A model selection criterion for model-based clustering of annotated gene expression data

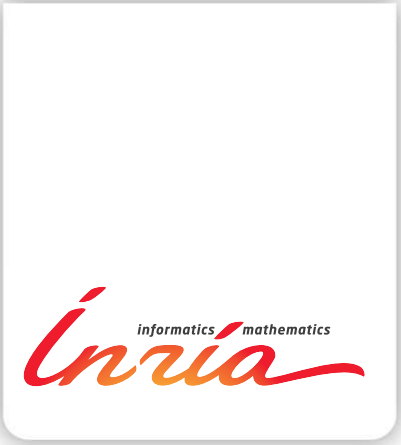
Mélina Gallopin, Gilles Celeux, Florence Jaffrézic, Andrea Rau

A model selection criterion for model-based clustering of annotated gene expression data

Mélina Gallopin, Gilles Celeux, Florence Jaffrézic, Andrea Rau

RESEARCH REPORT
N° ??
Novembre 2014
Project-Teams Select

ISSN 0249-6399 ISRN INRIA/RR--??--FR+ENG



A model selection criterion for model-based clustering of annotated gene expression data

Mélina Gallopin*†, Gilles Celeux‡, Florence Jaffrézic†§, Andrea Rau† §

Project-Teams Select

Research Report n° ?? — Novembre 2014 — 19 pages

Abstract: In co-expression analyses of gene expression data, it is often of interest to interpret clusters of co-expressed genes with respect to a set of external information, such as a potentially incomplete list of functional properties for which a subset of genes may be annotated. Based on the framework of finite mixture models, we propose a model selection criterion that takes into account such external gene annotations, providing an efficient tool for selecting a relevant number of clusters and clustering model. This criterion, called the Integrated Completed Annotated Likelihood (ICAL), is defined by adding an entropy term to a penalized likelihood to measure the concordance between a clustering partition and the external annotation information. The ICAL leads to the choice of a model that is more easily interpretable with respect to the known functional gene annotations. We illustrate the interest of this model selection criterion in conjunction with Gaussian mixture models on simulated gene expression data and on real RNA-seq data.

Key-words: Functional gene annotation, gene expression data, model-based clustering, model selection.

* Laboratoire de Mathématiques, UMR 8628, Bâtiment 425, Université Paris-Sud, F-91405, Orsay Cedex, France
† INRA, UMR 1313 Génétique animale et biologie intégrative, 78352 Jouy-en-Josas, France
‡ Inria Saclay Île-de-France, Bâtiment 425, Université Paris-Sud, F-91405, Orsay Cedex, France
§ AgroParisTech, UMR 1313 Génétique animale et biologie intégrative, 75231 Paris, France

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Un critère de sélection de modèle pour le clustering de données annotées d'expression de gènes par modèle de mélange

Résumé : En analyse de co-expression des données d'expression de gènes, on souhaite interpréter les groupes détectés en fonction d'informations externes, par exemple, à partir des listes potentiellement incomplètes d'annotations fonctionnelles des gènes. Dans le cadre des modèles de mélanges, nous proposons un critère de sélection de modèle prenant en compte ces annotations externes, fournissant ainsi un outil efficace pour sélectionner le nombre de groupes et le modèle les plus pertinents. Ce critèreICAL (Integrated Completed Annotated Likelihood) est défini par l'ajout d'un terme d'entropie à la vraisemblance pénalisée mesurant la concordance entre la partition des gènes et les annotations externes. Le critère ICAL conduit à choisir un modèle plus facilement interprétable par rapport aux annotations fonctionnelles disponibles. On illustre l'intérêt de cette méthode de sélection de modèle pour le modèle de mélange gaussien sur des données simulées et des données d'expression de gènes RNA-seq.

Mots-clés : Annotations fonctionnelles des gènes, données d'expression géniques, classification par modèle de mélange, sélection de modèle.

1 Introduction

Genome annotation broadly refers to the set of meta-data associated with the coding regions in the genome, typically including the identification of the location of each gene as well as a determination of the functions related to the gene product (e.g., protein or RNA). In particular, gene annotations correspond to known functions related to the gene product, including molecular functions, biological pathways, or the cellular location of the gene products. A variety of well-known unified databases have been constructed with known functional annotations collected from bibliographic sources across species, including the Gene Ontology (GO) (Ashburner et al., 2000), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) or the MSigDB (Molecular Signatures) databases (Liberzon et al., 2011). Although such databases contain a rich source of functional information about the genome in a large variety of species (e.g., *Arabidopsis thaliana*, human, rat, mouse, fly), our knowledge of functional annotations is often far from complete (Tipney and Hunter, 2010).

In recent years, substantial improvements in high-throughput technologies, such as microarrays (Schena et al., 1995) and more recently RNA sequencing (RNA-seq) (Mortazavi et al., 2008), have enabled the simultaneous measurement of the expression levels of tens of thousands of genes. A rich body of work is now available on the appropriate statistical analyses of such gene expression data, including the analysis of differential expression (Smyth, 2004; Anders and Huber, 2010) and co-expression analyses to identify groups of genes with similar profiles across several experimental conditions or over the course of time (Yeung et al., 2001; Rau et al., 2014). The latter is of particular interest in this work, as identifying genes that share the same dynamic patterns of expression may help identify groups of genes that are involved in similar biological processes and generate hypotheses about the functional properties of poorly characterized genes (Eisen et al., 1998; Jiang et al., 2004). Reviews and comparisons of different clustering methods for gene expression data may be found in Datta (2003).

In practice, annotation databases are often used to perform *a posteriori* validation and interpretation of co-expressed gene clusters through tests of functional enrichment (Steuer et al., 2006). Such functional annotation may instead be directly integrated into the clustering model itself. For example, Tari et al. (2009) incorporate GO annotations as prior knowledge in a fuzzy c-means clustering. Verbanck et al. (2013) proposed a clustering approach based on a distance defined conjointly on the similarity among expression profiles and that among functional profiles. Pan (2006) and Huang et al. (2006) proposed including gene annotation as prior information in a stratified mixture model. However, the inclusion of gene annotation directly in the model itself in this way may be questionable, particularly when they are also used to validate the gene clusters *a posteriori*. Moreover, as gene annotations tend to be incomplete, biases may be introduced if they are directly incorporated in the model, as unannotated genes (which represent those known to be unassociated with a given function as well as those of unknown function) may be erroneously separated from annotated genes.

