Cronfa - Swansea University Open Access Repository

Paper:

Baghfalaki, T., Ganjali, M. & Berridge, D. (2016). Missing value imputation for RNA-sequencing data using statistical models: A comparative study. *Journal of Statistical Theory and Applications, 15*(3), 221-236.

Journal of Statistical Theory and Applications, Vol. 15, No. 3 (September 2016), 221-236

Missing Value Imputation for RNA-Sequencing Data Using Statistical Models: A Comparative Study

Taban Baghfalaki

Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University
School of Biological Science, Institute for Research in Fundamental Sciences (IPM)
Tehran, Iran
t.baghfalaki@modares.ac.ir

Mojtaba Ganjali

Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University
School of Biological Science, Institute for Research in Fundamental Sciences (IPM)
Tehran, Iran
M-Ganjali@sbu.ac.ir

Damon Berridge

Farr Institute-CIPHER, College of Medicine, Swansea University
Swansea, Wales, U.K.,
d.m.berridge@swansea.ac.uk

Received 13 August 2013

Accepted 4 April 2016

RNA-seq technology has been widely used as an alternative approach to traditional microarrays in transcript analysis. Sometimes gene expression by sequencing, which generates RNA-seq data set, may have missing read counts. These missing values can adversely affect downstream analyses. Most of the methods for analysing the RNA-seq data sets require a complete matrix of RNA-seq data. In the past few years, researchers have been putting a great deal of effort into presenting evaluations of the different imputation algorithms in microarray gene expression data sets, However, these are limited works for RNA-seq data sets and a comparative study for investigating the performance of the missing value imputation for RNA-seq data is essential. In this paper, we propose the use of some parametric models such as Regression imputation, Bayesian generalized linear model, Poisson mixture model, EM approach , Bayesian Poisson regression, Bayesian quasi-Poisson regression and the Bootstrap version of two latter for single imputation of missing values in RNA-seq count data sets. The approaches are also applied for identifying differentially expressed genes in the presence of missing values. Multiple imputation, proposed by Rubin (1978), is also used for multiple imputation of missing RNA-seq counts. This approach allows appropriate assessment of imputation uncertainty for missing values. The performance of the single and multiple imputations are investigated using some simulation studies. Also, some real data sets are analyzed using the proposed approaches.

Keywords: Bayesian approach; Clustering analysis; EM algorithm; Missing data analysis; RNA-seq data set.

2000 Mathematics Subject Classification: 62P10, 60E99, 92D99

Published by Atlantis Press

1. Introduction

RNA-sequencing (RNA-seq) count data has became increasingly more popular than the traditional microarray data in gene expression analysis. The nature of microarray data sets is different to that of RNA-seq data. The microarray data are the measured intensities and they are continuous variables, but the RNA-seq data measure the gene expression by way of a count of reads. Count data need reliable statistical models which are different to those used for analyzing microarray data sets.

Analysing data in the presence of missing values has always been a challenging problem. The same as many other experimental data sets, RNA-seq data from DNA-sequencing technology often contain missing values (Chu, 2014). In spotted cDNA microarrays, each spot on the array is often assigned to a unique gene, so if missing value occurs in some spots, it would lead to the loss of information for that particular gene (Bras and Menezes, 2007). Ignoring the missing cases is not a good strategy to deal with missing values. There are many techniques for handling missing data in the literature (see, for example, Little and Rubin, 2002), where in general these approaches can be divided into two main categories: single imputation and multiple imputation (Little and Rubin, 2002).

Rubin (1976) provided a framework for handling incomplete data by introducing the important taxonomy of missing data mechanisms which consists of: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). A mechanism is said to be MCAR if missing values are independent of both unobserved and observed data, and MAR if, conditional on the observed data, the missing values are independent of the missing measurements and otherwise the missing process is termed MNAR.

In microarray gene expression data, there are many reasons for missing values, and these include hybridization failures, low resolution, artifacts on the microarray itself, image noise, corruption, and problems related to the spotting process (Yang et al., 2002; Moorthy et al., 2014; Hourani and Emary, 2009; Jornsten et al., 2005). There are many articles in the literature which discuss different approaches for imputing missing values in the microarray gene expression matrix (see, for examples, Troyanskaya et al., 2001; Oba et al., 2003; Kim et al., 2005; Sehgal et al., 2005; Zhang et al., 2008; Liew et al., 2010; Aittokallio, 2010; Moorthy et al., 2014)

Many algorithms for RNA-seq analysis require a complete data matrix as input. Therefore, in order to analyze the available data, missing values have to be imputed. Some approaches have been proposed for imputing missing values in RNA-seq data sets. Some examples are: replacing missing values with zeros, replacing missing values with averages, replacing missing values using the k nearest neighbor method (Chu, 2014) and replacing missing values with the SVD imputation method (Troyanskaya et al., 2001). Most of these are distribution free approaches which are proposed for handling missing values in microarray gene expression data sets, but they do not consider the countability property of RNA-seq data.

One appropriate distribution for modeling RNA-seq count data is the Poisson distribution (Marioni et al., 2008, Bullard et al., 2010 and Si et al, 2014). This distribution is used as the main distributional assumption for handling RNA-seq data set in this paper. Also, there are some references for considering overdispersion in RNA-seq count data (Lund et al., 2012). For imputation, our proposed model-based approaches are regression imputation, Bayesian generalized linear models, Poisson mixture model, EM approach, Bayesian Poisson regression, Bayesian quasi-Poisson regression, bootstrap Bayesian Poisson regression and bootstrap Bayesian quasi-Poisson regression. After imputation of missing values, we discuss identifying differentially expressed genes in the presence of missing values. Multiple imputation, which is an appropriate approach for considering the uncertainty of the sampling, is also discussed in this paper. We compare the performance of different approaches using some simulation studies and real data sets. All approaches discussed in this paper assume of handling a mechanism of MAR. In this paper, for simplicity, we only consider studies where the RNA-seq count data are generated under two conditions. The proposed approaches can be easily extended to more than two conditions.

