Une nouvelle méthode de réduction de dimension pour l'analyse de données RNA-Seq à l'échelle de cellule unique.

Svetlana Gribkova [1] & Davide Risso [2] &Fanny Perraudeau [2] & Jean-Philippe Vert [3] & Sandrine Dudoit [2]

[1] *Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot, 8 Place Aurélie Nemours, 75013 Paris*

[2] *Department of Statistics, University of California, Berkeley, 185 Li Ka Shing Center, Berkeley CA 94720-3370*

[3] *Centre de bio-informatique, École des Mines de Paris/U900 Institut Curie, 60 boulevard Saint-Michel 75006 Paris*

Résumé. Le séquençage de l'ARN à l'échelle de cellule unique est une technique biologique récente et révolutionnaire qui a permis de mesurer les expressions de gènes dans des cellules individuelles. L'hétérogénéité cellulaire transcriptomique joue un rôle important dans de nombreux processus biologiques tels que les transformations malignes ou les processus de développement de tissus. Les données RNA-Seq à l'échelle de cellule unique permettent d'étudier les structures de l'hétérogénéité des populations de cellules individuelles à partir de leurs transcriptomes. La réduction de dimension représente une étape importante dans l'analyse de ces données puisqu'elle permet de représenter les cellules par des points dans un espace de dimension faible, afin de visualiser et d'étudier ensuite la structure de leur population. Les distributions spècifiques de ces données de comptage avec excès de zéros rendent inefficaces les techniques standards de la réduction de dimension. Dans cet exposé, nous allons proposer une nouvelle méthode de réduction de dimension, adaptée à la structure des données, qui est basée sur la modélisation de celles-ci par des lois de comptage zéro-inflatées.

Mots-clés. RNA-Seq cellule unique, réduction de dimension, données de comptage zéro-inflatées.

Abstract. Single cell RNA-Seq is a recent and revolutionary technics which allows to measure gene expressions in individual cells. Cellular transcriptomic heterogeneity play an important role in numerous biological processes such as malignant transformations or development of tissues. Single cell RNA-Seq data allows to study the structures of heterogeneous populations of individual cells from their transcriptomes. A first step necessary for such analysis consists in reducing the dimension of data by representing individual cells as points in a low dimensional space. A very special form of distributions of single cell RNA-Seq count data characterized by excess of zeros, makes inefficient the standard technics of dimension reduction. In this talk we will propose a new method of dimension

reduction which is adapted to the structure of the data and is based on modeling by zero-inflated count distributions.

Keywords. Single cell RNA-Seq, dimension reduction, zero-inflated count distributions.

1 Structure du texte long

1.1 Données RNA-Seq à l'échelle de cellule unique

Nous nous intéressons au problème de réduction de dimension pour les données d'expression de gènes à l'échelle de cellule unique. Ce problème est devenu d'actualité depuis l'apparition relativement récente de la technologie de séquençage de l'ARN à l'échelle de cellule unique, qui permet de mesurer les expressions de gènes dans des cellules individuelles. L'hétérogénéité cellulaire joue un rôle important dans de nombreux processus biologiques. On peut par exemple citer la croissance d'embryo à partir de cellule unique, dont chaque étape est caractérisée par la naissance de nouvelles sous populations de cellules aux profils transcriptomiques distincts. Un autre exemple important est le développement de tumeurs dont la structure cellulaire est extrêmement hétérogène. L'analyse des facteurs expliquant l'hétérogénéité des populations cellulaires représente ainsi un outil important pour comprendre de nombreux phénomènes biologiques importants.

Jusqu'à récemment, pour la raison de sensibilité, le séquençage de l'ARN standard fonctionnait uniquement sur les échantillons biologiques composés de milliers de cellules. Le signal observé pour chaque gène représentait ainsi un mélange de signaux venant de toutes les cellules de l'échantillon. Les propriétés individuelles de celles-ci étant masquées, il était impossible de s'interroger sur la présence de l'hétérogénéité cellulaire et sur sa structure. Avec l'apparition de séquençage de l'ARN à l'échelle de cellule unique, voir par exemple Wang (2015), il est devenu possible de mesurer l'expression de gènes dans des cellules individuelles, et ce pour des milliers de cellules. Dans ces données d'expression, chaque observation correspond à une cellule individuelle et les variables observées sont les expressions de plusieurs milliers de ses gènes. Afin de visualiser et d'étudier la structure de l'hétérogénéité de la population de cellules séquencées, il est important de pouvoir réduire la dimension de données afin de représenter les cellules par des points dans un espace de dimension faible.

