Clustering high-throughput sequencing data with Poisson mixture models

Andrea Rau, Gilles Celeux, Marie-Laure Martin-Magniette, Cathy Maugis-Rabusseau

Clustering high-throughput sequencing data with Poisson mixture models

Andrea Rau , Gilles Celeux, Marie-Laure Martin-Magniette, Cathy Maugis-Rabusseau

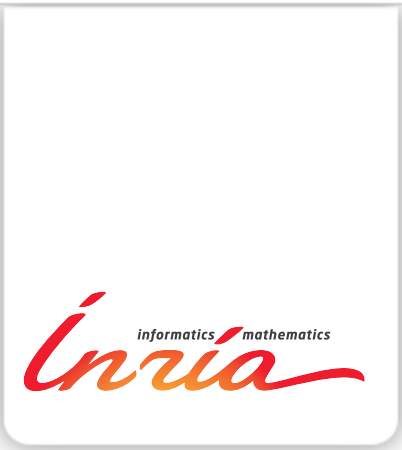
RESEARCH REPORT

N° 7786

November 2011

Project-Team Select

ISSN 0249-6399 ISRN INRIA/RR--7786--FR+ENG



Clustering high-throughput sequencing data with Poisson mixture models

Andrea Rau* †, Gilles Celeux†, Marie-Laure Martin-Magniette‡§, Cathy Maugis-Rabusseau¶

Project-Team Select

Research Report n° 7786 — November 2011 — 33 pages

Abstract: In recent years gene expression studies have increasingly made use of next generation sequencing technology. In turn, research concerning the appropriate statistical methods for the analysis of digital gene expression has flourished, primarily in the context of normalization and differential analysis. In this work, we focus on the question of clustering digital gene expression profiles as a means to discover groups of co-expressed genes. We propose two parameterizations of a Poisson mixture model to cluster expression profiles of high-throughput sequencing data. A set of simulation studies compares the performance of the proposed models with that of an approach developed for a similar type of data, namely serial analysis of gene expression. We also study the performance of these approaches on two real high-throughput sequencing data sets. The R package `HTSCluster` used to implement the proposed Poisson mixture models is available on CRAN.

Key-words: Mixture models, clustering, co-expression, RNA-seq, EM-type algorithms

* INRA, UMR 1313 GABI, Jouy-en-Josas, France
† INRIA Saclay - Île-de-France, Orsay, France
‡ UMR INRA 1165 - UEVE, ERL CNRS 8196, Unité de Recherche en Génomique Végétale, Evry, France
§ UMR AgroParisTech/INRA MIA 518, Paris, France
¶ Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

Classification de données de séquençage à haut-débit avec les modèles de mélange de Poisson

Résumé : De plus en plus, les études d'expression de gènes utilisent les techniques de séquençage de nouvelle génération, entraînant une recherche grandissante sur les méthodes les plus appropriées pour l'exploitation des données digitales d'expression, à commencer pour leur normalisation et l'analyse différentielle. Ici, nous nous intéressons à la classification non supervisée des profils d'expression pour la découverte de groupes de gènes coexprimés. Nous proposons deux paramétrisations d'un modèle de mélange de Poisson pour classer des données de séquençage haut-débit. Par des simulations, nous comparons les performances de ces modèles avec des méthodes similaires conçus pour l'analyse en série de l'expression des gènes (SAGE). Nous étudions aussi les performances de ces modèles sur deux jeux de donnnées réelles. Le package R `HTSCluster` associé à cette étude est disponible sur le CRAN.

Mots-clés : Modèles de mélange, classification, co-expression, RNA-seq, algorithmes de type EM

1 Introduction

The application of high-throughput sequencing (HTS), also commonly referred to as next generation sequencing (NGS), to the study of gene expression has revolutionized the scope and depth of understanding of the genome, epigenome, and transcriptome of dozens of organisms, including humans, mice, fruit flies, and plants such as maize and *Arabidopsis thaliana*. For example, the recent application of HTS technologies to sequence ribonucleic acid (RNA) content, such as messenger RNA (mRNA) and small RNA (sRNA), has rivaled the use of microarrays for transcriptomic studies for several reasons. This technique, referred to as RNA sequencing (RNA-seq), offers a way to quantify gene expression exactly by providing counts of transcripts. In addition, RNA-seq can provide information about the transcriptome at a level of a detail not possible with microarrays, including strand-specific, allele-specific, or gene isoform-specific expression, alternative splicing, and transcript discovery.

Although a variety of different technologies (e.g., the FLX pyrosequencing system of 454 Life Sciences and the Illumina Genome Analyzer of Solex) and protocols (e.g., single- and paired-end) exist for high-throughput sequencing studies, the same broad pre-processing steps are followed. Namely, if an appropriate genome sequence reference is available, reads are mapped to the genome or transcriptome (either directly or after first being assembled into contigs); otherwise, *de novo* assembly may be used. After alignment or assembly, read coverage for a given biological entity (e.g., a gene) is subsequently calculated. The quantification of gene expression in RNA-seq data is currently an active area of research (Trapnell *et al.*, 2010), and in this work, we focus on measures of digital gene expression (counts). These count-based measures of gene expression differ substantially from data produced with microarrays. For example, RNA-seq data are discrete, positive, and highly skewed, with a very large dynamic range (up to 5 orders of magnitude). In addition, the sequencing depth (i.e., the library size) and coverage vary between experiments, and read counts are known to be correlated with gene length (Oshlack and Wakefield, 2009; Łabaj *et al.*, 2011).

To date, most developments concerning the statistical analysis of RNA-seq data have dealt with the issues of experimental design (Auer and Doerge, 2010), normalization (e.g., Mortazavi *et al.*, 2008; Robinson and Oshlack, 2010), and the analysis of differential expression (e.g., Anders and Huber, 2010; Robinson and Smyth, 2007; Auer and Doerge, 2011). In this work, we focus on the question of clustering expression profiles for RNA-seq data. Identifying biological entities that share similar profiles across several treatment conditions, such as co-expressed genes, may help identify groups of genes that are involved in the same biological processes (Eisen *et al.*, 1998; Jiang *et al.*, 2004). Clustering analyses based on metric criteria such as the K-means algorithm (MacQueen, 1967) and hierarchical clustering (Ward, 1963), have been used to cluster microarray-based measures of gene expression as they are rapid, simple, and stable. However, such methods require both the choice of metric and criterion to be optimized, as well as the selection of the number of clusters. An alternative to such methods are probabilistic clustering models, where the objects to be classified (genes) are considered to be a sample of a random vector and a clustering of the data is obtained by analyzing the density of this vector (McLachlan *et al.*, 2004; Yeung *et al.*, 2001).

Presently, most proposals for clustering RNA-seq data have focused on variables (biological samples), rather than observations (biological entities, e.g., genes). For example, Anders and Huber (2010) perform a hierarchical clustering with a Euclidean distance of variables following a variance-stabilizing transformation, and Severin *et al.* (2010) cluster fourteen diverse tissues of soybean using hierarchical clustering with Pearson correlation after normalizing the data using a variation of RPKM. More recently, Witten (2011) discusses the clustering of variables using hierarchical clustering with a modified log likelihood ratio statistic as distance measure, based on two different parameterizations of a Poisson loglinear model. The model in this proposal is similar to those of Cai *et al.* (2004) and Kim *et al.* (2007) for the clustering of Serial Analysis of Gene Expression (SAGE) observations using a K-means algorithm (MacQueen, 1967) and a Poisson loglinear model.

In this work, like Cai *et al.* (2004) and Kim *et al.* (2007), we focus on the use of Poisson loglinear models for the clustering of count-based HTS observations; however, rather than using such a model to define a distance metric to be used in a K-means or hierarchical clustering algorithm, we make use of finite mixtures of Poisson loglinear models. This framework has the advantage of providing straightforward procedures for parameter estimation and model selection, as well as an a posteriori probability for each gene of belonging to each cluster. The rest of this paper is organized as follows. In Section 2, after a brief review of inference methods for finite mixture models, a Poisson mixture model approach for clustering RNA-Seq expression profiles is presented. In Section 3, the performances of the proposed models and those of Cai *et al.* (2004) and Kim *et al.* (2007) are evaluated in a simulation study. A clustering analysis is performed on real HTS data in Section 4, and Section 5 contains a discussion.

2 Methods

We first define the notation that will be used throughout this paper. Let Y_{ijl} be the random variable corresponding to the digital gene expression measure (DGE) for observation i ($i = 1, \ldots, n$) of condition j ($j = 1, \ldots, d$) in biological replicate l ($l = 1, \ldots, r_j$), with y_{ijl} being the corresponding observed value of Y_{ijl}. Note that we define an observation as a biological entity, such as a gene, and a variable as a replicate in a given condition. Let $q = \sum_{j=1}^{d} r_j$ be the total number of variables (all replicates in all conditions) in the data, such that $\mathbf{y} = (y_{ijl})$ is the $n \times q$ matrix of the DGE for all observations and variables, and $\mathbf{y}_i$ is the q-dimensional vector of DGE for all variables of observation i. Note that for microarray data, $\mathbf{y}_i \in \mathbb{R}^q$, and for RNA-seq data, $\mathbf{y}_i \in \mathbb{N}^q$. We use dot notation to indicate summations in various directions, e.g., $y_{\cdot jl} = \sum_i y_{ijl}$, $y_{i\cdot\cdot} = \sum_j \sum_l y_{ijl}$, etc. Finally, let s_{jl} represent the normalized library size for replicate l of condition j.

2.1 Finite mixture models

In the context of finite mixture models, the data $\mathbf{y}$ are assumed to come from g distinct subpopulations (clusters), each of which is modeled separately (McLachlan and Peel, 2000). The overall population is thus a mixture of these subpopulations. The

general form of a finite mixture model with g components is

$$f(\mathbf{y}; g, \mathbf{\Psi}_g) = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k) \tag{1}$$

where $\mathbf{\Psi}_g = (\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\theta}')'$, $\boldsymbol{\theta}'$ contains all of the parameters in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g$ assumed to be distinct, $f_k(\cdot)$ are the densities of each of the components and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)'$ are the mixing proportions, with $\pi_k \in (0, 1)$ for all k and $\sum_k \pi_k = 1$. The choice of family and parameterization of $f_k(\cdot)$ depends on the the nature of the observed data.

The mixture model in Equation (1) may also be thought of as an incomplete data structure model, with complete data

$$(\mathbf{y}, \mathbf{z}) = ((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$$

where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) = (z_{ik})$, and $z_{ik} = 1$ if observation i arises from group k and 0 otherwise. The latent variable $\mathbf{z}$ thus defines a partition $P = (P_1, \ldots, P_g)$ of the observed data $\mathbf{y}$ with $P_k = \{i | z_{ik} = 1\}$. There are two approaches to estimating the parameters of a finite mixture model and obtaining a clustering of the data: the estimation approach and the clustering approach.

