Eigen-Epistasis : une approche pour détecter les interactions entre gènes

Virginie Stanislas [1] & Cyril Dalmasso [1] & Christophe Ambroise [1]

[1] *UMR CNRS 8071/ Université d'Évry Val d'Essonne, Laboratoire de Mathématiques et Modélisation d'Évry (LaMME), 23 bvd de France, 91 037 Évry Cedex, France ; virginie.stanislas@math.cnrs.fr, cyril.dalmasso@genopole.cnrs.fr, christophe.ambroise@genopole.cnrs.fr*

Résumé. De nombreux travaux de recherche portent sur la détection et l'étude des interactions épistatiques dans les études d'association pangénomique GWAS (Genome Wide Association Study). La plupart des publications se focalisent sur des interactions de faible ordre entre marqueurs SNPs (single-nucleotide polymorphisms) ayant des effets principaux significatifs. Dans cet article nous proposons une nouvelle approche pour détecter l'épistasie à l'échelle des gènes sans filtrage systématique aux seuls gènes significatifs. Dans un premier temps nous calculons les variables d'interaction pour chaque paire de gènes en recherchant leur Eigen-Epistasis composantes chacune définie comme la combinaison linéaire des SNPs de chaque gène ayant la plus grande corrélation avec le phénotype. La sélection des effets significatifs est effectuée à l'aide d'une méthode de régression pénalisée basée sur le group lasso contrôlant le FDR (False Discovery Rate). La méthode est comparée à trois autres approches récemment proposées dans la littérature utilisant des données sous forme synthétisées et témoignant de bonnes performances dans des contextes variés. À partir d'une étude GWAS sur la spondylarthrite ankylosante, nous démontrons la puissance de notre approche en détectant de nouvelles interactions entre gènes.

Mots-clés. Études d'association pangénomique, Interactions gene-gene, Épistasie, Group Lasso

Abstract. A large amount of research has been devoted to the detection and investigation of epistatic interactions in genome-wide association study (GWAS). Most of the litterature focuses on low-order interactions between single-nucleotide polymorphisms (SNPs) with significant main effects. In this paper, we propose an original approach for detecting epistasis at the gene level, without systematically filtering on significant genes. We first compute interaction variables for each gene pair by finding its Eigen-epistasis Component defined as the linear combination of Gene SNPs having the highest correlation with the phenotype. The selection of the significant effects results from a penalized regression method based on group Lasso controlling the False Discovery Rate. The method is compared to three recent alternative proposals from the literature using synthetic

data and exhibits high performance in different settings. Using a genome-wide association study on ankylosing spondylitis cases, we demonstrate the power of the approach by detecting new gene-gene interactions.

Keywords. Genome-wide association study, Gene-gene interactions, Epistasis, Group Lasso

1 Introduction

Les études d'association pangénomique (GWAS) visent à trouver des marqueurs génétiques (SNPs) associés à un phénotype d'intérêt. Cependant, ces études permettent d'expliquer qu'une petite partie des variations phénotypiques observées dans les études familiales classiques. Les maladies complexes peuvent en partie résulter de structures génétiques complexes comme de multiples interactions entre les marqueurs, appelés épistasie qui ne peuvent pas être prises en compte par l'approche d'analyse univariée usuelle des GWAS. Au cours des dernières années, de nombreuses méthodes ont été proposées pour détecter l'épistasie, certaines considérant les interactions à l'échelle du gène plutôt qu'à celle du marqueur génétique.

Nous proposons ici une méthode group lasso qui prend en considération la structure en groupe de chaque gène, afin de détecter l'épistasie. Nous introduisons une nouvelle approche permettant de construire les variables d'interactions appelée Gene-Gene Eigen Epistasis (G-GEE). Nous comparons G-GEE avec trois différentes autres approche de modélisation des interactions inspirées de la littérature : l'analyse en composantes principales (ACP), les moindres carrés partiels (PLS) et l'analyse canonique des corrélations (CCA). Une approche Screen and Clean, l'adaptive ridge cleaning Bécu et al. (2015), est ensuite utilisée pour déterminer les p-valeurs de chaque groupe.

