Détection de profils conditionnels dans des matrices creuses pour la sélection génomique.

Mathieu Emily [1] & Alain Mom [2]

[1] *Agrocampus Ouest - IRMAR UMR CNRS 6625, 65, rue de Saint Brieuc, 35042 Rennes Cedex, France et mathieu.emily@agrocampus-ouest.fr*

[2] *Université Rennes 2 - IRMAR UMR CNRS 6625, Campus de Villejean, 35043 Rennes Cedex, France et alain.mom@univ-rennes2.fr*

Résumé. L'objectif de cet article est de proposer une méthodologie statistique pour détecter des profils conditionnels particuliers, appelés profils sparse-spécifiques. Ces profils correspondent à des signatures moléculaires caractérisant la présence d'une sélection génomique. L'approche proposée s'appuie sur une classification hiérarchique obtenue à partir d'une nouvelle dissimilarité appelée d_s^2. Par une approche théorique, appuyée par des simulations, nous montrons que d_s^2 est adaptée à la détection de profils *sparse-spécifiques*, notamment dans le cas de matrices de contingences creuses. L'application de notre méthodologie à un jeu de données traitant de la sélection génomique chez le chien domestique illustre également les avantages de notre dissimilarité d_s^2 par rapport à des dissimilarités classiques comme les distances du χ^2 et d_2^2.

Mots-clés. Dissimilarité, saut minimum, sélection génomique

Abstract. The aim of this work is to propose a new statistical framework to detect typical conditional profiles, called sparse-specific profiles. Such profiles are expected to be observed when searching for molecular signatures of selection in polymorphic data. Our approach relies on a hierarchical clustering based on a new dissimilarity d_s^2. Using theoretical analysis, enhanced by a simulation study, we show that d_s^2 is adapted to the detection of sparse-specific profiles, especially when dealing with sparse matrices. Furthermore, the application of our methology on a real dataset demonstrates the benefit of using d_s^2 compared with traditional dissimilarities, such as the χ^2 and the d_2^2 distances.

Keywords. Dissimilarity, single-linkage, genomic selection

1 Introduction

Pour étudier l'association entre deux variables qualitatives X et Y, il est courant de commencer par tester l'indépendance entre X et Y en s'appuyant sur la matrice de contingence croisant les effectifs des deux variables (Agresti, 2013). Lorsque l'hypothèse d'indépendence est rejetée, des analyses complémentaires permettent d'étudier les profils de X conditionnellement à Y, également appelés profils conditionnels. Ces études

approfondies permettent notamment le regroupement de profils conditionnels similaires, générant ainsi des "clusters" de profils.

Dans ce contexte, nous nous intéressons ici à la détection des profils conditionnels typiques de la manifestation de sélection génomique. Ces profils conditionnels d'intérêt sont caractérisés par deux aspects particuliers. Tout d'abord nous cherchons des profils conditionnels *sparse*, c'est-à-dire pour lesquels de nombreuses valeurs nulles sont attendues. D'autre part, les profils conditionnels d'intérêt sont supposés *spécifiques*, caractérisant le fait que les valeurs élevées pour ces profils seront faibles pour les autres profils. Par la suite, des profils conditionnels satisfaisant ces 2 caractéristiques seront appelés profils *sparse-spécifiques*. De plus, nous nous plaçons dans le cadre de matrices de contingence creuses. En effet, avec l'explosion des technologies haut-débit, notamment dans le domaine génomique, les profils observées comportent de nombreuses valeurs nulles.

Pour détecter des profils *sparse-spécifiques* à partir de matrices creuses, nous avons mis en place une procédure statistique s'appuyant sur une classification hiérarchique obtenue à partir d'une dissimilarité. Afin de séparer les profils conditionnels *sparse-spécifiques* des autres profils, nous proposons une nouvelle mesure de dissimiliarité, appelé d_s^2. Contrairement aux dissimilarités classiques, comme la distance de χ^2 ou la distance d_2^2, notre mesure d_s^2 présente l'avantage de tenir compte de façon équipondérée de la sparsité et de la spécificité des profils conditionnels.

L'article s'organise de la façon suivante. La section 2 est dédiée à la formalisation de la procédure que nous proposons. La section 3 est dédiée à la comparaison des performances de notre dissimilarité d_s^2, à celles obtenues avec les distances du χ^2 et d_2^2. Enfin, nous appliquons, dans la section 4, notre méthodologie à l'analyse d'un jeu de données génomiques construit pour étudier les signatures moléculaires de séléction génomique chez le chien domestique.