2.1.1 Estimation approach

In the estimation method, mixture parameters are estimated by computing the maximum likelihood estimate of the parameter $\mathbf{\Psi}_g$ and each observation is assigned to the cluster maximizing t_{ik}, that is, the conditional probability that observation i belongs to cluster k. The log likelihood is

$$L(\mathbf{\Psi}_g; \mathbf{y}, g) = \log \left[\prod_{i=1}^{n} f(\mathbf{y}_i; g, \mathbf{\Psi}_g) \right] = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{g} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k) \right].$$

The reference tool to derive maximum likelihood estimates (MLE) in a mixture model is the Expectation-Maximization (EM) algorithm of Dempster *et al.* (1977). After initializing the parameters $\mathbf{\Psi}_g^{(0)}$, the E-step at iteration b corresponds to computing the conditional probability that an observation i arises from the k^{th} component for the current value of the mixture parameters:

$$t_{ik}^{(b)} = \frac{\pi_k^{(b)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_k^{(b)})}{\sum_{m=1}^{g} \pi_m^{(b)} f_m(\mathbf{y}_i; \boldsymbol{\theta}_m^{(b)})}. \tag{2}$$

Then, in the M-step the mixture parameter estimates are updated to maximize the expected value of the completed likelihood, which leads to weighting the observation i for group k with the conditional probability $t_{ik}^{(b)}$:

$$\pi_k^{(b+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(b)}. \tag{3}$$

The updates of the parameters $\boldsymbol{\theta}_k^{(b+1)}$ depend on the likelihood equations defined by the densities $f_k(\cdot)$.

2.1.2 Clustering approach

In the clustering approach, the mixture parameters $\boldsymbol{\theta}$ and the underlying partition are estimated concurrently. The Clustering EM (CEM) algorithm (Celeux and Govaert, 1992) estimates both the mixture parameters and the labels by maximizing the completed likelihood:

$$L_C(\boldsymbol{\Psi}_g; \mathbf{y}, \mathbf{z}, g) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log \left[\pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)\right]. \tag{4}$$

In the E-step of the algorithm, the conditional probabilities $t_{ik}^{(b)}$ are computed as in Equation (2). In the C-step, each observation i is assigned to the component maximizing the conditional probability $t_{ik}^{(b)}$ (i.e., using the so-called maximum a posteriori (MAP) rule). Finally, in the M-step, the mixture parameter estimates are updated by maximizing the completed likelihood in Equation (4). Because it aims to maximize the complete likelihood, where the component label of each sample point is included in the data, the CEM may be seen as a K-means-like algorithm. As such, contrary to the EM algorithm, the CEM algorithm converges in a finite number of iterations, although it does provide biased estimates of the mixture parameters (see for instance McLachlan and Peel (2000), Section 2.21).

2.2 Poisson mixture model

Although a multivariate version of the Poisson distribution does exist (Karlis, 2003), it is difficult to implement, particularly for data with high dimensionality. For this reason, the variables are assumed to be independent conditionally on the components and an observation from the kth component follows a Poisson distribution:

$$\mathbf{y}_i | k \sim \prod_{j=1}^{d} \prod_{l=1}^{r_j} \mathcal{P}(y_{ijl}; \mu_{ijlk}),$$

where $\mathcal{P}(\cdot)$ denotes the standard Poisson density. This corresponds to the following Poisson mixture model (PMM):

$$f(\mathbf{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\pi}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} \prod_{l=1}^{r_j} \mathcal{P}(y_{ijl}; \mu_{ijlk})$$

where $\sum_k \pi_k = 1$ and $\pi_k \geq 0$ for all k. The unconditional mean and variance of Y_{ijl}, respectively, are:

$$\mathbb{E}[Y_{ijl}] = \sum_{k=1}^{g} \pi_k \mu_{ijlk}$$

and

$$\begin{aligned} \mathrm{Var}(Y_{ijl}) &= \mathbb{E}\left[\mathrm{Var}(Y_{ijl}|\mathbf{z}_i)\right] + \mathrm{Var}\left[\mathbb{E}(Y_{ijl}|\mathbf{z}_i)\right] \\ &= \sum_{k=1}^{g} \pi_k \mu_{ijlk} + \sum_{k=1}^{g} \pi_k \mu_{ijlk}^2 - \left(\sum_{k=1}^{g} \pi_k \mu_{ijlk}\right)^2. \end{aligned}$$

2.2.1 PMM-I parameterization

At this point, we consider two possible parameterizations for the mean μ_{ijlk} to model digital gene expression. In the first, called PMM-I, we consider

$$\mu_{ijlk} = w_i \lambda_{jk} \tag{5}$$

where w_i corresponds to the expression level of observation i (e.g., lowly to highly expressed) and $\boldsymbol{\lambda}_k = (\lambda_{1k}, \ldots, \lambda_{dk})$ corresponds to the clustering parameters that define the profiles of the genes in cluster k across all variables. Note that in this parameterization, different replicates (l and l') of a given observation i in a given condition j and cluster k have the same mean.

If the parameters w_i and λ_{jk} in Equation (5) are left unconstrained, the model is not identifiable. As such, we consider the constraint $\sum_j \lambda_{jk} r_j = 1$ for all $k = 1, \ldots, g$. The interpretation of this constraint is as follows: if each condition has one biological replicate ($r_j = 1$ for all j), the parameters λ_{jk} represent the percentage of total reads per gene that are attributed to each condition. When at least one condition has two or more replicates, λ_{jk} represents a down-weighted proportion; that is, if a particular condition has more than one replicate, λ_{jk} represents the proportion of counts taken up by a single replicate, and $r_j \lambda_{jk}$ represents the proportion taken up by all replicates in the condition.

For parameter estimation of the PMM-I model in (5), the E-step is done as shown in Equation (2), with $\boldsymbol{\theta}_k^{(b)} = (\mu_{ijlk}^{(b)}) = (\hat{w}_i \lambda_{jk}^{(b)})$. Note that due to the constraint mentioned above, $\hat{w}_i$ is calculated only once, such that $\hat{w}_i = y_{i\cdot\cdot}$, and thus is not indexed by the iteration b. For the M-step, $\boldsymbol{\pi}^{(b+1)}$ is estimated as shown in Equation (3). Then, after solving the likelihood equation subject to the aforementioned constraint, it is straightforward to show that

$$\hat{\lambda}_{jk}^{(b+1)} = \frac{\sum_i t_{ik}^{(b)} y_{ij\cdot}}{r_j \sum_i t_{ik}^{(b)} y_{i\cdot\cdot}}.$$

2.2.2 PMM-II parameterization

The assumption of equal means across replicates of a given gene i in condition j and cluster k is unlikely to be true, due in part to differences in library size. To address this issue, in the second parameterization, called PMM-II, we consider

$$\mu_{ijlk} = w_i s_{jl} \lambda_{jk} \tag{6}$$

where w_i and $\boldsymbol{\lambda}_k = (\lambda_{jk})$ are as before and s_{jl} is the normalized library size (a fixed constant) for replicate l of condition j. Note that different replicates (l and l') of a given gene i in condition j and cluster k no longer have the same mean, due to the presence of the library size parameter s_{jl}. As with other authors (Bullard *et al.*, 2010; Robinson and Oshlack, 2010; Anders and Huber, 2010), after estimating s_{jl} from the data, we consider this parameter to be a fixed, known constant in the model. One natural estimator for s_{jl} is $y_{\cdot jl}/y_{\cdot\cdot\cdot}$, although we will discuss alternatives in Section 3.

As in the model of Equation (5), a parameter constraint is necessary to allow for parameter estimation. We consider the constraint that $\sum_j \sum_l \lambda_{jk} s_{jl} = \sum_j \lambda_{jk} s_{j\cdot} = 1$ for all $k = 1, \ldots, g$. The interpretation of this constraint is similar to the previous one, the difference being that the clustering parameters λ_{jk} are

no longer equally shared among replicates to account for variations in library sizes.

As before, the E-step and estimates of $\boldsymbol{\pi}^{(b+1)}$ are calculated as shown in Equations (2) and (3), where $\boldsymbol{\theta}_k^{(b)} = (\mu_{ijlk}^{(b)}) = (\hat{w}_i s_{jl} \lambda_{jk}^{(b)})$ and $\hat{w}_i = y_{i\cdot\cdot}$ is calculated only once. The remainder of the M-step, subject to the above constraint, may be calculated to show

$$\hat{\lambda}_{jk}^{(b+1)} = \frac{\sum_i t_{ik}^{(b)} y_{ij\cdot}}{s_{j\cdot} \sum_i t_{ik}^{(b)} y_{i\cdot\cdot}}.$$

Finally, note that if $s_{jl} = s_j$ does not depend on l, the PMM-II model reduces to the PMM-I model since it induces that $s_{jl} = \frac{s_{j\cdot}}{r_j}$ for all l.

2.3 Implementation

To implement the Poisson mixture model, we must consider two important points: model selection (i.e., the choice of the number of clusters g) and the specification of initial parameter values $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ for the EM and CEM algorithms. For the former consideration, a reference penalized likelihood criterion for mixture models is the Bayesian information criterion (Schwarz, 1978):

$$\mathrm{BIC}(g) = \log f(\mathbf{y}; g, \hat{\boldsymbol{\Psi}}_g) - \frac{\nu_g}{2} \log(n)$$

where $\hat{\boldsymbol{\Psi}}_g$ are the maximum likelihood parameter estimates and ν_g is the number of free parameters in model with g clusters. Since BIC does not take into account the clustering purpose in assessing the number of clusters, it has a tendency to overestimate g regardless of the separation of clusters when applied to real data sets (Biernacki *et al.*, 2000). An alternative that focuses instead on the task of clustering is the Integrated Complete Likelihood (ICL) criterion (Biernacki *et al.*, 2000):

$$\mathrm{ICL}(g) = \mathrm{BIC}(g) + \mathrm{ENT}(g)$$

where

$$\mathrm{ENT}(g) \quad = \quad -\sum_{i=1}^{n} \sum_{k=1}^{g} \hat{z}_{ik} \log \hat{t}_{ik} \geq 0$$

is the entropy of the fuzzy clustering matrix, and $\hat{z}_{ik}$ is the MAP estimate. Because of this additional entropy term, the ICL favors models giving rise to data partitions with the greatest evidence. In general, the ICL appears to provide a stable and reliable estimate of g for real and simulated datasets from mixtures when the components do not overlap too much. However, as the ICL does not aim to discover the true number of mixture components, in some cases it can underestimate the number of components arising from mixtures with poorly separated components.

Parameter initialization is also an important consideration for finite mixture models, as different strategies can lead to different parameter estimates or slow convergence of the EM algorithm (McLachlan and Peel, 2000). To initialize parameter values for the EM and CEM algorithms, we use a so-called Small-EM strategy (Biernacki *et al.*, 2003). Five independent times, the following

procedure is used to obtain initial parameter values: first, a K-means algorithm (MacQueen, 1967) is run to partition the data into g clusters ($\hat{\mathbf{z}}^{(0)}$). Second, initial parameter values $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$ are calculated as follows:

$$\pi_k^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}^{(0)}$$

and

$$\lambda_{jk}^{(0)} = \frac{\sum_i y_{ij\cdot}\hat{z}_{ik}^{(0)}}{s_{j\cdot}\sum_i \hat{w}_i\hat{z}_{ik}^{(0)}}.$$

Third, five iterations of the EM algorithm are run. Finally, among the five sets of parameter values, we use $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\pi}}$ corresponding to the highest log likelihood or completed log likelihood to initialize the subsequent full EM or CEM algorithms, respectively.

