Bioinformatics, 34(8), 2018, 1329–1335
doi: 10.1093/bioinformatics/btx768
Advance Access Publication Date: 27 November 2017
Original Paper

Gene expression

Classifying next-generation sequencing data using a zero-inflated Poisson model

Yan Zhou[1], **Xiang Wan**[2,*], **Baoxue Zhang**[3] **and Tiejun Tong**[4,*]

[1]College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen 518060, China, [2]Department of Computer Science, and Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Kowloon Tong, Hong Kong, [3]School of Statistics, Capital University of Economics and Business, Beijing 100070, China and [4]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on July 18, 2017; revised on October 18, 2017; editorial decision on November 12, 2017; accepted on November 24, 2017

Abstract

Motivation: With the development of high-throughput techniques, RNA-sequencing (RNA-seq) is becoming increasingly popular as an alternative for gene expression analysis, such as RNAs profiling and classification. Identifying which type of diseases a new patient belongs to with RNA-seq data has been recognized as a vital problem in medical research. As RNA-seq data are discrete, statistical methods developed for classifying microarray data cannot be readily applied for RNA-seq data classification. Witten proposed a Poisson linear discriminant analysis (PLDA) to classify the RNA-seq data in 2011. Note, however, that the count datasets are frequently characterized by excess zeros in real RNA-seq or microRNA sequence data (i.e. when the sequence depth is not enough or small RNAs with the length of 18–30 nucleotides). Therefore, it is desired to develop a new model to analyze RNA-seq data with an excess of zeros.

Results: In this paper, we propose a Zero-Inflated Poisson Logistic Discriminant Analysis (ZIPLDA) for RNA-seq data with an excess of zeros. The new method assumes that the data are from a mixture of two distributions: one is a point mass at zero, and the other follows a Poisson distribution. We then consider a logistic relation between the probability of observing zeros and the mean of the genes and the sequencing depth in the model. Simulation studies show that the proposed method performs better than, or at least as well as, the existing methods in a wide range of settings. Two real datasets including a breast cancer RNA-seq dataset and a microRNA-seq dataset are also analyzed, and they coincide with the simulation results that our proposed method outperforms the existing competitors.

Availability and implementation: The software is available at http://www.math.hkbu.edu.hk/∼tongt.

Contact: xwan@comp.hkbu.edu.hk or tongt@hkbu.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput techniques have recently emerged as a revolutionary technology to replace hybridization-based microarrays for gene expression analysis (Mardis, 2008; Morozova *et al.*, 2009; Wang *et al.*, 2009). Due to the increased specificity and sensitivity of gene expression, next-generation sequencing data has become a popular choice in biological and medical studies. In particular, RNA-sequencing (RNA-seq) enables certain applications not achievable by microarrays, which targets for analyzing much less noisy data, such as the inference of differential expression (DE) between several conditions or tissues. Existing DE analyses include, but not limited to, edgeR (Robinson and Smyth, 2008; Robinson *et al.*, 2010), DESeq2 (Love *et al.*, 2014) and LFCseq (Lin *et al.*, 2014). RNA-seq data have clear advantage over microarray data to identify which type of diseases a new patient belongs to. With the reduced cost in sequencing technology, more and more researchers tend to use RNA-seq data to diagnose diseases (Lorenz *et al.*, 2014). For classification of microarray data, many discriminant methods are available in the literature, such as Diagonal Linear Discriminant Analysis (DLDA) and Diagonal Quadratic Discriminant Analysis (DQDA) in Dudoit *et al.* (2002). Huang *et al.* (2010) proposed two bias-corrected rules for DLDA and DQDA. Zhou *et al.* (2017a,b) proposed a bias-corrected geometric diagonalization method for regularized discriminant analysis. Different from microarray data, the RNA-seq reads are mapped on to the reference genome and are summarized as 'counts'. That is, we use a count number to measure the expression level of each gene in RNA-seq data. Unlike microarray data that follow a Gaussian distribution after normalization, RNA-seq data follow a discrete distribution, e.g. a Poisson distribution. As a consequence, the existing classification methods for microarray data may not provide a satisfactory performance or may not even be applicable for RNA-seq data.

There are a few discriminant methods developed for classifying RNA-seq data in the recent literature. Witten (2011) assumed a Poisson distribution for RNA-seq data and proposed a Poisson linear discriminant analysis (PLDA) for classification. Tan *et al.* (2014) conducted an extensive comparison study and concluded that, for classifying RNA-seq data, PLDA performed better or much better than the classification methods for microarray data. Dong *et al.* (2016) extended the Poisson model to the negative binomial model and developed a negative binomial linear discriminant analysis (NBLDA) in the presence of overdispersion in RNA-seq data. Note, however, that there may have excess zeros in real RNA-seq datasets, especially when the sequence depth is not enough or small RNA has the length of 18–30 nucleotides, such as microRNA. For instance, the cervical cancer dataset in Witten *et al.* (2010), which was also analyzed in Witten (2011) and Dong *et al.* (2016), contains about 47.6% zeros of all numerical values. Another dataset, the liver and kidney dataset in Marioni *et al.* (2008), also contains 45.5% zeros of all numerical values.

