BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1105–1111
doi:10.1093/bioinformatics/btt091

Sequence analysis

Advance Access publication February 21, 2013

Poisson factor models with applications to non-normalized microRNA profiling

Seonjoo Lee[1], Pauline E. Chugh[2], Haipeng Shen[3], R. Eberle[4] and Dirk P. Dittmer[2,*]

[1]Center for Neuroscience and Regenerative Medicine, The Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD 20892, [2]Department of Microbiology and Immunology, [3]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 and [4]Department of Veterinary Patholobiology, Center for Veterinary Health Sciences, Oklahoma State University, Stillwater, OK 74078, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Next-generation (NextGen) sequencing is becoming increasingly popular as an alternative for transcriptional profiling, as is the case for micro RNAs (miRNA) profiling and classification. miRNAs are a new class of molecules that are regulated in response to differentiation, tumorigenesis or infection. Our primary motivating application is to identify different viral infections based on the induced change in the host miRNA profile. Statistical challenges are encountered because of special features of NextGen sequencing data: the data are read counts that are extremely skewed and non-negative; the total number of reads varies dramatically across samples that require appropriate normalization. Statistical tools developed for microarray expression data, such as principal component analysis, are sub-optimal for analyzing NextGen sequencing data.

Results: We propose a family of Poisson factor models that explicitly takes into account the count nature of sequencing data and automatically incorporates sample normalization through the use of offsets. We develop an efficient algorithm for estimating the Poisson factor model, entitled Poisson Singular Value Decomposition with Offset (PSVDOS). The method is shown to outperform several other normalization and dimension reduction methods in a simulation study. Through analysis of an miRNA profiling experiment, we further illustrate that our model achieves insightful dimension reduction of the miRNA profiles of 18 samples: the extracted factors lead to more accurate and meaningful clustering of the cell lines.

Availability: The PSVDOS software is available on request.

Contact: ddittmer@med.unc.edu

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

Received on September 17, 2012; revised on January 16, 2013; accepted on February 18, 2013

1 INTRODUCTION

Gene expression profiling is at the center of targeted therapy and rapid disease diagnosis. High-throughput or NextGen sequencing has recently emerged as an alternative platform to hybridization-based microarrays for the purpose of gene transcription profiling. For example, Witten (2011) claims that NextGen sequencing is 'on track to replace microarray as the technology of choice' for characterizing gene expression.

*To whom correspondence should be addressed.

NextGen sequencing data have several features that create statistical challenges. First of all, sequencing data record the number of reads between a sample and a particular region of interest, which are naturally skewed non-negative counts with a large number of zeros. Second, the nature of the sequencing experiment, such as technical sequence lane capacity, can result in different samples with dramatically different total number of sequence reads, which suggest that the samples need to be normalized in a certain way. It is well established that for high-throughput sequencing data applications, Poisson distribution represents an appropriate choice (Chen *et al.*, 2008; Jiang and Wong, 2009; Srivastava and Chen, 2010). However, the predominant form of sequencing data analysis is to forcefully transform the data and then use statistical tools that were initially developed for microarray data, which are continuous and reasonably modeled using normal distributions (with or without transformation). This leads to sub-optimal results and prompted the development of Poisson-based methods (Bullard *et al.*, 2010; Marioni *et al.*, 2008; Witten, 2011).

Different from existing Poisson-based methods for analyzing sequencing data, we focus on dimension reduction, i.e. identifying low-dimensional features or factors in the data. Genetic studies usually involve a large number of genetic markers (e.g. thousands of genes) for a small group of samples (e.g. individuals, tumor samples or virus-infected cells), which face the 'curse-of-dimensionality'. Dimension reduction is thus a desirable (and sometimes necessary) pre-processing step, and the identified features can then be used as inputs for unsupervised clustering. In this article, we specifically take into account the Poisson nature of NextGen expression profiling count data and develop a new family of Poisson factor models for efficient dimension reduction on a collection of sequencing samples. Our approach extends the earlier work of Shen and Huang (2008) to automatically address the issue of sample normalization through the use of unknown offset parameters, which are simultaneously estimated along with the underlying factors. Model identification constraints are derived and incorporated in an efficient alternating estimation algorithm. We also introduce Poisson factor models to sequencing analysis and consider follow-up clustering analysis.

The rest of the article is organized as follows. We present our model in Section 2 and develop a computationally efficient estimation algorithm. A simulation study is reported in Section 3.1 to demonstrate the performance of our algorithm, as well as

several methodological details. In Section 3.2, we then illustrate its performance through our primary motivating application—micro RNA (miRNA) profiling. miRNAs are a class of 21–25 nt non-coding RNAs that are able to post-transcriptionally regulate gene expression (O'Hara *et al.*, 2008, 2009a, b). In addition to being heavily skewed, miRNA NextGen sequencing data are also sparse because in a typical cell, <1% of all known miRNAs are expressed, and 99% are not. The numerical illustrations suggest that our method results in accurate extraction of key features from the sequencing data, which then leads to more meaningful clustering of various experimental cell lines. We conclude the article with some discussion of future work in Section 4.