3 Simulations

In this section, we perform a set of simulation experiments in order to compare the performance of the two proposed parameterizations of the Poisson mixture model to that of the models proposed by Cai *et al.* (2004) and Kim *et al.* (2007) for clustering of serial analysis gene expression (SAGE) data, a sequence-based method for expression profiling which results in counts similar to RNA-seq data (albeit on a smaller scale).

3.1 Description of competing models

In the Cai *et al.* (2004) and Kim *et al.* (2007) models, observations are assumed to follow a Poisson distribution conditional on the cluster k, with mean $\mu_{ijk} = w_i\lambda_{jk}$, under the constraint that $\sum_j \lambda_{jk} = 1$ for all k. We note that this formulation is equivalent to the PMM-I model when each condition has only one replicate (i.e., $r_j = 1$ for all conditions j); however, unlike the PMM-I, this model is not able to account for more than one replicate within a given condition. Subsequently, a K-means algorithm is defined in the following manner:

1. At iteration $b = 0$, observations are randomly assigned to a given number of clusters, g, and maximum likelihood estimates of w_i and $\lambda_{jk}^{(0)}$ are calculated for each cluster.

2. Let $\hat{\boldsymbol{\mu}}_i^{(b)} = (\hat{\mu}_{ijk}^{(b)}) = (\hat{w}_i\hat{\lambda}_{jk}^{(b)})$. At iteration b, each observation i is assigned to the cluster with minimum deviation from the expected model, where the deviation is measured by one of the following criteria:

   PoisL (Cai *et al.*, 2004):

   $$L_{ik}^{(b)} = -\log f(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_i^{(b)})$$

   PoisC (Cai *et al.*, 2004):

   $$C_{ik}^{(b)} = \sum_{j=1}^{d} \left(y_{ij} - \hat{\mu}_{ijk}^{(b)}\right)^2 / \hat{\mu}_{ijk}^{(b)}$$

TransChisq (Kim *et al.*, 2007):

$$T_{ik}^{(b)} = \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} \left((y_{ij} - y_{ij'}) - (\hat{\mu}_{ijk}^{(b)} - \hat{\mu}_{ij'k}^{(b)}) \right)^2 / \left(\hat{\mu}_{ijk}^{(b)} + \hat{\mu}_{ij'k}^{(b)} \right)$$

3. Maximum likelihood estimates of λ_{jk} are calculated based on the updated cluster membership.

4. Go to step 2 until convergence.

Cai *et al.* (2004) and Kim *et al.* (2007) remark that although the PoisL and PoisC methods perform similarly in practice, the former is less practical in terms of running time and too slow to apply to large datasets; as such, the PoisL method is not included in the comparisons of the simulation study. The Gene Expression Analysis Application used to implement the remaining two methods may be found at the authors' website (`http://cell.rutgers.edu/gea`).

For all of these methods, one limitation is that the number of clusters must typically be pre-selected by the user. To put these methods on par with the PMM-I and PMM-II for model selection, we consider three different criteria. First, as recommended by Kim *et al.* (2007) we choose the number of clusters via the Gap statistic (Tibshirani *et al.*, 2001), which compares the change in within-cluster dispersion to that expected under an appropriate reference null distribution. Second, we consider the Caliński-Harabasz (CH) index (Caliński and Harabasz, 1974), a pseudo F-statistic that compares the between and within-cluster dispersion. Finally, we remark that estimating the PMM-I model with a single replicate per condition and equal cluster sizes ($\pi_1 = \ldots, \pi_g = 1/g$) via the CEM algorithm leads exactly to the Cai *et al.* (2004) method based on the likelihood criterion (PoisL). Since there is an underlying Poisson mixture model for the PoisL, PoisC, and TransChisq methods, we may also calculate the ICL for the PoisC and TransChisq methods using the framework of finite mixture models with equal proportions.

3.2 Simulation strategy

Given the proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ and the component distributions f_k, we draw the data according to the following scheme:

$$\begin{aligned} \mathbf{z}_i &\sim \mathcal{M}(\boldsymbol{\pi}) \\ \mathbf{y}_i &\sim \text{distribution of density } f_{\mathbf{z}_i} \end{aligned} \tag{7}$$

where $\mathcal{M}(\cdot)$ is the multinomial distribution. Using this strategy, we simulated 50 datasets for $n = 2000$ genes and $g = 4$ clusters under each of six simulation settings (Table 1) in the following manner. For each dataset, we simulated $w_i \sim \text{Exp}(1/\alpha)$, where α was taken to be the mean value of $y_{i\cdot\cdot}$ from one of two real HTS datasets (labeled A and B; see Table 2). For settings with equal library sizes (Settings 1 and 2), s_{jl} was set to be $1/\sum_j r_j$ for all j. For the others (Settings 3-6), s_{jl} was taken to be the observed value of $y_{\cdot jl}/y_{\cdot\cdot\cdot}$ in the A and B data, respectively. After fixing values of $\boldsymbol{\pi}$ and τ_{jk} as shown in Table 3, we set

$$\lambda_{jk} = \frac{\tau_{jk}}{\tau_{\cdot k} s_{j\cdot}}.$$

The values of τ_{jk}, and subsequently λ_{jk}, were chosen to yield low and high cluster separation. Finally, as described in Equation (7), we simulated the latent variable $\mathbf{z}_i$ from a multinomial distribution with parameters $\boldsymbol{\pi}$. Given $z_{ik} = 1$, we consider $y_{ijl}|k \sim \mathcal{P}(w_i s_{jl} \lambda_{jk})$. Because the sum of independent Poisson random variables is also Poisson, we have $y_{i\cdot\cdot}|k \sim \mathcal{P}(w_i \sum_j \sum_l s_{jl}\lambda_{jk})$. As such, given the sum $w_i = y_{i\cdot\cdot}$, we may simulate $\mathbf{y}_i|k$ from a multinomial distribution, with event probabilities $\mathbf{p}_k = (p_{jlk}) = (s_{jl}\lambda_{jk})$:

$$\mathbf{y}_i \sim \mathcal{M}(y_{i\cdot\cdot}, \mathbf{p}_k).$$

Each simulated dataset thus has a total of 2000 rows and either 8 (Settings 1-4) or 6 (Settings 5-6) columns.

Table 1: Simulation settings for six different scenarios ($g = 4$ clusters), including the number of conditions (d), the number of replicates per condition ($\mathbf{r} = (r_1, \ldots, r_d)$), variation in library size, cluster separation, and mean expression level w_i. See Table 2 for the specific values used for library size settings A and B.

Setting	d	$\mathbf{r}$	Library size	Cluster separation	Mean (α)
1	3	(1, 4, 3)	Equal	High	1640
2	3	(1, 4, 3)	Equal	Low	1640
3	3	(1, 4, 3)	A	High	1640
4	3	(1, 4, 3)	A	Low	1640
5	2	(4, 2)	B	High	1521
6	2	(4, 2)	B	Low	1521

Table 2: Library sizes (in percent) for the A and B library size simulation settings described in Table 1.

Setting	Condition	Library sizes (per replicate)
A	1	11.3
	2	(15.6, 7.1, 24.8, 16.5)
	3	(1.4, 2.8, 20.6)
B	1	(9.6, 8.4, 25.3, 20.5)
	2	(22.4, 13.8)

Table 3: Values for τ_{jk} for each of $g = 4$ clusters, by library size variation (equal, A, and B), and cluster separation (low and high).

Library size	Separation	$k = 1$	$k = 2$	$k = 3$	$k = 4$
		$\pi_1 = 0.1$	$\pi_2 = 0.2$	$\pi_3 = 0.3$	$\pi_4 = 0.4$
Equal and A	Low	(1, 3, 5)	(2, 4, 4)	(1, 5, 4)	(2, 5, 3)
	High	(1, 3, 5)	(5, 1, 3)	(3, 5, 1)	(5, 3, 1)
B	Low	(1, 3)	(2, 4)	(1, 5)	(2, 5)
	High	(1, 3)	(5, 1)	(3, 5)	(5, 3)

Model performance is assessed using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), the estimated number of clusters $\hat{g}$, and the misclassification error rate. An average oracle ARI is also calculated for comparison, based on the assignment of observations to components maximizing the conditional probability using the true values for the model parameters. For the PMM-I and PMM-II models, model selection is performed using the BIC and ICL criteria and parameter estimation is done using both the EM and CEM algorithms.

3.2.1 Library size normalization

A natural estimator for the normalized library size s_{jl} that is most often used is simply the column sums divided by the total sum, $y_{\cdot jl}/y_{\cdots}$ (referred to as the Total Count (TC) normalization). However, several authors have demonstrated that this choice can be biased in some cases when a small number of genes are very highly expressed (Robinson and Oshlack, 2010; Bullard *et al.*, 2010). In particular, the number of tags expected to map to a particular gene depends on the overall composition of the RNA population being sampled, in addition to the expression level and length of the gene. As such, in addition to the TC normalization, we also study whether alternative normalization approaches, including the quantile (Q) normalization of Bullard *et al.* (2010) and median ratio normalization (MedRatio) of Anders and Huber (2010), affect performance for the proposed clustering methods. In our simulations, no difference in model performance (ARI values and estimated number of clusters) was observed for the TC, Q, and MedRatio estimates of library size (data not shown), and in the remaining simulation results we discuss only the TC estimator for the PMM-II model.

3.3 Model comparisons

Table 4: Misclassification error rates (in %) using the PMM-I and PMM-II models (using the EM algorithm with the ICL model selection criterion) and the PoisC and TransChisq methods (with the ICL criterion) of Cai *et al.* (2004) and Kim *et al.* (2007), for each simulation setting.

	Simulation setting					
	1	2	3	4	5	6
PMM-I	**0.55**	2.89	5.39	19.95	1.65	14.52
PMM-II	**0.55**	**2.52**	**0.49**	**2.56**	1.34	**11.18**
PoisC	0.59	2.89	0.63	3.00	**1.22**	13.76
TransChisq	**0.55**	2.76	0.57	2.88	1.34	15.31

In comparing the PoisC and TransChisq methods to the proposed PMM-I and PMM-II on these simulated data, we aim to understand under what circumstances each method performs well, the appropriate criterion to be used for model selection, and the effect of varying library sizes on clustering performance for HTS data. Due to computational constraints, results for the PoisC and TransChisq methods are based on a subset of 10 of the simulated datasets for each setting.