In this paper, we propose a Zero-Inflated Poisson Logistic Discriminant Analysis (ZIPLDA) for RNA-seq data in the presence of excess zeros. To model the complex RNA-seq data with an excess of zeros, we take a two-step procedure to have a new discriminant classifier. We first address the complexity of the classifier caused by the mixture model, and then take into account the relation between the probability of zeros and the mean of the genes and the sequencing depth in the second step. By the above steps, we build a novel classifier to improve the class prediction for a future observation with RNA-seq data.

The remainder of the paper is organized as follows. In Section 2, we describe the problem of analyzing RNA-seq data with an excess of zeros, and present the motivation of our research work. In Section 3, we propose the ZIPLDA for RNA-seq data with an excess of zeros and also describe the estimation of parameters in details. In Section 4, simulation studies are provided to evaluate the performance of the new classifier via the comparison with the existing methods. In Section 5, we apply the proposed method to analyze two real next-generation sequencing datasets and compare its performance with the existing methods. Finally, we conclude the paper with some discussions and future directions in Section 6.

2 Data description and motivation

We first present our main motivation of the study through two real datasets as examples for classification.

The first example is the breast cancer RNA-seq dataset from The Cancer Genome Atlas (TCGA) with the link https://portal.gdc.cancer.gov/projects/TCGA-BRCA, a project that aims to offer a comprehensive overview of genomic changes involved in human cancer. This dataset includes 112 tumoral and 112 normal samples with measurements on 60 483 transcripts. The dataset generation and processing were described in The Cancer Genome Atlas Research Network (2014). There are about 45.6% zeros of all numerical values and 3477 out of 60 483 transcripts equal zero for all samples. After removing the 3477 transcripts, there are still about 42.3% zeros in the remaining transcripts. For details, one may refer to the first vertical bar in the left panel of Supplementary Figure S11.

The second example is a microRNA-seq dataset from the Gene Expression Omnibus (GEO) with access number GSE79017 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79017). The dataset was also released in Wolenski *et al.* (2017). MicroRNAs are small RNAs, with the length of 18–30 nucleotides, and play important regulatory roles in diverse biological processes (Birchler and Kavi, 2008; Stefani and Slack, 2008). The dataset has three classes, including 12 samples from liver, 18 samples from urine and 18 samples from plasma, with measurements on 832 microRNAs. There are about 66.1% zeros of all numerical values and 127 out of 832 microRNAs equal zero for all samples. Without the 127 microRNAs, there still exist about 59.9% zeros in the remaining microRNAs. For details, one may refer to the first vertical bar in the left panel of Figure 1.

For the above two datasets, PLDA may not provide a satisfactory performance due to the excess zeros. Specifically, since the expectation of the Poisson distribution for each gene of each sample is related to class, individual and gene, a direct issue is that there are many zeros but the expectation of the distribution is not close to zero and could be a large value in a gene with the PLDA method. For example, the expected expression of a gene, the red point in the right panel of Figure 1, is 68.8, and in which about 62.5% of observations of the gene are zeros. Note, however, that the probability of zeros in the Poisson distribution with expectation 68.8 is only $1.3\exp(-30) \approx 0$, which is much less likely to occur than a chance of 62.5%. Hence, it is not suitable to fit the data by a single Poisson

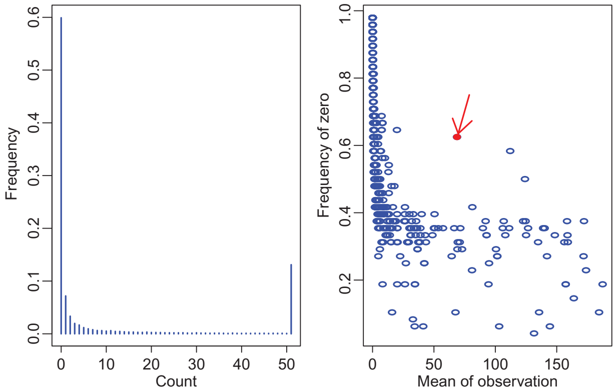

Fig. 1. The frequency of each observation and empirical probability of zero in the microRNA-seq dataset. The left panel is the frequency of each observation value (the last value is the frequency of larger than 50). The right panel is the empirical probability of zero in each microRNA, which is represented by one dot in the right panel

distribution for data with an excess of zeros. Furthermore, from the right panel of Supplementary Figure S11 and Figure 1, we find that the experiential probability of observing zeros in a gene is related to the mean of observation value. This calls for a new model for classifying RNA-seq data with an excess of zeros.

3 Materials and methods

Let K be the number of classes, and X_{ki_kg} denote the number of reads mapped to gene g in sample i of class k, $k = 1, \ldots, K$, $i_k = 1, \ldots, n_k$ and $g = 1, \ldots, G$. Specifically, there are n_k samples in class k. Let $n = \sum_{k=1}^{K} n_k$ be the total number of samples for all classes. Given a new sequence observation, $\boldsymbol{x}^* = (X_1^*, \ldots, X_G^*)^T$, the goal of classification is to predict which class label the observed $\boldsymbol{x}^*$ belongs to. Witten (2011) proposed a PLDA with the Poisson distribution to classify RNA-seq data. Dong *et al.* (2016) developed a NBLDA that uses the negative binomial distribution to count for the overdispersion in the data. In this section, we propose a new discriminant analysis for RNA-seq data by assuming that the data follow a mixture distribution.