2 METHODS

We first review standard factor models that can be used to analyze NextGen sequencing data, and then propose our Poisson factor model. Consider an $n \times m$ sequencing data matrix $\mathbf{Y} = (y_{ij})$ where the n rows correspond to samples (cell lines), the m columns correspond to the different genetic markers (e.g. miRNAs) and the entry y_{ij} records the read count of the jth miRNA from the ith cell line. Denote the ith row of $\mathbf{Y}$ as $\mathbf{y}_{(i)}^{\top} = (y_{i1}, \ldots, y_{im})$, referred to as the *count profile* of the ith sample.

2.1 Standard factor models

Consider the following K-factor model on the count profiles $\mathbf{y}_{(i)}$:

$$\mathbf{y}_{(i)} = \beta_{i1}\mathbf{f}_1 + \ldots + \beta_{ik}\mathbf{f}_k + \ldots + \beta_{iK}\mathbf{f}_K + \epsilon_{(i)}, \tag{1}$$

where $\mathbf{f}_k$ is the kth factor and β_{ik} is the corresponding score. For model identifiability, the factors are assumed to be orthonormal.

The factor model can be estimated through either principal component analysis (PCA) of $\mathbf{Y}$ or equivalently its singular value decomposition (SVD). PCA is a classical non-parametric linear dimension reduction technique that can be used for estimating the factor models, and the outcomes of PCA are often considered as inputs to unsupervised clustering analysis. PCA seeks to lower dimensions to a smaller number of components that capture most of the relevant structure in the data. It is particularly useful in identifying clusters of related samples, e.g. tumor subtypes based on gene expression levels, or in identifying clusters of co-regulated genes in a collection of different samples.

Not surprisingly, many investigations use PCA/SVD-based approaches in genetic studies to infer significant population or genetic structures (Holter *et al.*, 2000; Lee *et al.*, 2010; Liu *et al.*, 2003; Patterson *et al.*, 2006; Price *et al.*, 2006; Simon *et al.*, 2004; Witten *et al.*, 2010). Most of genetics modalities, such as microarray data, include systematic variations caused by non-biological sources, e.g. instrument error. For PCA to be more effectively, normalization or transformation is a common pre-processing step to reduce the non-biological variation (Bowtell and Sambrook, 2003; Cui *et al.*, 2003). Previous studies have used other variants of PCA for non-normal data, such as generalized PCA for exponential family (Collins *et al.*, 2002; Roy and Gordon, 2002) and sparse non-negative generalized PCA (Allen and Maletić-Savatić, 2011). We show that our approach augments the repertoire of tools for the analysis of extremely sparse multi-dimensional count data, such as those encountered in RNA and miRNA sequencing experiments.

Note that the raw RNAseq counts are often skewed. Hence, in practice, PCA/SVD is usually applied to transformed RNAseq data. For example, one option is to first normalize the data through the following cube-root transformation (Gentleman, 2005):

$$\widetilde{\mathbf{y}}_{(i)} = \frac{\sqrt[3]{\mathbf{y}}_{(i)} - median(\sqrt[3]{\mathbf{y}}_{(i)})}{IQR(\sqrt[3]{\mathbf{y}}_{(i)})/1.349}, \tag{2}$$

where IQR stands for the inter-quartile range. One can also use relative frequency profiles of miRNA-seq data where the miRNA count profile of each sample is divided by the total number of hit counts across all miRNA targets for that sample, i.e. the row count, and then apply SVD to the centered relative frequency data. Alternatively, one can apply quantile normalization (Bolstad *et al.*, 2003) before SVD. We refer to these methods as *SVD-Cuberoot*, *RSVD* and *QN-SVD*, respectively. Although their implementations are fairly straightforward, such transformations ignore the distributional nature of the data, and potentially can lose important features of the data (Witten, 2011). Our numerical comparisons (Section 3) show that they perform inferiorly to our proposed Poisson Singular Value Decomposition with Offset (PSVDOS) method.

2.2 Poisson factor models with offset

As discussed earlier, NextGen sequencing data exhibit special features that are not seen in hybridization microarray data, which create statistical challenges that need to be addressed. Later in the text we propose a new class of *Poisson factor models with offsets* to explicitly incorporate the special features: the Poisson count nature, the abundance of zero reads and the need for sample normalization.

2.2.1 Model We consider Poisson factor models within the generalized linear model framework and simultaneously incorporate normalization and dimension reduction. We assume that the read count y_{ij} is a Poisson random variable with rate λ_{ij}, and let $\mathbf{\Lambda} = (\lambda_{ij})$ denote the $n \times m$ hidden Poisson rate matrix. Specifically, we consider the following Poisson factor model:

$$\begin{cases} y_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \\ \lambda_{ij} = T_i\, p_{ij}, \\ \log(\lambda_{ij}) = \log(T_i) + \beta_{i1} f_{j1} + \ldots + \beta_{iK} f_{jK}, \end{cases} \tag{3}$$

where the scalar T_i is the offset parameter for the ith sample, p_{ij} is the normalized proportion of the jth miRNA in the ith *rate profile* $\lambda_{(i)}^{\top} = (\lambda_{i1}, \ldots, \lambda_{im})$, $\log(\cdot)$ is the canonical link function for Poisson variables used in generalized linear model, β_{ik} is the kth factor score for the ith rate profile and $\mathbf{f}_k = (f_{1k}, \ldots, f_{mk})^{\top}$ is the kth factor. For the identifiability of Model (3), the factors need to satisfy some constraints as discussed in the Supplementary Material.