This paper is organized as follows. In Section 2, we define the necessary notations and discuss different methods for the single imputation of RNA-seq data. In Section 3, identifying differentially expressed genes from RNA-seq data in the presence of missing values is discussed. Section 4 contains a multiple imputation approach for imputing missing RNA-seq data. Also, identifying differentially expressed genes from RNA-seq data in the presence of missing values by multiple imputation is discussed. Section 5 contains the results of some simulation studies for investigating the performance of each imputation approach. In Section 6, the proposed approaches are used for imputing missing values in two real data sets. In the final section, we present some discussion and conclusions.

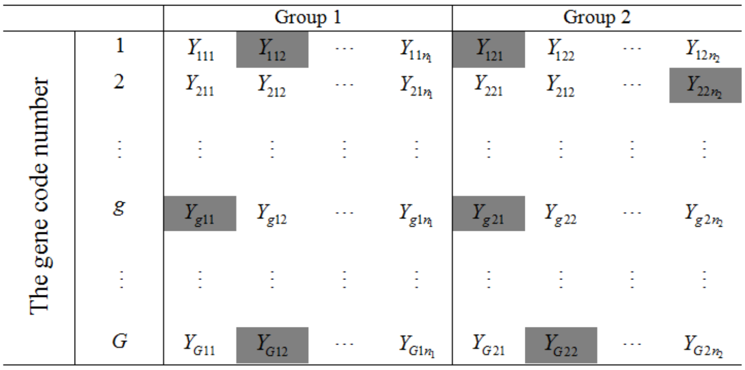
2. Different methods for imputation of missing values in RNA-seq data sets

We first define the notation that will be used throughout this paper and then introduce some methods of imputations.

The total number of genes is denoted by G, where the genes are generated under two conditions (or two groups). Each condition has n_i, $i = 1,2$, replications.

Let Y_{gir}, $g = 1,2,\cdots,G$, $r = 1,2,...,n_i$ and $i = 1,2$ be the read count for the g^{th} gene in the r^{th} iteration of condition i. Therefore, the table of read counts from an RNA-seq experiment is a $G \times (n_1 + n_2)$ matrix of non-negative integers.

We assume some genes have some missing values in their read counts. Therefore the vector of values $Y_g = (Y_{g11}, Y_{g12}, ..., Y_{g1n_1}, Y_{g21}, Y_{g22}, ..., Y_{g2n_2})'$ contains some missing values. Let Y_g be decomposed into two vectors Y_g^{obs} and Y_g^{mis}, where Y_g^{obs} denotes the observed read counts for gene g and Y_g^{mis} denotes the missing read counts for gene g. Figure 1 shows the notation which will be used in this paper and a possible pattern of missingness in the data set.



		Group 1				Group 2			
The gene code number	1	Y_{111}	Y_{112}	$\cdots$	Y_{11n_1}	Y_{121}	Y_{122}	$\cdots$	Y_{12n_2}
	2	Y_{211}	Y_{212}	$\cdots$	Y_{21n_1}	Y_{221}	Y_{212}	$\cdots$	Y_{22n_2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	g	Y_{g11}	Y_{g12}	$\cdots$	Y_{g1n_1}	Y_{g21}	Y_{g22}	$\cdots$	Y_{g2n_2}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	G	Y_{G11}	Y_{G12}	$\cdots$	Y_{G1n_1}	Y_{G21}	Y_{G22}	$\cdots$	Y_{G2n_2}

Fig. 1. Notation and a possible pattern of missingness (highlighted values are missing values) in the matrix of read counts.

Regression imputation

Given some missing values in the Y_{g^*}, only the K genes $Y^1_{g^*}$, $Y^2_{g^*}$,..., $Y^K_{g^*}$ that are most correlated with Y_{g^*} are included in the prediction model. None of the $Y^1_{g^*}$, $Y^2_{g^*}$,..., $Y^K_{g^*}$ is allowed to have a missing value. Also, let $x_{g^*} = (x_{g^*1}, x_{g^*2}, ..., x_{g^*(n_1+n_2)})$ be a vector of dummy variables such that $x_{g^*r} = 1$ if the r^{th} replication of gene g^* is under condition $i = 1$ and $x_{g^*r} = 0$, otherwise. Also, let $X_g = (Y^1_{g^*}, Y^2_{g^*}, ..., Y^K_{g^*}, x_g)$ be a $(n_1 + n_2) \times (K+1)$ matrix, which is used as the covariate matrix for the g^{th} gene. We consider the following model for the g^{*th} gene:

$$Y_{g^*}|\beta_{g^*} \sim Poisson(\mu_{g^*}), \tag{2.1}$$

where $log(\mu_{g^*}) = X'_{g^*}\beta_{g^*}$. Regression imputation (also known as conditional mean imputation) replaces missing values with predicted scores from a regression equation. We use the Spearman correlation for finding the K most correlated genes. Then, the regression model is fitted to the observed data. After fitting the regression model, predicted values of the missing values are used as imputed values for $Y^{mis}_{g^*}$.

EM approach

Here, an EM approach is used to estimate the parameters of regression model (2.1). Then the expectation of each missing value, given the final parameter estimates of the EM algorithm and the observed data, is used as the imputed value.

Poisson mixture model

A Poisson mixture model for clustered RNA-seq data was proposed by Rau et al. (2011). Let variables be independent, conditional on the component, and let an observation from the ℓ^{th} component follows Poisson distribution as follows:

$$Y_g|\ell \sim \prod_{i=1}^{2}\prod_{r=1}^{n_i} \mathscr{P}(y_{gir}; \mu_{gir\ell}),$$

where $\mathscr{P}(.)$ denotes the standard Poisson density with rate $\mu_{gir\ell}$. This corresponds to the following Poisson mixture model (PMM):

$$f(y_g; \mu_g, \pi) = \sum_{\ell=1}^{C} \pi_\ell \prod_{i=1}^{2}\prod_{r=1}^{n_i} \mathscr{P}(y_{gir}; \mu_{gir\ell}),$$

where $\pi = (\pi_1, \pi_2, ..., \pi_C)'$, $\pi_\ell \geq 0$ for all ℓ and $\sum_{\ell=1}^{C} \pi_\ell = 1$. Rau et al. (2011) used two parameterizations for $\mu_{gir\ell}$. Also, an EM approach is used for parameter estimation and clustering of each gene. After parameter estimation, one can use the Bayesian information criterion (Schwarz, 1978) and the specially developed Integrated Complete Likelihood (ICL) criterion (Biernacki et al., 2000) for finding a stable and reliable estimate of the number of clusters (C) for real and simulated datasets from mixtures when the components do not overlap too much. Also, an EM approach is used for parameter estimation. ICL represents a version of BIC that is penalized by a fuzzy classification entropy term: $ICL = BIC - 2\sum_{\ell=1}^{C} \pi_\ell \log(\pi_\ell)$. By construction, ICL is more conservative than BIC since instead of finding the optimal number of components, it aims at detecting the number of clusters. If all components are well-separated, the entropy term does not contribute to ICL and both criteria are equivalent (Celebi, 2015).