We first consider a comparison of the PMM-I and PMM-II models across simulation settings. Perhaps unsurprisingly, in settings with equal library sizes

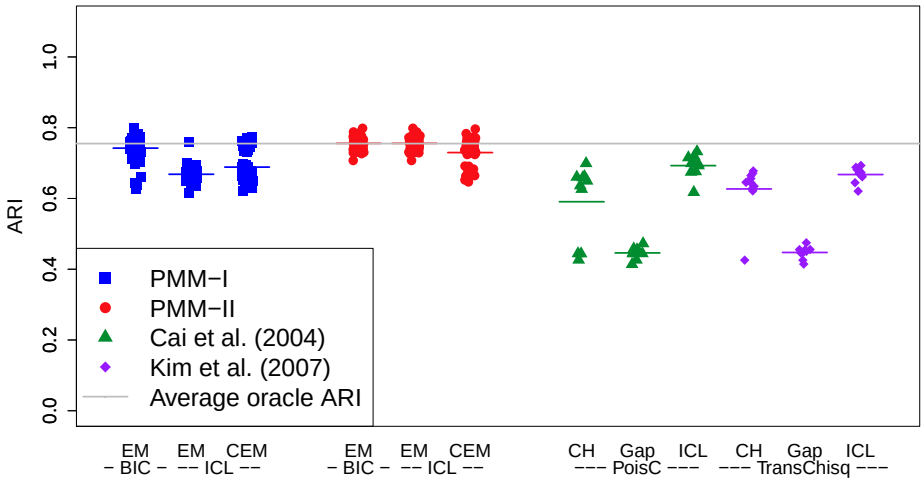


Figure 1: Adjusted rand index values for the PMM-I, PMM-II, PoisC, and TransChisq methods for data from simulation setting 6 (unequal library sizes B, with low cluster separation). The PMM-I and PMM-II results are presented for both the EM and CEM algorithms, with BIC and ICL model selection criteria. The PoisC and TransChisq methods are presented for the CH-index, the Gap statistic, and the ICL model selection criteria. Each point represents one simulated dataset, and the grey line indicates the average oracle ARI calculated across all fifty datasets.

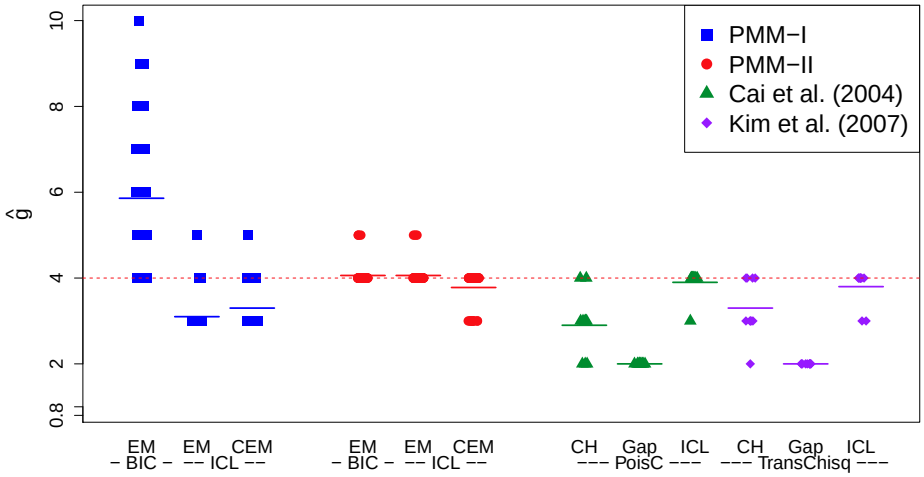


Figure 2: Estimated number of clusters ($\hat{g}$) for the PMM-I, PMM-II, PoisC, and TransChisq methods for data from simulation setting 6 (as above). The PMM-I and PMM-II results are presented for both the EM and CEM algorithms, with BIC and ICL model selection criteria. The PoisC and TransChisq methods results are presented for the CH-index, the Gap statistic, and the ICL model selection criterion. Each point represents one simulated dataset, and the red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

(Settings 1 and 2 in Table 1) the PMM-I and PMM-II display very similar performance (Table 4), whether clusters are simulated to be lowly or highly separated (see Appendix A for more details). In settings with unequal library sizes (Settings 3-6 in Table 1), greater differences between the PMM-I and PMM-II become apparent. With library sizes A (Settings 3 and 4), the PMM-II tends to have higher ARI values than the PMM-I, with the difference becoming more pronounced when clusters are poorly separated (see Figure 1 and Appendix A.1). In addition, the PMM-II consistently chooses the appropriate number of clusters (Figure 2), whereas the PMM-I tends to inflate this number (particularly for the BIC criterion). The advantage of the PMM-II method is further illustrated when comparing the misclassification error rates of all the models (Table 4), which are consistently equal to or smaller than those of the PMM-I. Of course, because the data in these settings were simulated based on the same model as the PMM-II, this advantage is to be expected.

For the PoisC and TransChisq methods, model selection across all settings appears to be highly dependent on the choice of model selection criterion. Regardless of library size variation or cluster separation, the Gap statistic recommended by Kim *et al.* (2007) consistently underestimates the number of clusters g, leading to much lower ARI values. Similarly, the CH index is unable to consistently identify g; as a result, the ARI of the PoisC and TransChisq tend to be lower than those of the PMM-I and PMM-II. On the other hand, the ICL criterion for the PoisC and TransChisq succeeds in correctly identifying the number of clusters g across all settings, and generally yields ARI values nearly as high as the PMM-II. Finally, there do not appear to be any significant differences between the performances of the PoisC and TransChisq methods.

We also consider the differences in using the BIC or ICL for model selection with the two PMM parameterizations, as well as the use of the EM and CEM algorithms. As before, with equal library sizes (Settings 1 and 2), no difference is observed between the two model selection criteria and between the two estimation algorithms. However, once systematic library size differences are introduced (Settings 3-6), we note that the BIC often leads to an overestimate of the number of clusters g within the PMM-I. Unsurprisingly, as the simulated data match the model of the PMM-II, the BIC criterion works well in all cases for this model.

These results suggest that in cases with small differences in library sizes (as in Settings 1 and 2) or with highly separated clusters (as in Settings 1, 3, and 5), the two proposed parameterizations of the PMM and the PoisC and TransChisq methods (with the ICL criterion) perform similarly. However, when large differences in library sizes are present with lowly separated clusters (as in Settings 4 and 6) the PMM-II model outperforms the other models for the performance criteria considered. Model selection for the PoisC and TransChisq methods clearly depends on the choice of criterion, and these simulations indicate that the CH-index and the Gap statistic should be avoided in this context. Finally, we also note that the PoisC and TransChisq methods do not take into account the presence of replicates in the data; rather than the condition-specific parameter of the PMM-I and PMM-II (λ_{jk}), these methods make use of a q-dimensional variable-specific parameter.

4 Data analysis

In this section, we apply the PMM-II model to two real high-throughput sequencing datasets: the Tag-seq experiment of Engström *et al.* (2010), and the yeast RNA-seq experiment of Nagalakshmi *et al.* (2008), both included as a Supplementary File in Anders and Huber (2010). We further describe each dataset in Sections 4.1 and 4.2. We note that both of these datasets exhibit high variability in the library sizes of each variable, suggesting the need for the parameterization of the PMM-II model. For both datasets, we ran the PMM-II model using the quantile estimator for s_{jl}, the Small-EM initialization strategy, and the EM algorithm for parameter estimation. We were unable to use the CEM algorithm for parameter estimation, as it often resulted in one or more empty clusters. This is a well-known issue for the CEM algorithm as the sample size tends to infinity, and one potential solution is to set the cluster proportions in the mixture to be equal (Bryant, 1991).

4.1 Tag-seq data

Tag sequencing (Tag-seq) data were obtained from two separate types of human tissue cultures: four replicates derived from glioblastoma-derived neural stem cells and two derived from non-cancerous neural stem cells (Anders and Huber, 2010). The number of reads per gene were summarized in a table that we retrieved as a Supplementary File from Anders and Huber (2010). After removing genes with a value of 0 for all variables, the data consist of counts for 18,752 genes. The counts range in size from 0 to 79,123 and the normalized library sizes of the six samples vary from 8.4% to 25.3%.

An advantage to using the framework of a model-based approach, such as the PMM-II, is the possibility to not only perform model selection, but also to determine whether a particular model is well-adapted to the data. Consider Figure 3, which indicates that as the number of clusters is increased (here, from $g = 2$ to 50), the BIC continues to increase as well. This is perhaps unsurprising, as the data are unlikely to truly follow a mixture of Poisson distributions. However, we note that the ICL criterion peaks for $\hat{g} = 12$, suggesting that this is the number of partitions with the greatest evidence in these data. In addition, a model-based framework allows for a consideration of the conditional probabilities of belonging to each cluster (see Appendix B.1). In these data, based on the choice of $\hat{g} = 12$ clusters, we find that about 38.5% of observations (7217) have a maximum estimated conditional probability greater than 90% (see Appendix B.1).

Finally, in Figure 4, we present a visualization of the estimated values of the per-cluster effects for the different types of human tissue in these data, $\hat{\boldsymbol{\lambda}}$, as well as the proportion of observations in each cluster, $\hat{\boldsymbol{\pi}}$. We can thus interpret the behavior of each cluster; for example, cluster 1 appears to be made up of a large proportion of observations (22.1% of the total) with expression largely balanced between the two types of tissue, while cluster 4 is made up of a small proportion of observations (2.6%) that tend to be more highly expressed in the non-cancerous cells than in the glioblastoma-derived cells.

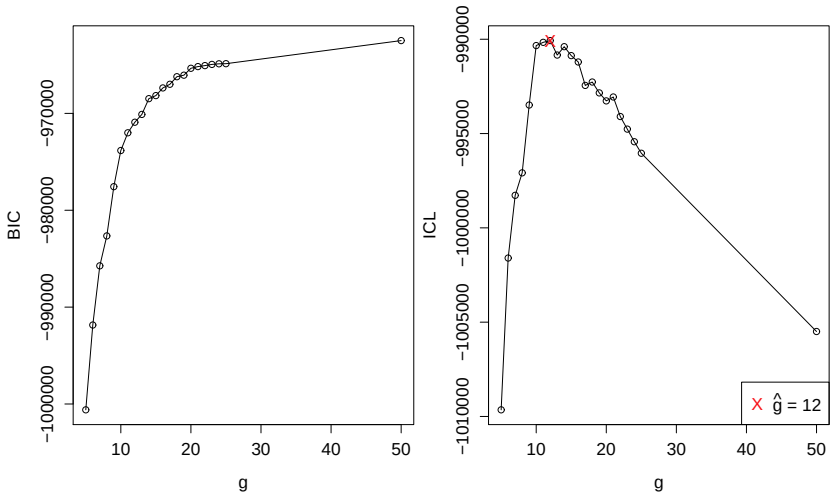


Figure 3: BIC (left) and ICL (right) values ($g = 2, \ldots, 50$) for the PMM-II parameterization using quantile normalization, the Small-EM initialization procedure, and the EM algorithm for estimation for the Tag-seq data. The displayed values for BIC and ICL are the maximum values calculated over 10 iterations at each g. The red X on the rightmost graph indicates the number of clusters selected by ICL: $\hat{g} = 12$.

4.2 Yeast RNA-seq data

Nagalakshmi *et al.* (2008) obtained RNA-seq data on replicates of *Saccharomyces cerevisiae* cultures, based on two different library preparation protocols (dT and RH), with three sequencing runs for each protocol (two technical replicates and one additional biological replicate); see Anders and Huber (2010) for additional details. After removing genes with a value of 0 for all sequencing runs, the data consist of counts for 6874 genes, ranging in size from 0 to 275,781. The normalized library sizes of the six sequencing runs vary from 10.4% to 23.1%.