One alternative to such approaches is to define a clustering model that accounts for external gene annotations without directly including them in the model itself. To this end model-based clustering provides a convenient framework, as it 1) allows for a large set of clustering models to be fit to the gene expression alone, and 2) facilitates the choice among this set a parsimonious model that simultaneously provides a good fit to the data and coherence with the external gene annotations. In this work, we address these points by proposing a model selection criterion that accounts for external gene annotations.

The rest of this paper is organised as follows. In Section 2, we present the context of model-based clustering and review classic model selection criteria. Our proposed annotated model selection criterion is presented in Section 3, and numerical illustrations of its behavior are presented on simulated data in Section 4 using Gaussian mixture models. Finally, we illustrate a co-expression analysis of real RNA-seq data in Section 5, and a discussion ends the paper.

2 Model-based clustering and model selection

Let $\mathbf{y}$ be the $(n \times q)$ matrix of observed gene expression, where n is the number of genes and q the number of biological samples. The vector $\mathbf{y}_i$ denotes the expression of gene i $(i = 1,\ldots,n)$ across the q samples. In the context of model-based clustering, the data $\mathbf{y}$ are assumed to be sampled from a finite mixture density of K random variables, each with parameterized density $\phi(\mathbf{y}_i;\mathbf{a}_k)$, $k = 1,\ldots,K$, where the mixture parameters $(\mathbf{a}_1,\ldots,\mathbf{a}_K)$ are all assumed to be distinct. The density of $\mathbf{y}$ may thus be written as

$$f(\mathbf{y};K,\theta_K) = \prod_{i=1}^{n}\sum_{k=1}^{K} p_k\phi(\mathbf{y}_i;\mathbf{a}_k), \tag{1}$$

where $\theta_K = (p_1,\ldots,p_{K-1},\mathbf{a}_1,\ldots,\mathbf{a}_K)$ are the parameters of the mixture model, and $(p_1,\ldots,p_K)$ are the mixing proportions with $p_k \in (0,1)$ for all k, $\sum_{k=1}^{K} p_k = 1$.

For parameter estimation, the mixture model in Equation (1) may be thought of as an incomplete data structure model where $\mathbf{z}$ is the $(n \times K)$ matrix of unknown mixture labels, where $z_{ik} = 1$ if gene i is from group k and 0 otherwise. Note that this matrix defines a partition of the genes.

Using the mixture labels z, the completed density of $\mathbf{y}$ may be written as follows:

$$f(\mathbf{y};K,\theta_K) = \prod_{i=1}^{n}\prod_{k=1}^{K}(p_k\phi(\mathbf{y}_i;\mathbf{a}_k))^{z_{ki}}. \tag{2}$$

The maximum likelihood estimate $\hat{\theta}_K$ of the mixture parameters is computed through the Expectation-Maximization algorithm (Dempster et al., 1977) by replacing the unknown labels $\mathbf{z}$ in Equation (2) with $\hat{\mathbf{z}}$, defined as:

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \arg\max_\ell \tau_{i\ell}(\hat{\theta}_K) = k \\ 0 & \text{otherwise,} \end{cases}$$

where $\tau_{i\ell}(\hat{\theta}_K)$ denotes the conditional probability given $\mathbf{y}_i$ of the ℓth mixture component under $\hat{\theta}_K$:

$$\tau_{i\ell}(\hat{\theta}_K) = \frac{\hat{p}_\ell\phi(\mathbf{y}_i;\hat{\mathbf{a}}_\ell)}{\sum_{t=1}^{K}\hat{p}_t\phi(\mathbf{y}_i;\hat{\mathbf{a}}_t)}.$$

In the context of model-based clustering, one important task is the choice of an appropriate model, most notably the revelant number of clusters K. To this end, a standard model selection criterion is the Bayesian Information Criterion (BIC) (Schwarz, 1978):

$$\mathrm{BIC}(K) = \log f(\mathbf{y};\hat{K},\hat{\theta}_K) - \frac{\nu_K}{2}\log(n),$$

where $\hat{\theta}_K$ is the maximum likelihood estimator of the mixture parameters and ν_K the number of free parameters in the model with K components. This criterion is an asymptotic approximation of the logarithm of the integrated likelihood:

$$f(\mathbf{y};K) = \int_{\theta_K} f(\mathbf{y};K,\theta_K)\pi(\theta_K)d\theta_K,$$

where $\pi(\theta_K)$ is a weakly informative prior distribution on θ_K.

An alternative to the BIC is the Integrated Completed Likelihood (ICL) criterion (Biernacki et al., 2000):

$$\mathrm{ICL}(K) = \mathrm{BIC}(K) - \mathrm{Ent}(K), \tag{3}$$

where $\mathrm{Ent}(K)$ is the estimated mean clustering entropy

$$\mathrm{Ent}(K) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}(\hat{\theta}_K)\log\tau_{ik}(\hat{\theta}_K) \geq 0. \tag{4}$$

Note that the ICL is a BIC-like approximation of the logarithm of the completed integrated likelihood:

$$f(\mathbf{y},\mathbf{z};K) = \int_{\theta_K} f(\mathbf{y},\mathbf{z};K,\theta_K)\pi(\theta_K)d\theta_K.$$

Because of the additional entropy term defined in Equation (4), the ICL favors models that lead to data partitions with the greatest evidence in terms of classification.