2.2.2 Maximum-likelihood estimation To estimate Model (3), we propose to maximize the corresponding Poisson likelihood. The offset parameters, the factors and their scores are all unknown, which makes direct likelihood maximization over all the unknown parameters challenging. Hence, we consider an alternating maximum-likelihood algorithm to estimate the parameters.

Assuming that the factors $\mathbf{f}_k$ are known, for each row i, we can estimate the offset parameter T_i and the factor scores β_{ik} by fitting a log-linear Poisson regression model with the ith count profile $\mathbf{y}_i$ as the response, the factors $\mathbf{f}_k$ as the covariates and the offset parameter T_i as the intercept. The parameters are estimated via iteratively re-weighted least squares (IRLS), which has nice convergence properties (McCullagh and Nelder, 1989). Then, given T_i and the factor scores β_{ik}, we estimate the factors $\mathbf{f}_k$ by fitting a log-linear Poisson regression model with the jth column count profile $\mathbf{y}_j = (y_{1j}, \ldots, y_{nj})^{\top}$ as the response, T_i as the fixed offset and the factor scores as the covariates. The identifiability constraints on the factors are incorporated into the IRLS algorithm.

The previous discussion suggests the following iterative algorithm for estimating the Poisson factor Model (3). We refer to the algorithm as *PSVDOS*, in the sense that the algorithm extends the SVD algorithm for fitting the standard factor Model (1) to incorporate Poisson distributions with offset parameters. Note that, although PCA and SVD make no distributional assumptions of the data $\mathbf{Y}$, there exist some theoretical

justifications for using PCA and SVD when the data are approximately normally distributed: the SVD estimates then are in fact the maximum-likelihood estimates of Model (1) (Gabriel and Zamir, 1979). The alternating algorithm increases the likelihood function at each iteration and is guaranteed to converge because of the convexity of the optimization criterion at every step. At the end of each iteration, we apply SVD to $\mathbf{BF}^\top$, where $\mathbf{B} = \{\beta_{ik}\}_{i=1,\ldots,N,k=1,\ldots,K}$ and $\mathbf{F} = (\mathbf{f}_1, \cdots, \mathbf{f}_K)$. This process ensures the uniqueness and the orthogonality of the updated components. The code is written in R (Chambers *et al.*, 1992), using the built-in *glm* function. In our numerical studies, the algorithm converges within 30 iterations on average.

The PSVDOS algorithm

Initialize

- Apply SVD to the row-centered $\log(\mathbf{Y})$ to obtain the first K right singular vectors: $\mathbf{v}_k$, $k = 1, \ldots, K$; Set $\mathbf{f}_k^{old} = \mathbf{v}_k$;

Iterate

(1) Fit n log-linear Poisson regression models with $\mathbf{y}_{(i)}$ as the response and $\mathbf{f}_k^{old} (k = 1, \ldots, K)$ as the covariates to obtain the estimates for T_i and the factor scores β_{ik}, denoted as T_i^{new} and β_{ik}^{new}; denote $\mathbf{B} = (\beta_1^{new}, \cdots, \beta_K^{new})$ with $\beta_k^{new} = (\beta_{1k}^{new}, \ldots, \beta_{nk}^{new})^\top$;

(2) Fit m log-linear Poisson regressions with $\mathbf{y}_j$ as the response, T_i^{new} as the fixed offset and $\beta_k^{new} (k = 1, \ldots, K)$ as the covariates to obtain the updated estimates for $\mathbf{f}_k$, denoted as $\mathbf{f}_k^{new}$; denote $\mathbf{F} = (\mathbf{f}_1^{new}, \cdots, \mathbf{f}_K^{new})$;

(3) Center each row of the matrix $\mathbf{BF}^\top$ and apply SVD to the row-centered matrix to obtain the first K left singular vectors $\mathbf{v}_k$; Set $\mathbf{f}_k^{new} = \mathbf{v}_k$;

(4) Repeat from Step 1 with $\mathbf{f}_k^{old} = \mathbf{f}_k^{new}$ until convergence.

We make three comments regarding the offset parameters and selection of the number of factors. First, the row-centering in Step 3 enforces the identifiability of the offset parameters. See Supplementary Materials for details. Second, sometimes it makes sense to assume the offsets as known from *a priori* knowledge. For example, one can treat the total read count of a sample as the offset. In that case, there is no need to update or estimate the offsets as part of the aforementioned PSVDOS algorithm. Finally, in practice, the number of factors K needs to be selected in a data-driven fashion. We propose to use the deviance reduction-based approach suggested by Shen and Huang (2008). More details are given in Section 3.1.

3 RESULTS

We illustrate the performance of our proposed PSVDOS method through a simulation study (Section 3.1) and an analysis of an miRNA-sequencing dataset (Section 3.2). We compare PSVDOS with five other SVD-based methods:

- SVD-Raw: first subtract each entry by the mean of each row, and then apply SVD to the row centered raw data;
- SVD-Cuberoot: first apply the transformation (2) to take cube-root of each entry, and then apply SVD to the transformed data;
- RSVD: first divide each entry by total count of each row, and then apply SVD to the obtained relative frequency data matrix;
- QN-SVD: first apply quantile normalization (Bolstad et al., 2003) to each row, and then apply SVD to the obtained normalized data matrix after row-centering;
- PSVD: apply Poisson SVD of Shen and Huang (2008) to the raw data, which ignores the existence of offsets.

