# Identifying G×E or G×G interactions in case of heteroscedasticity using mixture of regression models: application to GWAS

Marie Verbanck [1] & Boris Skrobek [1] & Loïc Yengo [1]

[1] *Laboratoire de Génomique et Maladies Métaboliques, Institut de Biologie de Lille, 1 rue du Professeur Calmette, 59000 Lille. marie.verbanck@good.ibl.fr*

**Résumé.** Nous proposons d'utiliser les modèles de mélange de régressions comme outil de détection d'interactions génotype×environnement ou génotype×génotype. En effet, la très grande majorité des outils utilisés dans la littérature est basée sur des modèles d'interaction. Or les modèles d'interaction permettent de détecter des interactions en supposant un modèle homoscédastique, contrairement aux modèles de mélanges de régressions. La stratégie du modèle de mélange de régressions est comparée au test de Levene très couramment utilisé pour détecter des SNP (single nucleotide polymorphism) en interaction à la fois sur des données simulées et sur un SNP (rs7202116) présentant des différences significatives de variances de l'IMC (indice de masse corporelle) entre ses génotypes. Le modèle de mélange de régressions fournit des résultats très prometteurs sur les données simulées puisqu'il est plus performant que le test de Levene dans un grand nombre de situations. De plus, sur les données d'une cohorte française de population générale composée de 4570 individus, le modèle de mélange de régressions détecte une interaction très significative pour le SNP rs7202116 contrairement au test de Levene.

**Mots-clés.** modèle de mélange de régressions, hétéroscédasticité, interaction génotype ×environnement, interaction génotype×génotype, GWAS

**Abstract.** We propose to use mixture of regression models as a tool to detect genotype×environment or genotype×genotype interactions. Indeed, the large majority of the tools used in the literature is based on interaction models. However, interaction models allow to detect interactions supposing an homoscedastic model, contrary to mixture of regression models. The strategy of mixture of regression models is compared with Levene's test which is well-used to detect interacting SNPs (single nucleotide polymorphisms) based both on simulated data and on a SNP (rs7202116) which shows significant differences of variance of the BMI (body mass index) between its genotypes. Mixture of regression models provide very promising results on simulated data since it outperforms Levene's test in a large number of situations. Moreover, on the data from a French general population cohort composed of 4570 individuals, mixture of regression models allow to detect a very significant interaction for the SNP rs7202116 contrary to Levene's test.

**Keywords.** mixture of regression models, heteroscedasticity, genotype×environment interaction, genotype×genotype interaction, GWAS

# 1 Context

Over the past decade, GWASs (Genome Wide Association Studies) have been massively used to identify SNPs (Single Nucleotide Polymorphisms) associated with complex traits. A recurring issue to all GWAS is the so-called missing heritability. The term missing heritability was introduced to qualify the surprising difference between the expected and observed proportion of variance explained by the SNPs uncovered through GWAS. The issue of missing heritability can be tackled through various aspects, one of them is to consider more complex statistical models, particularly to take into account interactions between SNPs (genotypes) and environmental factors or other genotypes. Therefore, an interacting SNP implies that the individuals can be split up into several sub-groups for which the effect of the SNP on the trait differs.

Besides, Sun et al. (2013) [1] propose to take into account the variability of complex traits across genotypes. Indeed, the classical strategies test for differences in genotypic mean and circumvent an eventual heterogeneity of variance across genotypes by supposing an interaction model (homoscedastic model). Therefore, we usually conceptualise interactions in terms of sub-groups for which the mean effect on the complex trait differs, without considering different variances of the trait in the sub-groups. However, if we take the example of a sub-group for which the SNP has a significant effect on the trait and a second sub-group for which the SNP has no effect, the variance of the second sub-group (control sub-group) is indeed expected to be smaller. That is why we propose to focus on the detection and characterisation of interacting SNPs in case of heteroscedasticity using mixture of regression models.

This communication first recalls the general framework of mixture of regression models. Then on the basis of a two-component model, we bring out the similarities and differences between classical interaction models and mixture of regression models. This formal comparison is performed to emphasise the interest of mixture of regression models to detect interactions in case of heteroscedasticity. Finally, we propose a simulation study as well as an application on GWAS data.