For imputing missing values in this scenario, at first the data set is clustered by the above-mentioned approach. Then, the missing values are imputed by arithmetic mean imputation of available genes in each cluster corresponding to the special iteration.

Bayesian Poisson regression

Bayesian Poisson regression modeling (Kleinke and Reinecke, 2011) uses an approach similar to that used by Rubin (1987). In this model, at first a Bayesian approach is used for the parameter estimation of model (2.1). Then, the Poisson imputation approach fits a Poisson model and calculates $\hat{\beta_g}$ and $\hat{V}(\hat{\beta_g})$, $g = 1,2,...,G$, the posterior mean and the posterior variance of β_g, respectively. Then, new parameters β_g^* are drawn from $N(\hat{\beta_g}, \hat{V}(\hat{\beta_g}))$. This approach does not impute deterministically, i.e. it does not directly return the fitted value. This would cause all imputations to lie directly on the regression line, which leads to an under-estimation of standard errors. The imputations are simulated from $y_{gi}^{mis} \sim Poisson(\mu_{gi}^{mis})$, where $log(\mu_{gi}^{mis}) = X_{gi}^{mis}\beta_g^*$.

However, the restriction of equality of the variance to the mean in the Poisson model is often violated in real life. Very often empirical data are overdispersed, which means that the variance is larger in comparison to the mean. Analyzing overdispersed data using ordinary Poisson regression model leads to an underestimation of the variation in the data (Kleinke and Reinecke, 2011). To produce parameter estimates and standard errors, some adjustments need to be made (McCullagh and Nelder, 1989). The imputation procedure proposed by Kleinke and Reinecke (2011) for overdispersed count data is based on quasi-Poisson regression.

The quasi-Poisson imputation approach is quite similar to Poisson imputation approach. The only difference to Poisson imputation is that it uses the quasi-Poisson family instead of the Poisson family and imputations are simulated from $Y_{gi}^{mis} \sim NB(\mu_{gi}^{mis}, \frac{\mu_{gi}^{mis}}{\delta - 1})$, where δ is the overdispersion parameter and $log(\mu_{gi}^{mis}) = X_{gi}^{mis}\beta_g^*$. As mentioned by Kleinke and Reinecke (2011), the negative binomial distribution is an appropriate model for handling overdispersed count data (see Hilbe, 2007).

Bootstrap Poisson regression

Sometimes drawing the parameter β_g^* from $N(\hat{\beta_g}, \hat{V}(\hat{\beta_g}))$ is implausible. Another approach in these cases is to draw a new bootstrap sample Y_g^* from the original data set Y_g B times and to fit the respective regression model to Y_g^* in order to obtain the parameter β_g^*. This new β_g^* is used for the Bayesian Poisson, Bayesian quasi-Poisson and negative binomial regression models. For more discussion about this approach, see Kleinke and Reinecke (2011).

2.1. *Bayesian generalized linear models*

Here, a Bayesian approach for generalized linear modeling with independent normal, t, or Cauchy prior distributions for the coefficients is used for imputing missing values. For this approach, generate an imputed matrix, apply the elementary functions iteratively to the variables with missingness in the data, randomly impute each variable and loop through until reaching approximate convergence. In other words, generate multiple imputations for incomplete data using iterative regression imputation. If the variables with missingness are a matrix Y with columns $Y_{(1)},...,Y_{(K)}$ and the fully observed predictors are X, this entails first imputing all the missing Y values using some crude approach (for example, choosing imputed values for each variable by randomly selecting from the observed outcomes of that variable); and then imputing $Y_{(1)}$ given $Y_{(2)},...,Y_{(K)}$ and X; imputing $Y_{(2)}$ given $Y_{(1)}, Y_{(3)},...,Y_{(K)}$ and X (using the newly imputed values for $Y_{(1)}$), and so forth, randomly imputing each variable and looping through until approximate convergence.

3. Identifying differentially expressed genes from RNA-seq data in the presence of missing values

In most studies on RNA-seq data and gene expression microarray data, the detection of differentially expressed genes is an important investigation. In this section, we discuss this subject in the context of imputation of missing values in RNA-seq data.

Some approaches exist for identifying differentially expressed genes. One way is to remove genes with missing values from the analysis. As any researcher would expect, this approach is not recommended. The other approach (see discussion after Section 2) is to choose a method of single imputation of missing values and then to use available approaches to identify differentially expressed genes from RNA-seq data. Some available packages in R such as DEGseq (Wang et al., 2010), easyRNAseq (Delhomme et al., 2012) and QuasiSeq (Lund et al., 2012) can be used for this purpose. One approach is to use a regression method for identifying differentially expressed genes as follows:

$$Y_{gir} \sim Poisson(\mu_{gir}), \tag{3.1}$$

where $\log(\mu_{gir}) = \beta_{0g} + \beta_{1g}x_{gir}$ and x_{gir} is an indicator function of belonging to each group (or being under each condition) for each gene. If $\beta_{1g} = 0$, then the condition effect is not significant and the gene is not differentially expressed.

The two latter approaches are more highly recommended than the first one. Another approach, which will be discussed in the next section, is the use of multiple imputation.

4. Multiple imputation

Throughout the last two decades, multiple imputation (Rubin, 1987; Schafer, 1997) has become more and more popular and has become one of the standard procedures to handle missing data (Schafer and Graham, 2002).

