l_1-Penalized model-based clustering for RNA-Seq count data

Ye Tian[a], Ruimin Sun[a], Heng Lian[a,*], Hua Liang[b]

[a] *Division of Mathematical Sciences, Nanyang Technological University, Singapore, 637371*

[b] *Department of Statistics, George Washington University, Washington, D.C. 20052, USA*

[*] *Corresponding author. Email: henglian@ntu.edu.sg*

Abstract

As the next generation RNA sequencing becomes the dominating technology for studying the gene expression profiles, downstream statistical analysis tools are needed urgently. Clustering samples is an important approach to reveal their relationships, such as for the discovery of new subtypes of cancer cells. To cluster high dimensional data, it is also of interest to select the variables (genes) informative for clustering. We proposed a penalized model-based method to select genes and perform clustering simultaneously. The negative binomial mixture model is developed which are suitable for the nonnegative and discrete count data. Moreover, our method can automatically determine the number of clusters using the Bayesian information criterion. Additionally, hybrid-hierarchical tree guided by the output from model-based clustering can be applied to visualize partial clustering structure in a hierarchical way.

Keywords: EM algorithm; Mixture model; Negative Binomial distribution; Penalized likelihood; RNA sequencing.

1 Introduction

The next generation RNA sequencing (RNA-Seq) (Wang *and others* (2009); Metzker (2010)) is currently a widely used technology to measure gene expressions. Briefly, a targeted RNA population is isolated and converted to a library of cDNA fragments. These cDNA fragments are then sequenced by high-throughput DNA sequencing approaches. After this process, millions of short reads (30-400 bp in length) can be obtained. Given the reference genome or transcriptome, these reads are aligned and pooled into regions, such as genes and exons. For convenience of reference, we assume that all such regions correspond to genes in this paper. The number of short reads mapped to each gene is then counted to quantify gene expression. The resulting RNA-Seq count data is nonnegative and discrete in nature.

With such data sets, researchers can use statistical methods to perform various downstream analysis, such as identifying differentially expressed (DE) genes (e.g., Anders and Huber (2010); Robinson *and others* (2010)) and classification or clustering of samples (e.g., Berninger *and others* (2008); Witten (2011)) and genes (e.g., Si *and others* (2014)). Note clustering can be performed on either genes or samples, or both simultaneously. In this article, we focus on the clustering of samples based on the variation of gene expression profiles across different tissues or subtypes of cancer cells. Due to the typically large number of genes compared to number of samples, this problem falls into the "large p small n" paradigm that has attracted a lot of attention recently in biostatistics (Guo *and others* (2010); Ma *and others* (2011); Wang *and others* (2012); Chen *and others* (2013); Hao and Zhang (2014)).

Although a lot of articles contribute to the clustering analysis with the use of the microarray data (e.g., Spellman *and others* (1998), Ramaswamy *and others* (1998), Yeung and Ruzzo (2001) and Pan and Shen (2007)), which was the most popular technology to quantify the gene expression before the advent of RNA-Seq, there are fewer methods proposed for the latter. Berninger *and others* (2008) developed a Bayesian method to compute the dissimilarity matrix in the clustering analysis

of small RNA cloning data from sequencing technology. Anders and Huber (2010) suggested to transform the data for stabilizing the variance and then compute the squared Euclidean distance using these transformed data. Robinson *and others* (2010) provided a clustering function using 500 features with highest variance in their edgeR software package. Witten (2011) modelled the RNA-seq count data with a Poisson log-linear model and computed the dissimilarity matrix by a modified log-likelihood using the power transformed data. These methods mentioned above are distance-based methods. The benefit of such methods is the ability to visualize the clustering results clearly by the related techniques, such as hierarchical clustering or multidimensional scaling. However, they lack a probabilistic interpretation for clustering and provide no statistically sound way of determining the number of clusters.

The model-based clustering (McLachlan and Peel, 2000) is a popular statistical approach to tackle these issues. This method allows soft allocation of samples to clusters. Moreover, with a well-defined likelihood, many criteria such as the Akaike Information Criterion (AIC) (Akaike, 1973), the Bayesian Information Criterion (BIC) (Schwartz, 1978) and Extended Bayesian Information Criterion (EBIC) (Chen and Chen (2008) and Chen and Chen (2012)), have been developed to perform model selection.

Another important issue related to the clustering of RNA-Seq count data is gene selection. Such data have large dimension P (the number of genes) and small sample size n and generally a lot of genes are noises in the sense they are not differentially expressed across different clusters. Accordingly, it is natural to perform gene selection when we conduct clustering. Besides that the set of selected genes can be of interest in itself, this will potentially improve clustering accuracy. To our knowledge, gene selection for RNA-Seq count data is not currently available for clustering, except to heuristically select those genes with larger variances in a preprocessing step (Robinson *and others*, 2010). We will apply the penalized model-based method to perform gene selection and clustering simultaneously. The penalized normal mixture model was

proposed in Pan and Shen (2007); Zhou *and others* (2009) for clustering microarray data. Khalili and Chen (2007) developed the penalized model-based method for mixture of regression models. To determine the number of clusters and the number of important genes, we use the Bayesian information criterion (BIC) as in Pan and Shen (2007). To model the RNA-Seq count data, Poisson model can work quite well if no biological replicates exist (Marioni *and others*, 2008); otherwise, the negative binomial (NB) model (Robinson *and others* (2010); Anders and Huber (2010)) can be applied because biological replicates may give rise to over-dispersion. In this paper, we focus on NB mixture model of which the Poisson mixture model is a special case. We compare our method with three competing methods: PoiClaClu (Witten, 2011), edgeR (Robinson *and others*, 2010) and DESeq (Anders and Huber, 2010) in simulation and four real data sets. Our proposed method can often get better clustering results. The proposed method can be executed in the R package PMixClus available at `https://github.com/TianYe00/PMixClus.git`.

