EXPLOITING BICLUSTERING FOR MISSING VALUE ESTIMATION IN DNA MICROARRAY DATA

K.O. Cheng, N.F. Law and W.C. Siu

Centre for Signal Processing, Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University

ABSTRACT

The missing values in gene expression data harden subsequent analysis such as biclustering which aims to find a set of coexpressed genes across a number of experimental conditions. Missing values are thus required to be estimated before biclusters detection. Existing estimation algorithms rely on finding coherence among expression values throughout the entire genes and/or across all the conditions. In view that both missing values estimation and biclusters detection aim at exploiting coherence inside the expression data, we propose to integrate them into a single framework. The benefits are twofold, the missing value estimation can improve bicluster analysis and the coherence in detected biclusters can be exploited for better missing value estimation. Experimental results show that the integrated framework outperforms existing missing values estimation algorithms. It reduces error in missing value estimation and facilitates the detection of biologically meaningful biclusters.

Index Terms— Gene expression, missing value estimation, biclustering.

1. INTRODUCTION

DNA microarrays [1] measure gene expressions of ten thousands of genes under hundreds of experimental conditions in parallel. Unfortunately, some data may be lost due to image corruption, dust or scratches on the slides and experimental errors. The missing entries should be filled in before subsequent analysis such as clusters detection because most existing analysis tools such as hierarchical clustering are applicable to complete datasets only. Although the missing data can be acquired by repeating the experiments, it is usually not feasible due to economical reason or sometimes limitations of samples. Thus, computational-based missing value estimation is necessary and crucial.

Existing missing value estimation methods can be divided into two classes: local and global approaches [2]. Local approaches such as local least square method [3, 4] identify similar genes or similar conditions to predict missing values. Global approaches such as Bayesian principal components analysis (BPCA) [5] exploit an overall covariance within genes or experimental conditions.

Recently, Friedland *et al.* [6] demonstrate that a combination of missing value estimation and clustering can achieve a better estimation of missing values. However, conventional clustering methods are global in nature in which one can either find similar genes under all conditions or find similar conditions for all genes. In reality, related genes co-express over a subset of conditions only. Unlike clustering, biclustering is able to group genes and conditions simultaneously [7, 8]. In this paper, missing value estimation and biclustering are combined in a single framework in which a model-based missing values estimation is introduced for missing value refinement inside the detected biclusters. More specifically, our method iterates between missing value estimation and bicluster detection so that local information found in biclusters can be used to provide a better estimation of missing values, and the accurately estimated gene expression data can in turn facilitate the biclusters detection. As a result, the accuracy of missing value estimation and biclustering can both be enhanced at the same time.

2. BACKGROUND

Our proposed algorithm is an iterative algorithm which performs missing value estimation and biclustering alternatively. In this section, biclustering and two popular imputation methods, LLS and BPCA, are presented.

2.1. Biclustering

Biclustering can identify homogeneous patterns known as biclusters which is a subset of rows having related expression values across a subset of columns [7, 8]. One useful bicluster model is the additive model [9, 10]. Denote b_{ij} as an expression value in a bicluster at position (i, j), the additive model can be described as,

$$b_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (1)$$

where μ is a constant, α_i is a row dependent factor, β_j is a column dependent factor and ε_{ij} is the noise term. As the identification of all biclusters is a NP-hard problem [7,9], an efficient biclustering algorithm called "BiVisu" that has polynomial-time complexity [10] is adopted in our proposed

framework for missing value estimation to reduce the computational time. It is freely available at http://www.eie.polyu.edu.hk/~nflaw/Biclustering/index.html.

2.2. LLS and BPCA imputation

The local least squares (LLS) imputation [3] is a popular local estimation method that makes use of correlation among similar genes. For each target gene which contains at least one missing value, k most similar genes are first selected from the gene expression matrix. Then the target gene is regressed on these k similar genes. The regression coefficients are calculated using least squares approach which can then be used to impute the missing values.

Bayesian principal component analysis (BPCA) [5] method is another popular imputation approach. It consists of two steps: 1) PCA is applied and coefficients of target genes are estimated to predict missing values; 2) the imputed missing values are used to obtain the new PCA and coefficients. These two steps are iterated until convergence. One of the main differences between LLS and BPCA is that LLS uses correlation in a subset of genes while BPCA considers the correlation in the whole expression matrix. However, both can achieve high performance among existing imputation algorithms.