2 Méthodologie

On considère n individus où $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ désigne le vecteur des traits phénotypiques. Pour chaque individu, des variants génétiques parmi G gènes sont considérés. Chaque gène est composé d'un nombre de SNPs p_g où $\sum_g p_g = p$. Dans ce qui suit, on suppose un modèle linéaire où le phénotype est considéré comme une variable aléatoire y_i dont l'espérance conditionnelle peut s'écrire comme une fonction des variables $\boldsymbol{X}_i$ et de leurs interactions $\boldsymbol{Z}_i$,

$$E[y_i|X] = \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{\gamma},$$

où

$$\boldsymbol{\beta} = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, ..., \beta_{1,p_1}}_{gene_1}, ..., \underbrace{\beta_{G,1}, ..., \beta_{G,p_G}}_{gene_G} \right)^T,$$

et $\boldsymbol{Z}_i$ est la i-ème ligne de la matrice d'interactions et $\boldsymbol{\gamma}$ un vecteur de paramètres de dimension appropriée.

L'effet principal de chaque gène est ainsi modélisé par la somme de tous les effets des SNPs. En ce qui concerne les effets d'interactions, nous calculons de nouvelles variables représentant l'interaction d'une paire de gènes donnée et définissons comme un groupe toutes les variables d'interactions liées à cette paire. La matrice d'interactions est ainsi structurée en $G(G-1)/2$ sous-matrices :

$$\boldsymbol{Z} = [\boldsymbol{Z}^{11} \cdots \boldsymbol{Z}^{rs} \cdots \boldsymbol{Z}^{G(G-1)/2}]$$

où $\boldsymbol{Z}^{rs}$ décrit l'interaction entre les deux gènes r et s. Le vecteur de paramètres $\boldsymbol{\gamma}$ est structuré en sous-vecteurs $\boldsymbol{\gamma}^{rs}$.

2.1 Modélisation des interactions avec Eigen-epistasis

L'idée générale de notre approche est de considérer la variable d'interaction entre les deux gènes r et s comme une fonction $f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)$ paramétrée par $\boldsymbol{u}$. une façon d'estimer $\boldsymbol{u}$ est de maximiser la corrélation entre la fonction d'interaction et le phénotype :

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cor(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)).$$

Si on considère que la fonction f est linéaire le problème devient facilement traitable avec une solution unique. Posons

$$\boldsymbol{Z}^{rs} = f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s) = \boldsymbol{W}^{rs}\boldsymbol{u},$$

où $\boldsymbol{W}^{rs} = \{X_{ij}^r X_{ik}^s\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$ et $\boldsymbol{u} \in \mathbb{R}^{p_r p_s}$ nous obtenons le problème suivant :

$$\max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||\hat{cor}[\boldsymbol{W}^{rs}\boldsymbol{u}, \boldsymbol{y}]||^2 = \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||\boldsymbol{u}^T \boldsymbol{W}^{rsT} \boldsymbol{y}||^2 = \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{W}^{rsT} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs}\boldsymbol{u} \quad . \qquad (1)$$

La solution $\boldsymbol{u}$ est le vecteur propre associé à la plus grande valeur propre de la matrice $\boldsymbol{W}^{rsT}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^{rs}$. Nous utilisons ensuite la projection de la matrice $\boldsymbol{W}^{rs}$ sur $\boldsymbol{u}$ comme variable d'interaction. Le vecteur Eigen-epistasis $\boldsymbol{Z}$ résultant est la combinaison linéaire de toutes les interactions SNP-SNP la plus corrélée avec le phénotype. Cette méthode permet de prendre en compte l'information phénotypique dans la construction des variables d'interactions.

2.2 Estimation des coefficients

Nous proposons un modèle groupe lasso pour estimer les paramètres du modèle. Un groupe peut être soit défini à partir de l'ensemble des SNPs d'un gène donné ou soit à partir de l'ensemble des termes d'interactions d'une paire de gène donnée.