The rest of the article is organized as follows. Section 2 presents the statistical models for clustering of count data and discuss some issues in implementation. Then the model selection criterion and a method to construct a partial hierarchical clustering guided by the ouput of model-based clustering is proposed. Section 3 contains our simulation studies and application of the method to four publicly available data sets. The paper is concluded in Section 4 with a discussion.

2 Methods

2.1 Model

Suppose that the read count data $\boldsymbol{x}$ contains n samples (rows) and P genes (columns). Let $\boldsymbol{x_j} = (x_{j1}, \ldots, x_{jP})$ denote the read counts of P genes in sample j for $j = 1, \ldots, n$. For mathematical simplicity, it is assumed that all genes are independent and $\boldsymbol{x_j}$ follows a finite mixture distribution $\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j}; \boldsymbol{\psi_{jk}})$, where f_k is the discrete

distribution for kth cluster with parameter vector $\boldsymbol{\psi_{jk}}$, and π_k is the mixing proportion satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The log-likelihood of the mixture model is

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log\{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j}; \boldsymbol{\psi_{jk}})\}, \tag{1}$$

where $\boldsymbol{\Theta} = \{(\pi_k, \boldsymbol{\psi_{jk}}) : k = 1, \ldots, K; j = 1, \ldots, n\}$. Denote $\mu_{jkp} = E(x_{jp})$ if sample j belongs to cluster k. To facilitate gene selection, one key development here is to decompose μ_{jkp} as $\mu_{jkp} = s_j \gamma_p \theta_{kp}$ subject to identifiability constraint $\sum_{j=1}^{n} s_j = n$, where s_j is the size factor for sample j (sequencing depth for sample j), γ_p is the average read counts of gene p over all samples and θ_{kp} represents cluster-specific effect for gene p. If $\theta_{kp} = 1$ for all k, it means that gene p is not differentially expressed across clusters and should be treated as noise variables. This observation suggests that we can perform gene selection by shrinking many of the θ_{kp} towards 1. Thus we propose to use the following LASSO penalty (Tibshirani, 1996):

$$P_\lambda(\boldsymbol{\Theta}) = \lambda \sum_{k=1}^{K} \sum_{p=1}^{P} |\log \theta_{kp}| \ ,$$

where $\lambda > 0$ is a tuning parameter. Accordingly, the penalized log-likelihood function for the mixture model is defined as,

$$\log L_P(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}) - P_\lambda(\boldsymbol{\Theta}) \ .$$

We now introduce the method based on the NB distribution. We assume $s_{jp}|j \in C_k \sim \text{NB}(\mu_{jkp}, \phi_p)$, where NB(.,.) is the negative binomial distribution with mean $\mu_{jkp} = s_j \gamma_p \theta_{kp}$ and variance $\mu_{jkp} + \phi_p \mu_{jkp}^2$, C_k contains the samples for cluster j, and ϕ_p is the gene-specific dispersion parameter. When $\phi_p = 0$, this reduces to the Poisson model. Thus the parameter set $\boldsymbol{\psi_{jk}}$ in the model (1) is $\boldsymbol{\psi_{jk}} = \{(s_j, \gamma_p, \theta_{kp}, \phi_p) : p = 1, \ldots, P\}$. The EM algorithm is applied to find the optimizer of the penalized

likelihood and the details are presented in the Appendix.

2.2 Initialization strategy

Many articles have already demonstrated the importance of the initialization strategy when the mixture model is fitted by EM algorithm (Seidel *and others*, 2000; McLachlan and Peel, 2000). For simplicity, we estimated the size factor s_j by median ratio method (Anders and Huber, 2010) and the value is fixed when running the EM algorithm. We use the K-means method to get initial class labels and the starting values of θ_{kp} and γ_p are obtained by simple moment estimator. For the initial values of ϕ_p. Lu *and others* (2005) proposed a dispersion estimator for the over-dispersed log-linear model by applying the goodness-of-fit statistic in the analysis of SAGE data. Li *and others* (2012) applied a similar idea to estimate the transformation for data exhibiting over-dispersion using the Poisson goodness-of-fit statistic. Following this idea, we use the NB goodness-of-fit statistic to obtain the starting values of ϕ_p. We divide genes into M groups according to the mean counts of genes and then estimate the dispersion parameters for each group of genes. In the mth group, the goodness-of-fit statistics is

$$\text{GOF}_{mp} = \sum_j \frac{(x_{jp} - \hat{\mu}_{k(j)p})^2}{\hat{\mu}_{k(j)p}(1 + \phi_m \hat{\mu}_{k(j)p})} \ ,$$

where $k(j)$ denotes the cluster identity for sample j. Since x_{jp}'s are independently NB distributed, the approximate distribution of GOF_{mp} is χ^2 with $(n-1)(P/M-1)$ degrees of freedom. In order to get rid of the outliers, we set $S_m = \{p : \text{GOF}_{mp} \text{ in } (\epsilon, 1-\epsilon) \text{ quantile of all } \text{GOF}_{mp}\}$, where $\epsilon \in (0, \frac{1}{2})$ is a fixed constant. Then $\hat{\phi}_m$ is estimated by

$$\sum_{p \in S_m} \text{GOF}_{mp} = (1-2\epsilon)(n-1)(P/M-1) \ .$$

We set $\epsilon = 0.25$ and divide genes into 10 groups. This initialization strategy borrows information from different genes within a group to deal with the problem of small

sample size typical for RNA-seq data.

