Transformation des données et comparaison de modèles pour la classification des données RNA-seq

Mélina Gallopin, Andrea Rau, Gilles Celeux, Florence Jaffrézic

TRANSFORMATION DES DONNÉES ET COMPARAISON DE MODÈLES POUR LA CLASSIFICATION DES DONNÉES RNA-SEQ

Mélina Gallopin [1,2,3] & Andrea Rau [2,3] & Gilles Celeux [4] & Florence Jaffrézic [2,3]

[1] *Laboratoire de Mathématiques UMR 8628, Université Paris-Sud 11, 91405, Orsay.*
melina.gallopin@math.u-psud.fr
[2] *Génétique Animale et Biologie Intégrative, UMR 1313, 78350, Jouy-en-Josas.*
[3] *AgroParisTech, UMR 1313, 75005, Paris.*
andrea.rau@jouy.inra.fr, florence.jaffrezic@jouy.inra.fr
[4] *Inria Saclay Ile-de-France, Projet select, Bât 425, Université Paris-Sud 11, 91405 Orsay.*
gilles.celeux@inria.fr

Résumé.
Les données d'expression issues du séquençage haut-débit (RNA-seq) sont des données de comptage très hétérogènes. Il est naturel de les représenter par des modèles basés sur des lois discrètes comme la loi de Poisson ou la loi binomiale négative. Mais des transformations simples des données peuvent permettre de se ramener à des modèles plus répandus fondés sur des lois gaussiennes. Nous montrons comment comparer objectivement les vraisemblances de ces modèles travaillant sur des données différentes. Nous nous focalisons pour mener ces comparaisons sur des problèmes de classification où les mélanges de Poisson et gaussiens peuvent être mis en compétition.

Mots-clés. Modèles de mélange, données RNA-seq, sélection de modèle, transformation des données, BIC.

Abstract.
High-throughput transcriptome sequencing data (RNA-seq) are made up of highly heterogeneous counts. Although they are often modeled with discrete distributions, including the Poisson and negative binomial distributions, Gaussian models on transformed data could alternatively be considered. We show how the likelihood of these different models can be objectively compared. We focus attention on the problem of clustering gene profiles, where Poisson mixtures on count data are compared with Gaussian mixtures on transformed data.

Keywords. Mixture models, RNA-seq data, model selection, data transformation, BIC.

1 Introduction

Les modèles de mélange ont été introduits en analyse des données d'expression de gènes par Yeung *et al.* (2001). Les données de puces à ADN, utilisées depuis le milieu des années 1990,

sont continues et bien modélisées par des mélanges de lois gaussiennes. Depuis la fin des années 2000, la technologie de séquançage à haut-débit révolutionne la manière de mesurer l'expression des gènes (*RNA sequencing*, ou RNA-seq), produisant des données discrètes et très hétérogènes. Un choix naturel de modélisation de ces données est un mélange de lois de Poisson, proposé par Rau *et al.* (2015) détaillé ci-dessous.

Un modèle de mélange de lois de Poisson pour les données RNA-seq On dispose d'une matrice de mesures d'expression de n gènes $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Pour chaque gène i ($i = 1, \ldots, n$), le vecteur $\mathbf{x}_i$ indique l'expression du gène pour les d conditions expérimentales j ($j = 1, \ldots, d$). On suppose que les données $\mathbf{x}$ sont la réalisation d'un mélange de K variables aléatoires de lois de Poisson de densité:

$$f(\mathbf{x}_i; K, \theta_K) = \sum_{k=1}^{K} p_k \prod_{j=1}^{d} \mathscr{P}(x_{ij}; \mu_{ijk}). \tag{1}$$

Les paramètres $(p_1, \ldots, p_K)$ sont les proportions de chaque composante du mélange et $\prod_{j=1}^{d} \mathscr{P}(x_{ij}; \mu_{ijk})$ est la densité d'un vecteur de d variables aléatoires indépendantes de lois de Poisson de moyennes respectives $\mu_{ijk} = w_i s_j \lambda_{jk}$ pour $k = 1, \ldots, K$. Les facteurs $s_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}$ corrigent un biais technique spécifique aux données RNA-seq. Sans cette correction, les mesures d'expression d'un gène pour deux conditions x_{ij} et $x_{ij'}$ différentes ne sont pas comparables. Les facteurs $w_i = x_{i\cdot}$ prennent en compte le niveau d'expression de chaque gène. Les paramètres $\lambda_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ correspondent aux profils d'expression des gènes de la composante k. Ainsi, le modèle classe les gènes en fonction de leur dynamique d'expression (λ_k) et non en fonction de leur niveau d'expression absolu (w_i). Les paramètres p_k et λ_{jk} sont estimés en maximisant la vraisemblance du modèle sous les contraintes $\sum_{k=1}^{K} p_k = 1$ et $\sum_{j=1}^{d} \lambda_{jk} s_j = 1$. L'implémentation de l'estimation des paramètres de ce modèle est proposée dans le package `HTSCluster` par Rau *et al.* (2015) (http://cran.r-project.org/web/packages/HTSCluster).

