Méthodes pour l'inférence post-clustering appliquées à l'expression génique

Benjamin Hivert, Denis Agniel, Rodolphe Thiébaut, Boris Hejblum

Méthodes pour l'inférence post-clustering appliquées à l'expression génique

Benjamin Hivert[1,2,3], Denis Agniel[4,5], Rodolphe Thiébaut[1,2,3,6] & Boris Hejblum[1,2,3]

[1] *Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France*
[2] *INRIA Bordeaux Sud Ouest, SISTM team Talence F-33400, France*
[3] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[4] *Rand Corporation, Santa Monica, CA 90401, USA*
[5] *Harvard Medical School, Boston, MA 02115, USA*
[6] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

benjamin.hivert@u-bordeaux.fr, denis.agniel@gmail.com, rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr

Résumé. L'analyse des données d'expression génique est souvent organisée autour de deux étapes successives : i) une classification non supervisée utilisant l'ensemble des gènes pour regrouper les unités d'observations (patients, échantillons ou cellules) en sous-groupes distincts et homogènes ; puis ii) l'analyse différentielle se faisant à l'aide de tests d'hypothèse visant à identifier quels gènes, c'est-à-dire quelles variables, sont différentiellement exprimés entre ces sous-groupes. Cependant, cette approche utilisant les même données lors des deux étapes ne permet pas de garantir un bon contrôle de l'erreur de type I à l'étape ii). Nous proposons deux méthodes d'inférence pour tenir compte de l'étape initiale de classification non supervisée lors de l'analyse différentielle et ainsi garantir un contrôle effectif de l'erreur de type I. La première méthode se base sur le concept d'inférence sélective tandis que la seconde repose sur une définition de la séparation de classes faisant uniquement intervenir les concepts d'unimodalité et de multimodalité. Nous avons évalué les performances des deux méthodes grâces à différentes simulations numériques, ainsi que dans une application sur un jeu de données réelles de faible dimension. Les méthodes proposées conduisent à des p-valeurs valides sous l'hypothèse nulle d'absence de différence entre les sous-groupes dans l'expression d'un gène sélectionné, indépendamment de la classification, tout en conservant une bonne puissance statistique. En grande dimension, cette inflation de l'erreur de type I peut-être contre-balancée par la dilution du signal utilisé pour la classification, à condition que les variables soient indépendantes. En revanche, en présence de corrélation (comme c'est le cas en pratique pour l'expression génique), des classes artificielles apparaissent alors que celles-ci ne sont pas séparables. Une adaptation des méthodes à ce contexte de grande dimension est donc nécessaire.

Mots-clés. Classification non supervisée, inférence sélective, analyse circulaire, génomique statistique, données de grande dimension ...

Abstract. The analysis of RNA-seq gene expression data is often organised around two successive steps : i) clustering using all of the genes to group the observation units (patients, cells, etc.) into separate and homogeneous subgroups ; then ii) differential analysis of individual genes using hypothesis testing to identify which genes, i.e. which variables, are differentially expressed between the subgroups. However, several subgroups constructed in i) can actually contain only units coming from the same homogeneous population : clustering will then artificially create differences between those spurious subgroups, leading to false positives in ii). We propose two inference methods to take into account the initial clustering step for differential analysis and thus guarantee an effective control of the type I error. This first method is based on the concept of selective inference while the second one use unimodality and multimodality to describe the separation between clusters. We evaluate the performance of both approaches in extensive numerical simulations as well as in an application to a real, low dimensional dataset. Both proposed methods lead to valid p-values under their null hypothesis of no difference between subgroups in expression at a selected gene independently of the clustering, while maintaining good statistical power. In high dimension, this type I error inflation can be overcome by the dilution of the clustering information, provided that the variables are independent. Yet, in the presence of correlation (as for gene expression), spurious clusters appear, even though they are not separable. An adaptation of the above methods to this high dimensional context is thus necessary.

Keywords. Clustering, selective inference, double-dipping, statistical genomics, high-dimensional data ...

