Procédure de test d'hypothèses composites pour l'analyse jointe de séries de probabilités critiques.

Annaïg De Walsche [1,2,*] & Franck Gauthier [2] & Alain Charcosset [2] & Tristan Mary-Huard [1,2]

[1] *UMR MIA Paris-Saclay, INRAE, AgroParisTech, Université Paris-Saclay, 91120 Palaiseau, France*

[2] *UMR Génétique Quantitative et Evolution - Le Moulon, INRAE, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France*

[*] *Corresponding author: annaig.de-walsche@inrae.fr*

Résumé. L'analyse jointe de résultats de différentes expériences pour identifier des configurations complexes est un objectif typique de l'intégration de données. On considère ici le cas d'une collection d'éléments $i = 1, ..n$ (par exemple des gènes) pour lesquels les hypothèses H_{0i}^q : « l'élément i n'a pas d'effet dans la condition q » ont été testées pour Q conditions. Chaque observation i consiste donc en un vecteur de Q probabilités critiques. L'objectif de l'analyse est alors d'identifier les éléments qui ont un effet dans toutes les conditions ou dans un sous-ensemble prédéfini de conditions. Les probabilités critiques doivent alors être combinées de manière flexible afin d'explorer des hypothèses complexes (appelées hypothèses composites), tout en contrôlant le taux de faux positif. Nous proposons une procédure de test d'hypothèses composites utilisant un modèle de mélange multivarié où chaque Q-uplet de probabilités critiques appartient à une des 2^Q classes caractérisée par une combinaison spécifique d'états de H_0^q et H_1^q. Notre méthode prend en compte la structure de dépendance entre les Q probabilités critiques, qui est modélisée dans les lois jointes conditionnelles à l'aide d'une fonction copule. L'inférence de ce modèle de mélange à 2^Q composantes est réalisée efficacement permettant son application à des cas où le nombre de marqueurs est en $\mathcal{O}(10^5)$, et où $Q = 20$. Elle consiste en deux étapes indépendantes : tout d'abord l'ajustement d'un modèle de mélange non paramétrique sur la distribution marginale de chacune des Q séries de probabilités critiques, puis l'estimation des proportions des composantes du modèle de mélange et des paramètres de copule via un algorithme EM. L'étape (E) est optimisée pour limiter l'empreinte mémoire de la procédure, passant de $O(n \times 2^Q)$ à $O(n + 2^Q)$. Des applications sur des données simulées ont été réalisées donnant des résultats concluants tant en termes de contrôle de faux positif et de puissance de détection qu'en terme d'efficacité de la méthode (temps de calcul et gestion de la mémoire). L'intérêt de la méthode est illustré par une analyse conjointe d'études d'association génétique afin de détecter des gènes pléiotropes parmi un ensemble de 14 troubles psychiatriques.

Mots-clés. Hypothèse composite, modèle de mélange, tests multiples, intégration des données, pléiotropie.

Abstract. Data integration often involves analysing results from different experiments to identify complex patterns. In this context, we consider a scenario where we have a collection of elements $i = 1, ...n$ (genes, for example) for which the hypotheses H_{0i}^q: "element i has

no effect in condition q" have been tested for Q conditions. Each observation i, therefore, consists of a vector of Q critical probabilities. The analysis aims to identify the elements that have an effect in all the conditions or a predefined subset of those conditions. The critical probabilities must then be combined flexibly to explore complex hypotheses (called composite hypotheses) while controlling the false positive rate. To achieve this, we need to combine the critical probabilities in a flexible way to explore complex hypotheses (called composite hypotheses) while controlling the false positive rate. We propose a composite hypothesis testing procedure based on a model where the Q-uplet of p-value associated with each gene/marker is distributed as a multivariate mixture where each of the 2^Q components corresponds to a specific combination of H_0^q and H_1^q states. Our method explicitly accounts for the dependence structure across p-value series through a copula function. The inference of this 2^Q component mixture model is performed efficiently, allowing its application to cases where the number of markers is $\mathcal{O}(10^5)$, and where $Q = 20$. The inference procedure consists of two independent steps: first, fitting a non-parametric mixture model to the marginal distribution of each Q series of p-values, then estimating the proportions of the mixture model components and the copula parameters using an EM algorithm. Step (E) is optimised to reduce the memory burden of the procedure from $O(n \times 2^Q)$ to $O(n + 2^Q)$. Applications on simulated data have been carried out, with conclusive results regarding false positive control and detection power and the method's efficiency (computation time and memory management). The interest in the method is illustrated by a joint analysis of genetic association studies to detect pleiotropic genes among 14 psychiatric disorders.