More recently, Baudry et al. (2014) proposed an ICL-like criterion that takes advantage of the potential explicative ability of external categorical variables $\mathbf{u} = (\mathbf{u}^1,\ldots,\mathbf{u}^R)$ where $u^r_{i\ell} = 1$ indicates that the gene i is in category ℓ for the r^{th} external categorical variable and 0 otherwise. The idea is to choose a classification $\mathbf{z}$ based on $\mathbf{y}$ that is coherent with $\mathbf{u}$. Assuming that $\mathbf{y}$ and $\mathbf{u}$ are conditionally independent given $\mathbf{z}$, the Supported Integrated Completed Likelihood (SICL) criterion is an asymptotic approximation of the logarithm of the integrated completed likelihood:

$$f(\mathbf{y},\mathbf{u},\mathbf{z};K) = \int f(\mathbf{y},\mathbf{u},\mathbf{z};K,\theta_K)\pi(\theta_K)d\theta_K.$$

The SICL criterion is defined as follows:

$$\text{SICL}(K) = \text{ICL}(K) + \sum_{r=1}^{R}\sum_{\ell=1}^{U_r}\sum_{k=1}^{K} n^r_{k\ell}\log\frac{n^r_{k\ell}}{n_{k\cdot}}, \tag{5}$$

where U_r is the number of levels of the variable $\mathbf{u}^r$,

$$n^r_{k\ell} = \text{card}\big\{i : z_{ik} = 1 \text{ and } u^r_{i\ell} = 1\big\},$$

and $n_{k\cdot} = \sum_{\ell=1}^{U_r} n^r_{k\ell}$. The last additional term in Equation (5) quantifies the strength of the link between the categorical variables $\mathbf{u}$ and the classification $\mathbf{z}$.

3 Taking genome annotations into account

As previously stated, the objective of this work is to make use of external gene annotations to choose a model for which clusters may be meaningfully interpreted both with respect to their expression profiles and their functional properties. To do so, we propose a novel model selection criterion that highlights the association between the clusters of expression profiles and the functional annotations associated with a subset of genes. Since gene annotations are binary variables (i.e., a gene is either annotated or unannotated), it may seem natural to directly use the SICL defined in Equation (5). However, in contrast to the situation considered by Baudry et al. (2014), gene annotation information is often incomplete. More precisely, for each of the G annotation terms, indexed by g, the available information $\mathbf{u}^g$ is as follows:

$$u^g_i = \begin{cases} 1 & \text{if gene } i \text{ is known to be implicated in function } g, \\ 0 & \text{if gene } i \text{ is not known to be implicated in function } g. \end{cases}$$

Note that $u^g_i = 0$ can indicate that information is missing (i.e., gene i has not yet been identified for annotation g) or that gene i is known to be unrelated to annotation g. As such, $u^g_i = 0$ does not represent the null level of variable and thus represents an incomplete binary variable. For this reason, the SICL criterion is not an appropriate measure of the link between an external annotation $\mathbf{u}^g$ and a classification $\mathbf{z}$, and a specific criterion must be defined to incorporate the gene annotation information into the model selection step. To this end, we propose the Integrated Completed Annotated Likelihood (ICAL) criterion as follows.

For each gene annotation $\mathbf{u}^g$, we first define the random matrix $\mathbf{b}^g$ of latent variables indicating the allocation of the annotations among the K clusters:

$$b_{ik}^g = \begin{cases} 1 & \text{with probability } p_k^g \text{ if } u_i^g = 1, \\ 0 & \text{if } u_i^g = 0. \end{cases} \tag{6}$$

Each row of the matrix $\mathbf{b}^g$ is a random vector following a multinomial distribution with parameters u_i^g and $(p_1^g, \ldots, p_K^g)$ if $u_i^g > 0$, and is the null vector $\mathbf{0}$ if $u_i^g = 0$.

For the sake of simplicity, we first derive ICAL when a single external annotation $\mathbf{b}^1$ is available. ICAL aims to select the clustering model that maximises the logarithm of the integrated annotated likelihood:

$$f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1; K) = \int_{\theta_K} f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1; K, \theta_K) \pi(\theta_K) d\theta_K. \tag{7}$$

As for the definition of the SICL, the variables $\mathbf{y}$ and $\mathbf{b}^1$ are assumed to be conditionally independent given $\mathbf{z}$. Using Bayes formula, we have

$$f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1; K, \theta_K) = f(\mathbf{y}, \mathbf{z}; K, \theta_K) f(\mathbf{b}^1 \mid \mathbf{y}, \mathbf{z}; K, \theta_K).$$

Note that since $\mathbf{y}$ and $\mathbf{b}^1$ are assumed to be independent given $\mathbf{z}$, the conditional distribution of $\mathbf{b}^1$ given $\mathbf{z}$ does not depend on $\mathbf{y}$ or the mixture parameters. Thus, as $f(\mathbf{b}^1 \mid \mathbf{y}, \mathbf{z}; K, \theta_K) = f(\mathbf{b}^1 \mid \mathbf{z}; K)$, it follows that:

$$\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1; K) = \log f(\mathbf{b}^1 \mid \mathbf{z}; K) + \log \int_{\theta_K} f(\mathbf{y}, \mathbf{z}; K, \theta_K) \pi(\theta_K) d\theta_K. \tag{8}$$

The last term in Equation (8) can be approximated with ICL(K) from Equation (3), and the first term may be approximated with

$$\log f(\mathbf{b}^1 \mid \hat{\mathbf{z}}, K) = \sum_{k=1}^{K} n_k^1 \log \frac{n_k^1}{n^1},$$

where $n^1 = \text{card}\{i : u_i^1 = 1\}$ and $n_k^1 = \text{card}\{i : \hat{z}_{ik} = 1 \text{ and } u_i^1 = 1\}$. Finally, an asymptotic approximation of the expression in (7) leads to the Integrated Completed Annotated Likelihood (ICAL) criterion:

$$\text{ICAL}(K) = \text{ICL}(K) + \sum_{k=1}^{K} n_k^1 \log \frac{n_k^1}{n^1}.$$

The generalization of this criterion to the case where $G > 1$ gene annotations are available is straightforward. The aim is now to maximize the logarithm of the integrated annotated likelihood:

$$\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K) = \log \int_{\theta_K} f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K, \theta_K) \pi(\theta_K) d\theta_K.$$

Assuming that $\mathbf{b}^1, \ldots, \mathbf{b}^G$ and $\mathbf{y}$ are conditionally independent given $\mathbf{z}$, we have

$$\log f(\mathbf{y}, \mathbf{z}, \mathbf{b}^1, \ldots, \mathbf{b}^G; K) = \log f(\mathbf{b}^1, \ldots, \mathbf{b}^G; \mathbf{z}, K) + \log \int_{\theta_K} f(\mathbf{y}, \mathbf{z}; K, \theta_K) \pi(\theta_K) d\theta_K.$$