2 Procédure statistique

La procédure statistique de détection de profils *sparse-spécifiques* que nous proposons s'appuie sur une classification hiérarchique des profils à partir d'une matrice de dissimilarité obtenue à l'aide d'une mesure appelée d_2^s. Dans cette section, nous commençons par introduire formellement la définition d'un profil *sparse-spécifique*. Puis nous définissons la dissimilarité d_s^2. Enfin nous détaillons l'ensemble de la procédure qui permet de sélectionner des profils *sparse-spécifiques*.

2.1 Formalisation d'un profil *sparse-spécifique*

Considérons tout d'abord une matrice de contingence composée de n individus (ou lignes) et k catégories (ou colonnes). A titre d'exemple, les données traitées à la section 4 comportent $n = 30$ races de chien pour un total de k haplotypes variant d'une région à

l'autre. Chaque individu est ainsi caractérisé par un vecteur de comptage $[n_{i1}, \ldots, n_{ik}] \in \mathbb{N}^k$, où n_{ij} est le nombre de fois où la catégorie j est observée pour l'individu i. Le profil conditionnel de l'individu i est défini par un vecteur à k dimensions $x_i = [p_1^i, \ldots, p_k^i]'$ où $p_j^i = n_{ij}/n_{i.}$ avec $n_{i.} = \sum_{j=1}^{k} n_{ij}$. L'ensemble des profils conditionnels est noté E.

Ces notations nous permettent de définir un profil *sparse-spécifique* comme suit :

Définition 1 *Un profil sparse-spécifique est caractérisé par un ensemble d'individus, noté A, et un ensemble de catégories, noté K. Les catégories dans K sont surreprésentées pour les individus de A et, de façon simultanée, les individus dans A possèdent des categories (presque) uniquement dans K. Un profil sparse-spécifique est ainsi défini par les deux caractéristiques suivantes :*

[Sparsité d'un profil] : pour $i \in A$ et $j \notin K$, p_j^i est petit.

[Spécificité d'un profil] : pour $i \notin A$ et $j \in K$, p_j^i est petit.

2.2 La dissimilarité d_s^2

Nous proposons dans cette section une nouvelle dissimilarité, notée d_s^2, qui est adaptée à la détection de profils *sparse-spécifiques*. d_s^2 accorde la même importance à la sparsité et à la spécificité d'un profil conditionnel. Plus précisément, d_s^2 est définie de la façon suivante :

Définition 2 $\forall x, y \in E$:

$$d_s^2(x, y) = \|x\|_2 \|y\|_2 d_\theta^2(x, y) \tag{1}$$

où

$$d_\theta^2(x, y) = 2\left(1 - \frac{\langle x, y\rangle_2}{\|x\|_2 \|y\|_2}\right) \tag{2}$$

est le carré de la distance angulaire entre x et y, $\langle , \rangle_2$ est le produit scalaire L_2 et $\|\|_2$ la norme L_2.

Le terme $\|x\|_2 \|y\|_2$ de l'équation 1 permet de quantifier la sparsité d'un profil tandis que le terme $d_\theta^2(x, y)$ évalue la spécificité entre les profils x et y. L'intérêt de la dissimilarité d_s^2 peut s'exprimer directement à partir de l'expression de d_2^2. Rappelons tout d'abord que $\forall (x, y) \in E, d_2^2(x, y) = ||x - y||_2^2 = \sum_{i=1}^{k} (x_i - y_i)^2$, nous pouvons alors remarquer que :

$$d_2^2(x, y) = d_s^2(x, y) + (\|x\|_2 - \|y\|_2)^2 \tag{3}$$

En rajoutant le terme $(\|x\|_2 - \|y\|_2)^2$ à l'expression de d_s^2, la dissimilarité d_2^2 donne un poids supplémentaire à la variation en sparsité entre les profils x et y. Ainsi, la dissimilarité d_2^2 est plus sensible à l'hétérogénéité des profils dans le sous-ensemble A et son complémentaire $\bar{A}$, en comparaison de d_s^2. Il est important de noter que, par définition, les profils appartenant à A sont supposés homogènes. Aucune hypothèse n'est toutefois faite sur $\bar{A}$. Par la suite nous insisterons donc sur l'impact de l'hétérogénéité de $\bar{A}$ sur la puissance de détection.

2.3 Procédure de détection de profils *sparse-spécifiique*

Pour détecter des profils *sparse-spécifiques*, nous proposons d'utiliser la mesure d'agrégation du saut minimum pour construire le dendrogramme obtenu à partir de la dissimilarité d_s^2. L'utilisation du saut minimum est particulièrement adaptée à notre problème puisqu'il permet de séparer les profils *sparse-spécifiques* du reste des profils.