# 2 Mixture of regression models versus interaction models

The general form of mixture of regression models is:

$$y_i = \sum_{k=1}^{K} Z_i \left( \beta_0^{(k)} + \beta_1^{(k)} x_i + \varepsilon_i^{(k)} \right) \tag{1}$$

with $Z_i \sim \mathcal{M}(\pi_1, ..., \pi_K)$, $\sum_{k=1}^{K} \pi_k = 1$ and $\varepsilon_i^{(k)} \sim \mathcal{N}(0, \sigma_k^2)$, $\forall k, i, i'$, $cov(\varepsilon_i^{(k)}, \varepsilon_{i'}^{(k)}) = 0$

Let us consider the simplest case where $K = 2$, then (1) is equivalent to:

$$\begin{aligned} y_i &= \beta_0^{(1)} Z_i + \beta_0^{(2)}(1 - Z_i) + \beta_1^{(1)} Z_i x_i + \beta_1^{(2)}(1 - Z_i)x_i + Z_i\varepsilon_i^{(1)} + (1 - Z_i)\varepsilon_i^{(2)} \\ y_i &= \beta_0^{(2)} + \left(\beta_0^{(1)} - \beta_0^{(2)}\right) Z_i + \beta_1^{(2)} x_i + \left(\beta_1^{(1)} - \beta_1^{(2)}\right) Z_i x_i + \left(Z_i\varepsilon_i^{(1)} + (1 - Z_i)\varepsilon_i^{(2)}\right) \\ y_i &= \gamma_0^{(2)} + \gamma_1 Z_i + \gamma_2 x_i + \gamma_3 Z_i x_i + \left(Z_i\varepsilon_i^{(1)} + (1 - Z_i)\varepsilon_i^{(2)}\right) \end{aligned} \quad (2)$$

Let us denote $\delta_i = Z_i\varepsilon_i^{(1)} + (1 - Z_i)\varepsilon_i^{(2)}$. Under the assumption that $\varepsilon_i^{(1)}$ and $\varepsilon_i^{(2)}$ follow the same distribution $\mathcal{N}(0, \sigma^2)$. If we consider that $\forall k,\ Z_i \perp\!\!\!\perp \varepsilon_i^{(k)}$ and $\forall k, k'\ \varepsilon_i^{(k)} \perp\!\!\!\perp \varepsilon_i^{(k')}$, then we can show that $\delta_i \sim \mathcal{N}(0, \sigma^2)$:

$$\begin{aligned} p(\delta_i) &= p(\delta_i|Z_i = 1)\mathbb{P}(Z_i = 1) + p(\delta_i|Z_i = 0)\mathbb{P}(Z_i = 0) \\ p(\delta_i) &= p(\delta_i|Z_i = 1)\left[\mathbb{P}(Z_i = 1) + \mathbb{P}(Z_i = 0)\right] \\ p(\delta_i) &= p(\delta_i|Z_i = 1) \end{aligned}$$

Thus (2) is equivalent to:

$$y_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 x_i + \gamma_3 Z_i x_i + \delta_i$$

In conclusion, when $Z_i$ is observed, mixture of regression models with homoscedastic components are strictly equivalent to interaction models. Therefore, mixture of regression models seem suitable to detect differences in phenotypic mean in case of heteroscedasticity as well as in case of homoscedasticity, contrary to the classical strategies which are based on interaction models.

# 3 Simulation study

We simulate a quantitative trait (denoted $y$) according to mixture of regression models: $y = Z(\beta_1^{(1)} g + e^{(1)}) + (1 - Z)(\beta_1^{(2)} g + e^{(2)}))$ with $Z$ a dummy variable and $g$ a vector of genotypes. In order to simplify the interpretation of the simulations, we assume $\beta_1^{(1)}$ to be equal to 0 and $e^{(1)} \sim \mathcal{N}(0, 1)$, that is $\sigma_1^2 = 1$. Therefore, the genotype has no mean effect on $y$ in the first sub-group which corresponds to a control group. Henceforth, the first sub-group will be called the control sub-group and the second one the effect sub-group. We vary the following parameters :

- $maf$ : minor allele frequency of the SNP (0.2 ; 0.5)
- $f_z$ : proportion of individuals in the control sub-group, it corresponds to $Z_i = 1$ (0.2 ; 0.5)
- $h^2$ : heritability of the trait in the effect sub-group (0 ; 0.1 ; 0.3), this allows to calculate $\beta_1^{(2)}$ (0 ; 0.47 ; 1.16)