Assuming in addition that $\mathbf{b}^1, \ldots, \mathbf{b}^G$ are independent and that gene annotations are missing at random, we can write

$$f(\mathbf{b}^1, \ldots, \mathbf{b}^G; \mathbf{z}, K) = \prod_{g=1}^{G} f(\mathbf{b}^g | \mathbf{z}, K), \tag{9}$$

leading to the generalized ICAL criterion:

$$\text{ICAL}(K) = \text{ICL}(K) + \sum_{g=1}^{G} \sum_{k=1}^{K} n_k^g \log \frac{n_k^g}{n^g}. \tag{10}$$

Comparing ICAL and SICL If we ignore the uncertainty associated with $u_i^g = 0$ (i.e., that gene i could either be unassociated with function g or that this information is missing), the SICL criterion could be considered to choose the model dimension K. In this case, using the notation from Section 2 and defining n_k the size of the cluster k , the SICL may be written as follows:

$$\mathrm{SICL}(K) = \mathrm{ICL}(K) + \mathrm{pen}_{\mathrm{SICL}},$$

where

$$\begin{aligned}\mathrm{pen}_{\mathrm{SICL}} &= \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k1}^g \log\frac{n_{k1}^g}{n_{k.}^g} + \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log\frac{n_{k0}^g}{n_{k.}^g},\\ &= \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k1}^g \log n_{k1}^g + \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log n_{k0}^g - G\sum_{k=1}^{K} n_k \log n_k.\end{aligned}$$

On the other hand, using the notation from Section 2 and defining $n_{.1}^g = \sum_{k=1}^{K} n_{k1}^g$, the ICAL may be written as follows:

$$\mathrm{ICAL}(K) = \mathrm{ICL}(K) + \mathrm{pen}_{\mathrm{ICAL}},$$

where

$$\begin{aligned}\mathrm{pen}_{\mathrm{ICAL}} &= \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k1}^g \log\frac{n_{k1}^g}{n_{.1}^g},\\ &= \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k1}^g \log n_{k1}^g - \sum_{g=1}^{G} n_{.1}^g \log n_{.1}^g.\end{aligned}$$

We note that the last term in the equation above is a constant independent of K. Finally, we can rewrite ICAL as a function of SICL:

$$\mathrm{ICAL}(K) = \mathrm{SICL}(K) - \sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log n_{k0}^g + G\sum_{k=1}^{K} n_k \log n_k + \text{constant}. \tag{11}$$

From Equation (11), we note that the SICL takes into account both modalities (0 and 1) of the external variables $\mathbf{u}$, while the ICAL discards the null modality (the $-\sum_{g=1}^{G}\sum_{k=1}^{K} n_{k0}^g \log n_{k0}^g$ term). Moreover, it can be seen that the ICAL penalises a large number of clusters, while the SICL does not (the $G\sum_{k=1}^{K} n_k \log n_k$ term). As such, the ICAL tends to select parsimonious models with a relatively small number of clusters, as compared to SICL.

It is also helpful to consider the behavior of the ICAL and SICL criteria in extreme conditions. If the number of clusters K equals 1, the ICAL penalty $\mathrm{pen}_{\mathrm{ICAL}}$ equals zero whereas SICL penalty $\mathrm{pen}_{\mathrm{SICL}}$ is not null ($\sum_{g=1}^{G} n_1^g \log\frac{n_1^g}{n} + \sum_{g=1}^{G} n_0^g \log\frac{n_0^g}{n}$). In contrast, if the number of clusters K is equal to the number of observations, with one gene per cluster, the SICL penalty $\mathrm{pen}_{\mathrm{SICL}}$ equals zero whereas the ICAL penalty is not null ($\sum_{g=1}^{G} n_1^g \log n_1^g$). In general, ICAL tends to merge clusters to group genes annotated for the same function, reducing the number of optimal clusters K with respect to the optimal number of clusters selected by ICL. SICL tends to split clusters in order to obtain clusters made up only of annotated genes, increasing the number of optimal clusters with respect to the optimal number of clusters selected by ICL. In other words, SICL tends to select more complex models than ICL while ICAL tends to favor more parsimonious models than ICL. Note that this behavior of ICAL and SICL are general trends, not rules: ICAL does not always merge clusters and SICL does not always split them since clusters for different solutions are not necessarily nested in each other.

4 Numerical illustrations

4.1 Simulation settings

To illustrate the behavior of the proposed ICAL criterion, we consider a numerical example. We simulate 200 observations from a mixture of four bivariate Gaussian distributions, 100 independent times (see parameters in Table 1). For a given model indexed by K, the estimation of parameters is performed with the R package `Rmixmod` (Biernacki et al., 2006; Lebret et al., 2013) for a Gaussian mixture model with diagonal variance matrix (that is, the $p_k L_k B_k$ model in the notation of the `Rmixmod` package, corresponding to clusters with variable proportions, variable volumes, variable shapes and vertical or horizontal orientation). We estimate the parameters for models with the number of clusters K varying from 1 to 10 and perform model selection to select the most appropriate number of clusters. Over the 100 replicated datasets, the BIC most frequently selects four clusters (81 times). Indeed, we note that these clusters correspond to the simulated Gaussian components. The ICL criterion selects either three (54 times) or four clusters (46 times), as it tends to merge the two similar components (1 and 2 from Table 1) .

Component	Mixing proportions	Component distribution
1	0.25	$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1.7\end{pmatrix}\right)$
2	0.25	$\mathcal{N}\left(\begin{pmatrix}0\\5\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1.7\end{pmatrix}\right)$
3	0.25	$\mathcal{N}\left(\begin{pmatrix}9\\8\end{pmatrix}, \begin{pmatrix}0.1 & 0\\0 & 1\end{pmatrix}\right)$
4	0.25	$\mathcal{N}\left(\begin{pmatrix}9\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 0.8\end{pmatrix}\right)$

Table 1: Parameters of simulated datasets: the first two components are close to one another while the third and fourth are clearly distinct from the first two and also distinct from each other.