2.3 Model selection and hybrid-hierarchical tree

For penalized model-based clustering, it is important to determine the number of clusters K and the regularization parameter λ. Several useful selection criteria can be applied in the model-based clustering, such as AIC, BIC and some modified versions. Here we use BIC (Pan and Shen (2007)) which is defined as

$$BIC = -2\log L(\hat{\Theta}) + \log(n) d_e \ ,$$

where $d_e = K + 2P + KP - 1 - q$ is the effective number of parameters and $q = \#\{(k,p) : \hat{\theta}_{kp} = 1\}$.

In many cases, we may want to visualize the hierarchical clustering structure. For this we use the hybrid-hierarchical (HH) tree guided by the result from penalized model-based clustering. The HH tree applies agglomerative clustering to the set of clusters obtained from model-based clustering. Thus it produces a partial hierarchical clustering containing only clusters coarser than output from model-based clustering. The method was first proposed in Karypis *and others* (1999); Vaithyanathan and Dom (2000) and later summarized and extended in Zhong and Ghosh (2003). Since one objective of clustering is to identify subtypes of cells, the partial hierarchical tree can be used to investigate how the subtypes organize themselves into coarser groups.

Let K_0 be the number of clusters selected by penalized model-based clustering. When building the HH tree, in the ith merging step, we have $K_0 - i + 1$ clusters $C_1, \ldots, C_{K_0 - i + 1}$. The distance between two clusters, C_a and C_b is defined by

$$D(C_a, C_b) = \log \frac{\prod_{j \in C_a} f_a(\boldsymbol{x_j}; \hat{\mu}_{ja}, \boldsymbol{\phi}) \prod_{j \in C_b} f_b(\boldsymbol{x_j}; \hat{\mu}_{jb}, \boldsymbol{\phi})}{\prod_{j \in C_c} f_c(\boldsymbol{x_j}; \hat{\mu}_{jc}, \boldsymbol{\phi})} \ ,$$

where $C_c = C_a \cup C_b$ and $\hat{\mu}_{ja}$, $\hat{\mu}_{jb}$ and $\hat{\mu}_{jc}$ are the MLE based on observations from

cluster C_a, C_b and C_c, respectively. The values of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_P)$ above assume the estimated values from the penalized model-based clustering. The HH tree is mainly used as a heuristic-based visualization tool to see how the clusters can be further grouped.

3 Numerical Evaluation and Comparison

We compare the results of our proposed method with those of PoiClaClu (Witten, 2011), edgeR (Robinson *and others*, 2010) and DESeq (Anders and Huber, 2010) and evaluate the performances in terms of clustering accuracy and gene selection. PoiClaClu measures the distance using the power transformed sequencing data based on the Poisson log-linear model. edgeR models the read counts with NB distribution and proposes a method to compute the distance matrix based on the 500 selected genes that have the largest dispersion across all samples. DESeq uses variance stabilizing transformation of count data based on the NB model and then computes the pairwise squared Euclidean distances.

3.1 Application to Real Data

We study the performances of PMixClus and the other three competing methods on four real data sets: Liver and Kidney (Marioni *and others*, 2008), MAQC-2 (Bullard *and others*, 2010), Yeast (Nagalakshmi *and others*, 2008) and Cervical Cancer (Witten *and others*, 2010). There are only technical replicates in the Liver and Kidney and MAQC-2 data sets. The Yeast data set has both technical and biological replicates while the Cervical Cancer has only biological replicates. Additionally, the MAQC-2 data set was generated from the MicroArray Quality Control consortium and hence we can use the data set from real time reverse-transcription PCR (qRT-PCR) as the gold standard to identify the DE genes. The details of these data sets can be found in the Appendix.

3.1.1 Clustering and Model Selection.

For Liver and Kidney and MAQC-2 data sets, all of the algorithms output the correct clustering when the number of clusters is specified to be $K = 2$. Furthermore, for the proposed PMixClust, $K = 2$ is indeed selected by BIC.

The clustering of Yeast and Cervical Cancer data sets is more challenging. For Yeast data set, the six samples fall into two known clusters (dT and RH). PMixClus and PoiClaClu obtain the same clustering results with $K = 5$. With a hierarchical representation (using HH tree for our method), all samples are correctly clustered at level $K = 3$ except for a RH biological replicate (Figure 2). In contrast, clustering from edgeR and DESeq looks worse based on the respective constructed trees.

For the Cervical Cancer data set, we again identify $K = 5$ by BIC and we build the HH tree with five leaves at the bottom. To make comparison with other methods, we cut all hierarchical trees constructed by different methods at level 5 (Figure 3). Visually, our method matches best with the known three clusters (color coded), followed by PoiClaClu. Figure 4 depicts the RI values of different methods when K varies from 2 to 5. The method PMixClus(HH) is based on the tree constructed by HH method, and the method PMixClus is based on applying our method with K fixed to a value between 2 and 5 when using the BIC to choose λ only. We also present the model-based clustering result when no gene selection is performed ($\lambda = 0$). From Figure 4 (b), it is clear that our proposed methods obtain better clustering results than the other three, even with no gene selection, except when $K = 2$ where PoiClaClu is only inferior to PMixClus(HH). Note in this example there are 3 known clusters.