De manière générale, les méthodes standards de la réduction de dimension supposent explicitement ou implicitement que la distribution de données peut être approchée par la loi normale. Cette hypothèse n'est pas satisfaite pour les données d'expression de gènes à l'échelle de cellule unique. Cela est dû à leur structure de données de comptage et, plus important encore, à la présence d'un nombre excessif de valeurs nulles (jusqu'à 60% des entrées sont des zéros). Ces zéros peuvent être observés en cas de gènes non exprimés mais aussi suite à des erreurs de détection dues à des limites de sensibilité de la méthode

de séquençage. Dans le deuxième cas, une valeur non nulle aurait dû être observée à la place de zéro. Plus l'expression de gène est faible, plus la probabilité d'observer un zéro à sa place est élevée. Quelque soit la nature de valeurs nulles, leur présence doit être prise en compte dans la modélisation des données. En effet, les facteurs de l'hétérogénéité cellulaire révélés par les méthodes standards qui ne tiennent pas compte de la structure des données, sont en général très liés à la variabilité de la quantité de valeurs nulles observées pour chaque cellule.

1.2 Modèle

Les données RNA-Seq à l'échelle de cellule unique sont généralement représentées sous forme de matrice de comptages dont l'entrée ij correspond à un nombre entier de "reads" qui a été observé pour le gène j dans la cellule i. Ce nombre est une mesure quantitative de l'expression de gène, c'est-à dire du nombre de molecules d'ARN que le gène j a produit dans la cellule i. Les observations pour chaque gène j représentent généralement un mélange d'un certain nombre de valeurs nulles et de comptages positifs élevés. Afin de prendre en compte cette structure des données, on modélise chaque entrée Y_{ij} comme une réalisation d'une variable aléatoire qui suit une loi de mélange donnée par la fonction de masse de la distribution binomiale négative zéro-inflatée:

$$f_{ZINB}(y;\mu_{ij},\theta_{ij},\pi_{ij}) = \pi_{ij}\delta_0(y) + (1-\pi_{ij})f_{NB}(y;\mu_{ij},\theta_{ij}), \quad \forall y = 1,2,\ldots,$$

où $f_{NB}(y;\mu_{ij},\theta_{ij})$ est la fonction de masse de la loi binomiale négative de moyenne $\mu_{ij} \geq 0$ et de paramètre de dispersion $\theta_{ij} > 0$:

$$f_{NB}(y;\mu_{ij},\theta_{ij}) = \frac{\Gamma(y+\theta_{ij})}{\Gamma(y+1)\Gamma(\theta_{ij})}\left(\frac{\theta_{ij}}{\theta_{ij}+\mu_{ij}}\right)^{\theta_{ij}}\left(\frac{\mu_{ij}}{\mu_{ij}+\theta_{ij}}\right)^{y}, \quad \forall y = 1,2,\ldots,$$

$\delta_0(\cdot)$ est la fonction de Dirac, et $\pi_{ij} \in [0,1]$ peut être interprété comme la probabilité que la valeur nulle est observée à la place d'un comptage non nul qui aurait dû être détecté. Elle modélise le phénomène de l'augmentation de la masse de zéro par rapport à celle d'un modèle binomiale négatif standard, d'où le nom de la loi binomiale négative zéro-inflatée.

Soit n le nombre de cellules, J le nombre de gènes et Y_{ij} la variable aléatoire modélisant le comptage observé pour le gène j (for $j = 1,\ldots,J$) dans la cellule i ($i = 1,\ldots,n$). La matrice des données a la forme suivante:

$$\mathbf{Y} = \begin{bmatrix} Y_{11} & \ldots & Y_{1J} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \ldots & Y_{nJ} \end{bmatrix}$$