We illustrate the potential utility of accounting for external gene annotations in model selection by simulating such annotations and performing model selection with the corresponding SICL and the ICAL criteria. We simulate three types of functional annotations: $\mathbf{u}_\mathrm{A}$, $\mathbf{u}_\mathrm{B}$ and $\mathbf{u}_\mathrm{C}$ (see Figure 1). The genes annotated for the first function $\mathbf{u}_\mathrm{A}$ are shared by the two closest mixture components (components 1 and 2 from Table 1). This annotation is designed to be *associated to the components* in the sense that it suggests the interest of merging the two clusters, as they share similar joint distributions and external annotations. The genes annotated for the second function $\mathbf{u}_\mathrm{B}$ are shared only by the two clearly distinct components (components 3 and 4 from Table 1). This annotation is designed to be *unassociated with the components*: although the components share a similar function, their joint distributions are too distinct to be merged from a modelling point of view. Finally, the genes annotated for the third function $\mathbf{u}_\mathrm{C}$ are randomly spread over the four components: meaning the annotation is *mixed* (half associated / half unassociated). For each function, we simulate the annotation using binomial random variables, with parameters fixed to yield on average 20 annotated genes over 200 possible genes.

4.2 Simulation results

All penalized criteria (BIC, ICL, SICL and ICAL) versus the number of clusters for one simulated dataset are displayed in Figure 2 (left). We note that the peak of ICAL is sharper for the three cluster solution than the peak of ICL. Over the 100 simulated datasets, ICAL selects three clusters 87 times, merging the two closest components 1 and 2 (Table 2). This three cluster solution is meaningful with respect to the

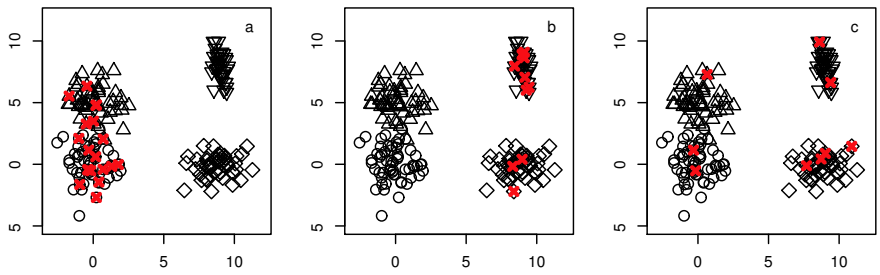


Figure 1: Illustration of a simulated dataset and three annotation patterns. For each figure, the 200 observations are drawn from a mixture of Gaussian bivariate components whose parameters are defined in Table 1: circles, triangles, inverted triangles and diamonds correspond to components 1, 2, 3 and 4. The three figures correspond to three annotation patterns: associated annotation $\mathbf{u}_A$ (left), unassociated annotation $\mathbf{u}_B$ (center) and mixed annotations $\mathbf{u}_C$ (right). For each annotation, the 20 annotated genes are represented by coloured bold crosses.

information provided by $\mathbf{u}_A$, as all annotated genes are attributed to the same cluster. In this case, the external information provided by the associated annotation $\mathbf{u}_A$ reinforces the model selection. Using the same pattern as $\mathbf{u}_A$ (annotations shared by components 1 and 2 only), we simulate twelve independent associated annotations using binomial random variables, each with parameters fixed to have on average 20 annotated genes over 200 possible genes. For this set of external annotations, the peak of the ICAL displayed in Figure 2 (right) is much sharper than the peak of ICL. Over the 100 simulated datasets, ICAL systematically selects a three cluster solution (Table 2). In contrast, SICL more frequently selects a four− or even five− cluster solution, as it leads to a preference of smaller clusters containing only annotated genes (i.e., a high specificity of annotation within each cluster). This demonstrates the utility of the ICAL criterion over the SICL, as it does not correctly take into account the specificity of gene annotation.

For unassociated annotation $\mathbf{u}_B$, the behavior of the information criteria versus the number of clusters for one simulated dataset is displayed in Figure 3 (left). We note that the ICAL criterion behaves similarly to the ICL. Over the 100 simulated datasets, ICAL, as ICL, leads to some uncertainty as to whether a three cluster solution (53 times) or a four cluster solution (47 times) is best (Table 2). In this case, the annotation $\mathbf{u}_B$ is not related to the components and has no impact on the resulting clustering, even if the number of annotations is increased to 12, each simulated with the same pattern as $\mathbf{u}_B$ as displayed in Figure 3 (right).

Finally, for the mixed annotation $\mathbf{u}_C$, ICAL most frequently selects three clusters (79 times) or four clusters (21 times). Because the annotation $\mathbf{u}_C$ is mixed, there is less evidence to merge the two clusters on the left than in the case of the informative annotation $\mathbf{u}_A$. Using the three types of annotations at the same time $(\mathbf{u}_A, \mathbf{u}_B, \mathbf{u}_C)$, the ICAL criterion almost systematically selects three clusters (Table 2).

The potential utility of accounting for external gene annotations in model selection is highlighted in the numerical results summarized in Table 2. First, these results illustrate that the SICL is not well-adapted to account for gene annotations in model selection; at best SICL behaves like ICL and at worst, erroneously splits clusters that should be merged. However, if the external information is associated to the components, even partially so, the use of the ICAL criterion improves model selection in terms of functional interpretability. If the external information is unassociated to the components, the ICAL criterion simply behaves like the ICL.