3.1.2 Gene Selection.

From Figure 4(b), we can tell that the results with gene selection (the red line and the blue line) are better than those without gene selection (the black line). In Figure 5 we display the ratios of correctly included genes to the total number of DE genes when the number of selected genes increases (obtained using different λ), as well as

the ratio of correctly excluded genes to the total number of non-DE genes. The BIC selected a model which includes 298 genes as noise (shown as solid triangle and solid circle in the figure).

3.2 Simulation Study

3.2.1 Simulation Setup.

In each data set, the read count $x_{jp}|j \in C_k$ for gene p in sample j belonging to cluster k follows the distribution $\text{NB}(s_j\gamma_p\theta_{kp}, \phi_p)$. The RNA-seq count data $\boldsymbol{x}$ with $P = 10000$ genes are generated from two clusters, each of which contains 10 samples. The size factor s_j is generated from $\text{Unif}(0.5, 1.7)$ and γ_p from $\text{Exp}(1/100)$. Among the 10000 genes, the first 3000 genes are differentially expressed between clusters. For some constant $z > 1$, in the first cluster, θ_{kp} is set to be z and $1/z$ for the first 1500 genes and the next 1500 genes, respectively. Similarly, for the second cluster, θ_{kp} is set to be $1/z$ and z, for the first 1500 genes and the next 1500 genes respectively. For the remaining 7000 genes, we set $\theta_{kp} = 1$. By the descriptions above, z represents the level of fluctuation between different clusters and we consider two values $z = e^{0.2}$ and $z = e^{0.5}$ in our simulations. For the dispersion parameter ϕ_p, we test four values: $\phi_p = 0.01$, $\phi_p = 0.1$ and $\phi_p = 0.5$, and $\phi_p = 1/(100 + \gamma_p)$, the last of which is similar to the setup used in Anders and Huber (2010) and Si *and others* (2014). For each setup, 50 data sets are generated.

3.2.2 Evaluation of Clustering.

We compare the clustering performance of PoiClaClu, edgeR, DESeq, and our proposed method PMixClus. Here we assume the true number of clusters is known, since the other three methods do not suggest a value for the number of clusters. Furthermore, for a fair comparison, we estimate the size factor s_j by median ratio method in all algorithms. To assess the clustering performance, we use the rand index

(RI) (Rand, 1971), which measures the similarity between the true clusters and the estimated clusters. The higher the RI value is, the more accurate is the estimated clustering .

When $z = e^{0.5}$ (larger differences between clusters), all methods can correctly assign the samples to the two clusters for all settings of dispersion parameter ϕ_p. For $z = e^{0.2}$, we compare the RI values of different methods in Table 1. All methods except for edgeR could achieve correct clustering results when dispersion value is not too high (last three settings of ϕ_p). When the over-dispersion is high ($\phi_p = 0.5$), PMixClus performs best among all methods.

3.2.3 Evaluation of Gene Selection.

To evaluate the performance of gene selection for the proposed method, we report noise features exclusion rate (NER), informative features exclusion rate (IER) and accuracy (ACC). NER is the ratio of the number of noise features excluded by a method to the number of true noise features. IER is the ratio of the number of informative features excluded to the number of true informative features. ACC is the proportion of true noise features and true informative features correctly found among all features.

In Table 2, we summarize the NERs, IERs and ACCs of our proposed method on simulated data sets with different settings of z and ϕ_p. We obtain a good balance of NER and IER when the over-dispersion was not too high. When over-dispersion is high, many informative features are falsely excluded. However this is not unexpected since it is hard for any algorithm to distinguish between differences in expression caused by different clusters and caused by over-dispersion. Note that we can still obtain the relative high RI values shown in Table 1 even with high dispersion.

3.2.4 Dispersion Parameter Estimation.

Dispersion parameters have a great effect on gene selection and clustering. However there are some difficulties that prevent us from obtaining accurate estimates of dispersion. Firstly, as a result of high costs of the experiment, the RNA-seq data normally has low sample size and hence it is challenging to estimate the dispersion accurately. Secondly, the fluctuation among different clusters may give rise to larger estimated values of dispersion for the DE genes. Thirdly, for the penalized mixture model, over-estimation can result from shrinking the mean parameters. We used a robust initial values of ϕ_p as explained in Section 2 that borrows information from multiple genes.

To illustrate the estimation of dispersion, we generate one data set from each simulation setup and plot the estimates of the dispersion parameter ϕ_p in Figure 1. It is worth noting that estimates obtained by our method come closer to the true value of ϕ_p when $\log(\gamma_p)$ increases and the fluctuation decreases. When the true dispersion is large ((a), (b), (e) and (f) in Fig. 1), the proposed method can get more accurate estimates.

3.2.5 Evaluation of Model Selection.

To select proper models for simulated data sets, we examine $K = 1, 2, 3, 4$ and a grid of values for λ. We apply BIC to obtain the optimal combination of λ and K. Table 3 shows that BIC can select the correct K in most cases ($K = 4$ never selected) except when both differences in clusters and dispersion is large. In the latter case, this is likely due to that differences in expression caused by dispersion can be confounded with differences between cluster.

4 Conclusions

In this work, we proposed the penalized model-based method to accomplish clustering analysis on RNA-seq count data. Typically these data sets have the characteristics of high dimension and low sample size. Moreover many of the features are noninformative about the cluster and hence should be automatically excluded in order to increase the accuracy of clustering. The proposed method has the desired ability of performing clustering and gene selection simultaneously. In addition, model-based approach allows us to apply BIC to determine the number of clusters.