Un modèle de mélange de lois gaussiennes sur données RNA-seq transformées Une alternative à ce modèle de mélange de lois de Poisson est un modèle de mélange de lois gaussiennes, classiquement utilisé pour les données de puces à ADN. Dans un cadre différent, celui de l'analyse différentielle d'expression de gènes, Law *et al.* (2014) ont proposé une transformation logarithmique des données RNA-seq afin d'utiliser les modèles linéaires gaussiens développés initialement pour l'analyse des données de puces. Dans le même esprit, on propose ici le même type de transformation pour l'utilisation d'un modèle de mélange de lois gaussiennes.

Les données $\mathbf{x}$ sont transformées de sorte que l'objectif de classification reste le plus proche de celui du modèle de mélange de Poisson précédent (modélisation de la dynamique d'expression entre conditions). Chaque comptage x_{ij} est divisé par le facteur $N_j = \sum_{i=1}^{n} x_{ij}/10^6$ afin de corriger le biais technique spécifique aux données RNA-seq. Le facteur N_j est le nombre de millions de comptages de la condition j. Il correspond au facteur s_j du modèle de mélange de lois de Poisson. Afin de modéliser la variation d'expression du gène, on compare le comptage normalisé x_{ij}/N_j à $m_i = \frac{1}{d} \sum_{j'=1}^{d} \frac{x_{ij'}}{Nj'}$, l'expression moyenne du gène i. Le facteur m_i correspond

au facteur w_i dans le modèle de mélange de lois de Poisson. On nomme cette transformation des données t:

$$t(x_{ij}) = \log(\frac{x_{ij}/N_j + 1}{m_i + 1}).$$

On modélise le vecteur des données transformées $\mathbf{y}_i = t(\mathbf{x}_i)$ par un mélange de K lois gaussiennes de densité:

$$g(\mathbf{y}_i; K, \eta_k) = \sum_{k=1}^{K} \tilde{p}_k \phi(\mathbf{y}_i; \nu_k, \Sigma_k). \tag{2}$$

Les paramètres $(\tilde{p}_1, \ldots, \tilde{p}_K)$ sont les proportions de chaque composante du mélange et $\phi(\mathbf{y}_i; \nu_k, \Sigma_k)$ est la densité d'une loi normale de dimension d de moyenne ν_k et de variance-covariance Σ_k. Une implémentation de l'estimation des paramètres de ce modèle est proposée dans le package `Rmixmod` par Lebret *et al.* (2013).

2 Transformation des données et comparaison de modèles

La vraisemblance du modèle de mélange de lois de Poisson s'écrit:

$$l_f(\mathbf{x}_1, \ldots, \mathbf{x}_n; K, \theta_K) = \prod_{i=1}^{n} f(\mathbf{x}_i; K, \theta_K).$$

La vraisemblance du modèle de mélange gaussien sur les données transformées s'écrit:

$$l_g(\mathbf{y}_1, \ldots, \mathbf{y}_n; K, \eta_K) = \prod_{i=1}^{n} g(\mathbf{y}_i; K, \eta_k).$$

De $\mathbf{y}_i = t(\mathbf{x}_i)$, on tire

$$g(\mathbf{y}_i; K, \eta_k) d\mathbf{y}_i = g(t(\mathbf{x}_i); K, \eta_k) t'(\mathbf{x}_i) d\mathbf{x}_i,$$

ce qui permet de réécrire la vraisemblance du modèle de mélange sur données transformées en fonction des données initiales $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$:

$$l_g(\mathbf{x}_1, \ldots, \mathbf{x}_n; K, \eta_K) = \prod_{i=1}^{n} g(t(\mathbf{x}_i); K, \eta_k) t'(\mathbf{x}_i).$$