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} \left(\sum_i (y_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{\gamma})^2 + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\boldsymbol{\gamma}^{rs}||_2 \right] \right) \quad ,$$

Afin d'améliorer la précision de l'estimation et d'obtenir des p-valeurs pour chacun des groupes sélectionnés, nous utilisons l'approche Screen and Clean proposée par Bécu et al. (2015). Cette approche est une méthode en deux étapes. Le modèle groupe lasso est premièrement ajusté sur la moitié des données. Les coefficients des groupes candidats sélectionnés par le modèle sont ensuite introduits dans un modèle de régression Ridge ajusté sur la seconde moitié des données avec une pénalité spécifique permettant de prendre en compte la structure en groupe. Des tests de permutation sont ensuite réalisés afin d'estimer la significativité des coefficients de régression pour chaque groupe.

3 Simulations

Une étude de simulation est réalisée afin d'évaluer la performance de G-GEE. Nous comparons notre approche à trois autres méthodes permettant de modéliser les interactions inspirées de la littérature. Les données sont simulées selon le design suivant.

3.1 Design

Les n lignes de la matrice des génotypes sont des échantillons i.i.d. issus d'un vecteur aléatoire multivarié $\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. La matrice de corrélations $\boldsymbol{\Sigma}$ est structurée en blocs correspondant aux différents gènes. Les valeurs de la matrice X_{ik} sont ensuite discrétisées en 0, 1 ou 2 à partir des fréquences génotypiques dérivées de l'équation de Hardy-Weinberg.

Les vecteurs de valeurs phénotypiques sont générées selon un modèle proposé par Wang et al. (2014) :

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s \right) + \epsilon_i, \tag{2}$$

où $\mathcal{C}$ représentent l'ensemble des SNPs causaux, $\mathcal{C}^2$ l'ensemble des interactions causales, et ϵ_i une variable aléatoire gaussienne. Pour chaque gène g, deux SNPs causaux sont considérés et un coefficient β_g est attribué à la somme standardisée de ces SNPs causaux. Cette même idée est reprise pour les interactions où tous les SNPs causaux d'une paire de gènes (r, s) causale sont multipliés deux à deux avec un coefficient γ_{rs} attribué à la somme standardisée de ces produits.

3.2 Autres méthodes pour modéliser les interactions

3.2.1 Analyse en Composantes Principales

Une manière de représenter l'interaction entre deux gènes r et s est d'utiliser le produit des composantes C^r et C^s obtenues à l'aide d'une Analyse en Composantes Principales (PCA) sur chacun des gènes. Plusieurs variables d'interactions peuvent être retenues pour représenter un même couple en fonction du nombre de composantes principales q choisies pour chacun des gènes. Dans ce contexte le terme d'interaction prend la forme suivante :

$$\boldsymbol{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}C_{ij}^{r}C_{ik}^{s}.$$

3.2.2 Analyse Canonique des Corrélation

L'Analyse Canonique des Corrélations (CCA) vise à trouver des combinaisons linéaires pour deux groupes de variables ayant une corrélation maximale. Dans notre contexte nous considérons chaque gène en tant que groupe de SNPs. Pour deux gènes r et s, nous définissons les nouvelles variables $\boldsymbol{A}^r$ et $\boldsymbol{B}^s$ qui sont des combinaisons linéaires des variables d'origine $\boldsymbol{X}^r$ et $\boldsymbol{X}^s$:

$$\begin{cases}\boldsymbol{A}^r = \boldsymbol{X}^r U^r, \\ \boldsymbol{B}^s = \boldsymbol{X}^s V^s\end{cases}$$

U^r et V^s sont les matrices dont les colonnes définissent les vecteurs de poids, obtenus par CCA. Nous proposons de représenter l'interaction d'un couple de gènes (r, s) par le produit des premier q couples de composantes obtenues par CCA :

$$\boldsymbol{Z}_i^{rsT}\boldsymbol{\gamma}^{rs} = \sum_{j=1}^{q}\gamma_{j}^{rs}A_{ij}^{r}B_{ij}^{s}.$$

3.2.3 Moindres carrés partiels

Wang et al. (2009) propose une méthode alternative pour représenter les interactions en utilisant une approche par moindres carrés partiels (PLS). Soit $(\boldsymbol{X}^r, \boldsymbol{X}^s)$ la matrice génotypique pour une paire de gènes (r, s) donnée. L'approche de Wang et al. (2009) calcule les composantes qui maximisent $cov^2(\boldsymbol{X}^r\boldsymbol{u}, \boldsymbol{T}\boldsymbol{v})$, avec $\boldsymbol{T} = (\boldsymbol{y}, \boldsymbol{X}^s)$ et $(\boldsymbol{u}, \boldsymbol{v})$ des vecteurs de poids. Cette approche permet de garder l'information phénotypique dans la construction des variables d'interactions.