En effet, le saut minimum est caractérisé par un effet de chaînage qui lui donne une tendance à l'agrégation plutôt qu'à la création de nouvelles classes, définissant ainsi des classes très longues. Ainsi, nous proposons de déterminer l'ensemble des profils conditionnels *sparse-spécifiques* comme le plus petit des 2 sous-ensembles qui se regroupent à la dernière étape de la classification hiérarchique ascendante.

3 Simulation

3.1 Procédure

Afin d'évaluer les performances de détection de profils *sparse-spécifiques*, nous proposons de comparer d_s^2 aux distances χ^2 et d_2^2 sur un ensemble de simulations. Nous concentrons notre analyse sur la simulation de 3 scénarios particuliers pour lesquels les individus dans A possèdent un profil conditionnel *sparse-spécifique*. Les trois scénarios de simulation se différencient alors par la structure des données dans $\bar{A}$.

Le premier scénario s'intéresse à la simulation de profils conditionnels *spécifiques* et *non-sparse* dans $\bar{A}$. Dans un second scénario, nous nous focalisons sur la simulation de profils homogènes dans $\bar{A}$. Enfin, dans un troisième scenario, nous simulons la présence de profils hétérogènes dans $\bar{A}$. Ces trois scénarios ont été choisi pour illustrer les caractéristiques observées dans le jeu de données traités à la section 4.

Pour étudier ces trois scénarios, nous avons tout d'abord utilisé un algorithme simple de simulation de table de contingence. L'ensemble des profils E a été coupé en trois sous-ensembles : A, B et C tels que $B \cup C = \bar{A}$. De même, l'ensemble des catégories a été scindé en 3 sous-ensembles : K, L_1 et L_2. Les matrices de contingence ont été obtenues en simulant des comptages issus de lois multinomiales dont les paramètres sont spécifiques à chaque scénario. Pour chaque scénario nous avons fixé d'une part les probabilités conditionnelles pour un individu du sous-ensemble A (resp. B, C) sachant le sous-ensemble K (resp. L_1, L_2) et d'autre part les cardinaux de chacun des sous-ensembles A, B, C, K, L_1 et L_2. L'ensemble de ces paramètres est résumé à la Table 1 (a). De plus, les valeurs des paramètres utilisés pour chaque scénario sont données dans la Table 1 (b).

3.2 Résultats

L'ensemble des résultats obtenus pour les trois scénarios de simulations est représenté à la Figure 1. Plus précisément, nous pouvons constater une perte de puissance de la

	$K(k_1)$	$L_1(\ell_1)$	$L_2(\ell_2)$
$A(n_A)$	p	p^*	p^*
$B(n_B)$	q_1^*	q_1	q_1^*
$C(n_C)$	q_2^*	q_2	q_2^*

(a)

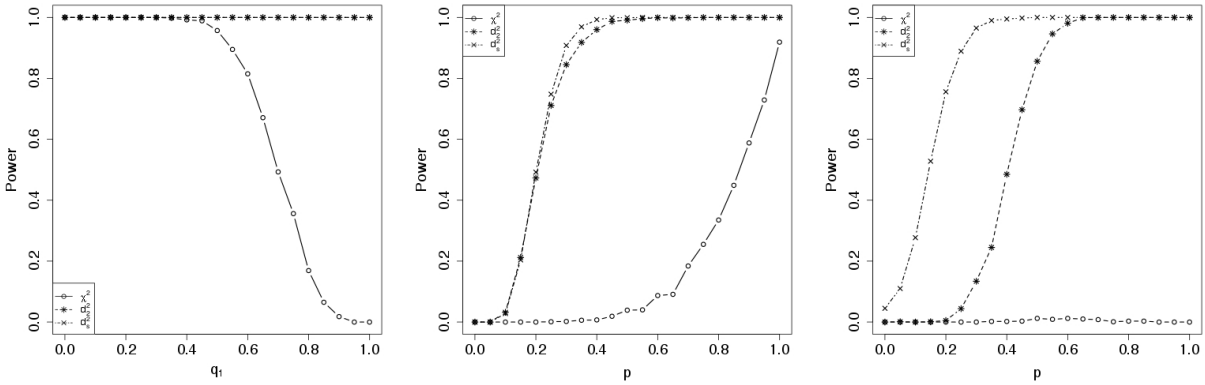
	p	q_1	q_2	n_A	n_B	k_1	ℓ_1
Scenario 1	1	x	0	1	1	1	100
Scenario 2	x	1/k	1/k	3	10	1	2
Scenario 3	x	0.45	0.1	3	10	1	2