We ran the PMM-II as previously described. As with the previous dataset, the BIC continuously increases as the number of clusters increases ($g = 2, \ldots, 30$), but the ICL peaks at $\hat{g} = 8$ clusters (see Figure 5). However, we note that there appears to be a plateau of ICL values for $g \in \{6, 7, 8, 9\}$. To determine what differences are present among these four models, we calculate the ARI between each pair, which indicates that the models $g = 6$ and $g = 9$ clusters yield divergent results from the initially selected model $g = 8$ (ARI values of 0.43 and 0.46, respectively). However, between models $g = 7$ and $g = 8$, there is nearly perfect accord between the label assignments (ARI value of 1.00); in fact, in Table 5 we see that the differences among these two models are largely due to the fact that cluster 7 in model $g = 7$ is split into two clusters in model $g = 8$. As such, the two models give very similar results, with the primary difference being that the latter is able to refine the results of the former (by splitting one of the clusters in two). Combined with the fact that this model has the maximum ICL value, we conclude that the appropriate number of clusters is in fact $\hat{g} = 8$.

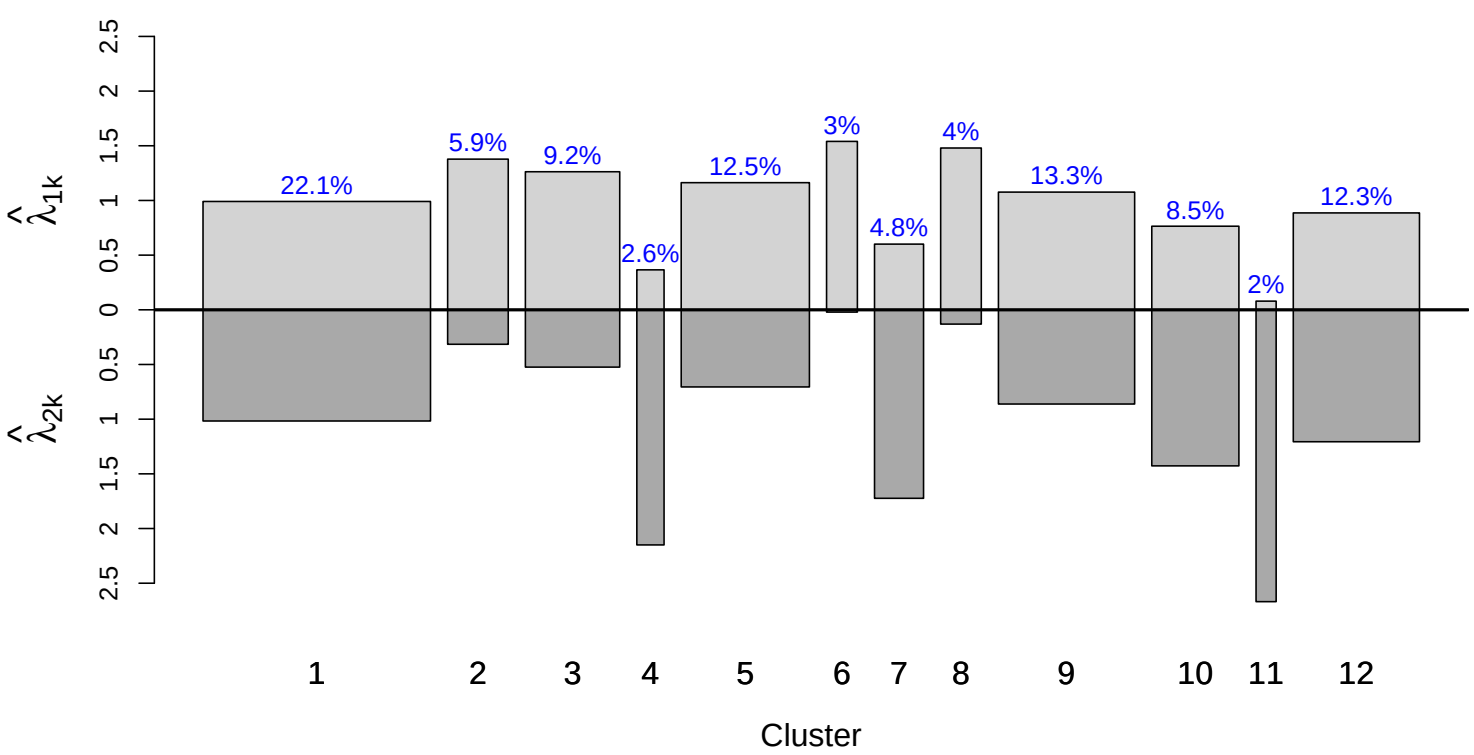


Figure 4: Visualization of values for $\hat{\boldsymbol{\lambda}}_k$ and $\hat{\pi}_k$ for $k = 1, \ldots, 12$ in the Tag-seq cluster analysis. Each bar represents one of the 12 clusters identified by the PMM-II model, where values above the horizontal axis (in light grey) represent the value of $\hat{\lambda}_{1k}$, and values below the horizontal axis (in dark grey) represent the value of $\hat{\lambda}_{2k}$ (the cluster-specific parameters for the glioblastoma-derived and non-cancerous neural stem cells, respectively). The width of each bar, and the percentages in blue above each, represent the corresponding value of $\hat{\pi}_k$.

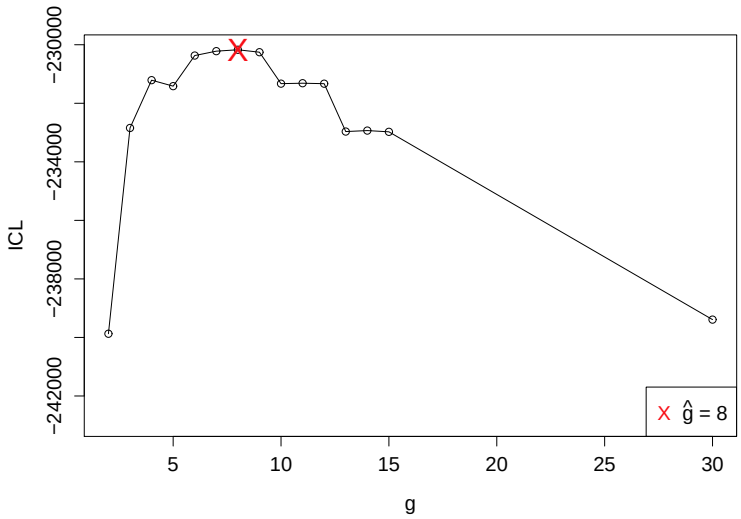


Figure 5: ICL values ($g = 2, \ldots, 50$) for the PMM-II parameterization using quantile normalization, the Small-EM initialization procedure, and the EM algorithm for estimation for the yeast RNA-seq data. The displayed values for ICL are the maximum values calculated over 10 iterations at each g. The red X indicates the value for g associated with the maximum ICL, $\hat{g} = 8$.

As in Section 4.1, we examine the conditional probabilities of belonging to each cluster. In these data, we find that about 37.1% of observations (2553) have a maximum estimated conditional probability greater than 90% (see Appendix B.2). However, unlike the previous dataset, we observe an accumulation of maximum conditional probability estimators with a value near 0.50; this indicates that for these observations, there is a great deal of uncertainty in assigning cluster membership. In considering the per-cluster maximum conditional probability estimators (Figure 6), it appears that most of these observations are assigned to cluster 3; as such, this particular cluster should be interpreted with care.

Finally, in Figure 7 we examine a visual representation for the per-cluster estimates for the different protocols in these data ($\hat{\boldsymbol{\lambda}}$) and the proportion of observations in each cluster ($\hat{\boldsymbol{\pi}}$). For example, cluster 5 appears to be a large block with largely balanced expression between the two protocols. On the other hand, cluster 4 (made up of 0.7% of the observations) appears to be more highly expressed in the dT protocol than the RH protocol, while cluster 2 (0.3% of the observations) displays the opposite behavior.

5 Discussion

In this work, we have proposed two potential parameterizations of a Poisson mixture model for clustering count-based HTS data profiles. As shown in the simulation study, the PMM-II model is clearly better able to model data with several replicates per condition (as in the two datasets considered in Section 4) than the PMM-I. This indicates the importance of modeling not only the condition-specific cluster effect λ_{jk} and gene expression level w_i, but also

Table 5: Correspondence between the models with $g = 7$ and $g = 8$ clusters for the yeast RNA-seq data, in terms of the number of genes identified in each cluster. For example, among the 256 observations in cluster 1 for the model with $g = 7$, 241, 3, and 12 fall into clusters 1, 5, and 7, respectively, in the model with $g = 8$.

	1	2	3	4	5	6	7	8
1	241				3		12	
7	2	19					28	
5			3926					
2				38				
3					748			
6						1554		
4								303

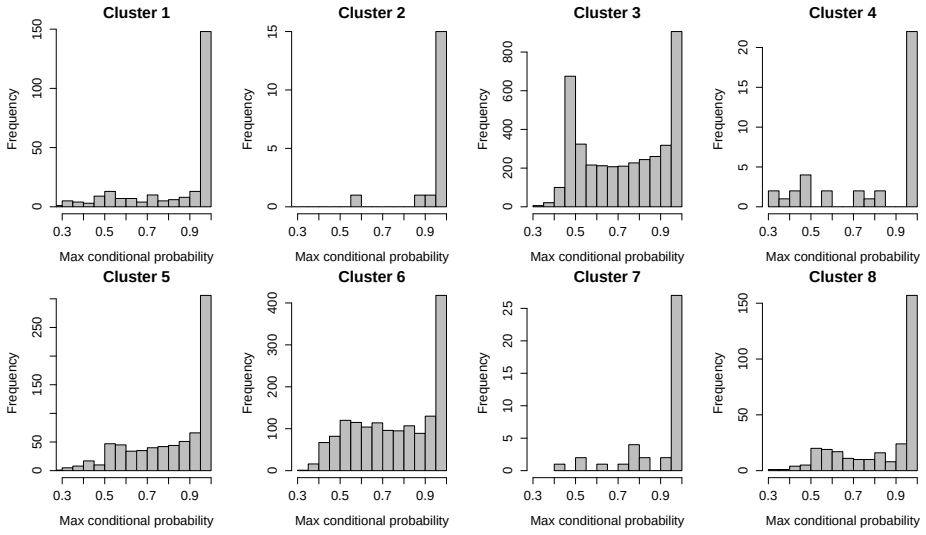


Figure 6: Per-cluster histogram of the maximum conditional probability estimators for the yeast RNA-seq data, using the PMM-II as described in the text.