Keywords. Composite hypothesis, mixture model, multiple testing, data integration, pleiotropy.

1 Introduction

Considérons une étude dont l'objectif est d'évaluer l'effet conjoint d'un traitement sur deux tissus différents. On cherche alors à définir une procédure de test qui rejette l'hypothèse H_0 "le traitement n'a pas d'effet conjoint", lorsque les deux hypothèses H_0^1 "le traitement n'a pas d'effet sur le tissu 1" et H_0^2 "le traitement n'a pas d'effet sur le tissu 2" sont fausses. Cela correspond à un cas particulier de test d'hypothèse composite où l'hypothèse composite H_0 à tester est $H_0^1 \cup H_0^2$. Une approche courante pour effectuer des tests d'hypothèses composites (THC) consiste à combiner les statistiques de test et/ou les probabilités critiques dérivées pour chacune des hypothèses marginales H_0^1 et H_0^2 en une seule statistique globale. Si les probabilités critiques associées à H_0^1 et H_0^2 peuvent usuellement être obtenues à l'aide de procédures statistiques classiques, construire une statistique de test adéquate (avec une distribution connue sous H_0) ainsi qu'une procédure de test valide (garantissant le contrôle du taux de faux positifs au niveau nominal requis) pour le test de l'hypothèse composite H_0 n'est pas trivial. Le test d'hypothèse composite de la forme $H_0^1 \cup H_0^2$ a été étudié dès le début des années 80, avec les travaux de [13], et son extension au cas $H_0^1 \cup ... \cup H_0^Q$ avec $Q \geq 2$ a été explorée par [2]. En génétique, le THC peut être utilisé pour analyser conjointement les résultats issus de plusieurs analyses d'association, réalisées à partir de panels non disjoints

(i.e. une partie des individus est commune aux différents panels). ([5, 16, 8]). Dans un tel cas le THC est réalisé au niveau du marqueur, ce qui entraîne un grand nombre d'hypothèses composites testées simultanément. Par ailleurs les différentes probabilités critiques collectées pour un même marqueur ne peuvent être considérées comme indépendantes du fait de la présence d'individus communs à tous les panels.

Nous proposons une approche de THC basée sur un modèle où le Q-uplet de probabilités critiques associé à chaque marqueur est issue d'un mélange multivarié où chacune des 2^Q composantes correspond à une combinaison spécifique d'états H_0^q et H_1^q. La méthode, appelée **qch_copula**, prend en compte la structure de dépendance entre les séries de probabilités critiques à l'aide d'une fonction copule. Nous montrons comment l'inférence d'un tel modèle de mélange à 2^Q composantes peut être réalisée efficacement, permettant son application à des cas où le nombre de marqueurs est en $\mathcal{O}(10^5)$, et où $Q = 20$. La procédure est illustrée sur des données simulées, ainsi que sur un exemple de détection de gènes pléiotropes parmi un ensemble de 14 troubles psychiatriques en analysant conjointement des études d'association génétiques. Les performances obtenues en termes de puissance de détection et de contrôle du taux d'erreur de type I sont très supérieures à celles des méthodes concurrentes.