The idea behind multiple imputation is that, for each gene with missing values, we impute, say M values, instead of just one value ($M = 5$ is suggested by Rubin, 1987). The M imputed values are used to create M complete data sets. Figure 2 presents a multiply-imputed RNA-seq data set, where each missing datum is replaced by a pointer to a vector of M values. This results in M completed data sets. The analysis of a multiply-imputed data set is quite direct. First, each data set completed by imputation is analyzed using the same complete-data method that would be used in the absence

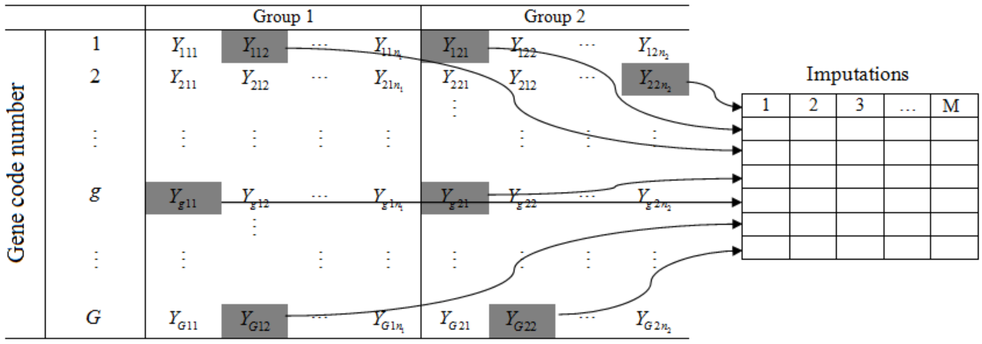

Group 1
Group 2
Gene code number
1
2
⋮
g
⋮
G
Y_{111} Y_{112} … Y_{11n_1} Y_{121} Y_{122} … Y_{12n_2}
Y_{211} Y_{212} … Y_{21n_1} Y_{221} Y_{222} … Y_{22n_2}
Y_{g11} Y_{g12} … Y_{g1n_1} Y_{g21} Y_{g22} … Y_{g2n_2}
Y_{G11} Y_{G12} … Y_{G1n_1} Y_{G21} Y_{G22} … Y_{G2n_2}
Imputations
1
2
3
…
M

Fig. 2. Matrix of RNA-seq data with M imputations for each missing gene.

of nonresponse. Let $\hat{\beta}_m$ and W_m, $m = 1,2,...,M$ be M complete-data estimates and their associated variances for an estimated parameter β, calculated from M repeated imputations under one model. The combined estimate is

$$\bar{\beta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m. \tag{4.1}$$

By averaging over M imputed data sets in (4.1), the efficiency of the estimate is increased. The variability associated with this estimate has two components (Little and Rubin, 2002): (1) the averaging within-imputation variance which is given by

$$\bar{V}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m,$$

and (2) the between-imputation component which is given by

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \bar{\beta}_M)^2. \tag{4.2}$$

The total variability associated with $\bar{\beta}_M$ is

$$T_M = \bar{V}_M + \frac{M+1}{M}B_M. \tag{4.3}$$

In order to use multiple imputation for imputing missing values after creating M complete matrices of RNA-seq read counts, as the scientific question is to identify differentially expressed genes, the model (3.1) is fitted to each complete RNA-seq matrix. Then the mean of $\beta^1_{1g}, \beta^2_{1g}, ..., \beta^M_{1g}$ is used as an estimate for β_{1g} (as mentioned previously it is denoted by $\bar{\beta}_g$). The variance of this estimate is given by (4.3).

5. Simulation studies

In this section, some simulation studies are performed in order to investigate the performance of the proposed imputation methods. An initial simulation study is performed for investigating the performance of the proposed single imputation approaches for imputing the missing RNA-seq counts and a second simulation study is performed for investigating the performance of single imputation approaches for detecting differentially expressed genes. The last simulation study is performed for investigating the performance of the multiple imputation approach for imputing missing RNA-seq counts and identifying differentially expressed genes.

5.1. *Simulation study 1: Performances of single imputation approaches for imputing missing values*

In order to investigate the performance of single imputation approaches, three different patterns of data are generated and the data sets are merged to obtain one new collected data set. We consider $n_1 = n_2 = 20$ iterations, and $G = 3000$ RNA-seq read counts are generated. 1000 genes are generated from each pattern. The Monte Carlo simulation is iterated $N = 1000$ times. Figure 3 shows one set of generated data for each cluster (pattern) and the final merged data set. 10% of genes are randomly selected and for them 10%, 20% and 30% of counts are missing and different imputation methods

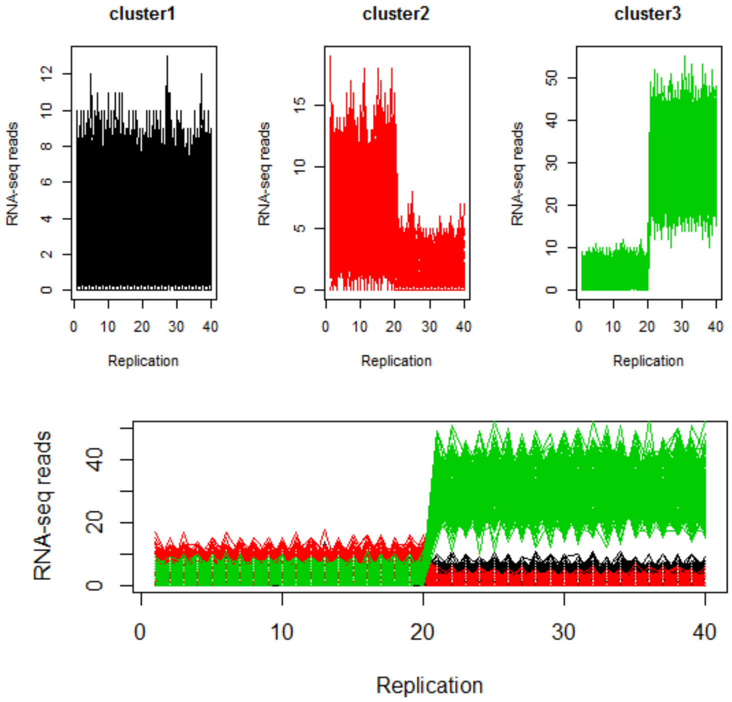


Fig. 3. Pattern of generated RNA-seq data in simulation study 1.