A shortcoming of the proposed method is that computationally it is much slower than other methods. Table 4 compares the computational time of different methods on the four real data sets. For our method, 20 values of λ and $K = 1, \ldots, 6$ is used for parameter search. The algorithm is stopped when the relative change in penalized log-likelihood is less than 10^{-6}.

The l_1 penalty is applied in the mixture model to select genes. However it may penalize the large values excessively (Zou, 2006). This bias can also lead to larger estimates of dispersion and further result in inaccuracy of the gene selection. Some other penalties may be applied to tackle this problem, such as hard thresholding penalty or SCAD penalty (Fan and Li (2001)). Nevertheless, these penalties can complicate numerical computation significantly. Consequently to design better penalties in the NB mixture model for clustering with fast numerical implementation will be our future research topics.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. and Caski, F. (editors), *In Proc. of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado. pp. 267–281.

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

Berninger, P., Gaidatzis, D., van Nimvegen, E. and Zavolan, M. (2008). Computational analysis of small RNA cloning data. *Methods* **44**, 13–21.

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

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., Hunkapiller, K., Jensen, R. V., Knight, C. R., Lee, K. Y., Ma, Y., Maqsodi, B., Papallo, A., Peters, E. H., Poulter, K., Ruppel, P. L., Samaha, R. R., Shi, L., Yang, W., Zhang, L. *and others.* (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology* **24(9)**, 1115–1122.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. and Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**(2), 244–258.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* **94**, 759–771.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* **22**, 555–574.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood

and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fraly, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing* **20**, 270–281.

Frazee, A. C., Langmead, B. and Leek, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* **12**(449).

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**, 70–84.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66**(3), 793–804.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association* **109**, 1285–1301.

Karypis, G., Han, E.-H. and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computers* **32(8)**, 68–75.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**, 1025–1038.

Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13(3)**, 523–538.

Lu, J., Tomfohr, J. K. and Kepler, T. B. (2005). Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6**(165).

Ma, S., Huang, J. and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* **12**(4), 763–775.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18**, 1509–1517.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York:Wiley.

Metzker, M. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**, 31–46.

Nagalakshmi, U., Wong, 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* **302**, 1344–1349.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8**, 1145–1164.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M, Latulippe, E., Mesirov, J, Poggio, T., Gerald, W., Loda, M., Lender, E. *and others*. (1998). Multiclass cancer diagnosis using tumor gene expression signature. *PNAS* **9**, 3273–3975.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expresion data. *Biostatistics* **26**, 139–140.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics* **6**, 461–464.

Seidel, W., Mosler, K. and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* **52**, 481–487.

Si, Y., Liu, P., Li, P. and Thomas, P. B. (2014). Model-based clustering for RNA-seq data. *Bioinformatics* **30**, 197–205.

Spellman, P. T., Sherlock, G., Iyer, V. R., Zhang, M., Anders, K., Eisen, M. B., Broun, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273–3975.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Vaithyanathan, S. and Dom, B. (2000). Model-based hierarchical clustering. In: *In Proc. 16th Conf. Uncertainty in Artificial Intelligence*. UAI. pp. 599–608.

Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**(2), 353–360.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics* **5**, 2493–2518.

Witten, D. M., Tibshirani, R., Gu, S., Fire, A. and Lui, W. (2010). Ultra-high through-put sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumous and matched controls. *BMC Biology* **8**(58).

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.

Zhong, S. and Ghosh, J. (2003). A unified framework for model-based clustering. *Journal of Machine Learning Research* **4**, 1001–1037.

Zhou, H., Pan, W. and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics* **3**, 1473–1496.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

APPENDIX

A.1 EM Algorithm

We use EM algorithm (Dempster *and others*, 1977; McLachlan and Peel, 2000; Fraly, 1998; Pan and Shen, 2007; Si *and others*, 2014) to estimate the parameters. The complete-data penalized log-likelihood is given by

$$\log L_{c,P}(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{k=1}^{K} z_{kj}\{\log\pi_k + \log f_k(\boldsymbol{x_j};\boldsymbol{\psi_{jk}})\} - \lambda\sum_{k=1}^{K}\sum_{p=1}^{P}|\log\theta_{kp}|, \qquad (A.1)$$

where $z_{kj} = 1$ if sample j belongs to cluster k and $z_{kj} = 0$ otherwise. z_{kj} is treated as missing data in the EM algorithm.

A.1.1 E-Step.

We need to compute the conditional expectation of penalized log-likelihood (A.1) for the complete data with respect to z_{kj} given data $\boldsymbol{x}$. On the $(m+1)$th iteration, this conditional expectation is

$$\begin{aligned} Q_P(\boldsymbol{\Theta};\hat{\boldsymbol{\Theta}}^{(m)}) &= E_{\hat{\boldsymbol{\Theta}}^{(m)}}(\log L_{c,P}(\boldsymbol{\Theta})|\boldsymbol{x}) \\ &= \sum_k\sum_j \hat{\tau}_{kj}^{(m)}[\log\pi_k + \log f_k(\boldsymbol{x_j};\mu_{jk},\boldsymbol{\phi})] - \lambda\sum_{k=1}^{K}\sum_{p=1}^{P}|\log\theta_{kp}| \qquad (A.2) \end{aligned}$$

where $\hat{\tau}_{kj}^{(m)}$ is the posterior probability that the sample j comes from kth cluster given the estimates of other parameters from previous iterations:

$$\hat{\tau}_{kj}^{(m)} = E_{\hat{\boldsymbol{\Theta}}^{(m)}}(z_{kj}|\boldsymbol{x}) = \frac{\hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j};\hat{\mu}_{jk}^{(m)},\hat{\boldsymbol{\phi}}^{(m)})}{\sum_{k=1}^{K}\hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j};\hat{\mu}_{jk}^{(m)},\hat{\boldsymbol{\phi}}^{(m)})} \ .$$

A.1.2 M-Step.