2 Hypothèse composite

On considère une collection d'éléments $i = 1, ...n$ (par exemple des gènes ou des SNP) dont l'effet a été testé dans $q = 1, ..., Q$ conditions. Nous désignons par H_0^q (resp. H_1^q) l'hypothèse nulle (resp. alternative) correspondant au test q $(1 \leq q \leq Q)$ et considérons l'ensemble $\mathcal{C} := \{0, 1\}^Q$ de toutes les combinaisons possibles d'hypothèses nulles et alternatives parmi les Q. Pour une configuration donnée $c := (c_1, ..., c_Q) \in \mathcal{C}$, l'hypothèse conjointe $\mathcal{H}^c$ se définit comme suit

$$\mathcal{H}^c := \left(\bigcap_{q:c_q=0} H_0^q \right) \cap \left(\bigcap_{q:c_q=1} H_1^q \right)$$

Etant donné deux sous-ensembles complémentaires $\mathcal{C}_0$ et $\mathcal{C}_1$ de $\mathcal{C}$ nous définissons les hypothèses composites nulle $\mathcal{H}_0$ et alternative $\mathcal{H}_1$ telles que :

$$\mathcal{H}_0 := \bigcup_{c \in \mathcal{C}_0} \mathcal{H}^c, \quad \mathcal{H}_1 := \bigcup_{c \in \mathcal{C}_1} \mathcal{H}^c$$

L'objectif est ici de tester $\mathcal{H}_0$ par rapport à $\mathcal{H}_1$ pour chaque élément i $(1 \leq i \leq n)$.

3 Modèle

On désigne par P_i^q la probabilité critique obtenue pour le test q sur l'élément i. Et on définit le z-score : $Z_i^q = -\Phi^{-1}(P_i^q)$, où Φ représente la fonction de répartition de la loi Gaussienne

standard. On note $Z_i := (Z_i^1, ..., Z_i^Q)$ le vecteur contenant les z-scores de l'élément i.

A chaque élément i est associé un vecteur $L_i := (L_i^1, ..., L_i^Q) \in \mathcal{C}$, où L_i^q est la variable binaire étant égale à 0 si H_{0i}^q est vraie et 1 si H_{1i}^q est vraie. En supposant que les éléments sont indépendants, chaque vecteur de z-scores est issu d'un modèle de mélange avec 2^Q composantes défini comme suit :

$$Z_i \sim \sum_{c \in \mathcal{C}} w_c \psi^c. \tag{1}$$

où ψ^c est la loi de Z_i conditionnellement à $L_i = c$ et $w_c = \Pr\{L_i = c\}$.

Le modèle de mélange (1) implique 2^Q distributions multivariées ψ^c devant être estimées. Dans la suite, nous faisons l'hypothèse que les fonctions de répartition Ψ_c associées ont la forme suivante :

$$\Psi_c^\theta(Z_i) = C_\theta(F_{c_1}^1(Z_i^1), ..., F_{c_q}^q(Z_i^q), ..., F_{c_Q}^Q(Z_i^Q))$$

où F_0^q (resp. F_1^q) est la fonction de répartition marginale de Z_i^q conditionnellement à $L_i^q = 0$ (resp. $L_i^q = 1$) et C_θ est une fonction copule de paramètre θ modélisant la structure de dépendance entre les Q z-scores. Les distributions ψ_c s'écrivent alors:

$$\psi_\theta^c(Z_i) = c_\theta \left(F_{c_1}^1(Z_i^1), ..., F_{c_Q}^Q(Z_i^Q) \right) \prod_{q:c_q=0} f_0^q(Z_i^q) \prod_{q:c_q=1} f_1^q(Z_i^q) \tag{2}$$

où f_0^q (resp. f_1^q) est la fonction densité marginale de Z_i^q conditionnellement à $L_i^q = 0$ (resp. $L_i^q = 1$). Ainsi, seules $2Q$ fonctions de répartition univariées $F_0^1, ..., F_0^Q, F_1^1, ...F_1^Q$, les probabilités w_c ainsi que le paramètre θ sont à estimer. Nous considérons ici la copule gaussienne, qui permet de spécifier des corrélations différentes pour chaque paire (q, q') de conditions.

Les paramètres du modèle sont les distributions $f_1^1, ...f_1^Q$, les proportions du mélange w_c et le paramètre de la copule θ. Pour réduire le temps de calcul de la procédure d'inférence, nous proposons de la diviser en deux étapes :

1. Ajuster un modèle de mélange sur chaque ensemble de z-scores $\{Z_i^q\}_{1 \leq i \leq n}$ afin d'obtenir une estimation de chaque distribution alternative f_1^q.