are used for imputing the missing RNA-seq read counts. We consider the normalized root mean square error (NRMSE; Oba et al., 2003) for comparing the performance of approaches; the lower the NRMSE the more accurate is the chosen imputation algorithm. Let Y_{gir}^{real} and Y_{gir}^{imp} be the real and imputed values for gene g in the i^{th} group and r^{th} iteration, respectively. Then NRMSE is defined as:

$$NRMSE = \sqrt{\frac{\sum_{g=1}^{G}\sum_{i=1}^{2}\sum_{r=1}^{n_i}(Y_{gir}^{imp} - Y_{gir}^{real})^2}{\sum_{g=1}^{G}\sum_{i=1}^{2}\sum_{r=1}^{n_i}(Y_{gir}^{real})^2}}.$$

The results of this simulation study are given in Table 1. The Poisson mixture model performs better than other approaches, because of the clustered nature of the generated data (note that the number of clusters in this approach are considered by the smallest value of the ICL criterion). The performances of other approaches are also well. The means and variances of NRMSE tend to increase as the proportion of missingness increases.

5.2. *Simulation study 2: Performance of single imputation for identifying differentially expressed genes*

We generate $G = 4000$ genes such that $G_1 = 2000$ genes are generated from a Poisson distribution with rate 3 for both groups. These genes are not differentially expressed. $G_2 = 1000$ genes are

Table 1. Results of simulation study (mean and standard error of NRMSE values) investigating performance of single imputation approaches (the lowest values are highlighted in bold).

Expected proportion of missingness	10%	20%	30%
Method	Est.(S.E.)	Est.(S.E.)	Est.(S.E.)
Regression imputation	0.0170(0.0022)	0.0265(0.0059)	0.0343(0.0115)
Bayesian generalized linear model	0.0233(0.0028)	0.0379(0.0155)	0.0545(0.0698)
Poisson mixture model	**0.0139(0.0013)**	**0.0199(0.0013)**	**0.0230(0.0013)**
EM approach	0.0165(0.0017)	0.0245(0.0021)	0.0295(0.0035)
Bayesian Poisson regression	0.0306(0.0090)	0.0444(0.0134)	0.0545(0.0274)
Bayesian quasi-Poisson regression	0.0217(0.0052)	0.0302(0.0090)	0.0373(0.0124)
Bootstrap Bayesian Poisson regression	0.0292(0.0088)	0.0532(0.0484)	0.3157(1.6080)
Bootstrap Bayesian quasi-Poisson regression	0.0217(0.0119)	0.0367(0.0217)	0.1338(0.6067)

generated with rates 1 and 3 for the two groups and $G_3 = 1000$ genes are generated with rates 6 and 1 for the two groups.

Also $n_1 = n_2 = 20$. We have randomly selected 10% of genes. Then 10% and 20% of the read counts of each gene are randomly removed. Different approaches are used for imputing missing values. Then the following regression model is used for identifying differentially expressed genes:

$$Y_{gir} \sim Poisson(\mu_{gir}), \tag{5.1}$$

where $\log(\mu_{gir}) = \beta_{0g} + \beta_{1g} x_{gir}$ such that

$$x_{gir} = \begin{cases} 1 \text{ if r}^{\text{th}} \text{ replication of gene g is under the first condition (belongs to the first group)} \\ 0 \qquad\qquad\qquad\qquad \text{o.w.} \end{cases}$$

In other words, $x_{g1r} = 1$ and $x_{g2r} = 0$, $g = 1,2,...,G$, $r = 1,2,...,n_i$ and $i = 1,2$. The results of this simulation study are given in Table 2. Table 2 indicates that there is little to choose between the various approaches in term of the mean and variance of TPR, and that the proportion of missingness has little effect on the results.

For comparing the results, true positive rate (TPR), false positive rate (FPR) and true discovery rate (TDR) are used.

In order to record a case identified by the regression model as being differentially expressed, we define the following indicator variables:

$$I_{gk}^{Method} = \begin{cases} 0 \; if \; \hat{\beta}_{1g} \text{ is not statistically significant} \\ 1 \qquad\qquad o.w. \end{cases},$$

such that $I_{gk}^{Method} = 1$ if $p - value_{gk} < \alpha$. For $g = 1,2,...,G$ and $k = 1,2,...,N[= 1000]$ also, for real scenarios

$$I_{gk}^{Real} = \begin{cases} 0 \; \beta_{1gk} = 0 \\ 1 \quad o.w. \end{cases}.$$

Let ω be the ratio of differentially expressed genes. The true positive rate, the false positive rate and the true discovery rate for the k^{th} iteration can be calculated as follows:

$$TPR_k = \frac{\sum_{g=1}^{G} I_{gk}^{Real} I_{gk}^{Method}}{G \times \omega},$$

$$FPR_k = \frac{\sum_{g=1}^{G}(1 - I_{gk}^{Real})I_{gk}^{Method}}{G \times (1-\omega)},$$

$$TDR_k = \frac{\sum_{g=1}^{G} I_{gk}^{Real} I_{gk}^{Method}}{\sum_{g=1}^{G} I_{gk}^{Method}}, \ k = 1,2,..,N.$$

Averaging these rates across all iterations, we have:

$$T\bar{P}R = \frac{1}{M}\sum_{k=1}^{M} TPR_k, \tag{5.2}$$

$$F\bar{P}R = \frac{1}{M}\sum_{k=1}^{M} FPR_k, \tag{5.3}$$

$$T\bar{D}R = \frac{1}{M}\sum_{k=1}^{M} TDR_k. \tag{5.4}$$

The results are given in Table 2. This table only shows the means of TPR and their standard errors. The results show that the performance of all the single imputation methods are well. Also, the means and standard errors of $F\bar{P}R$ and $T\bar{D}R$ for all the methods are 0 and 1, respectively.