The comparison will illustrate the shortcomings of ignoring the Poisson count nature of the data, as well as the necessity of incorporating sample-specific scaling effect through offsets. Both numerical studies suggest that PSVDOS performs the best. We also show that our data-driven approach can select the number of underlying factors accurately and stably, and that the PSVDOS algorithm can estimate the offset parameters accurately.

3.1 Simulation study

Data generation We generate a synthetic miRNA-seq dataset according to Model (3). The Poisson rate matrix $\mathbf{\Lambda}$ follows

$$\log(\mathbf{\Lambda}) = \log(\mathbf{T}) + \log(\mathbf{P}) = \log(\mathbf{T}) + \mathbf{USV}^\top,$$

where the offset matrix $\mathbf{T} = (T_1, \ldots, T_n)^\top \otimes \mathbf{1}_m^\top$, and the proportion matrix $\mathbf{P}$ (in log-scale) follows a four-factor SVD model, with the $n \times 4$ left singular vector matrix $\mathbf{U}$, the $m \times 4$ right singular vector matrix $\mathbf{V}$ and the 4×4 diagonal singular value matrix $\mathbf{S}$ containing the four positive singular values as its diagonal entries.

We simulate $n = 40$ different samples measured on $m = 200$ miRNAs. Figure 1a displays the true offsets in the log-scale, which are generated as follows: in the log-scale, they fall into six clusters with distinct cluster means (gray dotted lines) and are uniformly distributed around the cluster mean within each cluster; the overall mean across the clusters is 4 (solid line). Each offset's cluster membership is displayed with different colors and markers.

The four diagonal elements of $\mathbf{S}$, i.e. the singular values, are $s_1 = 70, s_2 = 25, s_3 = 15, s_4 = 1$. The four columns of $\mathbf{U}$ and $\mathbf{V}$, i.e. the left and right singular vectors, are plotted in the columns of Figure 1b and c, respectively. Each row in Figure 1b corresponds to one sample, which depicts four different clustering patterns in the samples. Figure 1c shows the heat map of the right singular vectors in the 200×4 matrix $\mathbf{V}$ or the miRNA factor profiles, embedded with certain clustering patterns. Each column indicates one miRNA factor profile, and each row corresponds to one miRNA. Additional details regarding the data generation can be found in the Supplementary Materials.

Figure 2a displays the true $\log(\mathbf{P})$ used in the simulation using a blue (negative)–red (positive) color coding. The color bar on the left side of the heat map shows the clustering membership of the rows, which fall into six clusters as indicated by the six colors (red, cyan, green, blue, magenta and yellow). The clustering pattern is obtained by applying the complete linkage hierarchical clustering analysis on the rows (Eisen *et al.*, 1998; Wilkinson and Friendly, 2009). Similarly, the color bar above the heat map shows the cluster membership of the miRNAs, as represented by the color coding and the corresponding clustering dendrogram plot.

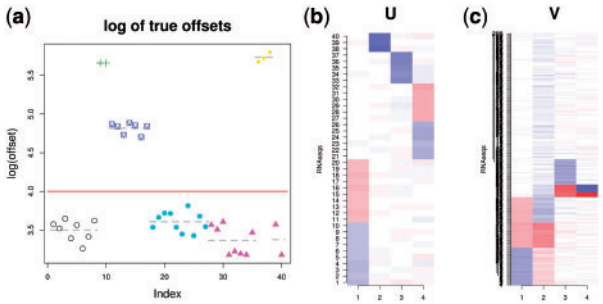

Fig. 1. Plots of simulation parameters. (**a**) True offsets in the log-scale, forming six clusters. (**b**) True left singular vectors **U**, indicating sample clusters. (**c**) True right singular vectors **V**, indicating miRNA clusters

Low-rank approximation and clustering All six methods are applied to the synthetic data to extract the first four SVD components, which result in best-rank-four approximation for the underlying signal, as plotted in Panels b–g of Figure 2. We superimpose the hierarchical clustering results based on the rows and columns, respectively, of the estimated signal matrices. To better illustrate the accuracy of the clustering, we use the ordinal cluster membership contained in log(**P**) to color code the rows and the columns in the dendrogram plots. PSVDOS (Panel b) offers the best approximation to the true signal, and it gives the most accurate clustering result, which is as expected, as the dataset was designed in a way that takes advantage of the unique features of PSVDOS.

Figure 2c shows that PSVD incorrectly separates red and purple clusters, and it cannot separate the blue and green clusters. As the original PSVD algorithm does not take into account abundance sequence depth, the variation of the total number of reads was reflected in the first components. SVD-Raw in Figure 2d performs badly overall, as it is driven by absolute abundance. As the data were generated from Poisson distribution, exponential relationship is naturally imposed. Thus, in the absence of any normalizing previous transformation, SVD cannot recover the underlying pattern. SVD-Cuberoot in Figure 2e improves on SVD-Raw by reducing the influence of estimated sample/miRNAs, but the clustering result of the rows differs significantly from the truth signal. RSVD (Fig. 2f) fails to separate the blue and green clusters. QN-SVN in Figure 2g performs well for bi-clustering, but worse than the PSVDOS. This indicates that quantile normalization is not sufficient to overcome the extreme differences in sequence depth, even though it does take into account abundance to improve the clustering results. Heatmaps of the estimated singular vectors can be found in the supplementary document.