On the $(m+1)$th M-step, we firstly get the estimator of π_k:

$$\hat{\pi}_k^{(m+1)} = \sum_j \hat{\tau}_{kj}^{(m)}/n, \qquad k = 1, \ldots, K.$$

In this paper, the size factor s_j is computed by median ratio method (Anders and Huber, 2010). Then we maximize (3) with respect to θ_{kp}, γ_p and ϕ_p for $p = 1, \ldots, P$ and $k = 1, \ldots, K$. Since it is hard to jointly maximize over these parameters, we maximize each paramter in turn with others fixed. For the Poisson model (this special case is available in the R package although not discussed in the paper), the maximizers can be found in closed form. For the NB model, we apply the Newton Raphson (NR) algorithm which is similar to that used in Goeman (2010) to compute the maximizers.

A.2 Real Data Sets.

Liver and Kidney compared the expression of 22925 genes between a liver sample and a kidney sample from a human male. Seven technical replicates were generated for each sample. We extracted five replicates, which had the same library preparation (at the 3 pM concentration), for each sample. We focused on the 18228 genes whose total gene counts over all samples are not less than 5. The data set can be downloaded from a supplementary file in Marioni *and others* (2008).

MAQC-2 is the mRNA-seq data set related to MicroArray Quality Control Project, comparing two types of biological samples (Brain and UHR). There were seven technical replicates with one specific library preparation for each biological sample. A subset of genes (around one thousand) were assayed by qRT-PCR (Canales *and others* (2006)). Based on the fold changes of genes in qRT-PCR data, we selected 188 genes from the subset, including 141 DE genes (fold change > 2) and 47 non-DE genes (fold change < 0.2). Then we replicated the 47 non-DE genes 5 times

so that the ratio of DE genes to non-DE genes is more reasonable. The sequencing data set can be downloaded from `http://bowtie-bio.sourceforge.net/recount` (Frazee *and others*, 2011) and the qRT-PCR data can be downloaded from `www.ncbi.nlm.nih.gov/geo` with GEO Accession GSE5350.

Yeast is the RNA-seq data set, comparing the replicates of *Saccharomyces cerevisiae* cultures. Three replicates were tested under each of two library preparation, oligo(dT) (dT) and random hexamers (RH). Specifically, there was one original replicate, one technical replicate and one biological replicate under each library preparation protocol. We focused on the 6710 genes whose total gene counts over all samples are at least 3. The data set can be downloaded from a supplementary file in Anders and Huber (2010).

Cervical Cancer is the microRNA (miRNA), 18-30 nucleotides in length, sequencing data set which were used to compare cervical cancer tissues and normal tissues. This data set included 29 samples from each of cervical cancer tissues and 29 from each of normal tissues with 714 miRNA. Among cervical cancer tissue samples, there are 21 squamous cell carcinomas (SCC), 6 adenocarcinomas (ADS) and 2 unclassified. We excluded two unclassified samples from analysis. We focused on the 636 genes whose total gene counts over all samples are at least 5. The data set can be downloaded from a supplementary file in Witten *and others* (2010).

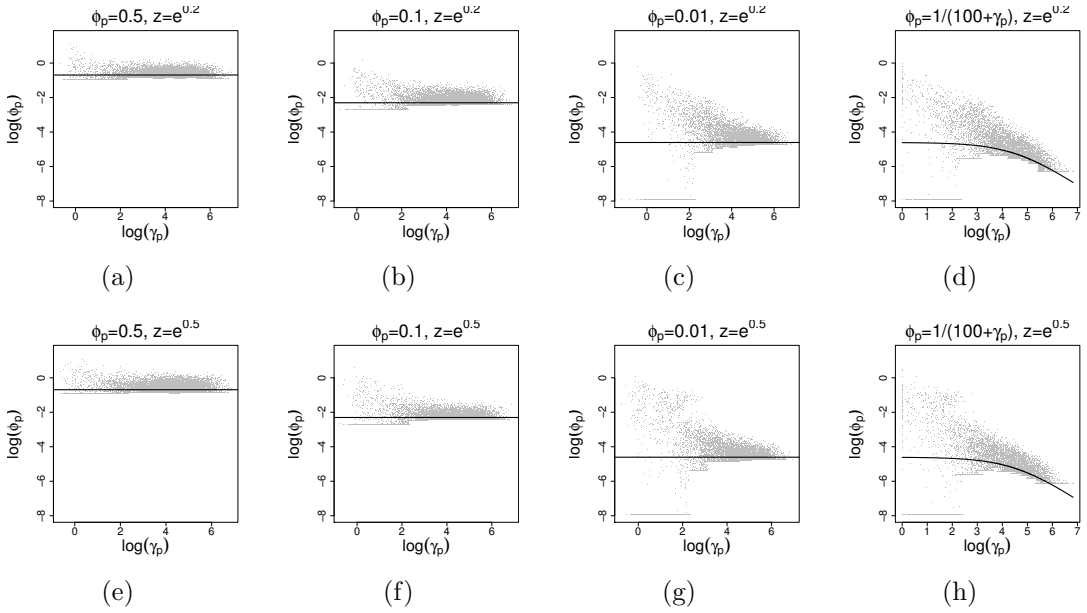


Figure 1: This figure displays the $\log(\phi_p)$ versus the true values of $\log(\gamma_p)$. The grey dots are the estimates from our proposed method with fixed $K = 2$. The black line represents the true dispersion value.