Les deux modèles peuvent alors être comparés par un critère de vraisemblance pénalisée comme le BIC:

$$\mathrm{BIC}_f(\mathbf{x}_1, \ldots, \mathbf{x}_n; K) = \sum_{i=1}^{n} \log f(\mathbf{x}_i; K, \hat{\theta}_K) - \frac{\nu_f}{2} \log(n),$$

$$\mathrm{BIC}_g(\mathbf{x}_1, \ldots, \mathbf{x}_n; K) = \sum_{i=1}^{n} \log g(\mathbf{y}_i; K, \hat{\eta}_k) + \sum_{i=1}^{n} \log t'(\mathbf{x}_i) - \frac{\nu_g}{2} \log(n).$$

Les quantités $\hat{\theta}_K$ et $\hat{\eta}_K$ sont les estimateurs du maximum de vraisemblance des paramètres des modèles respectifs, ν_f et ν_g sont les nombres de paramètres des modèles respectifs. Le modèle s'ajustant le mieux aux données est le modèle maximisant le critère BIC associé. Cette prise en compte de la transformation appliquée aux données dans le calcul du BIC a été utilisée auparavant dans un autre domaine par Thomas *et al.* (2008).

3 Illustrations sur des données simulées

Afin d'illustrer la comparaison de modèles proposée, nous simulons des données sous le modèle de mélange de lois de Poisson détaillé à l'équation (1), en fixant le nombre de conditions expérimentales d à 3, le nombre de gènes n à 5000, les facteurs de normalisation $(s_1, s_2, s_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ et les facteurs w_i à partir d'un jeu de données réelles en sélectionnant aléatoirement trois conditions expérimentales et n gènes parmi les gènes du jeu de données ayant au moins 20 comptages par gènes. Les valeurs des w_i varient ainsi de 20 à 1 800 000 comptages. Nous fixons ensuite le nombre de classes $K = 4$, les proportions de chaque classe $(p_1, p_2, p_3, p_4) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ et les paramètres λ_{jk} pour $j = 1,2,3$ et $k = 1,2,3,4$ tels que $\sum_j \lambda_{jk} s_j = 1$:

$$\lambda = \begin{pmatrix} 1.5 & 1 & 0.5 & 1.5 \\ 0.5 & 1.5 & 1 & 1 \\ 1 & 0.5 & 1.5 & 0.5 \end{pmatrix}.$$

La figure 1 (gauche) illustre les comptages simulées transformés pour la condition 1 versus la condition 2. Conformément au résultat attendu, la figure 1 (droite) montre que le BIC du modèle de mélange gaussien, ajusté pour la transformation des données est inférieur au BIC du modèle de mélange de lois de Poisson pour un nombre de classes supérieur ou égal à 4.

4 Données réelles

Pour deux jeux de données RNA-seq, nous effectuons la classification des gènes à l'aide du modèle de mélange de Poisson sur les données de comptage brutes, et à l'aide du modèle de mélange gaussien sur les données transformées. Sultan *et al.* (2008) ont analysé l'expression des gènes dans les cellules humaines embryonnaires du rein (HEK293T) et dans les cellules de la lignée Ramos B en séquençant deux réplicats biologiques dans chaque type de cellule par la technologie RNA-seq. Après avoir supprimé les gènes peu exprimées, nous effectuons la classification des 4959 gènes restants. Mach *et al.* (2014) ont analysé les différences d'expression entre trois tissus (le duodenum, le jejunum et l'ileum) de l'intestin grêle de quatre porcelets sains. Après avoir sélectionné les gènes différentiellement exprimés entre ces trois tissus à l'aide d'un modèle linéaire généralisé basé sur une loi négative binomiale, on effectue la classification des 4021 gènes restants. Nous constatons sur la figure 2 que le modèle s'ajustant le mieux aux données est différent pour les deux jeux de données.

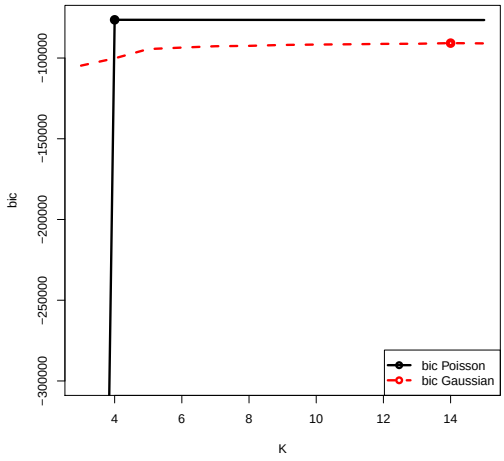

Figure 1: *A gauche*, comptages simulés sous un modèle de mélange de lois de Poisson et t-transformés pour la conditions 1 versus la condition 2. Les différentes couleurs correspondent aux quatre classes simulées. *A droite,* BIC du modèle de mélange de Poisson et BIC du modèle de mélange gaussien ajusté pour la transformation de données t pour un nombre de classes variant de 3 à 15. Le point sur chaque courbe BIC indique le nombre de classes sélectionné par chacun des modèles.