Le problème de la réduction de dimension dans notre contexte consiste à expliquer les observations de J variables (gènes) de la matrice $\{Y_{ij}\}$ par un nombre restreint de facteurs

connus et/ou latents. Pour apprendre ces facteurs, au lieu d'utiliser les modèles basés sur l'hypothèse de normalité des données, nous supposerons que la matrice des observations est une version bruitée, avec la loi de bruit binomiale négative zéro-inflatée, d'une matrice sous-jacente $\{\mu_{ij}\}$ de rang faible. La contrainte de faible rang est introduite explicitement dans le modèle en utilisant la factorisation suivante de la matrice:

$$\begin{aligned} \log(\mu_{ij}) &= \left(X\beta_\mu + (V\gamma_\mu)^T + W\alpha_\mu + O_\mu\right)_{ij} , & (1) \\ \text{logit}(\pi_{ij}) &= \left(X\beta_\pi + (V\gamma_\pi)^T + W\alpha_\pi + O_\pi\right)_{ij} , & (2) \\ \log(\theta_{ij}) &= \zeta_j , & (3) \end{aligned}$$

où l'on a utilisé la fonction logistique:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

- X est une matrice de taille $n \times M$ connue contenant les observations de M variables explicatives connues qui décrivent les cellules et $\beta = (\beta_\mu, \beta_\pi)$ est une matrice de taille $M \times J$ de paramètres de régression à inférer. X peut inclure à la fois les variables qui induisent la variabilité d'intérêt, par exemple les types de cellules, et les variables qui induisent la variabilité "technique" telles que les mesures de la qualité des échantillons ou encore les labels de réplicats techniques/biologiques. Une colonne constante $\mathbf{1}_n$ égale à 1 fait souvent partie du modèle et permet d'introduire les intercepts spécifiques aux gènes.

- V est une matrice de taille $J \times L$ connue contenant les observations de L variables explicatives connues qui décrivent les gènes, par exemple la longueur d'un gène ou encore le GC-content, et $\gamma = (\gamma_\mu, \gamma_\pi)$ est une matrice de taille $L \times n$ de paramètres de régression à inférer. V peut également inclure une colonne constante $\mathbf{1}_J$ pour inclure les intercepts spécifiques aux cellules, tels que le facteur de taille de la librairie de séquençage.

- W est une matrice de taille $n \times K$ inconnue qui correspond à K facteurs latents qui expliquent l'hétérogénéité cellulaire. Ces facteurs peuvent être liés à la fois à la variabilité de "nuisance" qui ne représente pas d'intérêt biologique, et à des sources de variabilité d'intérêt, par exemple les types de cellules inconnus. La matrice $\alpha = (\alpha_\mu, \alpha_\pi)$ de taille $K \times J$ contient les paramètres de régression. Les matrices W et α dans notre modèle sont analogues aux composantes principales de l'ACP et doivent être inférées à partir des données.

- O_μ and O_π sont les matrices des offsets connues de taille $n \times J$.

- $\zeta \in R^J$ est un vecteur de paramètres de dispersion à l'échelle logarithmique. On suppose que le paramètre de dispersion varie selon le gène mais et que, pour un gène donné, il reste le même pour toutes les cellules.

1.3 Estimation du modèle

Les matrices X, V, O_μ, O_π et l'entier K sont les paramètres connus pris en entrée par l'algorithme. Les paramètres à inférer sont $\beta = (\beta_\mu, \beta_\pi)$, $\gamma = (\gamma_\mu, \gamma_\pi)$, W, $\alpha = (\alpha_\mu, \alpha_\pi)$, et ζ. Les observations sont les entrées de la matrice Y de taille $n \times J$. La fonction de la log-vraisemblance des données s'écrit comme suit:

$$\ell(\beta, \gamma, W, \alpha, \zeta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \ln f_{ZINB}(Y_{ij}; \mu_{ij}, \theta_{ij}, \pi_{ij}) \,,$$

où μ_{ij}, θ_{ij}, et π_{ij} dépendent de $(\beta, \gamma, W, \alpha, \zeta)$ à travers le modèle décrit dans la section précédente. Afin d'inférer les valeurs des paramètres inconnus, on maximise la log-vraisemblance pénalisée par un terme permettant d'éviter le surajustement du modèle et d'améliorer la stabilité numérique du problème d'optimisation. Plus précisément, on cherche à résoudre le problème suivant:

$$\max_{\beta,\gamma,W,\alpha,\zeta} \{\ell(\beta, \gamma, W, \alpha, \zeta) - Pen(\beta, \gamma, W, \alpha, \zeta)\} \,,$$

avec

$$Pen(\beta, \gamma, W, \alpha, \zeta) = \frac{\epsilon_\beta}{2}\|\beta^0\|^2 + \frac{\epsilon_\gamma}{2}\|\gamma^0\|^2 + \frac{\epsilon_W}{2}\|W\|^2 + \frac{\epsilon_\alpha}{2}\|\alpha\|^2 + \frac{\epsilon_\zeta}{2}V(\zeta) \,,$$