(b)

Table 1: *Résumé des paramètres utilisés pour la simulation de profils conditionnels. Le cardinal de chaque ensemble est donné entre parenthèses tel que* $n_A + n_B + n_C = n = 30$ *et* $k_1 + \ell_1 + \ell_2 = k = 200$. *La valeur* x *dans la Table (b) signifie que nous avons étudié la puissance en fonction de ce paramètre.*

distance de χ^2 dans le scénario 1 lorsque le paramètre q_1 augmente. Il apparaît donc que la distance de χ^2 est très sensible à la présence de profils *spécifiques* et *non-sparse* dans $\bar{A}$. Cette constatation est renforcée par la faible puissance obtenue par la distance de χ^2 pour les scénarios 2 et 3.

D'autre part, la comparaison des résultats des scénarios 2 et 3 illustrent l'intérêt d'utiliser la dissimilarité d_s^2 par rapport à la dissimilarité d_2^2. En effet, lorsque la structure des profils conditionnels est homogène dans $\bar{A}$, d_s^2 et d_2^2 ont des performances similaires. Par contre, une structure hétérogène dans les profils conditionnels de $\bar{A}$ entraine une perte de puissance pour d_2^2 relativement à d_s^2.



Scénario 1 Scénario 2 Scénario 3

Figure 1: *Puissance empirique pour* χ^2, d_2^2 *et* d_s^2 *en fonction de* q_1 *(a),* p *lorsque les individus de* $\overline{A}$ *sont homogènes (b) puis divisés en 2 groupes (c).*

4 Analyse de données en sélection génomique chez le chien domestique

Dans cette section, nous nous intéressons à la détection de régions génomiques associées à un événement de sélection chez le chien domestique. Pour illustrer l'intérêt de notre méthode, nous avons utilisé six régions génomiques connues pour leur implication dans des événements de sélection. Pour chacune de ces six régions, nous avons extrait, les

haplotypes de 456 chiens répartis en 30 races (Lequarré, (2011)). Ainsi, les données de chaque région sont résumées en une matrice de contingence avec $n = 30$ races en lignes et les différents haplotypes observés en colonne. Les caractéristiques générales des six régions, comme le numéro du chromosome, le gène associé au trait phénotypique sélectionné ainsi que la(les) race(s) sous sélection génomique sont résumées dans la Table 2.

Pour chacune des régions, nous avons appliqué la méthode de la section 2 en utilisant les trois dissimilarités : d_s^2, d_2^2 et χ^2. Nous pouvons remarquer que la dissimilarité d_s^2 permet d'identifier correctement le (les) race(s) sous sélection génomique pour cinq régions sur six. D'autre part, l'utilisation du χ^2 confirme la faible puissance identifiée dans les simulations de la Section 3. Enfin, les résultats obtenus pour la Région 6 sont également consistent avec les résultats des simulations. En effet, dans le cas de la Région 6, nous observons une hétérogénéité forte des profils conditionnels d'haplotypes pour l'ensemble des races non-sélectionnées, ensemble correspondant à $\bar{A}$ dans les simulations.

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
Chromosome	1	13	13	13	18	27
Gène	*HMGA2*	*RSPO2*	*HSA2*	*HSA2*	*Fgf4*	*KRT71*
Trait	Brachi cephalie	Fourni	Peau plissée	Fièvre périodique	Chondro-dysplasie	Bouclé
Race(s) sélectionnée(s)	EBD	BoT, IrW, JRT, StP et TYo	ShP	ShP	Dac et TYo	StP
n_A	1	5	1	1	2	1
k_1	1	2	2	2	1	1
k	377	367	124	228	44	55
Détection	d_s^2 d_2^2		d_s^2 d_2^2 χ^2	d_s^2 d_2^2	d_s^2 d_2^2	d_s^2

Table 2: *Résumé des six régions génomiques utilisées pour valider la méthode proposée. La dernière ligne, appelée Détection, nous donne les dissimilarités qui ont détecté le signal de sélection avec succès.*

Les résultats obtenus ouvrent donc des perspectives intéressantes dans la détection de régions génomiques sous sélection. De plus, avec l'explosion des données disponibles dans de nombreux domaines, les tableaux de contingence traités sont souvent creux. Les travaux présentés dans cet article démontrent l'intérêt de développer des stratégies et des métriques adaptées à la problématique traitée.

Bibliographie

[1] Agresti, A. (2013). Categorical Data Analysis. Wiley, New- York, third edition.
[2] Lequarré, *et. al.* (2011). Lupa: A european initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. *The Veterinary Journal* 1**89**, 155-159.