3.1 Zero-inflated Poisson logistic discriminant analysis

For excess zeros in a sample, it is not suitable to assume X_{kig} follows a Poisson or negative binomial distribution. In such situations, we consider the following zero-inflated Poisson distribution for RNA-seq data. That is,

$$X_{ki_kg} \sim \begin{cases} \delta_{\{0\}} & p_{ki_kg} \\ \text{Poisson}(\mu_{ki_kg}) & (1 - p_{ki_kg}), \end{cases} \tag{1}$$

where $\delta_{\{0\}}$ denotes the point mass at zero, μ_{ki_kg} is the expectation for gene g in sample i_k in class k and p_{ki_kg} is the probability of $\delta_{\{0\}}$ in gene g of sample i_k in class k. Since μ_{ki_kg} is related to class, individual and gene, we assume $\mu_{ki_kg} = d_{kg} s_{i_k} \lambda_g$, where d_{kg} allows the gth gene to be differentially expressed between classes, s_{i_k} is used to identify individual in the kth class and $\lambda_g = \sum_{k=1}^{K} \sum_{i_k=1}^{n_k} X_{i_kg}$ is the total count of short reads of gene g. We know that the expectation of X_{ki_kg} should be zero if it follows $\delta_{\{0\}}$ distribution, but the reverse statement is not true. When we observe $X_{ki_kg} = 0$, we have considered two scenarios: (i) there is no expression in the gth gene; (ii) the gene is expressed but we cannot observe the signal. Therefore, the probability of X_{ki_kg} can be written as

$$P(X_{ki_kg}) = \begin{cases} p_{ki_kg} + (1 - p_{ki_kg}) e^{-\mu_{ki_kg}} & X_{ki_kg} = 0 \\ (1 - p_{ki_kg}) \dfrac{\mu_{ki_kg}^{X_{ki_kg}}}{(X_{ki_kg})!} e^{-\mu_{ki_kg}} & X_{ki_kg} > 0. \end{cases} \tag{2}$$

We have $E(X_{ki_kg}) = (1 - p_{ki_kg})\mu_{ki_kg}$ and $\text{Var}(X_{ki_kg}) = \mu_{ki_kg}(1 - p_{ki_kg})(1 + p_{ki_kg}\mu_{ki_kg})$. Lambert (1992) introduced the zero-inflated Poisson regression to model count data with an excess of zeros. This model assumes that the outcomes are generated from two processes, where the first process is to model the zero inflation by including a proportion $1 - p$ of excess zeros and a proportion $p \exp(-\lambda)$ of zeros coming from the Poisson distribution, and the second process models the nonzero counts using the zero-truncated Poisson model. We also note that there are similar models for analyzing the zero-inflated data in the literature, such as Ridout *et al.* (1998) and Mouatassim and Ezzahid (2012). Recently, Liu *et al.* (2016) also proposed a zero-inflated Poisson model for analyzing the transposon sequencing (Tn seq) data with an excess of zeros. They modeled the probability of the insertion at the location of the gene with independent Bernoulli distributions, and this probability may vary across different locations depending on the genomic sequence or other potential covariates that may affect the chance of insertion with a logistic regression. In this paper, we consider a similar logistic relation between the probability of zeros and the mean of the genes with the sequencing depth. That is,

$$\log\left\{\frac{P(X_{ki_kg} = 0)}{1 - P(X_{ki_kg} = 0)}\right\} = \alpha + \beta_1\left(\frac{N_{ki_k}}{N_{1i_1}}\right) + \beta_2\mu_{ki_kg}. \tag{3}$$

Note that $P(X_{ki_kg} = 0)$ in model (3) can be replaced with $p_{ki_kg} + (1 - p_{ki_kg})e^{-\mu_{ki_kg}}$. Then with some simplifications, we have

$$1 - p_{ki_kg} = \frac{1}{(1 - e^{-\mu_{ki_kg}})(1 + e^{\alpha + \beta_1\left(\frac{N_{ki_k}}{N_{1i_1}}\right) + \beta_2\mu_{ki_kg}})}, \tag{4}$$

where N_{ki_k} is the total sequencing depth of the i_kth sample in class k and α, β_1 and β_2 are the intercept and coefficients of N_{ki_k}/N_{1i_1} and μ_{ki_kg}.