3.3 Résultats

On considère un scénario simple avec 6 gènes pour 600 individus reproduit sur 1000 simulations. Chaque gène est composé de 6 SNPs. On considère une interaction causale

entre le gène 3 et le gène 4 et deux effets principaux pour les gènes 1 et 2. La même valeur est attribuée aux coefficients ($\beta_g = \gamma_{rs} = 2, \forall g, r, s$). On compare ensuite la puissance de chacune des méthodes sous différentes valeurs (allant de 0.05 à 0.7) pour le coefficient de détermination r^2.

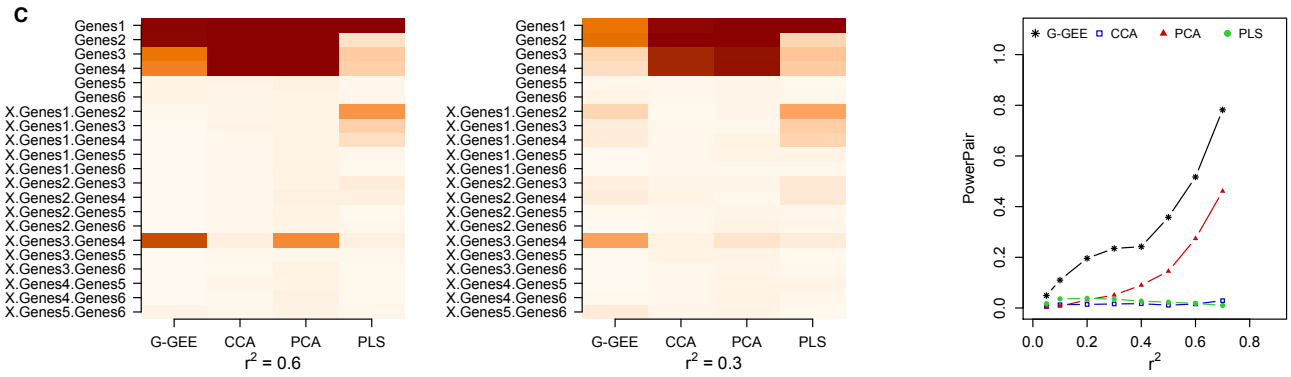

FIGURE 1 – Les deux premières figures représentent la part du nombre de fois où la variable est considérée comme significative sur le nombre total de simulations pour un r^2 donné. La dernière figure représente la puissance de chaque méthode en fonction du r^2.

Dans l'ensemble la méthode G-GEE témoigne d'une bonne puissance pour détecter les interactions par comparaison aux autres approches. La méthode PLS se caractérise par un manque de puissance pour détecter les interactions. Le contexte est plus favorables pour cette méthode lorsque les effets principaux des gènes en interaction sont également présents (données non présentées ici). Les méthodes basées sur l'ACP et l'ACC détectent surtout les effets principaux. Elles rencontrent des difficultés dans la détection des effets d'interactions souvent confondus comme des effets principaux.

Bibliographie

[1] Bécu JM, Grandvalet Y, Ambroise C, Dalmasso C (2015), *Beyond Support in Two-Stage Variable Selection*, ArXiv150507281 Stat .

[2] Wang X, Zhang D, Tzeng JY (2014), *Pathway-Guided Identification of Gene-Gene Interactions*, Annals of Human Genetics 78 :478–491.

[3] Wang T, Ho G, Ye K, Strickler H, Elston RC (2009), *A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped*, Genet Epidemiol 33 :6–15.