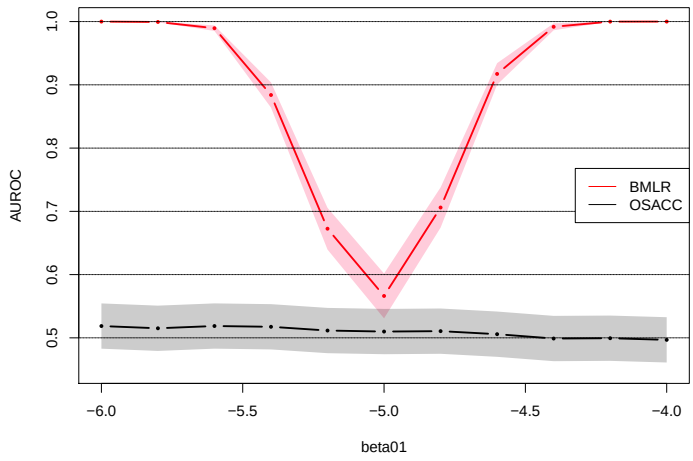


Figure 5: Comparison between BMLR and OSACC according to β_{01} where $\beta_{02} = -5.0$; $\beta_{11} = 3.0$ and $\beta_{12} = 2.5$.