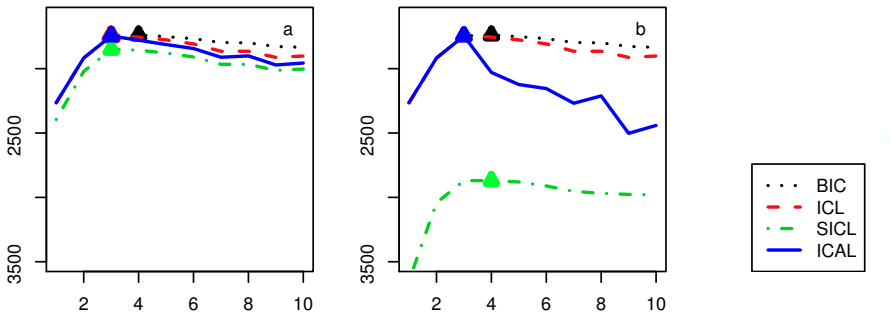


Figure 2: BIC, ICL, SICL and ICAL information criteria versus the number of clusters on one simulated dataset for the informative annotations: $\mathbf{u}_{\mathrm{A}}$ (left) and $\mathbf{u}_{\mathrm{A}}^{1}, \ldots, \mathbf{u}_{\mathrm{A}}^{12}$ (right). Triangles indicate the maximum value attained by each criterion.

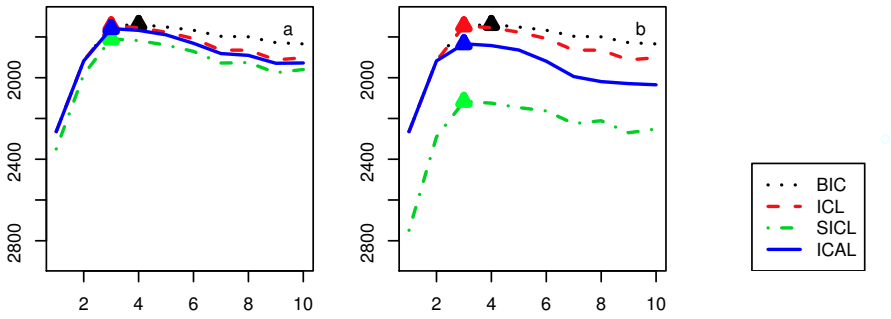


Figure 3: BIC, ICL, SICL and ICAL information criteria versus the number of clusters on one simulated dataset for the non informative annotations: $\mathbf{u}_{\mathrm{B}}$ (left) and $\mathbf{u}_{\mathrm{B}}^{1}, \ldots, \mathbf{u}_{\mathrm{B}}^{12}$ (right). Triangles indicate the maximum value attained by each criterion.

		K	1	2	3	4	5	6	7	8	9	10
		BIC			19	81	2					
		ICL			54	46						
Associated annotations	$\mathbf{u}_A$	SICL			53	47						
		ICAL			87	13						
	$\mathbf{u}_A^1, \ldots, \mathbf{u}_A^{12}$	SICL			29	69	2					
		ICAL			100							
Unassociated annotations	$\mathbf{u}_B$	SICL			53	47						
		ICAL			53	47						
	$\mathbf{u}_B^1, \ldots, \mathbf{u}_B^{12}$	SICL			53	47						
		ICAL			53	47						
Mixed annotations	$\mathbf{u}_C$	SICL			49	51						
		ICAL			79	21						
Multiple annotations	$\mathbf{u}_A, \mathbf{u}_B, \mathbf{u}_C$	SICL			48	52						
		ICAL			97	4						

Table 2: Number of simulated datasets for which each model ($K = 1, \ldots, 10$) was selected by BIC, ICL, SICL and ICAL for several external annotations over 100 independent datasets simulated with parameters detailed in Table 1. The model most commonly selected for each criterion is highlighted in red.

5 RNA-seq data analysis

5.1 Presentation of the RNA-seq data and clustering settings

Mach et al. (2014) analyzed transcriptome differences in the small intestine of healthy piglets to better understand their immune response. The expression of 24924 genes across 12 samples was measured using RNA-seq, corresponding to 3 different tissues (the duodenum, the jejunum and the ileum), each sequenced for 4 different healthy piglets. We performed a differential analysis using a negative binomial generalized linear model as implemented in the `edgeR` package version 3.4.2 (Robinson et al., 2010). We identified 4021 genes as differentially expressed among any of the tissues after controlling the false discovery rate (FDR) below the level 0.05 with the approach of Benjamini and Hochberg (1995). For the following co-expression analysis, we restrict our attention to this set of differentially expressed genes.

Prior to the co-expression analysis, we applied the voom transformation to stabilize the markedly unequal variabilities typical of RNA-seq data (Law et al., 2014), where library size normalization factors were computed on the full set of genes. Subsequently, Gaussian mixture models were estimated for the transformed data using the `Rmixmod` package version 2.0.2 (Biernacki et al., 2006) for a number of clusters from 1 to 50. For each model, we used a *small EM* strategy for initiation and repeated estimation 10 times.

5.2 Presentation of functional annotation data

The Molecular Signatures Database (Liberzon et al., 2011) was built by the Brain Institute and provides collections of annotated gene sets for use with the Gene Set Enrichment Analysis software (Subramanian et al., 2005). The Molecular Signatures Database (MSigDB) contains collections of gene sets from several sources: positional gene sets, curated gene sets from online pathway databases, motif gene sets, computational gene sets, GO gene sets, oncogenic canonical pathways and immunologic signatures. We used the Canonical Pathways (CP) gene sets collection, compiling 1320 canonical representations of biological processes curated by domain experts from online metabolic and signaling pathways databases such as the KEGG (`http://www.genome.jp/kegg`), BioCarta (`http://www.biocarta.com`) and Reactome databases (`http://www.reactome.org`).

Among the 1320 CP in the database, 1131 are represented among the 4021 differentially expressed genes. We select the CPs for which annotated genes are overrepresented in the set of differentially expressed genes with respect to the set of non-null genes using a Fisher's exact test. Since a test is performed for every possible annotation (i.e., each CP), we select those whose adjusted p-value is less than 0.05, after applying a Bonferroni correction for multiple testing. This procedure yields 10 CPs of interest, as described in Table 3.

5.3 Model selection

We compare the results of model selection performed by the four different criteria presented in Sections 2 and 3: BIC selects 28 clusters, ICL and SICL select 23 clusters while ICAL selects 20 (see Figure 4).