Table 2. Results of the simulation study (mean and standard error of TPR values) investigating the performance of single imputation approaches.

Expected proportion of missingness	10%	20%
Method	Est.(S.E.)	Est.(S.E.)
Real data without missingness	0.9509(0.0151)	0.9507(0.0147)
Regression imputation	0.9497(0.0151)	0.9480(0.0150)
Bayesian generalized linear model	0.9480(0.0153)	0.9438(0.0158)
Poisson mixture model	0.9470(0.0149)	0.9509(0.0145)
EM approach	0.9497(0.0151)	0.9479(0.0146)
Bayesian Poisson regression	0.9447(0.0152)	0.9381(0.0155)
Bayesian quasi-Poisson regression	0.9454(0.0156)	0.9401(0.0158)
Bootstrap Bayesian Poisson regression	0.9486(0.0153)	0.9409(0.0155)
Bootstrap Bayesian quasi-Poisson regression	0.9473(0.0157)	0.9383(0.0151)

5.3. *Simulation study 3: Performance of multiple imputation approaches*

In this section, the generated data set of Section 5.2 is considered and the performance of multiple imputation approaches for imputing missing values and detecting differentially expressed genes is investigated. For this purpose, $M = 5$ different methods are chosen for imputation.

In this simulation study, two expected rates of missingness (20% and 30%) for 10% of genes are considered. At first, for each imputation method, model (4.1) is fitted to the data set and $\beta_{1g}, g = 1,2,...,G$ is estimated for each method. We denote them, for each imputation method, as $\hat{\beta}_{1g}^{m}, g = 1,2,...,G$ and $m = 1,2,...,M[= 5]$. Then the results are combined in order to obtain the multiple imputation results as follows:

$$\begin{aligned}
\hat{\beta}_{1g}^{mi} &= \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_{1g}^{m}, \\
\bar{V}_M(\hat{\beta}_{1g}^{mi}) &= \frac{1}{M}\sum_{m=1}^{M}V(\hat{\beta}_{1g}^{m}), \\
B_M(\hat{\beta}_{1g}^{mi}) &= \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_{1g}^{m}-\hat{\beta}_{1g}^{mi})^2, \\
T_M(\hat{\beta}_{1g}^{mi}) &= \bar{V}_M(\hat{\beta}_{1g}^{mi}) + \frac{M+1}{M}B_M(\hat{\beta}_{1g}^{mi}).
\end{aligned}$$

We use NRMSE, $T\bar{P}R$, $F\bar{P}R$ and $T\bar{D}R$ for reporting the results of this simulation study which are summarized in Table 3.

The results show the well-performance of the multiple imputation methods. Also, the means and standard errors of $F\bar{P}R$ and $T\bar{D}R$ for all the methods are 0 and 1, respectively.

Table 3. Results of the simulation study (mean and standard error of NRMSE and TPR values) for investigating the performance of multiple imputation approaches.

Expected proportion of missingness	10%	20%
criterion	Est.(S.E.)	Est.(S.E.)
NRMSE	0.0159(0.0018)	0.0301(0.0123)
TPR	0.9474(0.0133)	0.9458(0.0135)

6. Applications

6.1. *Pickrell et al.'s (2010) and Montgomery et al.'s (2010) RNA-seq data set*

6.1.1. *Single imputation*

These RNA-seq data (3050 gene and 129 RNA sample) were collected by Pickrell et al. (2010) and Montgomery et al. (2010). The data set contains RNA-seq from 60 CEU individuals and 69 lymphoblastoid cell lines derived from unrelated Nigerian individuals that have been extensively genotyped by the International HapMap Project.

We removed the genes that had zeros for all samples. For these genes, there is no possibility of differential expression. Thus 1587 genes were analyzed in this study.

There are some real RNA-seq data sets containing missing values. Two examples are fibroblast data set and fruit flies data set. These can be found in Chu (2014). So, imputation of the missing values is an important matter which will gives us the ability to use the available software to analyze the data set.

In our data set, we have randomly selected 30% of genes. Then, we randomly removed 15% of values from the chosen genes. We use all of the single imputation approaches, which were described in Section 2. The results of NRMSE (see Section 5.1) can be found in the second column of Table 4. For regression imputation and other regression-based imputation, in the first stage, the Spearman correlation (Becker et al., 1988) is computed for pairwise complete observations and $K = 5$ genes

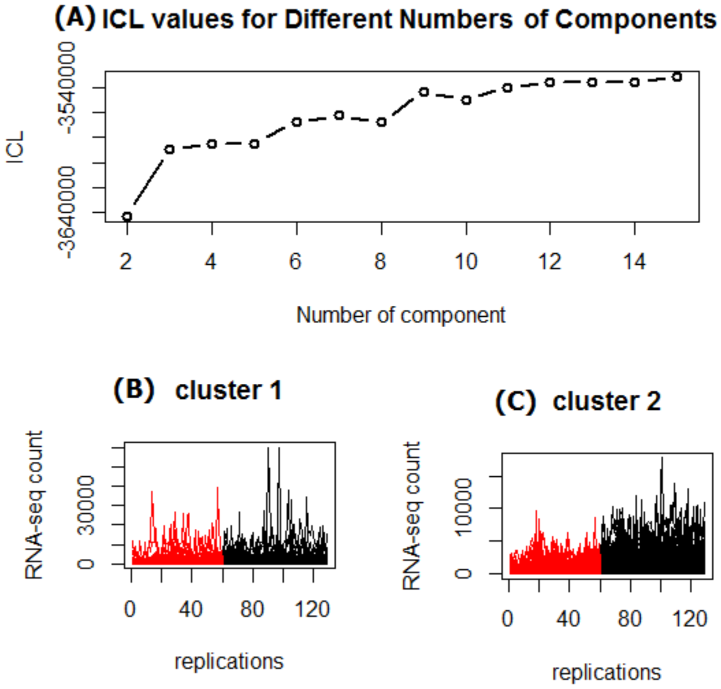


Fig. 4. ICL criterion and two clusters in Pickrell et al.'s (2010) and Montgomery et al.'s (2010) data set.

with highest correlation are selected as explanatory variables. Also, the following indicator variable is used to consider the group effect:

$$x_{gr} = \begin{cases} 1 \text{ if the r}^{\text{th}} \text{ replication of gene g is under the first condition (belongs to the first group)}, \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

In the Poisson mixture model, different numbers of components are considered. Panel (A) of Figure 4 shows ICL values versus the number of components in the Poisson mixture model. This panel shows that $C = 2$ components have the smallest value of ICL. Panels (B) and (C) present the data for two detected clusters. Table 4 shows that the regression-based imputation approaches, especially the EM approach, Bayesian Poisson regression and Bootstrap Bayesian Poisson regression, are the best approaches for imputing missing values in this data set.