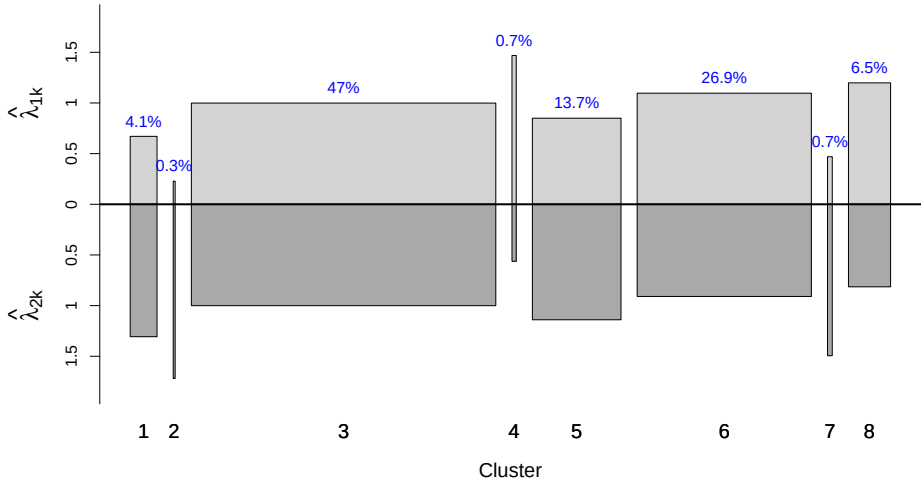


Figure 7: Visualization of values for $\hat{\boldsymbol{\lambda}}_k$ and $\hat{\pi}_k$ for $k = 1, \ldots, 8$ for the yeast RNA-seq data. Each bar represents one of the eight clusters identified by the PMM-II model, where values above the horizontal axis (in light grey) represent the value of $\hat{\lambda}_{1k}$, and values below the horizontal axis (in dark grey) represent the value of $\hat{\lambda}_{2k}$ (the cluster-specific parameters for the dT and RH protocols, respectively). The width of each bar, and the percentages in blue above each, represent the corresponding value of $\hat{\pi}_k$.

the variability among replicates (through the library size parameter s_{jl}). The advantage of considering a model-based framework for a co-expression analysis is further confirmed by the success of the ICL model selection criterion for the PoisC and TransChisq methods in the simulation study. We expect that as HTS experiments continue to increase in number and size (i.e., more than two conditions) and focus shifts from analyses of differential expression, the proposed Poisson mixture model will be a valuable tool for studying co-expression of HTS observations.

In addition to studying the performance of several methods on clustering simulated data, we also compared the EM and CEM algorithms for parameter estimation, as well as the BIC and ICL criteria for model selection, within the PMM-I and PMM-II models. Although little difference was observed between the EM and CEM algorithms for simulated data, the CEM algorithm often yielded one or more empty clusters for real data. As previously noted, this is a well-known issue for the CEM algorithm as the sample size n increases (Bryant, 1991). This limitation can be remedied by setting the cluster proportions to be equal, but such a constraint is seemingly unrealistic for clustering count-based HTS data profiles. For the BIC and ICL criteria, when the model matches that used to simulate data (e.g., the PMM-I for simulation settings 1 and 2 and the PMM-II for all settings), little difference is observed between the two. However, when the model does not match that of the data (e.g., the PMM-I for simulation settings 3-6 and both models for the real data analyses), the BIC tends to overestimate the number of clusters, while the ICL tends to select the number of partitions with the greatest evidence in the data. Thus, it may be preferable to use the ICL rather than the BIC when the goal is to reveal

interesting clustering structures in count-based HTS data profiles.

As previously mentioned, the PMM-I and PMM-II models share some similarities with the approaches proposed by Cai *et al.* (2004); Kim *et al.* (2007) and Witten (2011), although there are several key differences. First, clustering in both the PoisC and TransChisq models uses a K-means algorithm based on a χ^2-statistic rather than a finite mixture model. In addition, both of these models provide variable-specific (and not condition-specific) clustering parameters, thus ignoring replicates within the data. On the other hand, Witten (2011) makes use of a hierarchical clustering procedure based on a dissimilarity matrix calculated by fitting Poisson loglinear models separately for each pair of observations, rather than fitting a global model as proposed here. Rather than imposing interpretable parameter constraints as we have done for the PMM-I and PMM-II, Witten (2011) also makes use of a two-step estimation procedure. Although the model in Witten (2011) can be reconfigured to cluster genes rather than variables, it is computationally demanding to calculate and use a distance matrix for a matrix of dimension $(n \times n)$.

Finally, we note that in analyses of differential expression, several authors have focused on the use of Negative Binomial (NB) rather than Poisson models, due to the observed presence of overdispersion among replicates for a fixed gene in a fixed condition (Anders and Huber, 2010; Robinson and Smyth, 2007). While the latter assumes that the mean and variance of a fixed observation in a fixed sample are equal, the former includes an extra parameter in the model as a variance inflation parameter. At first glance, it may thus seem natural to apply a finite mixture of NB distributions to the task of clustering HTS observations. In fact, Si *et al.* (2011) recently attempted to do just this, where the per-gene overdispersion parameter ϕ_i is estimated from the data using a quasi-likelihood approach and treated as fixed in the mixture (rather than being estimated at each iteration of the EM or CEM algorithms).

It is also important to consider whether they are well-adapted to the context of a co-expression analysis, in which the goal is to obtain clusters of observations with similar expression profiles across variables. Namely, under the assumption of conditional independence of the variables given the component, a NB mixture model assumes overdispersion on the *coordinates* (all observations in a fixed variable), which does not equate with the overdispersion on the *observations* (all variables in a fixed observation), as is done in differential analyses. In fact, the clusters identified by the PMM-II model in real data do not appear to display a larger variability than that expected under the theoretical Poisson model (see Figures 19 and 21 in Appendix B); instead, it seems that the Poisson mixture model accurately represents the data, with the exception of low count proportions $(y_{ijl}/y_{...})$. A more useful direction for future research may be to define a mixture allowing for overdispersion and underdispersion among weakly and highly expressed observations, respectively.

Acknowledgements

We thank the members of the Statistics for Systems Biology (SSB) working group for their helpful and insightful comments. AR was funded as a postdoctoral researcher at Inria Saclay - Île-de-France for the duration of this work.

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(R106), 1–28.

Auer, P. L. and Doerge, R. W. (2010). Statistical design and analysis of RNA-Seq data. *Genetics*, **185**, 1–12.

Auer, P. L. and Doerge, R. W. (2011). A two-stage Poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology*, **10**(26), 1–26.

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., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, **41**(1), 561–575.

Bryant, P. G. (1991). Large-Sample Results for Optimization-Based Clustering Methods. *Journal of Classification*, **8**, 31–44.

Bullard, J. H., Purdom, E. A., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC B*, **11**(94).

Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C., and Wong, W. H. (2004). Clustering analysis of SAGE data using a Poisson approach. *Genome Biology*, **5**, R51.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, **3**(1), 1–27.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, **14**(3), 315–332.

Dempster, A. P., 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)*, **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. *PNAS*, **95**(25), 14863–14868.

Engström, P., Tommei, D., Stricker, S., Smith, A., Pollard, S., and Bertone, P. (2010). Transcriptional characterization of glioblastoma stem cell lines using tag sequencing.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

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.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, **30**(1), 63–77.

Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.-B., Feldman, L. J., and Huang, H. (2007). Measuring similarities between gene expression profiles through new data transformations. *BMC Bioinformatics*, **8**(29).

Łabaj, P. P., Leparc, G. G., Linggi, B. E., Markillie, L. M., Wiley, H. S., and Kreil, D. P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**(ISMB), i383–i391.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, number 1, pages 281–297. Berkely, University of California Press.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

McLachlan, G., Do, K.-A., and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Wiley-Interscience.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5**(7), 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, **4**(14).

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(R25).

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**(21), 2881–2887.

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

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., Muehlbauer, G. J., Nelson, R. T., Grant, D., Specht, J. E., Graham, M. A., Cannon, S. B., May, G. D., Vance, C. P., and Shoemaker, R. C. (2010). RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biology*, **10**(160).

Si, Y., Liu, P., Li, P., and Brutnell, T. (2011). Model-based clustering for RNA-seq data. Technical Report 11, Iowa State University.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, **63**(2), 411–423.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5), 511–518.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236–244.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics*, **In press**.

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), 977–987.

A Additional simulation results

In this section, we give additional results from the simulation study presented in Section 3. In the following, we present stripcharts of the Adjusted rand index (ARI) values and choice of g for the PMM-I, PMM-II, PoisC, and TransChisq methods for data from simulation setting 1, 2, 3, 4, and 5 (see Table 1 for a full description of each of these simulation settings). Note that the results for setting 6 are displayed in the text in Figures 1 and 2. The PMM-I and PMM-II results are presented for both the EM and CEM algorithms, with BIC and ICL model selection criteria. The PoisC and TransChisq methods are presented for the CH-index, Gap statistic, and ICL model selection criteria. Each point represents one simulated dataset.

A.1 ARI results

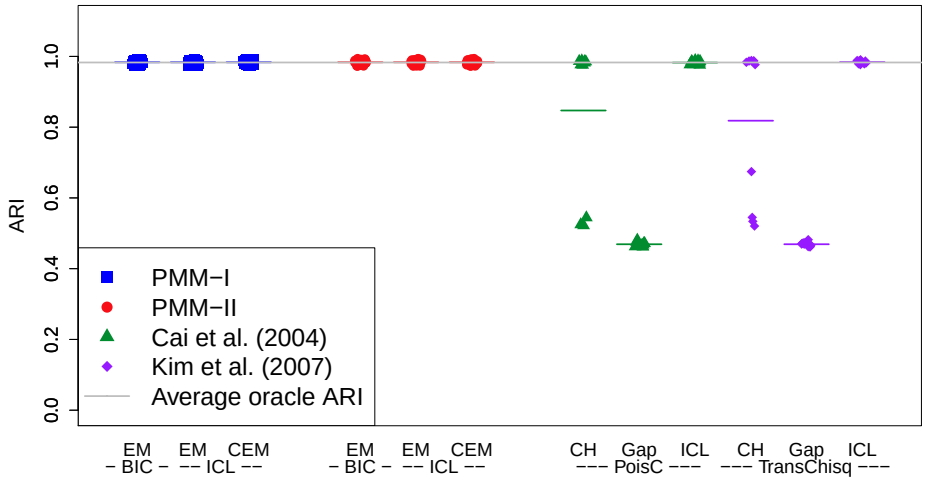


Figure 8: ARI values, by model, for Setting 1 (equal library sizes and high separation). The grey line indicates the average oracle ARI calculated across all fifty datasets.