où $(\epsilon_\beta, \epsilon_\gamma, \epsilon_W, \epsilon_\alpha, \epsilon_\zeta)$ sont les paramètres positifs de la régularisation, β^0 et γ^0 désignent les matrices β and γ sans les lignes correspondantes aux intercepts, $\|\cdot\|$ est la norme matricielle de Frobenuis, donnée par:

$$\|A\| = \sqrt{tr(A^T A)},$$

et

$$V(\zeta) = \frac{1}{J-1} \sum_{i=1}^{J} \left(\zeta_i - \frac{1}{J} \sum_{j=1}^{J} \zeta_j \right)^2$$

est la variance empirique associée au vecteur ζ. Le terme ajouté pénalise ainsi les grandes valeurs de paramètres estimés, à une exception faite pour les intercepts qui ne sont pas pénalisés, et pour les paramètres de dispersion dont on pénalise la variabilité à travers les gènes, au lieu de pénaliser leur valeur absolue.

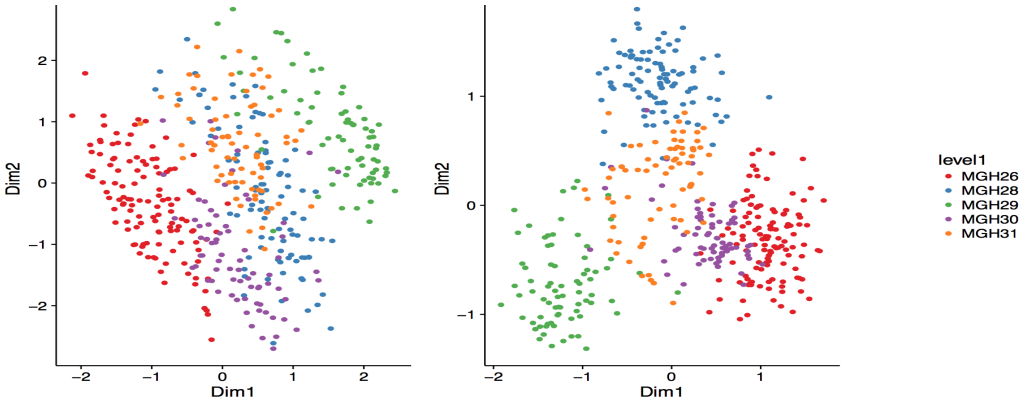
Pour résoudre ce problème d'optimisation, nous proposons un algorithme itératif basé sur la descente de gradient alternée et parallélisée afin de réduire le temps de calcul.

2 Illustration sur les données réelles

Nous avons étudié la performance de notre méthode sur plusieurs jeux de données réelles et simulées. A titre d'exemple, nous présentons ici un résultat de l'analyse, par notre

méthode, d'un jeu de données réelles de RNA-Seq à l'échelle de cellule unique. Les données portent sur les expressions de gènes mesurées à l'échelle de cellule unique, pour 430 cellules provenant des 5 patients différents atteints de glioblastome, un cancer de cerveau très agressif et hétérogène dans sa composition cellulaire. Ces données proviennent de la publication de Patel (2014).

Nous avons comparé la performance de notre méthode à celle de l'analyse en composantes principales, le critère étant leur capacité de distinguer les cellules venant de patients différents. Nous avons appliqué notre méthode, avec le nombre de facteurs latents $K = 2$ et les matrices connues X et V réduites aux intercepts, et d'autre part, l'ACP en gardant deux premières composantes. La figure suivante montre les projection des cellules sur les deux premières axes principales de l'ACP (à gauche) et leurs coordonnées selon les deux colonnes de la matrice W estimée (à droite). Les couleurs différentes correspondent aux cinq patients, dont les identifiants sont donnés sur la légende. On constate que les clusters de cellules correspondant aux patients différents sont bien visibles sur la projection obtenue par notre méthode contrairement à celle obtenue par l'ACP.



Bibliographie

[1] Wang et al. (2015), Advances and application of single-cell sequencing technologies, *Molecular cell*, Vol. 58, 598–609 (2015).
[2] Patel, A. et al. (2014), Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, *Science*, Vol. 344, Issue 6190, pp. 1396–1401.