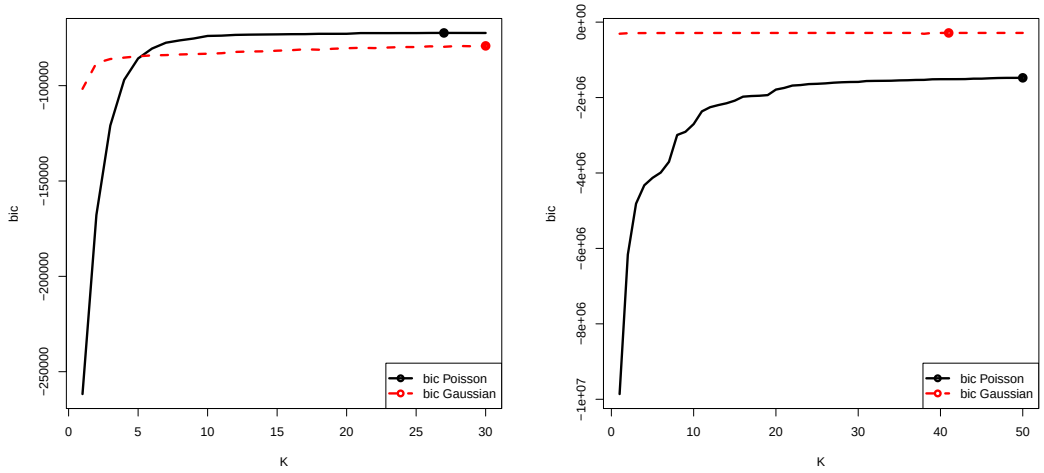

Figure 2: BIC du modèle de mélange de Poisson sur données brutes et du modèle de mélange gaussien sur données transformées pour les données de Sultan *et al.* (2008) (gauche) et Mach *et al.* (2004) (droite).

5 Discussion

A l'aide de la comparaison proposée, on peut ainsi déterminer si une transformation des données fournit un meilleur modèle. Dans cet exposé, on illustrera l'intérêt pratique de la comparaison de modèles sur plusieurs jeux de données réelles, en comparant notamment les méthodes de sélection de modèles pour le choix du nombre de classes associées aux différents modèles (modèles de mélange de Poisson, modèles de mélange gaussien) ainsi que les classifications obtenues. Nous verrons notamment que la transformation logarithmique des données que nous avons proposée fournit souvent des modèles beaucoup plus convaincants que le modèle de mélange de Poisson. Toutefois, il n'est pas sûr que l'usage du critère BIC soit toujours le plus adéquat compte tenu des objectifs de classification. Une comparaison des modèles par le critère ICL, proposé par Biernacki *et al.* (2000), pourrait être considérée.

Bibliographie

[1] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. et Ruzzo, W. L. (2001), Model-based clustering and data transformations for gene expression data, *Bioinformatics* 17 (10), 977-987.
[2] Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.L. et Celeux, G. (2015), Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models, *Bioinformatics*, doi: 10.1093/bioinformatics/btu845.
[3] Law C.W., Chen, Y., Shi, W. et Smyth, G.K. (2014), Voom: precision weights unlock linear model analysis tools for RNA-seq read counts, *Genome Biology*, 15:R29.
[4] Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. et Govaert, G. (2013), Rmixmod: The R package of the model-based unsupervised, supervised and semi-supervised classification Mixmod library, *Journal of Statistical Software* (in revision).
[5] Thomas, I., Frankhauser, P., et Biernacki, C. (2008), The Fractal Morphology of the Built-Up Landscape, *Landscape of Urban Plan*, Vol. 84, No. 2, pp. 99-115.
[6] Sultan, M. *et al.* (2008), A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome, *Science*, 321, 956 .
[7] Mach, N. *et al.* (2014) Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PLoS ONE* 9(2): e88515.
[8] Biernacki, C., Celeux, G. et Govaert,G. (2000), Assessing a mixture model for clustering with the integrated classification likelihood, *IEEE Transaction on PAMI*, 22, 719-725.