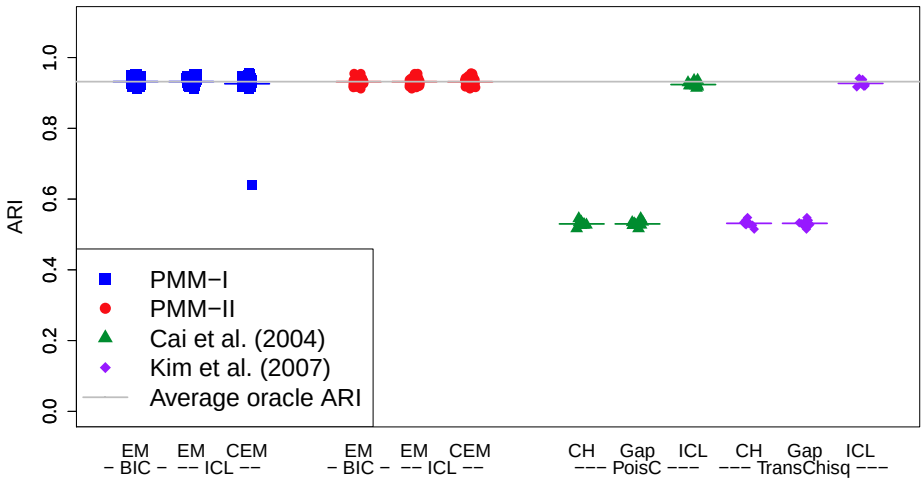


Figure 9: ARI values, by model, for Setting 2 (equal library sizes and low separation). The grey line indicates the average oracle ARI calculated across all fifty datasets.

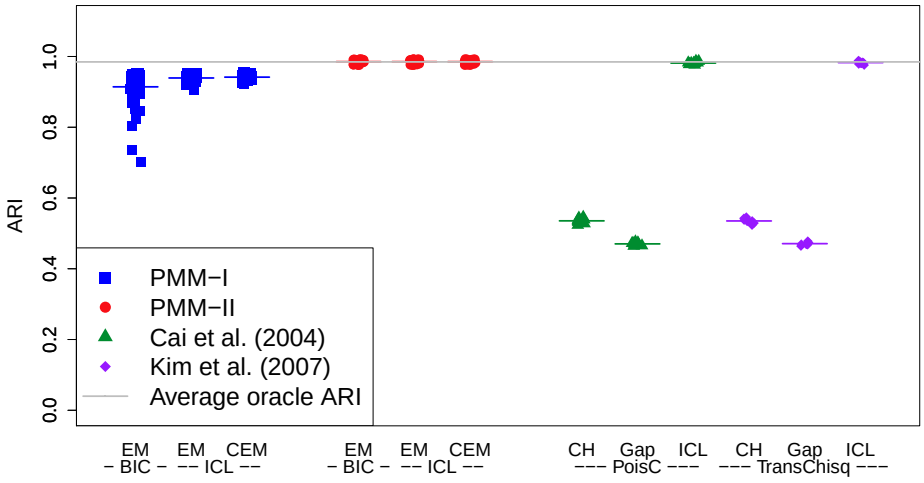


Figure 10: ARI values, by model, for Setting 3 (library sizes A and high separation). The grey line indicates the average oracle ARI calculated across all fifty datasets.

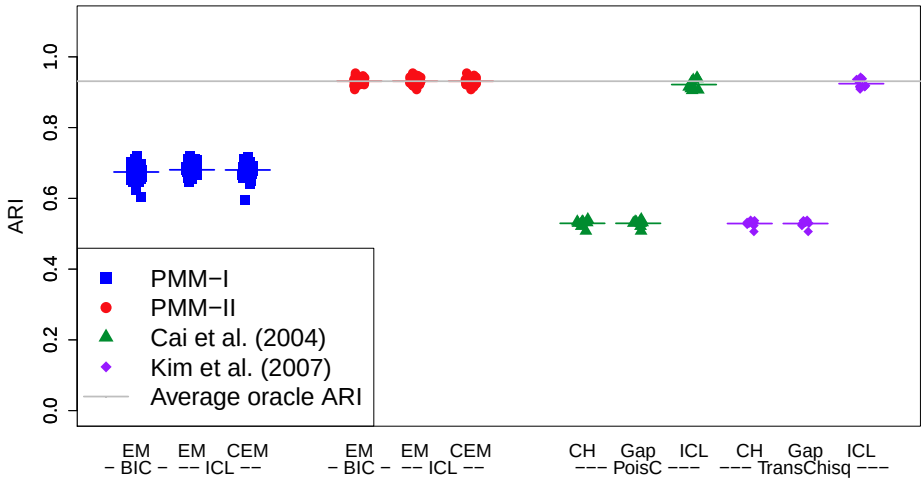


Figure 11: ARI values, by model, for Setting 4 (library sizes A and low separation). The grey line indicates the average oracle ARI calculated across all fifty datasets.

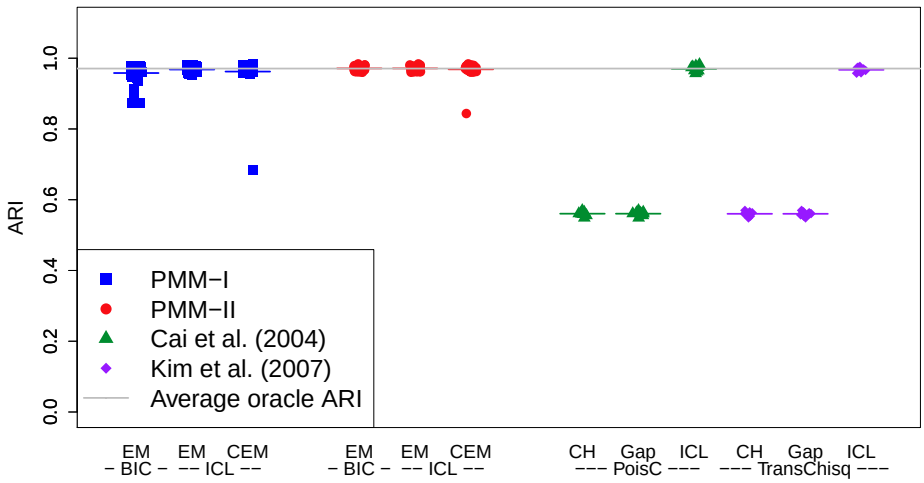


Figure 12: ARI values, by model, for Setting 5 (library sizes B and high separation). The grey line indicates the average oracle ARI calculated across all fifty datasets.

A.2 Choice of g results

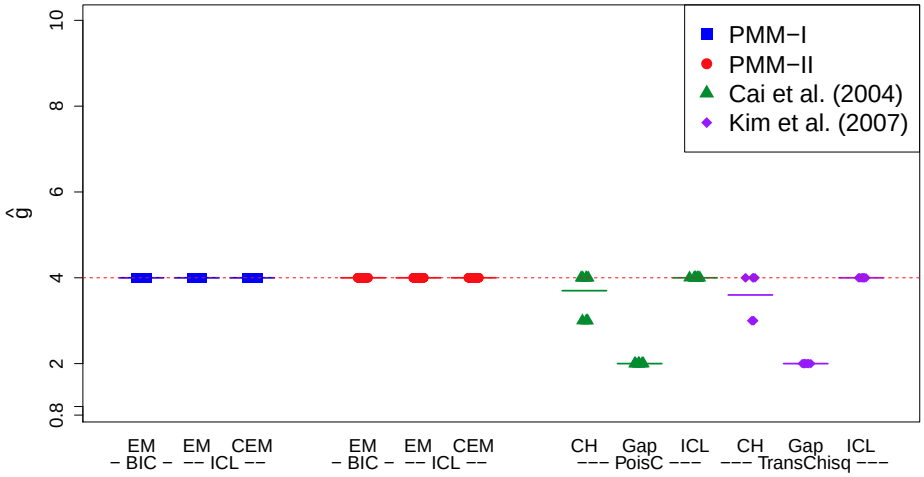


Figure 13: $\hat{g}$ values, by model, for Setting 1 (equal library sizes and high separation). The red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

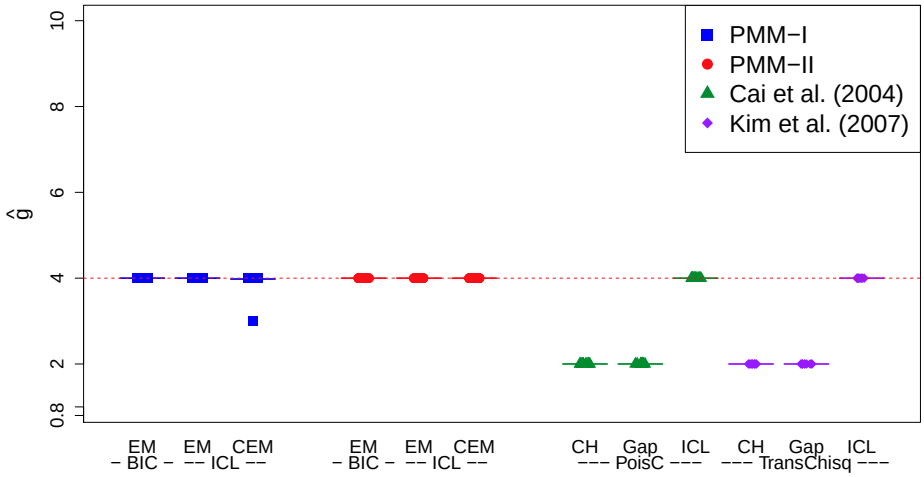


Figure 14: $\hat{g}$ values, by model, for Setting 2 (equal library sizes and low separation). The red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

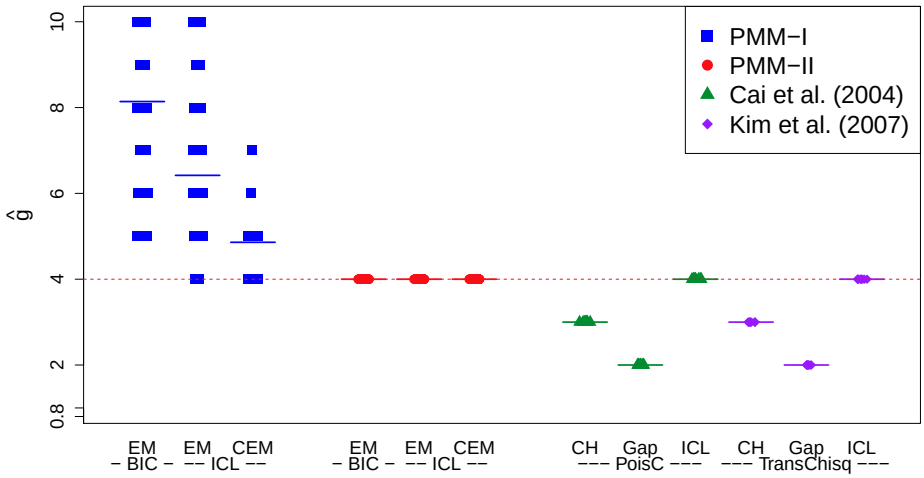


Figure 15: $\hat{g}$ values, by model, for Setting 3 (library sizes A and high separation). The red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