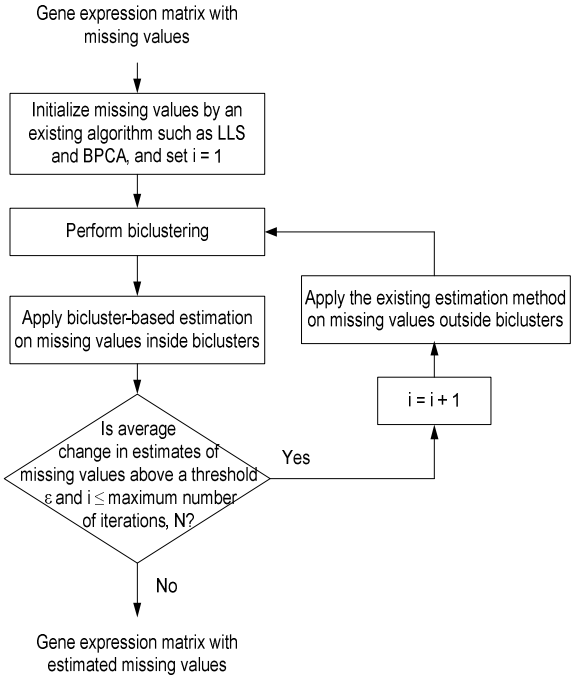

Fig. 1. Flow chart of the proposed algorithm.

3. PROPOSED ESTIMATION METHOD

Since missing value estimation and biclustering are interrelated, an integrated approach for these two processes is proposed. In particular, missing value estimation and biclusters detection are applied iteratively. The procedure of the proposed algorithm is summarized in Fig. 1. First, either LLS or BPCA is used to impute all missing values in the expression matrix. Then, biclusters are detected using the BiVisu software. For all the detected biclusters, the bicluster model parameters are estimated (will be discussed in Section 3.1) and used to estimate the missing values that are inside the detected biclusters. If the average change in missing value estimation falls below a threshold or the maximum number of iterations is reached, the process is terminated. Otherwise, missing values outside the detected biclusters will be re-estimated and biclusters detection will be performed again.

3.1. Bicluster-based Estimation

An approach to obtain least square estimates of the additive model (1) is to minimize the following term,

$$\sum_i \sum_j \left(b_{ij} - \mu - \alpha_i - \beta_j\right)^2 \tag{2}$$

Equivalently, a linear regression model can be used, i.e.,

$$b_{ij} = \alpha_1 p_{ij}^{\alpha 1} + \cdots + \alpha_m p_{ij}^{\alpha m} + \beta_1 p_{ij}^{\beta 1} + \cdots + \beta_n p_{ij}^{\beta n} + \mu + \varepsilon_{ij} \tag{3}$$

where m and n are the numbers of rows and columns of the bicluster, $\{p_{ij}^{\alpha i'}\}$ and $\{p_{ij}^{\beta j'}\}$ are indicator variables given by,

$$p_{ij}^{\alpha i'} = \begin{cases} 0, i \neq i' \\ 1, i = i' \end{cases} \text{ and } p_{ij}^{\beta j'} = \begin{cases} 0, j \neq j' \\ 1, j = j' \end{cases} \tag{4}$$

In order to obtain a unique solution, the following two constraints are imposed,

$$\sum_i \alpha_i = 0 \text{ and } \sum_j \beta_j = 0 \tag{5}$$

By using equations (3) and (5), it can be shown that

$$b_{ij} = \alpha_1 x_{ij}^{\alpha 1} + \cdots + \alpha_{m-1} x_{ij}^{\alpha m-1} + \beta_1 x_{ij}^{\beta 1} + \cdots + \beta_{n-1} x_{ij}^{\beta n-1} + \mu + \varepsilon_{ij} \tag{6}$$

where

$$x_{ij}^{\alpha i'} = p_{ij}^{\alpha i'} - p_{ij}^{\alpha m},\ i' = 1,2,\ldots,m\text{-}1$$
$$x_{ij}^{\beta j'} = p_{ij}^{\beta j'} - p_{ij}^{\beta n},\ j' = 1,2,\ldots,n\text{-}1 \tag{7}$$

Equation (7) can be expressed in matrix form as,

$$\mathbf{b} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon} \tag{8}$$

where $\mathbf{b}$ is a $mn\times1$ vector which is constructed by concatenating columns of the bicluster one by one, $\mathbf{w} = (\alpha_1,\ldots,\alpha_{m-1},\beta_1,\ldots,\beta_{n-1},\mu)^T$ forms a $(m+n-1)\times1$ vector of the model parameters, $\mathbf{X}$ is a $mn\times(m+n\text{-}1)$ data matrix of variables $\{ x_{ij}^{\lambda s}: \lambda = \alpha \text{and } s = i' = 1, \ldots, m\text{-}1 \text{ or } \lambda = \beta \text{ and } s = j' = 1, \ldots, n\text{-}1\}$ and $\boldsymbol{\varepsilon}$ is a $mn\times1$ vector of noise terms. The least square solution of equation (8) is given by,