1 Introduction

Le séquençage de l'ARN est une technologie mesurant l'expression génique à différentes échelles qui permet une meilleure compréhension des mécanismes biologiques. L'analyse de ces données d'expression génique repose souvent sur deux grandes étapes : i) une étape de classification non supervisée et ii) une étape d'analyse différentielle. La classification non supervisée a pour but de regrouper les observations afin de former des sous-groupes (des classes) homogènes et séparés. Il est néanmoins toujours possible de construire des classes à l'aide d'une méthode de classification même si toutes les observations proviennent en réalité d'une même population homogène. Dans ce cas les différences entre les classes obtenues ne s'expliquent pas par la présence d'un processus biologique séparant ces groupes d'observations, mais simplement par un mauvais partitionnement des observations. Ce problème est particulièrement dérangeant dans le cadre de l'analyse des données d'expression génique puisque la construction des classes va permettre de formuler les hypothèses de test durant l'étape d'analyse différentielle. Durant cette étape, on va chercher à identifier les gènes séparant les classes (*i.e* les gènes s'exprimant de manière différentielle entre les classes). Cependant, si les deux classes considérées contiennent uni-

quement des observations provenant en réalité d'une même population, alors les gènes qui seront identifiés comme différentiellement exprimés ne le seront pas pour des raisons biologiques, mais simplement comme une conséquence de la classification non supervisée qui aura artificiellement forcée des différences entre les observations qu'elles contiennent. Ainsi, les méthodes d'inférence classiques ne contrôlent plus l'erreur de type I en particulier puisque que les hypothèses de tests ne sont pas construites a-priori, mais à l'aide des mêmes données que celles utilisées pour le test d'hypothèse. Ce problème d'inférence post-classification pour les données RNA-seq est devenu l'un des problèmes majeurs liés à leur analyse [Lähnemann et al., 2020] et fait l'objet de développements méthodologiques récents [Zhang et al., 2019, Gao et al., 2020].

Nous avons donc pour objectif de proposer de nouvelles méthodes d'inférence post-classification permettant d'identifier les gènes différentiellement exprimés entre des classes construites à partir des données à l'aide d'une méthode de classification non supervisée. Pour cela, nous proposons deux méthodes d'inférence permettant de tenir compte de l'étape de classification non supervisée dans l'analyse différentielle : a) nous étendons le travail de Gao et al. [2020] au cas uni-variable qui s'appuie sur le concept d'inférence sélective [Tibshirani et al., 2016] où l'on conditionne sur la classification dans le test statistique ; b) avec une définition plus restrictive d'un sous-groupe (utilisant les notions d'unimodalité et de multimodalité pour caractériser respectivement l'homogénéité au sein d'un sous-groupe et la séparation entre deux sous-groupes) nous étudions la séparabilité de deux classes selon une variable d'intérêt par un test de multimodalité.

2 Méthodes

Soit $\boldsymbol{X}$ une matrice $n \times p$. Sur cette matrice $\boldsymbol{X}$, on applique une méthode de classification non supervisée C afin de construire $C(\boldsymbol{X})$, une partition des n observations en K classes disjointes $C_1, \ldots, C_K$. On souhaite alors étudier la capacité d'une variable $\boldsymbol{X}_g \in \mathbb{R}^n$ de $\boldsymbol{X}$ à séparer deux classes estimées C_k and $C_{k'}$ grâce à la méthode de classification C.

2.1 Inférence sélective post-classification

Une première solution permettant d'étudier la séparation de C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ est de tester une différence de moyenne entre les individus de C_k et ceux de $C_{k'}$ sur $\boldsymbol{X}_g$. En supposant que la variable $\boldsymbol{X}_g$ suit une distribution gaussienne : $\boldsymbol{X}_g \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, il est alors possible d'étudier la séparation de C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ en considérant les hypothèses :

$$\mathcal{H}_0 : \overline{\mu}_g^{C_k} = \overline{\mu}_g^{C_{k'}} \qquad \text{vs} \qquad \mathcal{H}_1 : \overline{\mu}_g^{C_k} \neq \overline{\mu}_g^{C_{k'}} \tag{1}$$

où $\overline{\mu}_g^{C_l} = \frac{1}{|C_l|} \sum_{i \in C_l} \mu_{gi}$ pour $l \in \{k, k'\}$.