Given a new observation $\boldsymbol{x}^* = (X_1^*, \ldots, X_G^*)^T$, we hope to identify which class label the observed $\boldsymbol{x}^*$ belongs to. Let π_k be the proportion of observing a sample from the kth class. Throughout the paper, we set $\pi_k = n_k/n$ so that $\sum_{k=1}^{K} \pi_k = 1$. Here, we propose a Zero-Inflated Poisson Logistic Discriminant Analysis (ZIPLDA) to assign $\boldsymbol{x}^*$ to the class with label $\text{argmin}_k d_k(\boldsymbol{x}^*)$, where $d_k(\boldsymbol{x}^*)$ is the discriminant score defined as

$$\begin{aligned} d_k(\boldsymbol{x}^*) = &\sum_{g=1}^{G} I_{(x_g^*=0)} \log\left(\hat{p}_{kg}^* + (1 - \hat{p}_{kg}^*)e^{(-d_{kg}s^*\lambda_g)}\right) \\ &-\sum_{g=1}^{G} I_{(x_g^*>0)} d_{kg}s^*\lambda_g + \sum_{g=1}^{G} I_{(x_g^*>0)} \log(1 - \hat{p}_{kg}^*) \\ &+\sum_{g=1}^{G} I_{(x_g^*>0)} x_g^* \log(d_{kg}) + \log\pi_k + C, \end{aligned} \tag{5}$$

where $\hat{p}_{kg}^*$ and s^* are related to $\boldsymbol{x}_g^*$ and C is a constant independent of k. Our proposed discriminant score $d_k(\boldsymbol{x}^*)$ is derived from the Bayes rule, that is,

$$P(y^* = k \,|\, \boldsymbol{x}^*) \;\propto\; f_k(\boldsymbol{x}^*)\pi_k, \tag{6}$$

where y^* indicates the class of $\boldsymbol{x}^*$, f_k is the probability density function of the sample in class k, and π_k is the prior probability that one sample comes from class k. Under the condition $y^* = k$, the zero-inflated Poisson density of $X_g^* = x_g^*$ is

$$\begin{aligned} P(X_g^* = x_g^* \,|\, y^* = k) = &\left(\hat{p}_{kg}^* + (1 - \hat{p}_{kg}^*)e^{-\mu_{kg}^*}\right)^{I_{(x_g^*=0)}} \\ &\left((1 - \hat{p}_{kg}^*)\frac{(\mu_{kg}^*)^{x_g^*}}{(x_g^*)!}e^{-\mu_{kg}^*}\right)^{I_{(x_g^*>0)}}, \end{aligned} \tag{7}$$

where $\mu_{kg}^* = d_{kg}s^*\lambda_g$. Then, we take $d_k(\boldsymbol{x}^*) = \log(P(X_g^* = x_g^* \,|\, y^* = k))$ as the discriminant score.

From formula (5), when $\hat{p}_{kg}^* \to 0$, it follows that

$$\log\left(\hat{p}_{kg}^* + (1 - \hat{p}_{kg}^*)e^{(-d_{kg}s^*\lambda_g)}\right) \to d_{kg}s^*\lambda_g.$$

Therefore, when $\hat{p}_{kg}^* \to 0$, the ZIPLDA score reduces to the PLDA score as

$$\begin{aligned} \log P(y^* = k \,|\, \boldsymbol{x}^*) \approx &\sum_{g=1}^{G} x_g^* \log d_{kg} - \sum_{g=1}^{G} s^*\lambda_g d_{kg} \\ &+\log\pi_k + C. \end{aligned} \tag{8}$$

And accordingly, the ZIPLDA classifier reduces to the PLDA classifier when there are no zero values.

3.2 Parameter estimation

There are several unknown parameters in (5), including $\hat{p}^*_{kg}$, s^* and d_{kg}. In this section, we describe the procedures of estimating those parameters for practical use.

3.2.1 Zero distribution probability estimation

As specified in (1) and (5), the probability of the zero distribution is a very important parameter in the ZIPLDA method. In order to estimate $\hat{p}^*_{kg}$, we first estimate α, β_1 and β_2. Liu *et al.* (2016) proposed an EM algorithm (Dempster *et al.*, 1977) to estimate the parameters α, β_1 and β_2. In this paper, we choose to maximize the log-likelihood to estimate the parameters directly with the relationship of formula (4). Given the estimation of α, β_1 and β_2, p_{ki_kg} can be estimated with formula (4). For a new sequence observation, we can obtain the zero distribution probability in every gene of each class.

3.2.2 Size factor estimation

Due to the characteristics of next-generation sequencing technology, different experiments may have different total reads, i.e. the sequencing depths. To make the gene expression levels comparable, we perform a normalization step first. The normalization step aims to adjust the systematic technical effects and reduce the noise in the data as well. For the size factor s^*, various methods have been proposed in the literature (Anders and Huber, 2010; Bullard *et al.*, 2010; Dillies *et al.*, 2013; Robinson and Oshlack, 2010; Zhou *et al.*, 2017a,b). We consider to estimate the size factor s_{i_k} for the training data and the size factor s^* for the test data using three different methods: total count, DESeq2 and Upper quartile. These methods were first introduced in the PLDA method by Witten (2011). In our simulation studies, we note that there is little difference in the performance of classification among the three methods. Therefore, we simply choose the total count method, in which the size factor s_{i_k} for the training data is estimated by

$$\hat{s}_{i_k} = \frac{\sum_{g=1}^{G} X_{i_kg}}{\sum_{k=1}^{K} \sum_{i_k=1}^{n_k} \sum_{g=1}^{G} X_{i_kg}},$$

and the estimation of size factor s^* for the test data is

$$\hat{s}^* = \frac{\sum_{g=1}^{G} X^*_g}{\sum_{k=1}^{K} \sum_{i_k=1}^{n_k} \sum\limits_{g=1}^{G} X_{i_kg}}.$$