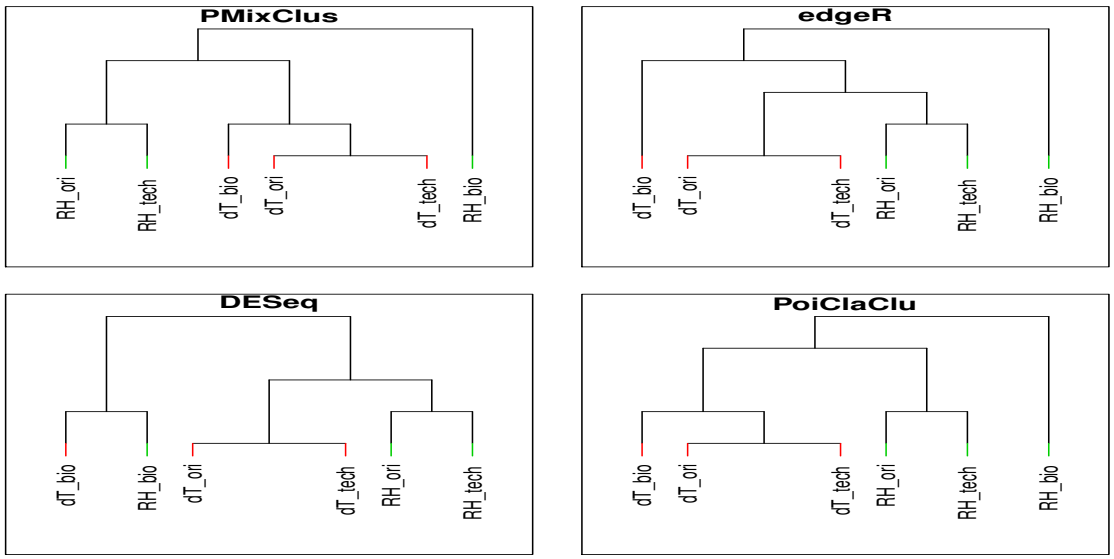


Figure 2: The hierarchical cluster dendrograms of Yeast data set are plotted based on the analysis of PMixClus, PoiClaClu, edgeR and DESeq. The dT samples and RH samples are in red and green respectively.

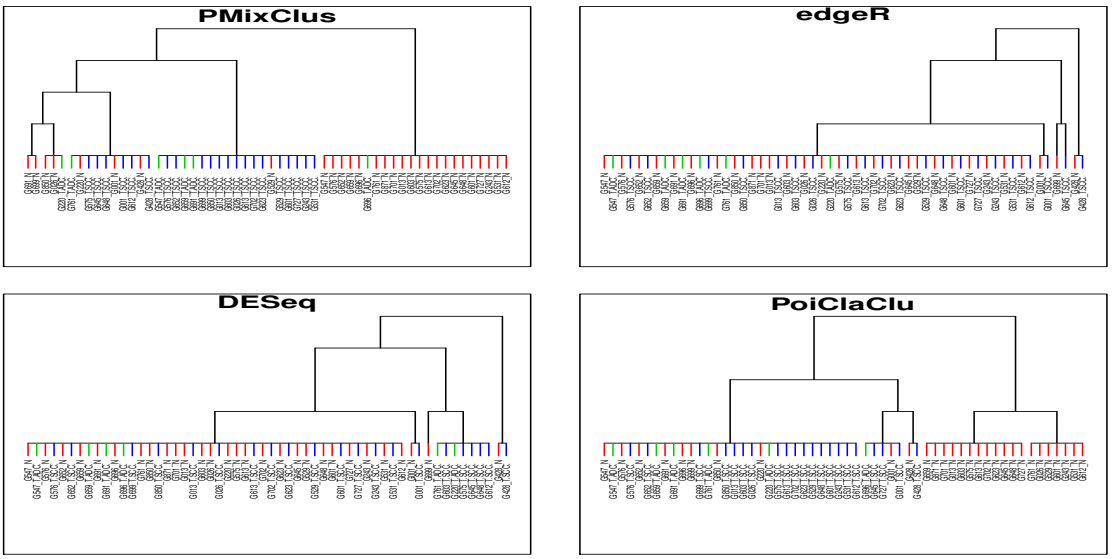
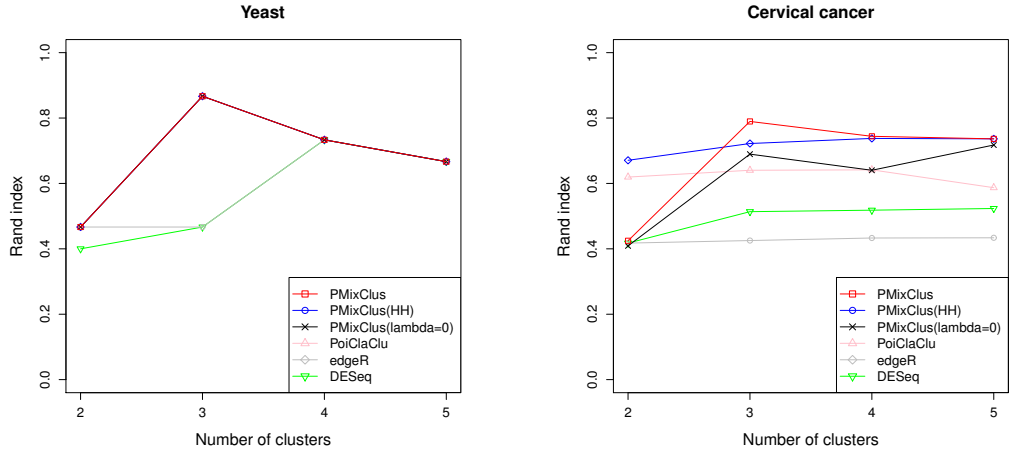


Figure 3: The hierarchical cluster dendrograms of Cervical Cancer data set are plotted based on the analysis of PMixClus, PoiClaClu, edgeR and DESeq. The dendrograms of PoiClaClu, edgeR and DESeq are from cutting the hierarchical tree at the fifth level. The ADC, SCC and normal samples are in green, blue and red respectively.