Selection of number of factors and offset parameter estimation We use the deviance reduction plot as suggested by Shen and Huang (2008) to choose the number of underlying factors, which is a likelihood-based extension of the screen plot. Figure 3a displays the deviance reduction by the number of factors for one particular simulated dataset (dotted line), which shows how much additional data 'variability' can be explained by every extra factor. The elbow of the deviance reduction plot suggests that four factors are sufficient for modeling the data. To

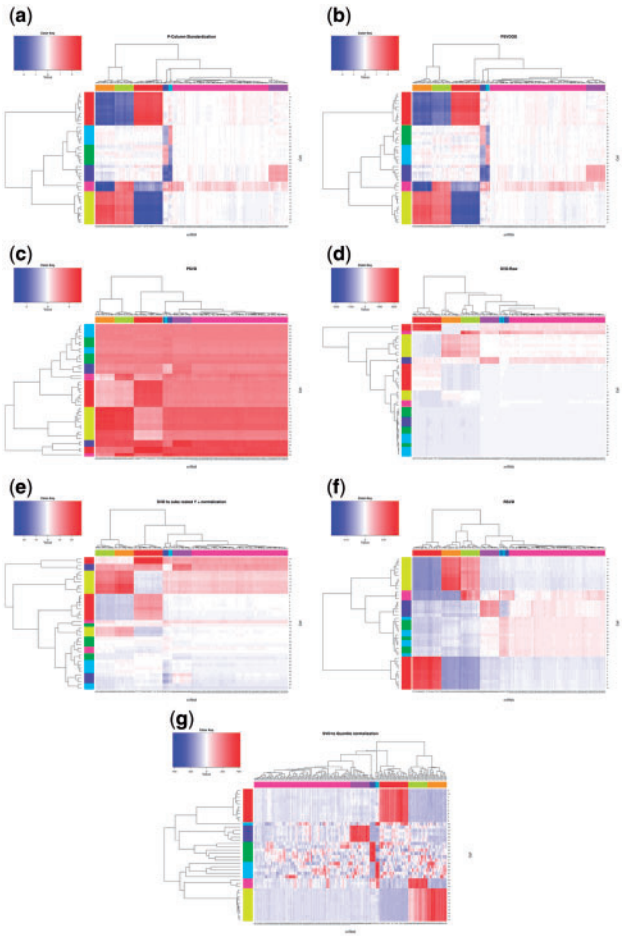

Fig. 2. Heatmaps of the truth and various estimates: (**a**) True log(**P**); (**b**) PSVDOS; (**c**) PSVD; (**d**) SVD-Raw; (**e**) SVD-Cuberoot; (**f**) RSVD; and (**g**) QN-SVD. Hierarchical clustering is performed on the rows (or samples) and the column (or miRNAs) separately. The color bars next to the rows (or columns) represent the original clustering membership of each row (or column). PSVDOS performs best in approximating the true signal and clustering

better illustrate the ignorable contribution of the additional factors, a zoomed-in version of the plot is included in the Supplementary Material.

This way of selecting the number of factors is stable. We repeated the simulation 100 times and obtained the deviance reduction plot for each simulation run. The horizontal gray lines represent the pointwise 95% intervals of the 100 obtained deviance reduction plots (gray lines) by the numbers of factors. The deviance reduction plots all have an elbow when the model includes four factors.

We now demonstrate that PSVDOS can accurately estimate the offset parameters, the T_i's and their overall mean (in log-scale). For each of the 100 simulation runs, we applied PSVDOS with four factors (as suggested by the deviance reduction plot) to obtain the estimates for $\log(T_i)$. For each $i = 1, \ldots, 40$, we calculated the 2.5 and 97.5% quantiles of $\log(\hat{T}_i) - \log(T_i)$, which provide the empirical 95% confidence intervals (CI) for the differences, as depicted by the vertical lines

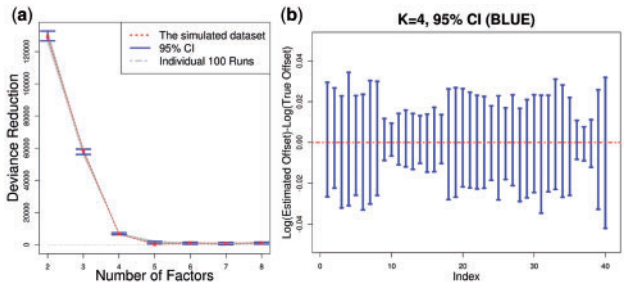

Fig. 3. Simulation study. (**a**) Deviance reduction by the number of factors, suggesting four factors; (**b**) 95% CI of $\log(\hat{T}_i) - \log(T_i)$

of Figure 3b. All CIs contain the value zero (the dotted line). Similar 95% CI is plotted for the overall mean in the Supplementary Materials.

We also investigated how the bias of the offset estimates depends on the number of factors K included in the model. As shown in the Supplementary Materials, the estimates seem to be biased when K is under-estimated, i.e. <4, whereas the estimates become less biased as K increases; eventually when K is ≥4, the offset estimates do not have any bias.