3.2.3 Class differences estimation

As in Witten (2011), $\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg})/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g)$ is used to distinguish the different classes. Here, $\hat{d}_{kg}$ indicates the under or over-expressed degree relative to the baseline for gene g in class k. To handle the small signals such as $\sum_{i_k=1}^{n_k} X_{i_kg} = 0$, we put a Gamma($\beta$, β) prior on $\hat{d}_{kg}$ in the above formula. This results in the posterior mean as $\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg} + \beta)/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g + \beta)$. Generally, $\sum_{i_k=1}^{n_k} X_{i_kg}$ and $\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g$ are much larger than 1. Therefore, the parameter β is not sensitive in our study. In our model, we consider $\beta = 1$ throughout the simulations.

4 Simulation studies

4.1 Simulation design

In this section, we assess the performance of the proposed ZIPLDA via a number of simulation studies. We compare our method with PLDA in Witten (2011) and other popular classification methods for high-dimensional data, including NBLDA in Dong *et al.* (2016), the support vector machines (SVM) classifier in Meyer *et al.* (2014) and the k nearest neighbors (kNN) classifier in Ripley (1996). In our experiments, we use the R packages 'PoiClaClu' for PLDA and 'e1071' for SVM, where the latter one can be downloaded from https://cran.r-project.org/web/packages/e1071/index.html. We also consider the number of nearest neighbors as 1, 3 or 5 for kNN.

We first generate the data from the negative binomial distribution:

$$X_{ki_kg} \sim \text{NB}(d_{kg}s_{i_k}\lambda_g, \phi), \tag{9}$$

and then set $X_{ki_kg} = 0$ with probability p_{ki_kg}, which is related to $d_{kg}s_{i_k}\lambda_g$ and the sequence depth. In each simulation study, we compare the misclassification rates by changing one parameter and fixing the others.

In the simulation studies, we consider both the binary classification with $K = 2$, and the multiple classification with $K = 3$. For a fair comparison with PLDA, the parameters s_{i_k}, λ_g and d_{kg} are set as the same as those in Witten (2011). Specifically, the size factors s_{i_k} are from the uniform distribution on [0.2, 2.2], the λ_g values are from the exponential distribution with expectation 25, and the log d_{kg} values are from $N(0, \sigma^2)$. We also set p_{ki_kg} as a random variable following a uniform distribution on [0, 1] for each sample. In each experiment, we generate n (the summation of all classes) samples as the training set and generate another n samples as the test set.

We first consider the binary classification with $K = 2$. In Study 1, we fix $\phi = 0.001$ and $\sigma = 0.2$ and consider the case that the number of features $p = 100$ or 1000, 20% or 40% of which are differentially expressed (DE) between the two classes. Then we compare the misclassification rates of all methods with different sample sizes, $n = 8$, 16, 24, 40 and 64, for two classes. In Study 2, we investigate the performance of the methods when the proportions of differentially expressed genes are 0.2, 0.4, 0.6, 0.8 and 1.0 with fixed sample size $n = 8$ or 20. In this study, we also set $\phi = 0.001$, $\sigma = 0.2$ and $p = 100$ or 1000. In Study 3, we test the performance of all methods with the different numbers of features, including $p = 20$, 40, 60, 100, 200, 500 and 1000. We fix $\phi = 0.001$ and $\sigma = 0.2$, and consider the case that the sample size $n = 8$ or 40, and 40 or 80% of features are differentially expressed. In Studies 4 and 5, we compare all the methods with different settings of ϕ and σ. The dispersion parameter σ ranges from 0.001 to 1 in Study 4, which represents the overdispersion from very slight to very high. In Study 5, the parameter ϕ changes from 0.05 to 0.65, in which a larger ϕ means a larger difference.

Next, we consider the multiple classification with $K = 3$. Studies 6 to 10 are conducted to investigate the performance of the methods. All parameters are kept the same as those in the binary classification except for the sample sizes. Considering at least four samples in each class, we set $n = 12$, 24, 36, 60 and 96 in these studies, respectively.

Our last study is to investigate the robustness of the proposed method when the data include no excess of zeros, that is, when the data follow exactly the Poisson distribution. Studies 11 to 13 are designed to assess the performance of the methods. The settings of the three studies are kept the same as the first three studies of the binary classification with $K = 2$, except for the distribution of the data.

4.2 Simulation results

For each simulated data, we use the misclassification rate for evaluation, which is computed by repeating the simulation 1000 times and taking an average over all the simulations. We report the misclassification rates along with various parameters in Figures 3–5. More simulation results can be seen in Supplementary Figures S1 and S2 for the binary classification, in Supplementary Figures S3–S7 for the multiple classification with $K = 3$, and in Supplementary Figures S8–S10 for the Poisson distribution, respectively.