- $\rho = \frac{\sigma_2^2}{\sigma_1^2}$ : ratio of the residual variances. Knowing that $\sigma_1^2 = 1$, when $\rho$ is lower than 1, the control sub-group has a higher residual variance than the effect sub-group, whereas it is the contrary when $\rho$ is greater than 1 (0.1 ; 0.3 ; 0.8 ; 1 ; 3 ; 7 ; 10)

We propose to confront mixture of regression models to Levene's test which is a well-known test to detect interacting SNPs (Struchalin et al., 2010) [2]. Mixture of regression models are adjusted using the `Flexmix` R package (Grün and Leisch, 2008) [3] and the number of components is chosen according to the BIC criterion. We compare the power of the methods to detect the interaction. To do so, for each scenario, we consider the proportion of simulated data sets for which Levene's Test has a p-value lower than a certain threshold $\alpha$ (10%). Mixture of regression models provide as many p-values associated with the genotype as the number of detected sub-groups. Thus we consider the proportion of data sets for which mixture of regression models detect the correct number of sub-groups and the p-value associated with the genotype is lower than another threshold $\alpha*$ (5%) in the effect sub-group (lower p-value).

*Note: to determine the thresholds $\alpha$ and $\alpha*$ which allow a fair comparison of both methods, we based our choice on the results obtained under the null hypothesis which corresponds to $h^2 = 0$. Results for $h^2 = 0$ are represented on Figure 1. Therefore, for $\alpha = 10\%$ and $\alpha* = 5\%$, the type-I error rate is below 10% both for Levene's test and mixture of regression models.*

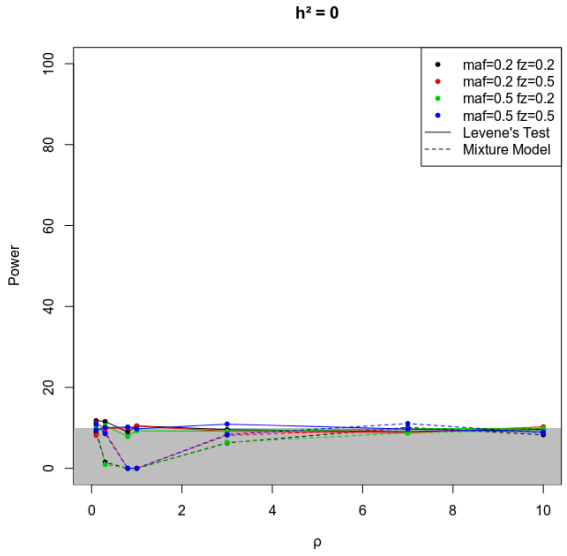


Figure 1: Results under the null hypothesis (with 3000 individuals) for Levene's test (solid lines) and mixture of regression models (dashed lines), with on the x axis the ratio of the residual variances and on the y axis the power of the methods. The colors represent various combinations of maf, $f_z$ and $h^2$.