3.2 Analysis of miRNA-sequencing data

We first illustrate the application of PSVDOS to NextGen-based miRNA expression profiles of different virus-infected samples (the validity of our approach will be further established through a simulation study in Section 3.1.). The miRNA-sequencing data were collected on a series of samples that were infected with human and non-human primate herpesviruses. One such virus, the monkey B virus, is fatal to humans, whereas its relative, the human herpes simplex virus, only causes cold sores in 99% of the infections. These experiments (to be described elsewhere) were designed to identify novel biomarkers that can differentiate between harmless and fatal exposures. Because viral infection is deduced based on the transcription profile of hot miRNA, this indirect approach is particularly useful if the infecting virus is not known or even entirely novel.

In this particular application, we want to test the hypothesis (i) that the cells are infected by a virus of the family α-herpesviruses, and (ii) does the infecting virus have biological consequences similar to herpes simplex virus, in which case, the patient develops minor skin ulcerations, i.e. cold sores, or to monkey B virus, in which case the patient dies within 7 days. PSVDOS sensitively revealed meaningful clusterings among the samples, as well as the corresponding miRNA markers that can be potentially used to differentiate the samples.

A key characteristic of this dataset is the even greater sparsity of the data and more limited depth of the count data compared with the simulated dataset in Section 3.1 because of the expense associated with comprehensive NextGen sequencing. This makes for an important practical role. PSVDOS outperformed other methods under generous experimental constraints. The ratio of samples to targets was 40:200 or 1:5. Under those experimental parameters, PSVDOS recovered the 'true' data structure. As there are >2000 human miRNAs, a corresponding experiment would require at least 400 samples. Such large number of samples can be obtained and sequenced only by large consortia, e.g. Hudson *et al*. (2010). Most published studies use 1:50 ratios, for which we expect the practical benefit of PSVDOS compared with previous methods to be even more pronounced.

Data description Briefly, the Illumina small RNA kit v1.5 was used to establish cDNA libraries of small RNAs from human fibroblasts infected with either human or non-human primate α-herpesviruses. These were the human herpes simplex virus 1 and 2 (HSV-1, HSV-2) and their homologous primate viruses for baboons, squirrel monkeys and macaques. The small RNAs libraries were then sequenced using the Illumina platform for single-end sequencing. As additional known controls, we used HUVEC and CHME cells, which were either mock infected or infected with an irrelevant, widely divergent virus, namely, West Nile virus (WNV). We used a slightly different method of isolation and purification for each cDNA library of small RNAs, as we tried different versions of the manufacturer's kit. This is clearly not an ideal experimental design, as it introduces additional technical variation, but one that is typical for experimental science. Further samples were lymphoma (PEL), human tonsil and another cell line. The resulting reads for each sample were aligned to a human miRNA database (Kozomara and Griffiths-Jones, 2011).

The read counts for each of 398 miRNAs were obtained for each sample of cells infected with human and non-human primate α-herpesviruses (HSV-1, HSV-2, HVP2, SQHV, BV, SA8 and ChHV) and other control and infected cell lines and tissues (PEL-A, PEL-B, fibroblasts, tonsil, Control1, Control2, Control3, CHMEinfected, CHME5mock, HUVECinfected and HUVECmock). As the dataset was sparse indeed, in that many miRNAs have zero or small number of reads, only miRNAs with the total count over all the cell lines >15 are included in the analysis, which means that 265 miRNAs are used in the analysis reported later in the text.

For these 265 miRNAs, the total counts of each cell line are displayed in Table 1, along with the number (%) of miRNAs with zero reads. Note that 50% of miRNAs had zero counts. This is typical for the biology of miRNA expression. The group of HSV-infected cells have ~3000–4000 total counts, the cells in the other group have mildly varying total counts, ~30 000–80 000 and others have total counts >900 000. The heterogeneity among the total counts, i.e. wide spread, is one of the features that PSVDOS is designed to accommodate.

Dimension reduction and clustering As suggested by the deviance reduction plot (not shown here, also see Section 3.1), we extracted three factors using the methods except PSVD, which could only produce estimates for the first two factors. The legend in Figure 4 lists the cell lines numbered and colored according to their correct grouping. Panels (af) then display the scatter plots of the extracted factor scores for each method, respectively, using the corresponding numbers and colors provided in the legend. We then applied complete linkage hierarchical clustering (Eisen *et al*., 1998; Wilkinson and Friendly, 2009) to the cell lines based on the dimension reduced data, setting the number of clusters as *six*. The dendrograms plots are displayed on top of the corresponding scatter plots, where the leaves are colored according to the clustering results, whereas the labels are colored according to

Table 1. Total counts of 18 samples and the numbers of miRNAs with zero counts

Group	ID	Total counts[a]	No. of miRNAs with 0 counts
HSV	1. HSV-1	3736	128 (48.30%)
	2. HSV-2	4621	135 (50.94%)
	3. HVP2	2526	149 (56.23%)
	4. SQHV	3266	127 (47.92%)
	5. BV	3031	154 (58.11%)
	6. SA8	2580	136 (51.32%)
	7. ChHV	3083	142 (53.58%)
Other	8. CHMEinfected[b]	2 217 495	74 (27.92%)
	9. CHME5mock[b]	4 379 718	73 (27.55%)
	10. PEL-A	77 494	116 (43.77%)
	11. PEL-B	61 041	130 (49.06%)
	12. HUVECinfected[b]	908 577	101 (38.11%)
	13. HUVECmock[b]	1 564 148	97 (36.60%)
	14. Fibroblasts	1973	143 (53.96%)
	15. Tonsil	35 263	125 (47.17%)
	16. Contol1	41 807	125 (47.17%)
	17. Control2	46 819	137 (51.70%)
	18. Control3	35 185	137 (51.70%)