Study 1 investigates the effect of different sample sizes for the binary classification. Figure 2 shows that the misclassification rates of all methods have decreased with an increasing number of sample sizes. ZIPLDA performs significantly better than the other methods in all settings, especially for small number of genes. Figure 3 shows that the misclassification rates of all methods are decreased with an increasing number of differentially expressed genes. ZIPLDA shows its superiority over the other methods in Study 2. Study 3 shows the impact of the number of genes on the misclassification rate. From Figure 4, we note that an increasing number of genes will lead to a lower misclassification rate and the proposed method again outperforms the other methods, especially for small sample size. Supplementary Figure S1 investigates the misclassification rate of different overdispersion ϕ. In this study, ZIPLDA is again better than PLDA, NBLDA, SVM and kNN. Supplementary Figure S2 shows the performance of all methods with different σ for two classes. From the results, it is evident that an increasing σ will lead to a lower misclassification rate and ZIPLDA is the best classifier.

Supplementary Figures S3–S7 also display the similar results for multiple classification with $K = 3$. In summary, the experimental results show that ZIPLDA performs significantly better than the other methods in all settings. We also note that, among PLDA, NBLDA, SVM and kNN, PLDA often performs better than SVM, NBLDA and kNN in most settings. Supplementary Figures S8–S10 show that ZIPLDA performs nearly the same as PLDA when the data follow the Poisson distribution without excess zeros. This demonstrates that our proposed ZIPLDA is a robust method for classification.

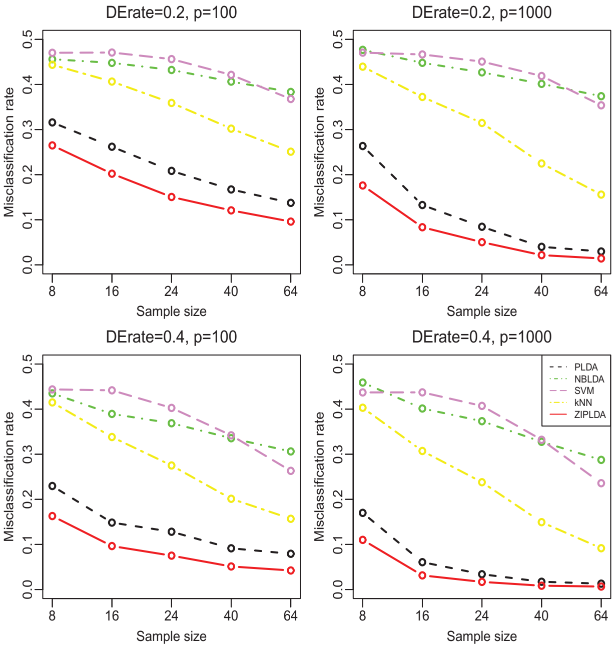

Fig. 2. The misclassification rates of all methods with different sample sizes for two classes (Study 1). Here, $\phi = 0.001$ and $\sigma = 0.2$ for all plots. The left panels have 100 features with different DE rates. The right panels have 1000 features with different DE rates

5 Application to real data

Now we apply the proposed ZIPLDA to analyze the two datasets in Section 2, the breast cancer dataset and the microRNA-seq dataset, and compare its performance with the existing methods.

Note that the majority of genes are not differentially expressed and they are irrelevant for class distinction. As one example, we

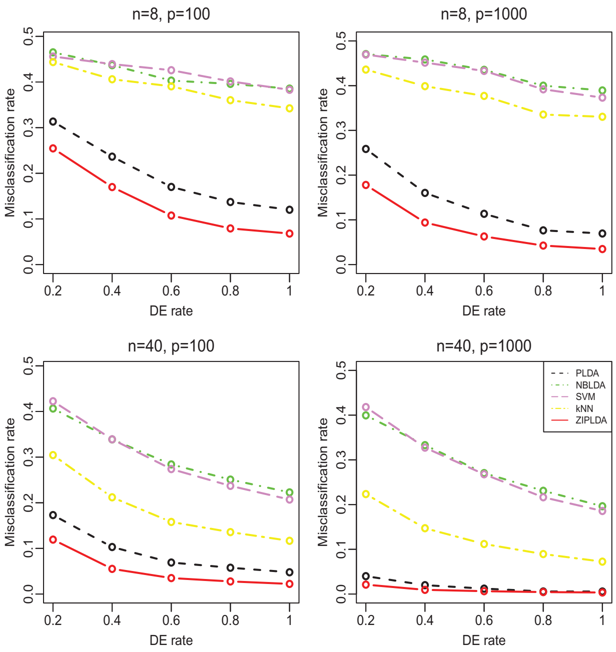

Fig. 3. The misclassification rates of all methods with different DE rates for two classes (Study 2). Here, $\phi = 0.001$ and $\sigma = 0.2$ for all plots. The left panels have 100 features with different sample sizes. The right panels have 1000 features with different sample sizes