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{b} \tag{9}$$

Bicluster model parameters are estimated for all the detected biclusters. Due to estimation error, some missing entries may be wrongly included in a bicluster. These destroy the bicluster homogeneity and thus affect the estimation accuracy of model parameters. To have a reliable

estimation, an outlier detection method is adopted which is based on covariance ratio statistics [11]. In particular, the covariance ratio of the *t*th observation is defined by,

$$CVR_t = \left|(\mathbf{X}_t^T \mathbf{X}_t)^{-1} s_t^2\right| \Big/ \left|(\mathbf{X}^T \mathbf{X})^{-1} s^2\right| \quad (10)$$

where $|A|$ dentoes the cardinality of A, s^2 is the estimated variance calculated from all observations, s_t^2 and $\mathbf{X}_t$ are the estimated variance and variable matrix excluding the *t*th observation. If $CVR_t < 1 - 3k^*/N$, where $k^* = m + n - 1$ is the number of free parameters and $N = mn$ is the total number of observations, the removal of the *t*th observation can reduce the variance significantly. Thus, the *t*th observation is omitted in bicluster model estimation in equation (9).

4. EXPERIMENTAL RESULTS

Experiments have been performed to evaluate the proposed integrated framework. LLS [3] and BPCA [5] were integrated in the proposed framework for initialization and estimation of missing values outsides biclusters. The proposed algorithm using LLS will be referred as *BA-LLS* while that using BPCA will be referred as *BA-BPCA*. The experiments consisted of two real datasets: yeast cell cycle expression data measured using α factor based method (yeast_alpha) and cdc15 based method (yeast_cdc15) [12]. The sizes of yeast_alpha and yeast_cdc15 were 4489×18 and 4381×24 respectively.

Missing values were randomly assigned in the matrix so that their estimates can be compared with the true values for evaluation. For each dataset, the test was repeated five times. Furthermore, five missing rates: 1%, 5%, 10%, 15% and 20% were considered. The estimation accuracy was measured by the normalized root mean square error (NRMSE) defined as

$$\text{NRMSE} = \sigma^{-1}\sqrt{\sum (a - \hat{a})^2 / m} \quad (11)$$

where a is the true value, $\hat{a}$ is the estimate of a, m is the total number of entries in the whole matrix and σ is the standard deviation of the matrix. The lower the NRMSE, the higher the estimation accuracy is.

4.1. Performance in NRMSE

In the experiments, biclusters are required to have at least 5 rows and 4 columns. The maximum number of iterations was set to 25. The NRMSE for the real datasets yeast_alpha and yeast_cdc15 are provided in Tables 1 and 2 respectively. As can be seen in Table 1, there is no difference in NRMSE for the algorithms with and without bicluster-based imputation at 1% missing rate. This means that the information in the detected biclusters cannot help to improve the missing value estimation. However, improvement is found for missing rate exceeding 1%. The NRMSE of *BA-LLS* is lower than that of LLS by 4.82% and 3.36% at 10% and 20% missing rates respectively. The NRMSE of *BA-BPCA* is lower than that of BPCA by 1.52% and 1.70% at 10% and 20% missing rates respectively. The improvement achieved by using the biclusters information demonstrates that the information is useful for missing value estimation. Similar observations can also be found for the real dataset yeast_cdc15 provided in Table 2.

Table 1. Comparison of NRMSE of LLS, BPCA and the proposed bicluster-based algorithms *BA-LLS* and *BA-BPCA* on yeast_alpha.

Missing rate	1%	5%	10%	15%	20%
LLS	0.0189	0.0901	0.1950	0.2557	0.3124
BA-LLS	0.0189	0.0894	0.1856	0.2456	0.3019
Improvement	0%	0.78%	4.82%	3.95%	3.36%
BPCA	0.0207	0.0932	0.1706	0.2366	0.2937
BA-BPCA	0.0207	0.0931	0.1680	0.2316	0.2887
Improvement	0%	0.11%	1.52%	2.11%	1.70%

Table 2. Comparison of NRMSE of LLS, BPCA and the proposed bicluster-based algorithms *BA-LLS* and *BA-BPCA* on yeast_cdc15.