6.1.2. *Multiple imputation*

In this section, the missing values are imputed initially using $M = 5$ different single imputation approaches and the mean of them are considered as the final missing value imputation. The value of NRMSE using this approach is 0.0500. Then, the regression model (5.1) is fitted to the real data set and the imputed data set. Detected differentially expressed genes for real data sets are considered as real differentially expressed genes and the criteria introduced in Section 5.2 are computed. In this case, TPR and TDR are equal to 1 and FPR is equal to 0, which means the well performance of the approaches.

Table 4. Comparison of direct single imputation approaches (by NRMSE) for Pickrell et al.'s (2010) and Montgomery et al.'s (2010) RNA-seq data set and Marioni et al.'s (2008) RNA-seq data set.

Method	P& M's data set	M's data set
Regression imputation	0.0763	0.0063
Bayesian generalized linear model	0.0650	0.0061
Poisson mixture model	0.1349	0.1452
EM approach	0.0586	0.0057
Bayesian Poisson regression	0.0568	0.0088
Bayesian quasi-Poisson regression	0.0763	0.0103
Bootstrap Bayesian Poisson regression	0.0566	0.0094
Bootstrap Bayesian quasi-Poisson regression	0.0767	0.0114

6.2. *Marioni et al.'s (2008) RNA-seq data set*

6.2.1. *Single imputation*

RNA-seq counts (5000 genes and 14 DNA samples) are extracted from Marioni et al.'s (2008) data. These are the total RNA counts from liver and kidney samples of a single human male. To assess technical variance within and between runs, each sample was sequenced in seven lanes (for kidney: R1L1Kidney, R1L3Kidney, R1L7Kidney, R2L2Kidney, R2L4Kidney, R2L6Kidney, R2L8Kidney; and for liver: R1L2Liver, R1L4Liver, R1L6Liver, R1L8Liver, R2L1Liver, R2L3Liver, R2L7Liver).

As the data are in the form of counts, a Poisson model may be appropriate. The difference between this data set and previous data set is in the number of replications. The number of replications in this data set is seven for each group. We have randomly selected 20% of genes. Then, we randomly removed 15% of iterations from the chosen genes. Also, 2672 genes with non-zero RNA-seq counts are considered. The NRMSE results can be found in the third column of Table 4. For regression imputation and other regression-based imputation, the Spearman correlation for pairwise complete observations is computed and $K = 3$ genes with the highest correlation are selected as explanatory variables.

In the Poisson mixture model, different numbers of components are considered. Then, using the ICL criterion, two clusters are detected in the data set (see Figure 5). Table 4 shows that the regression-based imputation approaches, especially the EM approach, Bayesian Poisson regression and Bootstrap Bayesian Poisson regression, are the best performing approaches.

6.2.2. *Multiple imputation*

As in sub-section 6.1.2, the missing RNA-seq counts are imputed by $M = 5$ imputation methods and NRMSE equals 0.0799. Also, after fitting the regression model to the real data set, and using multiple imputation approaches for identifying differentially expressed genes, TPR and TDR are found to be equal to 1 and FPR to be equal to 0.

7. Discussion

In this paper, some single imputation approaches for imputing missing values of RNA-seq data sets were proposed. Most of them were based on Poisson regression. The Bayesian and non-Bayesian frameworks were also considered for implementation of the imputation approaches. The single

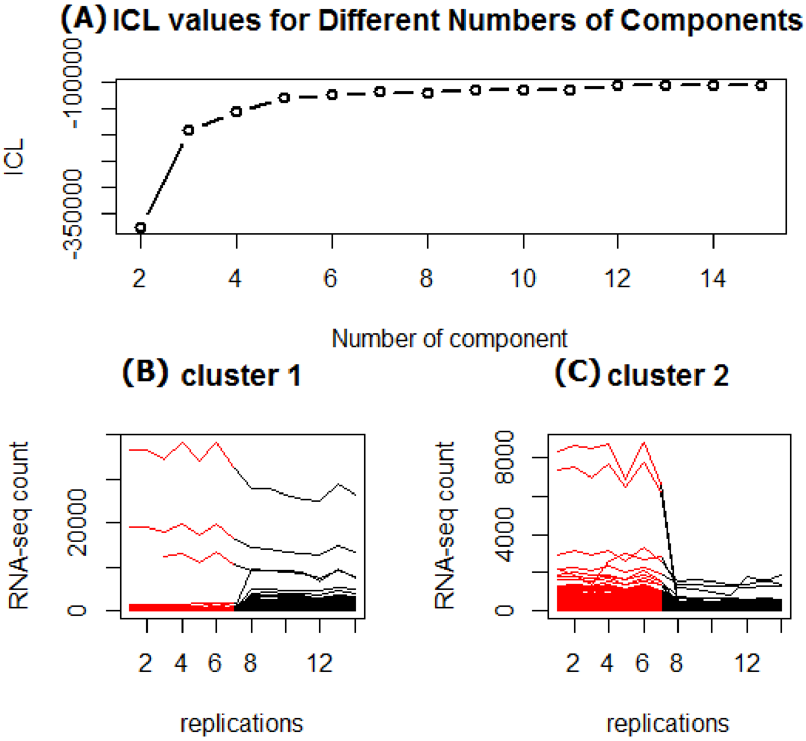


Fig. 5. ICL criterion and two clusters in Marioni et al.'s (2008) dataset.

imputation approaches do not take into account the uncertainty of the sampling. Thus, we also considered multiple imputation approaches for imputing the missing values.