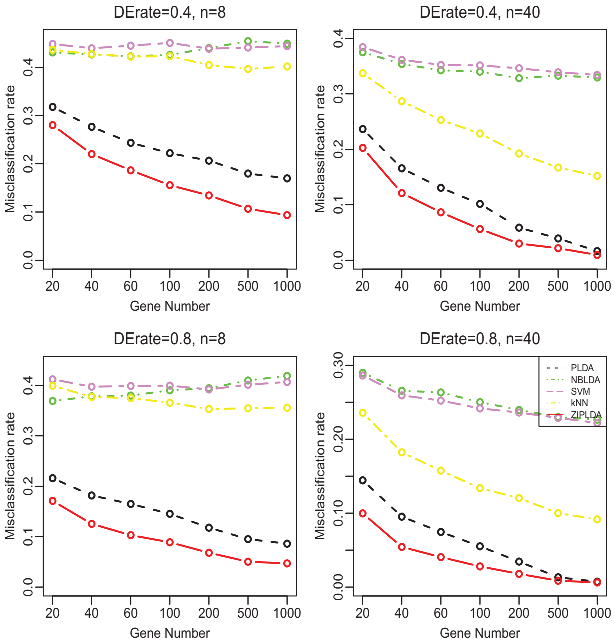

Fig. 4. The misclassification rates of all methods with different gene numbers for two classes (Study 3). Here, $\phi = 0.001$ and $\sigma = 0.2$ for all plots. The top panels have a same DE rate 0.4 with different sample sizes. The bottom panels have a same DE rate 0.8 with different sample sizes

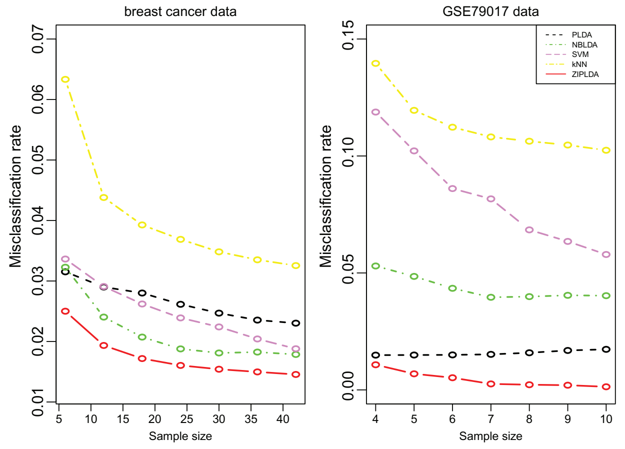

Fig. 5. The misclassification rates of all five methods for the breast cancer dataset and the microRNA-seq dataset, respectively

observe in Figure 3 that the large rate of irrelevant genes for class distinction will reduce the accuracy of the classifiers. To improve the classification performance and save the computation time, we conduct a gene selection to remove the irrelevant genes as the first step. There are two common approaches to remove the irrelevant genes for class distinction. The first one is based on the test method, such as edgeR (Robinson and Smyth, 2007; Robinson *et al.*, 2010), to select differentially expressed genes. The second one is to select the top differential genes by computing the ratio of the sum of squares between groups to within groups for each gene (Dudoit *et al.*, 2002). These two ways are comparable for the binary classification. But for the multiple classification, the first one needs the pairwise comparison and a new criterion to choose the top differentially expressed genes while the second one can be easily implemented by calculating the ratio for gene *j* using the following formula:

$$\mathrm{BW}(j)=\frac{\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\bar{x}_{k.j}-\bar{x}_{..j})^2}{\sum_{k=1}^{2}\sum_{i=1}^{n_k}(x_{kij}-\bar{x}_{..j})^2},$$

where $\bar{x}_{..j}$ is the averaged expression values across all samples and $\bar{x}_{k.j}$ is the averaged expression value across samples belonging to class k. The top p genes are selected with the largest BW ratios. Therefore, we use the ratio of the sum of squares between groups to that of within groups to select top differentially expressed genes in this paper.

To compare the performance of all classification methods, we randomly draw one half of the samples from each class to build the training set, and regard the other half as the test set. We repeat this procedure 1000 times and report the average misclassification rates for each method. We select the top 50 microRNAs from the microRNA-seq dataset and the top 200 transcripts from the breast cancer dataset (the dimensionality of the breast cancer dataset is much higher than that of the microRNA-seq dataset) and show the results in Figure 5. From the figure, it is clear that the performance of ZIPLDA is consistently better than those of the other methods for different sample sizes. Among them, NBLDA performs the best in the breast cancer dataset and PLDA performs the best in the microRNA-seq dataset. *k*NN performs the worst in the two datasets.

We further compare these five methods with respect to the different number (P) of selected genes using the two datasets and display the results in Supplementary Tables S1 and S2. From the results, we note that ZIPLDA outperforms the other methods for different P values. The misclassification rates of *k*NN are higher than all other methods for different P values. From the real data analysis, we conclude that ZIPLDA is a robust method and performs better than other methods in the presence of excess zeros.