L'hypothèse du test (*i.e* la séparation des deux classes) étant établie à partir des données elles mêmes et non a priori, un moyen d'obtenir des tests statistiques valides est

d'utiliser les concepts d'inférence sélective [Fithian et al., 2014, Tibshirani et al., 2016]. Selon ces concepts, puisque les hypothèses ont été construites en utilisant les données, il est nécessaire de conditionner sur cette utilisation des données dans la définition de la p-valeur. Ainsi, nous définissions notre p-valeur par :

$$p_g^{C_k,C_{k'}} \equiv \mathbb{P}_{H_0}\left(\left|\overline{X}_g^{C_k} - \overline{X}_g^{C_{k'}}\right| > \left|\overline{x}_g^{C_k} - \overline{x}_g^{C_{k'}}\right| \mid C_k, C_{k'} \in C(\boldsymbol{X})\right) \tag{2}$$

En particulier, nous conditionnons sur le fait que les deux classes d'intérêt ont été estimées (à l'aide de la méthode de classification non supervisée C) via les mêmes données que celles servant à faire le test. En utilisant les travaux de Gao et al. [2020], il est possible de ré-écrire cette p-valeur (2) comme :

$$p_g^{C_k,C_{k'}} = \mathbb{P}_{H_0}\left(|\phi_g| > \left|\overline{x}_g^{C_k} - \overline{x}_g^{C_{k'}}\right| \mid \phi_g \in S_g\right) \tag{3}$$

où $S_g = \{\phi_g : C_k, C_{k'} \in C\left(\boldsymbol{X}\left(\phi_g\right)\right)\}$ et $\phi_g = \left|\overline{X}_g^{C_k} - \overline{X}_g^{C_{k'}}\right| \overset{H_0}{\sim} \mathcal{N}\left(0, \tau_g^2\right)$. $\boldsymbol{X}(\phi_g)$ est une version perturbée des données selon la statistique de test ϕ_g où seulement la $g^{\text{ème}}$ variable est impactée par cette perturbation. Cette p-valeur définie en (3) est ensuite calculée par une approche de Monte-Carlo. La variance τ_g^2 de la statistique de test ϕ_g est le seul paramètre à estimer et cette estimation se fait de manière à respecter l'hypothèse nulle d'absence de séparation entre les deux classes.

Ce conditionnement sur l'évènement de clustering peut aboutir à une perte de puissance statistique dans certains cas. Pour pallier ce problème, nous proposons une extension de ce test d'inférence sélective basée sur l'agrégation des p-valeurs issues du test sur l'ensemble des clusters adjacents entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$.

2.2 Test de multimodalité

Pour s'affranchir des limites du test d'inférence sélective (temps de calculs importants dus à l'estimation de Monte-Carlo, perte de puissance statistique dans certains cas), nous proposons également un test de multimodalité. Les notions d'unimodalité et de multimodalité ont déjà été utilisées pour caractériser respectivement la distribution de données avec et sans une structure en classes [Kalogeratos and Likas, 2012, Siffer et al., 2018]. Nous proposons d'utiliser ces deux notions afin d'évaluer la présence d'un continuum dans la distribution de $\boldsymbol{X}_g$ entre C_k et $C_{k'}$. En effet, si un tel continuum n'existe pas pour aller de C_k à $C_{k'}$ (en passant par toutes les adjacentes se trouvant entre ces deux), alors cela signifie que C_k et $C_{k'}$ sont séparées sur $\boldsymbol{X}_g$.

Tester la présence de ce continuum entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ revient alors à tester l'unimodalité de cette variable en ne considérant que les individus se trouvant dans l'ensemble des classes adjacentes entre C_k et $C_{k'}$. Le test d'unimodalité que nous utilisons est le DipTest [Hartigan et al., 1985], un test non-paramétrique se basant sur la fonction de répartition des données. En pratique, le DipTest compare la fonction de répartition

empirique des données à celle de la distribution uniforme, considérée comme la pire des distributions unimodales d'après la statistique Dip.

Cette seconde approche est en fait très proche de celle basée sur l'inférence sélective. En effet, les perturbations des données sur lesquelles repose notre première approche sont un moyen d'étudier la possibilité à créer un continuum entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$. Avec cette seconde approche, nous testons directement la présence de ce continuum dans les données à l'aide d'un test de multimodalité.