(a) (b)

Figure 4: This figure displays the RIs by cutting the hierarchical tree at corresponding levels, except for PMixClus and PMixClus($\lambda = 0$). The curves for PMixClus and PMixClus($\lambda = 0$) show the RIs resulting from using different fixed number of clusters in our method.

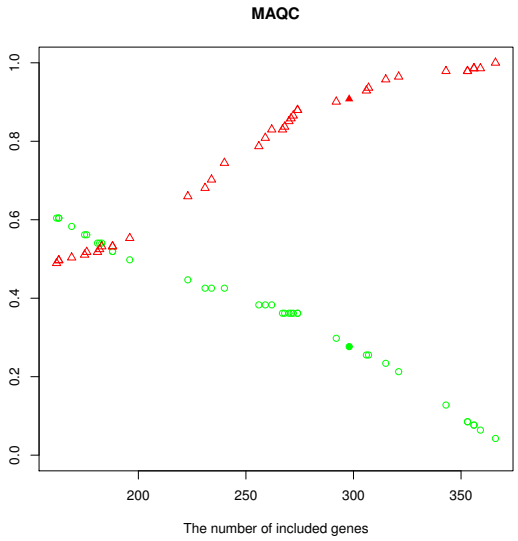


Figure 5: Plot for the the true positive ratio (number of correctly included DE genes/number of DE genes, red triangular) and the true negative ratio (number of correctly excluded non-DE genes/number of non-DE genes, green circle).

Table 1: Mean values of RIs and standard errors over 50 simulated data sets with the level of fluctuation $z = e^{0.2}$.

ϕ_p	PMixClus	PoiClaClu	edgeR	DESeq
0.5	0.990(0.074)	0.760(0.204)	0.502(0.094)	0.851(0.233)
0.1	1(0)	1(0)	0.490(0.060)	1(0)
0.01	1(0)	1(0)	0.751(0.258)	1(0)
$1/(100+\gamma_p)$	1(0)	1(0)	0.671(0.242)	1(0)

Table 2: Mean values and standard errors of NERs, IERs and ACCs over 50 simulated data sets when we use the true $K = 2$ or select K by BIC.

z	ϕ_p	NER(K=2)	IER(K=2)	ACC(K=2)	NER(BIC)	IER(BIC)	ACC(BIC)
$e^{0.2}$	0.5	0.985(0.042)	0.946(0.035)	0.706(0.019)	0.985(0.010)	0.935(0.031)	0.709(0.004)
	0.1	0.999(0.000)	0.960(0.005)	0.712(0.001)	0.945(0.051)	0.600(0.209)	0.781(0.039)
	0.01	0.839(0.009)	0.162(0.013)	0.839(0.005)	0.817(0.078)	0.158(0.021)	0.824(0.049)
	$1/(100+\gamma_p)$	0.790(0.028)	0.181(0.022)	0.799(0.014)	0.790(0.028)	0.181(0.022)	0.799(0.014)
$e^{0.5}$	0.5	0.966(0.113)	0.882(0.269)	0.712(0.026)	0.752(0.168)	0.180(0.181)	0.772(0.095)
	0.1	0.771(0.026)	0.041(0.083)	0.828(0.008)	0.750(0.055)	0.029(0.004)	0.817(0.038)
	0.01	0.698(0.011)	0.072(0.008)	0.767(0.007)	0.698(0.011)	0.072(0.008)	0.767(0.007)
	$1/(100+\gamma_p)$	0.628(0.049)	0.084(0.092)	0.714(0.009)	0.621(0.013)	0.071(0.007)	0.714(0.008)

Table 3: Frequencies of the number of clusters K selected by BIC from 50 simulated data sets. The mean values of selected λ are also reported.

		$K=1$		$K=2$		$K=3$	
z	ϕ_p	Freq	λ	Freq	λ	Freq	λ
$e^{0.2}$	0.5	10	0(0)	36	7.71(0.93)	4	7.39(0)
	0.1	1	0(0)	48	20.09(0)	1	14.07(1.11)
	0.01	0	–	45	20.09(0)	5	17.81(3.12)
	$1/(100+\gamma_p)$	0	–	50	21.82(1.57)	0	-
$e^{0.5}$	0.5	2	0(0)	2	10.31(0)	46	4.84(0.96)
	0.1	1	0(0)	48	7.39(0)	1	7.39(0)
	0.01	0	–	50	14.39(0)	0	–
	$1/(100+\gamma_p)$	1	0(0)	49	14.39(0)	0	–

Table 4: Computational time for different methods on the real data sets.

	PMixClus	PoiClaClu	edgeR	DESeq
Liver and Kidney	2h	4s	1s	<1s
MAQC-2	1h	2s	1s	<1s
Yeast	1.5h	1s	<1s	<1s
Cervical Cancer	1h	2s	1s	<1s