6 Discussion

Classification of different disease types with RNA-seq data is of great importance in medical research, such as disease diagnosis and drug discovery. In this paper, we propose a zero-inflated Poisson logistic discriminant analysis (ZIPLDA) for RNA-seq data in the presence of excess zeros. Our proposed work has two main contributions: (i) addressing the case of excess zeros in the classifier and (ii) modeling the relation between the probability of zeros and the mean of the genes and the sequencing depth. In detail, we consider a mixture distribution with a point mass at zero and a Poisson distribution for the remaining data, and a logistic regression for fitting the relation between the probability of zeros and the mean of the genes and the sequencing depth.

In simulation studies, we consider both the binary classification and the multiple classification. For a fair comparison with PLDA, we essentially follow the simulation settings in Witten (2011) except for the probability of zeros. Simulation results show that our proposed method performs much better than, or at least as well as, the existing competitors in most cases. In real data analysis, we analyze two real next-generation sequencing datasets with an excess of zeros. The results indicate that ZIPLDA performed well in a wide range of settings in comparison with the existing methods.

Considering the abundant study of classification for microarray data, our study is just a pilot work on the study of classification for RNA-seq data. Although the proposed method has largely improved the existing literature for the classification of RNA-seq data, many problems remain to be solved, such as very high overdispersion in RNA-seq data, e.g. when ϕ is larger than 5. In such situations, the proposed classification method may not provide the optimal performance in practice. In view of this, we plan to develop new classification methods, including a mixture distribution with a point mass at zero and a negative binomial distribution, for analyzing RNA-seq data to further improve our current work.

Funding

Yan Zhou's research was supported by the National Natural Science Foundation of China (Grant No. 11701385), National Statistical Research Project (Grant No. 2017LY56), the Doctor Start Fund of Guangdong Province (Grant No. 2016A030310062) and the National Social Science Foundation of China (Grant No. 15CTJ008). Xiang Wan's research was supported by the Hong Kong RGC grant HKBU12202114 and the National Natural Science Foundation of China (Grant No. 61501389). Baoxue Zhang's research was supported by the National Natural Science Foundation of China (Grant No. 11671268). Tiejun Tong's research was supported by the Hong Kong Baptist University grants FRG1/16-17/018 and FRG2/16-17/074, the Health and Medical Research Fund (Grant No. 04150476) and the National Natural Science Foundation of China (Grant No. 11671338).

Conflict of Interest: none declared.

References

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

Birchler,J.A. and Kavi,H.H. (2008) Slicing and dicing for small RNAs. *Science*, **320**, 1023–1024.

Bullard,J.H. *et al*. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

Dempster,A.P. *et al*. (1977) Maximum likelihood estimation from incomplete data via the EM Algorithm. *J. R. Stat. Soc. Ser. B*, **9**, 1–38.

Dillies,M.A. *et al*. (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinf*., **14**, 671–683.

Dong,K. *et al*. (2016) NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinformatics*, **17**, 369.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Huang,S. *et al*. (2010) Bias-corrected diagonal discriminant rules for high-dimensional classification. *Biometrics*, **66**, 1096–1106.

Lambert,D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lin,B. *et al*. (2014) LFCseq: a nonparametric approach for differential expression analysis of RNA-seq data. *BMC Genomics*, **15**, S7.

Liu,F. *et al*. (2016) A zero-inflated Poisson model for insertion tolerance analysis of genes based on Tn-seq data. *Bioinformatics*, **32**, 1701–1708.

Lorenz,D.J. *et al*. (2014) Using RNA-seq data to detect differentially expressed genes. In: Datta,S. and Nettleton,D (eds.) *Statistical Analysis of Next Generation Sequencing Data*, Springer, New York, pp 25–49.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*., **15**, 550.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet*., **9**, 387–402.

Marioni,J.C. *et al*. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*., **18**, 1509–1517.

Meyer,O. *et al*. (2014) *Support Vector Machines on Large Data Sets: Simple Parallel Approaches*. Springer, New York.

Morozova,O. *et al*. (2009) Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet*., **10**, 135–151.

Mouatassim,Y. and Ezzahid,E.H. (2012) Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. *Eur. Actuarial J*., **2**, 187–204.

Ridout,M. *et al*. (1998) *Models for count data with many zeros*. In: International Biometric Conference, Cape Town.

Ripley,B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge, New York.

Robinson,M.D. and Smyth,G.K. (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, **9**, 321–332.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

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

Stefani,G. and Slack,F.J. (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol*., **9**, 219–230.

Tan,K.M. *et al*. (2014) Classification of RNA-seq data. In: Datta,S. and Nettleton,D (eds.) *Statistical Analysis of Next Generation Sequencing Data*, Springer, New York, pp 219–246.

The Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202–209.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.

Witten,D.M. *et al*. (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*, **8**, 58.

Witten,D.M. (2011) Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat*., **5**, 2493–2518.

Wolenski,F.S. *et al*. (2017) Identification of microRNA biomarker candidates in urine and plasma from rats with kidney or liver damage. *J. Appl. Toxicol*., **37**, 278–286.

Zhou,Y. *et al*. (2017a) A hypothesis testing based method for normalization and differential expression analysis of RNA-Seq data. *PLoS One*, **12**, e0169594.

Zhou,Y. *et al*. (2017b) GD-RDA: a new regularized discriminant analysis for high dimensional data. *J. Comput. Biol*., **24**, 1099–1111.