The approximate correspondences between clusters in the ICAL and ICL solutions are displayed in Table 4. Although the result of the former is not perfectly nested in the latter, in many cases the attribution of genes to clusters in the ICAL solution is a result of collapsing or partially collapsing several clusters from the ICL solution. This suggests that the ICL favors a slightly more complex solution, as expected; we next investigate whether the more parsimonious solution of the ICAL appears to be coherent given the set of CP used.

For the ICL and ICAL solutions, we examine associations between clusters and CP using Fisher's exact test. Significant p-values are summarized in Table 5. The ICAL criterion yields a clustering that maximizes the number of genes annotated in each cluster for each CP while still only grouping genes

CP	Name	DE genes	Total genes
1	Reactome metabolism of lipids and lipoproteins	141	480
2	Reactome transmembrane transport of small molecules	124	415
3	Reactome hemostasis	99	468
4	Reactome SLC mediated transmembrane transport	73	243
5	Reactome phospholipid metabolism	54	200
6	Reactome fatty acid triacylglycerol and ketone body metabolism	53	170
7	KEGG PPAR signaling pathway	34	71
8	KEGG ECM receptor interaction	34	86
9	Reactome transport of inorganic cations anions and amino acids oligopeptides	33	96
10	KEGG peroxisome	31	80

Table 3: Number of genes annotated for each canonical pathway (CP): among the 4021 differentially expressed (DE) genes and among the full CP gene set collection of the MSigDB database.

that share sufficiently similar expression profiles. For example, we note that CP8 is associated with two different clusters in the ICL solution, while it is associated with a single cluster in the ICAL solution; similarly, CP10 is associated with three clusters in the ICL solution and only two clusters in the ICAL solution. On the other hand, although clusters 10 and 17 in the ICAL solution both share annotations for CP10, these clusters are not collapsed into one using the proposed criterion, as their expression dynamics are too different. As such, the ICAL solution appears to enable the identification of more biologically interpretable clusters than the ICL, while still ensuring that the clustered genes share sufficiently similar expression dynamics.

Finally, we note that the ICAL solution exhibits two clusters of particular interest with respect to the biological processes studied: Cluster 5 (379 genes) is associated with CP3 (reactome homeostasis, p=0.0002) and CP8 (KEGG ECM receptor interaction, p=0.00001). Cluster 10 (297 genes) is associated with CP1 (reactome metabolism of lipids and lipoproteins, p=0.002), CP6 (reactome fatty acid triacylglycerol and ketone body metabolism, p=0.005) and CP10 (KEGG peroxisome, p=0.0001), all of which correspond to fatty acid metabolism. Both clusters 5 and 10 contain unknown genes that may be good candidates for follow-up studies to determine whether they may be implicated in the corresponding canonical pathways.

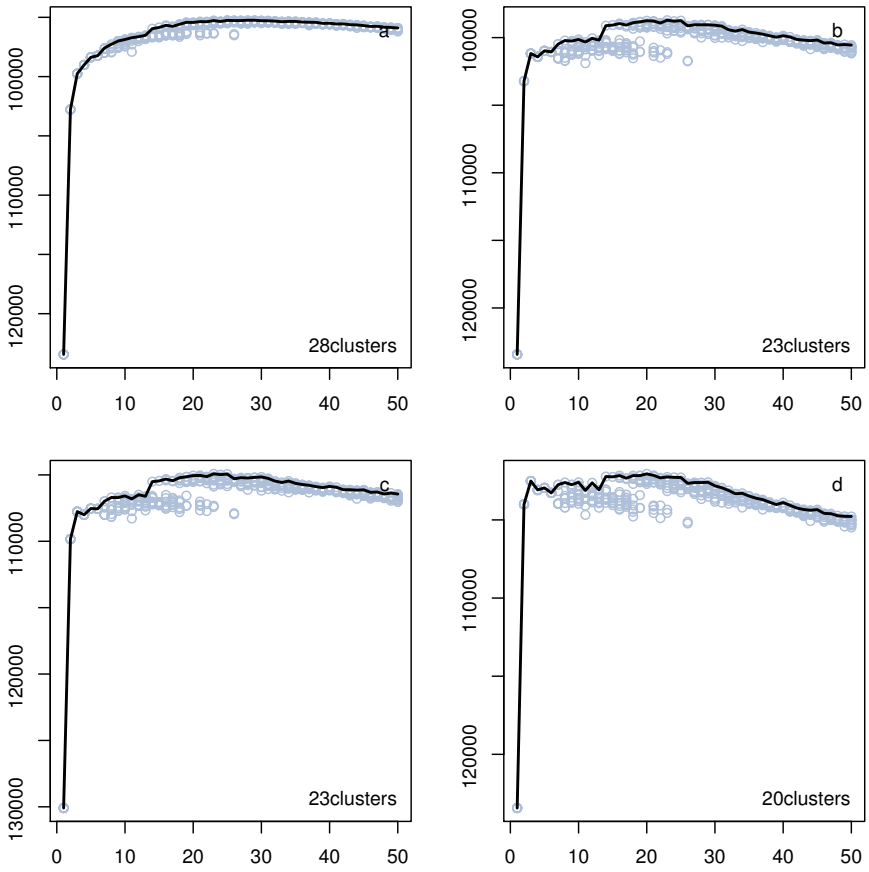


Figure 4: BIC, ICL, SICL and ICAL information criteria (respectively a, b, c and d) versus the number of clusters for the pig RNA-seq data for 10 independent initializations, represented by 10 grey circles for each number of clusters K. Solid lines link the maximum of the criteria over the 10 initializations over the collection of models. The red crosses correspond to the maximum of the criteria, which correspond to the selected models.