[a]Total counts also represent sequence depths. [b]Samples that were prepared using the Illumina small RNA library preparation kit v1.0 (versus kit v1.5). Herpes virus sample abbreviations are HSV-1 (human herpesvirus 1), HSV-2 (human herpesvirus 2), HVP2 (herpesvirus papio 2), SQHV (squirrel monkey herpesvirus), BV (macaque herpes B virus), SA8 (simian agent 8), ChHV (chimpanzee herpesvirus).

the legend, i.e. the correct grouping. Hence, a miss match between the colors of the leaf and the label would mean wrong clustering.

The plots suggest that PSVDOS correctly clusters all the HSV-infected cells together and groups the other cells according to their true subtypes. The fibroblasts sample is clustered with the HSV-infected samples. This actually represents the expected biological results, as the HSV samples represent fibroblasts that were infected with the different herpes simplex viruses, whereas the other samples stem from different tissues of origin. This is also apparent from the similarity in terms of the numerical summaries in Table 1 between the HSV samples and the fibroblasts sample. Generally, miRNA profiles tend to be linked to tissue of origin.

By contrast, the other five methods give inferior clustering results for these samples. PSVD (Panel b) gives the second best clustering performance; however, it groups one control sample together with the tonsil sample. SVD-Raw (Panel c) lacks clear clustering. SVD-Cuberoot (Panel d), RSVD (Panel e) and QN-SVD (Panel f) split the control group samples into different clusters; in addition, the HUVEC and CHME5 cells that are infected with the same virus are not clustered together using the RSVD method. Taken together, these clustering results reveal that PSVDOS outperforms the other methods and reveals more accurate hierarchical clustering results from different sample groups of NextGen sequencing data.

Furthermore, we applied a separate cluster analysis to the eight HSV-infected cells, including the seven herpesvirus and the one fibroblast samples. The clustering results are included in Figure 4 of the Supplementary Material, which show that

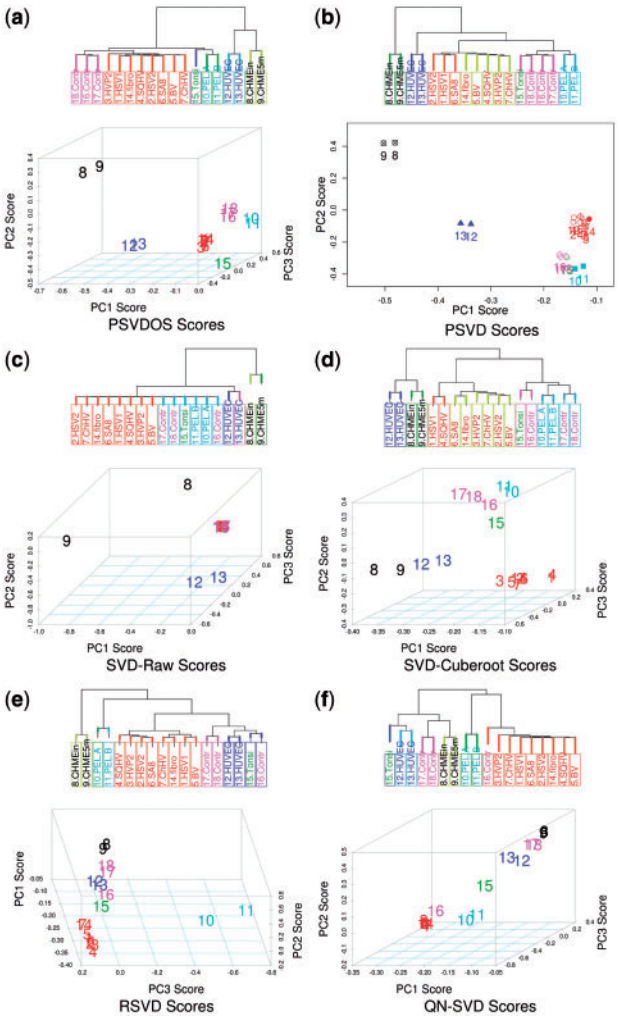

Fig. 4. Analysis of miRNA-seq data: scatter plots of the leading factor scores. PSVDOS clusters the different cell line groups perfectly

(i) PSVDOS and PSVD separate fibroblast from the other cells, whereas the other methods place fibroblast close to HSV samples; (ii) all approaches, except RSVD and QN-SVD, confirm the finding of Ohsawa *et al.* (1999) that HVP2 is more similar to BV than to HSV1 or HSV2.

4 DISCUSSION

NextGen sequencing-based mRNA and miRNA expression profiling is rapidly gaining popularity and may eventually replace other methods; however, it yields a completely different data structure, compared with hybridization-based microarray experiments. Microarray-based expression profiling data can, with some difficulty, be transformed into a dataset that has a normal distribution and is amenable to statistical tools for pattern discovery and classification; NextGen sequencing-based expression data cannot.