The performance of the approaches were investigated by some simulation studies and true real data sets. Also, the performance of the proposed approaches was investigated for their ability to identify differentially expressed genes.

When, choosing one proposed approach for a real data set, one can remove all the genes with missing RNA-seq count, then randomly remove some of the genes in the new data set. After that, the proposed approaches may be used for imputing the missing values, and by using the proposed criteria, one may select some of the imputation approaches for imputing the real RNA-seq counts. Therefore, different imputation approaches should be used to find the best performing methods in different data sets. Due to the nature of the data, different methods may perform differently (a training-testing approach).

Note that, if the data set be only under one condition (that is, in the notation of Section 2, $i = 1$) then, in the regression-based approaches, the covariate which describes the effect of condition will have to be removed from the model. Also, if there are more than two conditions (that is, in the notation of Section 2, $i = 1, 2, \ldots, d$ such that $d > 1$), one will have to define more than one dummy variable in order to represent the effect of conditions.

The majority of the available software for missing data analysis in R are useful for imputation of RNA-seq data sets with missing values. Some of the methods in this paper can be performed by the available packages of R. For example, the poisson mixture model can be implemented using the “PoisMixClus” function of the “HTSCluster” package and the “mi” package can be used to implements the Bayesian generalized linear models.

In this paper, we have only considered data with missing at random mechanism. One can try to find some approaches for imputation of RNA-seq counts assuming a missingness not at random. Also, as count data may include excess zeros, one may try to find appropriate approaches such as zero inflated Poisson and zero inflated negative binomial for imputing missing values in these kinds of RNA-seq data sets.

Acknowledgments

The research presented in this study was carried out on the High Performance Computing Cluster supported by the Computer Science department of Institute for Research in Fundamental Sciences (IPM). We would like to thank the Iranian National Science Foundation (INSF BS1393103 and BS13940202) for their support.

References

[1] Aittokallio T. (2010), Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinform*, **11**, 253–264.

[2] Bras, L. and Menezes, J. (2007). Improving cluster-based missing value estimation of DNA microarray data. *Biomolecular engineering*, **24**(2):273.

[3] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*. Pacific Grove, CA, USA: Wadsworth & Brooks/Cole.

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

[5] 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.

[6] Celebi, M. E. (2015), *Partitional Clustering Algorithms*. Springer, New York.

[7] Chu, Man-Kee Maggie (2014), Statistical methods for the analysis of RNA sequencing data, *Electronic Thesis and Dissertation Repository*. The University of Western Ontario.

[8] Delhomme N, Padioleau I, Furlong EE, Steinmetz LM. (2012), easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics*, **28**(19), 2532-2533.

[9] Gower, J. C. (1971), A general coefficient of similarity and some of its properties, *Biometrics* 27, 857-874.

[10] Hilbe, J. M. (2007), *Negative binomial regression.* Cambridge: Cambridge University Press. ISBN 9780521198158.

[11] Hourani, M., and El Emary, I. (2009). Microarray missing values imputation methods: Critical analysis review. *Computer Science and Information Systems*, **6**(2), 165-190.

[12] Jornsten, R., Wang, H., Welsh, W. and Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, **21**(22), 4155-4161.

[13] Kim H, Golub GH, Park H. (2005), Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* **21**, 187-198.

[14] Kleinke, K., and Reinecke, J. (2011), COUNTIMP A multiple imputation package for incomplete count data (Tcehnical Report). Bielefeld: Bielefeld University, Faculty of Sociology.

[15] Liew WC, Law NF, Yan H. (2010), Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform*, **12**, 498-513.

[16] Lund S, Nettleton D, McCarthy D, Smyth G, (2012), Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol*, **11**, Article 8.

[17] Little, R. J. A., Rubin, D. B. (2002), *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

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

[19] McCullagh, P. and J. A. Nelder. (1989), *Generalized Linear Models.* Second ed. London: Chapman and Hall.

[20] Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET (2010), Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, 773–777.

[21] Moorthy, K., M. Saberi Mohamad, and S. Deris (2014), A review on missing value imputation algorithms for microarray gene expression data. Current *Bioinformatics*, **9**(1), 18-22.

[22] Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A (2003), Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088-2096.

[23] Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010), Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*, 768-772.

[24] Rau, A., Celeux, G., Martin-Magniette, M.-L., Maugis-Rabusseau, C (2011), Clustering high-throughput sequencing data with Poisson mixture models. *Inria Research Report*, 7786.

[25] Rubin, D. B. (1976), Inference and missing data. *Biometrika*, **63**, 581–592.

[26] Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse (with discussion). In Proc. Survey Research Methods Section. 2034. *Amer. Statist. Assoc.*, Alexandria, VA.

[27] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

[28] Sande, I. G. (1983), Hot-deck imputation procedures. In Incomplete Data in Sample SWV VS, Volume 3, Proceedings of the Symposium. W. G. Madow and I.. Ourin (eds.). New York: Academic Press, 334-350.

[29] Sehgal MSB, Gondal I, Dooley LS. (2005), Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, **21**, 2417-2423.

[30] Schafer, J. L. (1997), *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall.

[31] Schafer, J. L., Graham, J. W. (2002), Missing data: Our view of the state of the art. *Psychological Methods*, **7**, 147-177.

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

[33] Sedransk, J., and Titterington, D. M. (1984), Mean imputation and random imputation models. Unpublished technical report.

[34] Si Y, Liu P, Li P, Brutnell, T. (2014), Model-based clustering of RNA-seq data. *Bioinformatics*, .

[35] Troyanskaya, o., Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman, (2001), Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**(6), 520–525

[36] Wang L, Feng Z, Wang X, Wang X, Zhang X. (2010), DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136-138.

[37] Yang YH, Buckley MJ, Dudoit S, Speed TP. (2002), Comparison of methods for image analysis on cDNA microarray data. *J Comput Graph Stat*, **11**, 108-136.

[38] Zhang X, Song X, Wang H, Zhang H. (2008), Sequential local least squares imputation estimating missing value of microarray data. *Comput Biol Med*, **38**, 1112-1120.