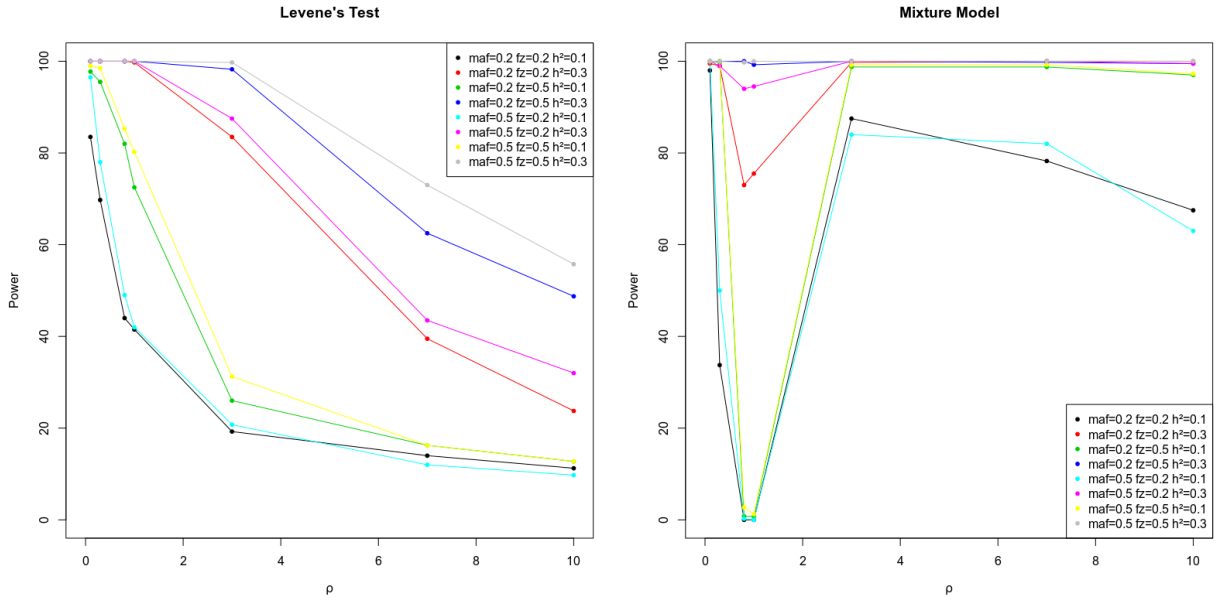


Figure 2: Results of the simulation study (with 3000 individuals) for Levene's test (left) and mixture of regression models (right), with on the x axis the ratio of the residual variances and on the y axis the power of the method. The colors represent various combinations of maf, $f_z$ and $h^2$.

Results of the simulation study are represented on Figure 2. What is particularly striking is the totally different behaviours of Levene's test and mixture of regression models. The power of Levene's test clearly decreases when $\rho$ increases, whereas the power of mixture of regression models is higher when the heteroscedasticity is strong ($\rho \ll 1$ or $\rho \gg 1$), which was expected. Not surprisingly, both methods are more efficient when the heritability is high and when the number of individuals in the sub-groups is balanced. Globally, mixture of regression models are very competitive compared to Levene's test: except for situations of homoscedasticity, mixture of regression models systematically detect a higher proportion of interactions. Moreover, in real data the variance of a control sub-group is expected to be smaller than the variance of an effect sub-group, therefore situations where $\rho$ is greater than 1 are much more plausible: mixture of regression models neatly outperform Levene's test when $\rho \gg 1$.

In addition, when the heteroscedasticity is strong, the p-value associated with the genotype in the effect sub-group detected with mixture of regression models is even smaller than the p-value associated with the genotype in the linear model in a majority of situations. It is particularly striking when $\rho$ equals 0.1, where mixture of regression models provide a smaller p-value than the linear model in over 90% of the situations.

# 4 Application

In the context of a meta-analysis involving 170,000 individuals, Yang et al. (2012) [4] showed that a SNP (rs7202116) within the FTO gene locus, known to be implicated in obesity, shows an interaction in terms of variability of the BMI (Body Mass Index). In other words, the variance of the BMI is significantly different across the three genotypes. Although the problematic is slightly different, the SNP in question seemed to be a good candidate to apply our methodology. We analysed the same SNP in 4,570 individuals from a French general population cohort.

Firstly, the p-value associated with Levene's test on the BMI of the 4,570 individuals is of 0.52 (0.56 when the test is performed on the residuals of a linear regression on the BMI taking into account the age and the sex). Therefore Levene's test does not detect any interaction. Then, a mixture of regression model is adjusted for the BMI taking into account the SNP as well as the sex and the age. Mixture of regression models do detect an interaction and the model with the lowest BIC coefficient is the model with three sub-groups. In addition, it is worth mentioning that the p-value associated with the SNP is of $2.0 \times 10^{-3}$ for the linear regression model taking into account the age and the sex, whereas the three p-values measuring the effect of the SNP (one for each sub-group) are respectively of 0.95, 0.17 and $5.7 \times 10^{-4}$, therefore the smallest p-value of mixture of regression models is lower than the p-value of the linear regression model.

Thus, the results on the SNP rs7202116 at the FTO locus are very encouraging. Mixture of regression models are a promising tool to detect genotype×environment or genotype×genotype interactions in case of heteroscedasticity.

# References


[1] Xiangqing Sun, Robert Elston, Nathan Morris, and Xiaofeng Zhu (2013), What is the significance of difference in phenotypic variability across SNP genotypes?, *The American Journal of Human Genetics*, 93(2):390–397.

[2] Maksim V. Struchalin, Abbas Dehghan, Jacqueline CM Witteman, Cornelia van Duijn, and Yurii S. Aulchenko (2010), Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations, *BMC Genetics*, 11(1):92.

[3] Bettina Grün and Friedrich Leisch. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, 2008.

[4] Jian Yang, Ruth J. F. Loos, Joseph E. Powell,... , Michael E. Goddard, and Peter M. Visscher (2012) FTO genotype is associated with phenotypic variability of body mass index, *Nature*, 490(7419):267–272.