2. Estimer les proportions w_c de chaque configuration c et le paramètre de la copule θ à l'aide d'un algorithme EM, après avoir incorporé les estimations $\hat{f}_1^q$.

La distribution marginale des z-scores Z_i^q associés au q-ième test peut être déduite du modèle (1) combiné à la définition du ψ_c (2). On a

$$Z_i^q \sim \pi_0^q f_0^q + (1 - \pi_0^q) f_1^q, \tag{3}$$

où f_0^q est la fonction densité de Z_i^q conditionnellement à $L_i^q = 0$ et f_1^q sa fonction densité conditionnellement à $L_i^q = 1$.

Par définition des z-scores Z_i^q, leur distribution nulle (c'est-à-dire la distribution conditionnelle à $L_i^q = 0$) est la loi Gaussienne standard. On a ainsi $f_0^q = \phi$ pour tous les q, où

ϕ représente la fonction de densité gaussienne standard. Pour estimer les proportions nulles π_0^q, utilisons l'estimateur introduite par Storey [15]. Les distributions alternatives f_1^q sont estimées de manière non paramétrique à l'aide d'une méthode d'estimation à noyau.

Les estimations de $\hat{f}_1^q$ peuvent être utilisées pour calculer le $\hat{F}_1^q$. Ces estimations peuvent être directement introduites dans le modéle de mélange, de sorte que les seules quantités à estimer sont les proportions w_c des 2^Q composantes de mélange et le paramètre de la copule θ. L'inférence peut être réalisée efficacement à l'aide d'un EM standard.

Notre méthode permet d'obtenir des estimations des probabilités *a posteriori* d'appartenir à une configuration $c \in \mathcal{C}_1$, pouvant être utiliser comme statistique de test, de la manière suivante :

$$\hat{\tau}_i = \sum_{c \in \mathcal{C}_1} \widehat{\Pr}\{L_i = c | Z_i; \hat{\theta}\},$$

4 Illustration : Identification de gènes pléiotropes dans une étude portant sur 14 troubles psychiatriques

Afin d'illustrer notre méthode, nous avons analysé conjointement 14 troubles psychiatriques à partir de résultats d'études d'association génétique issues du Psychiatric Genomics Consortium (PGC). Les troubles étudiés sont l'anorexie mentale [4], l'anxiété [10], l'autisme [6], les troubles liés à la consommation d'alcool [12], le trouble obsessionnel compulsif [1], la bipolarité [14], la schizophrénie [7], le stress post-traumatique [9], le syndrome de la Tourette [18], la consommation de cannabis [11], la dépression majeure [17] et le trouble déficitaire de l'attention avec ou sans hyperactivité [3]. Les études d'association génétiques ont été réalisées sur $n = 17,425$ gènes, et l'objectif de l'analyse est d'identifier des gènes pléiotropes, c'est-à-dire associés simultanément à plusieurs troubles. Ici, nous nous sommes intéressés particulièrement aux gènes associés à au moins 8 troubles différents. L'hypothèse composite nulle correspondante pour le gène i est alors :

$$\mathcal{H}_{0i} : \{\text{Le gène } i \text{ est associé à au plus 7 troubles.}\}$$

Avec un seuil $\alpha = 0.05$, notre procédure a identifé 17 gènes associés à au moins 8 troubles psychiatriques, et suggère la présence de 2 "hubs" de gènes très impliqués dans les troubles psychiatriques sur les chromosomes 3 et 17 (Figure 1).