Expression patterns of miRNAs represent examples of the most skewed expression data that are encountered in the

biological literature. The reasons for this extreme distortion are both technical and biological, as only a handful of miRNAs tend to dominate the miRNA population within cells. It offers a practical justification to develop statistical methods for extreme data. Profiling miRNAs is a novel approach to query cell status and to classify samples, in our case, different viral infections. Hence, there exists an urgent need to develop appropriate clustering and classification approaches that take into account the particulars of these data structures.

PSVD is a factorization method to perform dimension reduction, specifically for Poisson count data. It has been applied for data dimension reduction in non-biological applications, such as call center data (Shen and Huang, 2008). In this article, we proposed an extended approach (PSVDOS) to improve unsupervised clustering of RNAseq data by incorporating offset parameters, so that necessary normalization of miRNA sequencing data can be automatically accounted for.

Using simulated data, as well as an even sparser experimental example set, that highlights variation and limitations of real-world NextGen sequencing data, we show that PSVDOS was superior to other approaches that separate normalization from dimension reduction, such as SVD on cube-rooted or relative frequency or the raw counts. Those normalization methods are commonly used to eliminate variation because of technical imperfection. PSVDOS correctly clustered samples on the basis of NextGen-derived miRNA expression profiles. This new approach should help the analysis of NextGen-based RNAseq data in general and miRNA-based classification of experimental and clinical samples in particular.

There are several future research directions worth pursuing. The current Poisson factor Model (3) uses row-specific (or cell line-specific) offset parameters T_i to normalize the cell lines. More generally, the offset parameters can be allowed to depend on both the cell line and the miRNA, e.g. denoted as T_{ij} for the ith cell line and the jth miRNA. One can then impose some two-way analysis of variance model on the offsets to model effects of cell lines and miRNAs, such as

$$\log(T_{ij}) = \mu + \alpha_i + \gamma_j,$$

subject to some identifiability constraints, such as $\sum_i \alpha_i = \sum_j \gamma_j = 0$. Interaction terms can also be included if necessary.

Our framework makes use of the Poisson distributional nature of the sequencing data through the Poisson likelihood function. The likelihood approach is general and flexible enough that it can be extended to model other distributions. For example, researchers have noticed that some sequencing count data exhibit overdispersion with respect to Poisson distribution and propose to use negative-binomial distribution instead. See, among others, Anders and Huber (2010) and Robinson and Oshlack (2010) in the context of supervised clustering. An interesting direction for future research is to develop an appropriate factor model to address the overdispersion.

Funding: NIH (DE018304, CA019014 to D.P.D.) (in part); NIH/NIDA (1 RC1 DA029425-01); the Xerox Foundation UAC Award; NSF (CMMI-0800575, DMS-1106912 to H.S.) (in part) and PHS (P40 RR12317 to R.E.).

Conflict of Interest: none declared.

REFERENCES

Allen,G. and Maletić-Savatić,M. (2011) Sparse non-negative generalized PCA with applications to metabolomics. *Bioinformatics*, **27**, 3029–3035.

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

Bolstad,B. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bowtell,D. and Sambrook,J. (2003) *DNA Microarrays: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Pr, Cold Spring Harbor, New York, USA.

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

Chambers,J. *et al.* (1992) *Statistical Models in S*. Chapman & Hall, London.

Chen,W. *et al.* (2008) Mapping translocation breakpoints by next-generation sequencing. *Genome Res.*, **18**, 1143–1149.

Collins,M. *et al.* (2002) A generalization of principal component analysis to the exponential family. *Adv. Neural Inf. Process. Syst.*, **1**, 617–624.

Cui,X. *et al.* (2003) Transformations for cDNA microarray data. *Stat. Appl. Genet. Mol. Biol.*, **2**, Article 4.

Eisen,M. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gabriel,K. and Zamir,S. (1979) Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, **21**, 489–498.

Gentleman,R. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Holter,N. *et al.* (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.

Hudson,T. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Jiang,H. and Wong,W. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D152–D157.

Lee,M. *et al.* (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 1087–1095.

Liu,L. *et al.* (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

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

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models*, Vol. 37. Chapman & Hall/CRC, London, UK.

O'Hara,A. *et al.* (2008) Gene alteration and precursor and mature microRNA transcription changes contribute to the miRNA signature of primary effusion lymphoma. *Blood*, **111**, 2347–2353.

O'Hara,A. *et al.* (2009a) Pre-micro RNA signatures delineate stages of endothelial cell transformation in Kaposi sarcoma. *PLoS Pathogens*, **5**, e1000389.

O'Hara,A. *et al.* (2009b) Tumor suppressor microRNAs are underrepresented in primary effusion lymphoma and Kaposi sarcoma. *Blood*, **113**, 5938–5941.

Ohsawa,K. *et al.* (1999) Herpesvirus papio 2: alternative antigen for use in monkey b virus diagnostic assays. *Comp. Med.*, **49**, 605–616.

Patterson,N. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Price,A. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

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

Roy,N. and Gordon,G. (2002) Exponential family PCA for belief compression in POMDPs. *Adv. Neural Inf. Process. Syst.*, **15**, 1667–1674.

Shen,H. and Huang,J. (2008) Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Ann. Appl. Stat.*, **2**, 601–623.

Simon,R. *et al.* (2004) *Design and Analysis of DNA Microarray Investigations*. Springer, New York.

Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.

Wilkinson,L. and Friendly,M. (2009) The history of the cluster heat map. *Am. Stat.*, **63**, 179–184.

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

Witten,D. *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 Biol.*, **8**, 58.