ICAL clusters	ICL clusters
Cluster 1	$\frac{1}{2}$ **9**
Cluster 2	15
Cluster 3	10
Cluster 4	11
Cluster 5	$\frac{1}{2}$ **5** + **12** + $\frac{1}{2}$**20** + $\frac{1}{2}$**22**
Cluster 6	**2** + $\frac{1}{2}$**5** + $\frac{1}{3}$**18**
Cluster 7	8
Cluster 8	**4** + **9** + **16**
Cluster 9	**3** + $\frac{1}{2}$**6**
Cluster 10	**7** + $\frac{1}{2}$**6**
Cluster 11	**13** + $\frac{1}{4}$ **20** + $\frac{1}{2}$**22**
Cluster 12	23
Cluster 13	**7** + **17** + $\frac{1}{3}$ **18**
Cluster 14	21
Cluster 15	19
Cluster 16	1
Cluster 17	**14** + $\frac{1}{3}$ **18**
Cluster 18	1
Cluster 19	$\frac{1}{4}$**20** + $\frac{1}{2}$ **5**
Cluster 20	16

Table 4: Approximate composition of the 20 clusters of the ICAL solution with respect to the 23 clusters of the ICL solution. Lines in bold correspond to clusters of the ICAL solution that are formed by several clusters or parts from clusters of the ICL solution. For example, Cluster 5 of the ICAL solution is approximately made of Clusters 12 and parts of Clusters 5, 20 and 22 of the ICL solution.

(a) **ICL solution**

	size	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP 8	CP9	CP10
Cluster 2	58		$\star$	$\star$	$\star$						
Cluster 5	203								$\star$		
Cluster 6	47										$\star\star$
Cluster 7	258	$\star$					$\star$				$\star$
Cluster 8	96					$\star\star$					
Cluster 10	287									$\star$	
Cluster 14	225										$\star\star$
Cluster 22	144			$\star\star$					$\star\star\star$		

(b) **ICAL solution**

	size	CP1	CP2	CP3	CP4	CP5	CP6	CP7	CP 8	CP9	CP10
Cluster 3	297									$\star$	
Cluster 5	379			$\star\star$					$\star\star\star$		
Cluster 6	156		$\star\star$	$\star$							
Cluster 7	92					$\star$					
Cluster 10	267	$\star$					$\star\star$				$\star\star$
Cluster 17	235										$\star\star$

Table 5: Table of associations between clusters and CP for the ICL solution (a) and the ICAL solution (b). Associations are detected using Fisher's exact tests: the number of stars indicates the value of the p-value ($\star$ below 0.01, $\star\star$ below 0.001, $\star\star\star$ below 0.0001).

6 Discussion

In this paper, we present a novel way to incorporate functional annotations into model-based clustering of gene expression data. To this end, we develop a model selection criterion, the Integrated Completed Annotated Likelihood (ICAL) which is designed to select the model that jointly maximises the goodness-of-fit to the data and the association of clusters and annotations. From a biological point of view, the ICAL criterion aims to select models with more interpretable clusters than those selected by BIC or ICL. It is important to note that the functional annotations are not directly included in the clustering model and are only used to select the best model. This approach is a good compromise between two opposite strategies: including functional annotations directly in the clustering model (Morlini, 2011) or excluding them altogether and using them only to validate clusters *a posteriori*. Since we do not include annotations in the clustering model, we detect associations between annotations and clusters with a stronger evidence than if we had included the external annotations in the clustering model. In particular, the ICAL criterion is a good way to include prior biological expertise without according it too much importance, which is a good balance between what can be observed in the data and what experts expect to see in the data.

As illustrated in numerical simulations, the performance of the ICAL is highly dependent on the quality of the information provided. Selecting the appropriate annotations to include in model selection is an important step and should be performed by an expert. We also suggest the use of gene annotation databases that are curated manually by experts, such as the gene sets collection from the MSigDB database (Liberzon et al., 2011).

In this work, we applied the ICAL using the framework of Gaussian mixture models, but the extension to other mixture models is straightforward; including Poisson (Rau et al., 2014) or Dirichlet multinomial mixture models (Holmes et al., 2012). In addition, this model selection strategy may be useful for other types of data which may also be associated with incomplete external annotations (e.g., sociology, marketing).

7 Acknowledgements

We thank Jordi Estellé for providing the RNA-seq data in section 5 and giving insights on the external annotations and clustering results.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–9.

Baudry, J.-P., Cardoso, M., Celeux, G., Amorim, M. J., and Ferreira, A. S. (2014). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*, 1(1):1–20.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),*, 57(1):286–300.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster analysis and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, 51:587–600.

Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466.

Dempster, A., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39(1):1–38.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–8.

Holmes, I., Harris, K., and Quince, C. (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS ONE*, 7(2):e30126.

Huang, D., Wei, P., and Pan, W. (2006). Combining gene annotations and gene expression data in model-based clustering: weighted method. *Omics : a journal of integrative biology*, 10(1):28–39.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., and 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.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–40.

Mach, N., Berri, M., Esquerré, D., Chevaleyre, C., Lemonnier, G., Billon, Y., Lepage, P., Oswald, I. P., Doré, J., Rogel-Gaillard, C., and Estellé, J. (2014). Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PloS ONE*, 9(2):e88515.

Morlini, I. (2011). A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification*, 6(1):5–28.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–8.

Pan, W. (2006). Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801.

Rau, A., Martin-Magniette, M.-L., Maugis-Rabusseau, C., and Celeux, G. (2014). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *(submitted)*.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–70.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3.

Steuer, R., Humburg, P., and Selbig, J. (2006). Validation and functional annotation of expression-based clusters based on gene ontology. *BMC Bioinformatics*, 7:380.

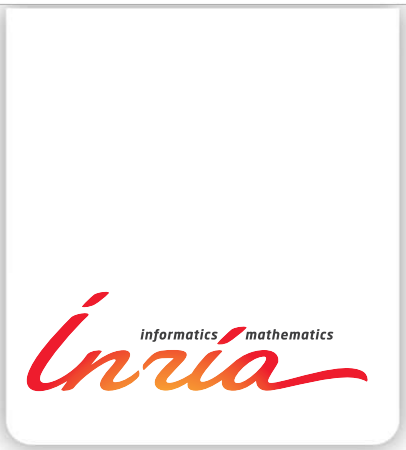
Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–50.

Tari, L., Baral, C., and Kim, S. (2009). Fuzzy c-means clustering with prior biological knowledge. *Journal of Biomedical Informatics*, 42(1):74–81.

Tipney, H. and Hunter, L. (2010). An introduction to effective use of enrichment analysis software. *Human Genomics*, 4(3):202.

Verbanck, M., Lê, S., and Pagès, J. (2013). A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, 14:42.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, a. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–87.



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399