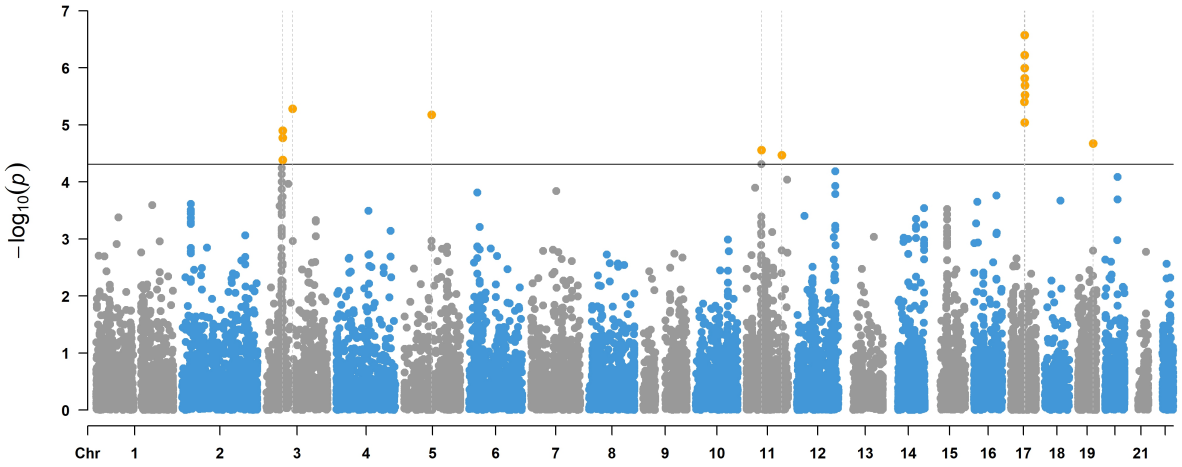

Figure 1: -log10(p-values) du test d'hypothèse composite le long des chromosomes. Les gènes significatifs sont coloriés en orange.

Bibliographie

References

[1] Paul D. Arnold et al. "Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis". In: *Molecular Psychiatry* 23 (5 May 2018), pp. 1181–1181.

[2] Roger L Berger. "Multiparameter Hypothesis Testing and Acceptance Sampling". In: 24 (4 1982), pp. 295–300.

[3] Ditte Demontis et al. "Discovery of the first genome-wide significant risk loci for attention-deficit/hyperactivity disorder". In: *Nature genetics* 51 (1 Jan. 2019), p. 63.

[4] Laramie Duncan et al. "Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa". In: *American Journal of Psychiatry* 174 (9 Sept. 2017), pp. 850–858.

[5] Kevin J. Gleason et al. "Primo: Integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits". In: *Genome Biology* 21 (1 Sept. 2020), pp. 1–24.

[6] Jakob Grove et al. "Identification of common genetic risk variants for autism spectrum disorder". In: *Nature genetics* 51 (3 Mar. 2019), p. 431.

[7] Max Lam et al. "Comparative genetic architectures of schizophrenia in East Asian and European populations". In: *Nature genetics* 51 (12 Dec. 2019), pp. 1670–1678.

[8] T. Mary-Huard et al. "Querying multiple sets of p-values". In: (2021).

[9] Caroline M. Nievergelt et al. "International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci". In: *Nature Communications 2019 10:1* 10 (1 Oct. 2019), pp. 1–16.

[10] T. Otowa et al. "Meta-analysis of genome-wide association studies of anxiety disorders". In: *Molecular psychiatry* 21 (10 Oct. 2016), p. 1391.

[11] Joëlle A. Pasman et al. "GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia". In: *Nature neuroscience* 21 (9 Sept. 2018), p. 1161.

[12] Sandra Sanchez-Roige et al. "Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts". In: *American Journal of Psychiatry* 176 (2 Feb. 2019), pp. 107–118.

[13] Michael E Sobel. "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models". In: *Source: Sociological Methodology* 13 (1982), pp. 290–312.

[14] Eli A. Stahl et al. "Genome-wide association study identifies 30 Loci Associated with Bipolar Disorder". In: *Nature genetics* 51 (5 May 2019), p. 793.

[15] John D. Storey. "A Direct Approach to False Discovery Rates". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64 (3 Aug. 2002), pp. 479–498.

[16] Ziqiao Wang and Peng Wei. "IMIX: a multivariate mixture model approach to association analysis through multi-omics data integration". In: *Bioinformatics* 36 (22-23 Apr. 2021), pp. 5439–5447.

[17] Naomi R. Wray et al. "Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression". In: *Nature genetics* 50 (5 May 2018), p. 668.

[18] Dongmei Yu et al. "Interrogating the genetic determinants of Tourette's syndrome and other tiC disorders through genome-wide association studies". In: *American Journal of Psychiatry* 176 (3 Mar. 2019), pp. 217–227.