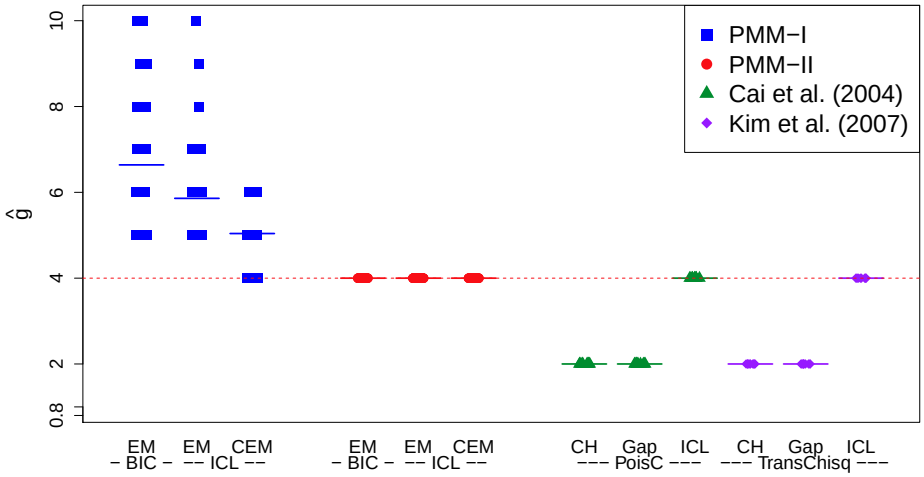


Figure 16: $\hat{g}$ values, by model, for Setting 4 (library sizes A and low separation). The red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

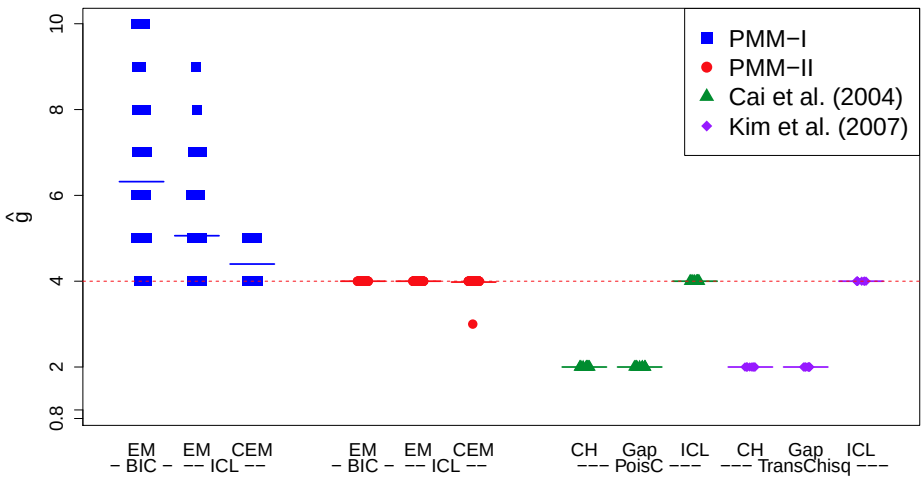


Figure 17: $\hat{g}$ values, by model, for Setting 5 (library sizes B and high separation). The red dotted line indicates the true number of clusters in the simulated data, $g = 4$.

B Additional results from real data analysis

In this section, we provide complementary results and graphics from the analyses of the Tag-seq and yeast RNA-seq data described in Section 4 of the text.

B.1 Tag-seq data analysis

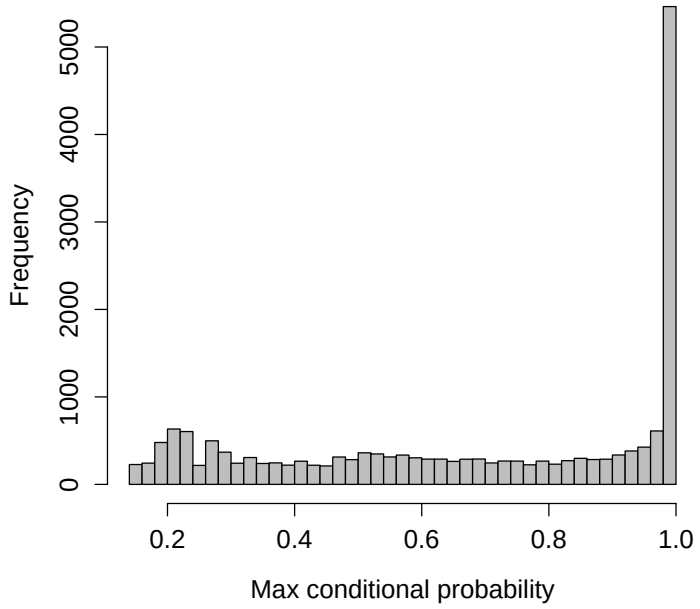


Figure 18: Histogram of the maximum conditional probability estimators for the observations in the Tag-seq data, using the PMM-II model as described in the text. About 38.5% of observations (7217) have an estimated maximum conditional probability greater than 90%.

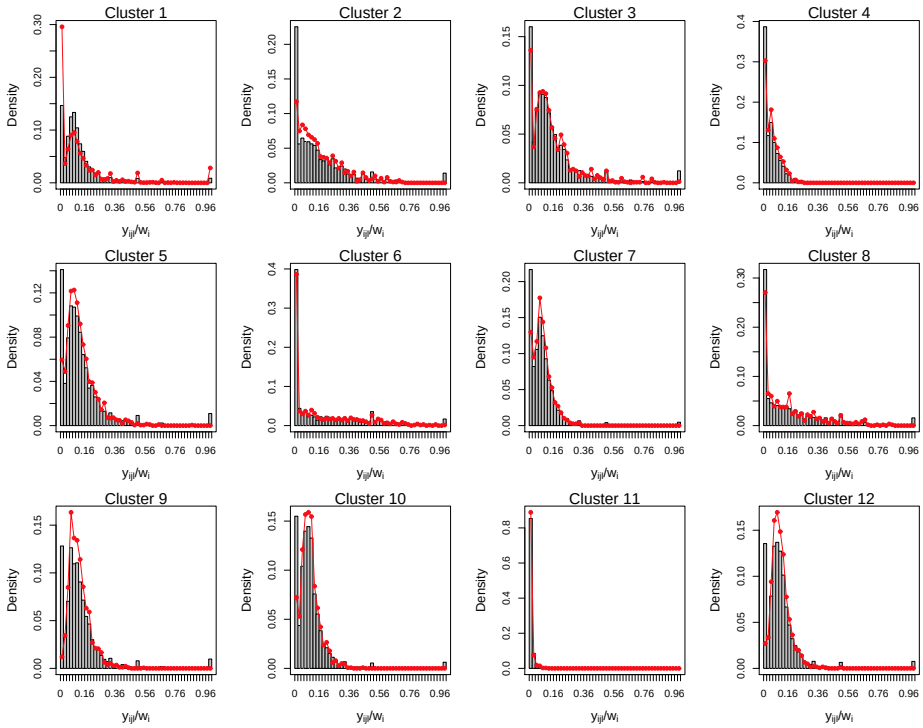


Figure 19: Example of weighted histograms for variable 1 of the Tag-seq data. The grey histogram represents the observations y_{i11} in the first replicate of the first condition (glioblastoma-derived cells), weighted by their a posteriori probability of belonging to each of the 12 clusters ($\hat{t}_{ik}$ from Equation 2). The red line represents an empirical reference distribution calculated as follows: for each cluster k, after calculating the density $\mathcal{P}(\hat{w}_i s_{jl} \hat{\lambda}_{jk})$ for all observations with $\hat{t}_{ik} > 0.90$, we calculate the proportion of these densities falling in each bin of the weighted histogram.

B.2 Yeast RNA-seq data analysis

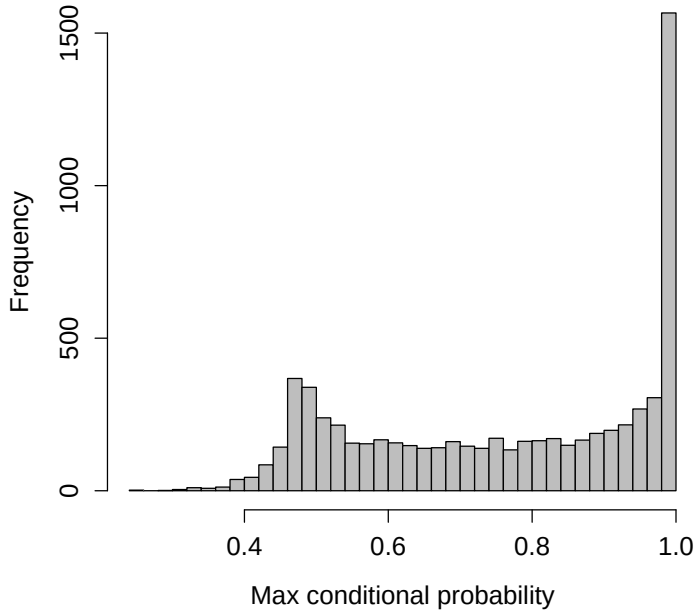


Figure 20: Histogram of the estimated maximum conditional probability estimators for the yeast RNA-seq data, using the PMM-II model as described in the text. About 37.1% of observations (2553) have an estimated maximum conditional probability greater than 90%.

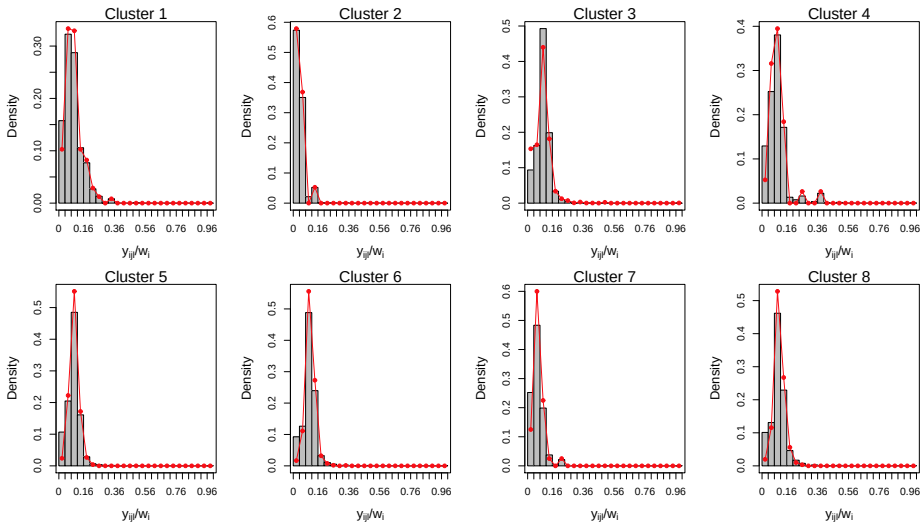
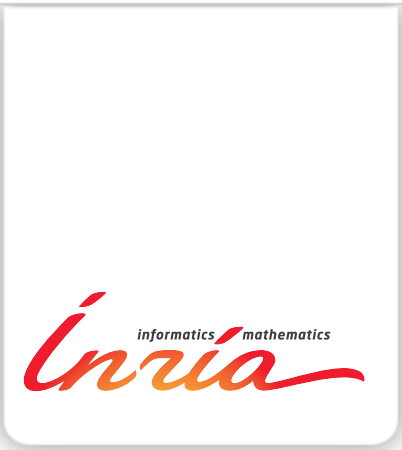


Figure 21: Example of weighted histograms for variable 1 of the Tag-seq data (see Figure 19 for a full description of this graph).



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Parc Orsay Université
4 rue Jacques Monod
91893 Orsay Cedex

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

ISSN 0249-6399