3 Résultats

3.1 Simulations numériques

Nous avons étudié le comportement des trois tests que nous proposons (le test d'inférence sélective, son extension par combinaison de p-valeurs ainsi que le test de multimodalité) en comparaison avec un test classique, le t-test, dans le cadre de la recherche des variables séparant des pairs de classes. Pour cela, nous avons généré des données selon deux scénario : i) un scénario sous $\mathcal{H}_0$ dans lequel les données sont simulées comme provenant toutes d'une même distribution gaussienne et ii) un scénario sous $\mathcal{H}_1$ dans lequel les données sont simulées telles qu'il y ait des vrais sous-groupes (Figure 1A). Dans les deux cas, nous avons estimé un nombre de classes correspondant au vrai nombre de sous-groupes présents sous $\mathcal{H}_1$. Alors que sous $\mathcal{H}_1$, l'algorithme de classification non supervisée ne fait que découvrir la structure des données en identifiant les vrais sous-groupes d'observations présents dans les données, sous $\mathcal{H}_0$, il force des différences entre des sous-groupes d'observations afin de construire des classes. Cela se traduit par des p-valeurs pour le t-test anormalement significatives, témoignant d'un mauvais contrôle de l'erreur de type I. Nos tests parviennent bien à corriger ce problème garantissant un contrôle effectif de l'erreur de type I en cas de mauvais partitionnement des données. Sous $\mathcal{H}_1$, nos tests restent suffisamment puissants pour détecter la séparation entre deux classes (Figure 1B).

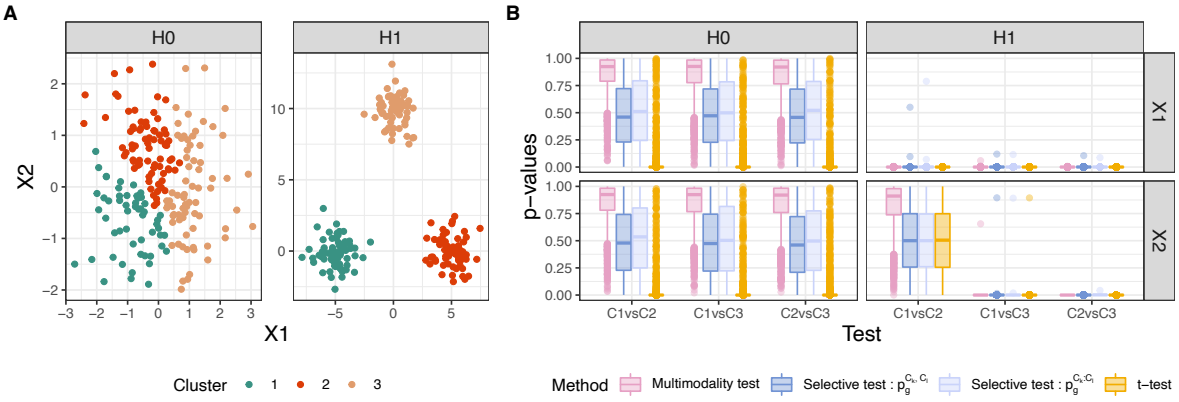

FIGURE 1 – Distributions des p-valeurs des tests proposés sous deux scénario différents

3.2 Grande dimension

Le problème en grande dimension s'avère différent. En effet, l'impact de la classification non supervisée sur la distribution des variables n'est pas le même qu'en petite dimension. Pour commencer, la grande quantité de bruit par rapport à un vrai signal relatif à une structure en classes rend la tâche de classification elle-même plus compliquée. En l'absence de réels sous-groupes d'observations dans les données, seulement un faible nombre de variables seront impactées par l'algorithme de classification, et au vue du grand nombre de variables observées. La correction pour la multiplicité des tests suffit alors à garantir un contrôle effectif de l'erreur de type I des méthodes d'inférence classiques. Cependant, une corrélation entre les variables peut amplifier le nombre de variables touchées par l'algorithme de clustering : le problème d'inférence post-classification demeure alors toujours présent.

Références

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1) :1–35, 2020.

Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4) :383–392, 2019.

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *arXiv preprint arXiv :2012.02936*, 2020.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514) :600–620, 2016.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv :1410.2597*, 2014.

Argyris Kalogeratos and Aristidis Likas. Dip-means : an incremental clustering method for estimating the number of clusters. *Advances in neural information processing systems*, 25 :2393–2401, 2012.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Are your data gathered ? In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2210–2218, 2018.

John A Hartigan, Pamela M Hartigan, et al. The dip test of unimodality. *Annals of statistics*, 13(1) :70–84, 1985.