Missing rate	1%	5%	10%	15%	20%
LLS	0.0195	0.1157	0.1815	0.2397	0.2896
BA-LLS	0.0195	0.0918	0.1719	0.2306	0.2790
Improvement	0%	20.66%	5.29%	3.80%	3.66%
BPCA	0.0203	0.0949	0.1672	0.2278	0.2782
BA-BPCA	0.0203	0.0927	0.1611	0.2216	0.2724
Improvement	0%	2.32%	3.65%	2.72%	2.08%

4.2. Performance evaluation on biclusters detection

The biclustering analysis using BiVisu was performed on the yeast_alpha datasets at 10% missing rate. The first 50 largest biclusters were considered in the analysis based on GO annotation. As shown in Fig. 2, the proposed algorithm *BA-LLS* has apparent improvement over LLS. The percentage of overrepresented biclusters in the case of *BA-LLS* is close to that obtained from the dataset with no missing value. For example, at the significance level of 1×10^{-4}, the percentage of overrepresented biclusters for *BA-LLS* and LLS are 42.8% and 36.8% respectively. It is 44.0% for the dataset with no missing value. Therefore, the better estimated gene expression matrix can positively affect the quality of the detected biclusters.

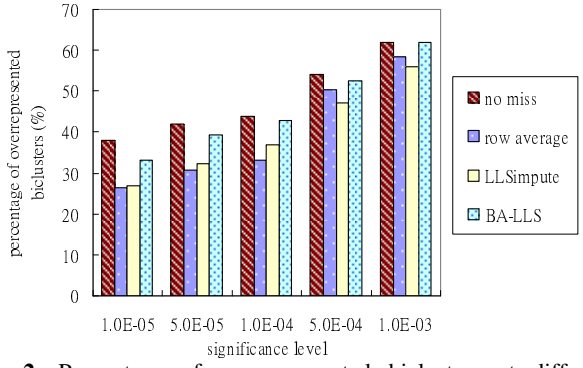

Fig. 2. Percentage of overrepresented biclusters at different significance levels.

4.3. Convergence evaluation

To study the convergence of the proposed integrated approach in missing value estimation and biclustering, mean absolute difference (MAD) of the consecutive estimates of missing values in the first 25 iterations using *BA-LLS* on the real dataset yeast_alpha at 20% missing rate is plotted in Fig. 3(a). It can be seen that the MAD generally decreases with the iteration. Fig. 3(b) shows the corresponding plot of NRMSE. It shows that the average estimation error becomes stable after 10 iterations. Hence, the maximum number of iterations was set to 25.

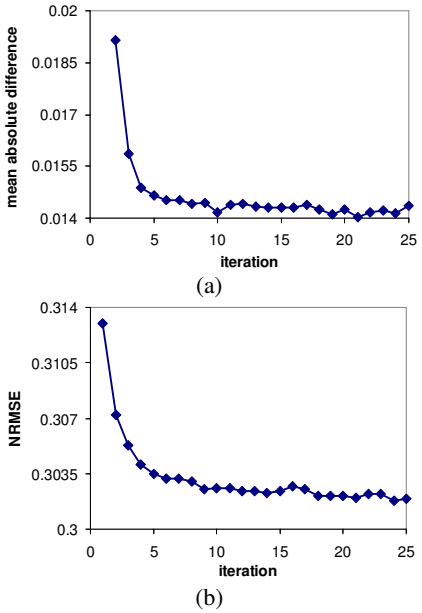

Fig. 3. Plots of (a) MAD and (b) NRMSE between consecutive estimates using BA-LLS on yeast_alpha at 20% missing rate.

5. CONCLUSIONS

In this paper, an integrated framework of biclustering and missing value estimation is proposed so that the two constituent processes can interact constructively and improve each other. In particular, our algorithm iterates between missing value estimation and biclustering. The missing values within the detected biclusters are estimated using the biclusters model while those not in the detected biclusters are estimated using existing algorithms such as LLS or BPCA. In this way, coherence inside biclusters is used to refine the missing values estimation. At the same time, an accurately imputed gene expression matrix facilitates the process of biclusters detection. Experimental results show that the proposed framework can consistently improve both missing value estimation and biclustering analysis.

6. ACKNOWLEDGMENT

This work is supported by the Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, the Hong Kong Polytechnic University.

7. REFERENCES

[1] D.J. Lockhart and E.A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, vol. 405, pp.827 – 836, June 2000.

[2] A.W.C. Liew, N.F. Law and H. Yan, "Missing Value Imputation for Gene Expression Data: Computational Techniques to Recover Missing Data from Available Information", *Briefings in Bioinformatics*, Vol. 12, No. 5, 498-513, 2011.

[3] H. Kim, G.H. Golub and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol.21, no.2, pp.187 – 198, Jan. 2005.

[4] K.O. Cheng, N.F. Law and W.C. Siu, "Iterative Bicluster-based Least Square Framework for Missing Values Estimation", *Pattern Recognition*, Vol. 45, 1281-1289, 2012.

[5] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol.19, no.16, pp.2088 – 2096, Nov. 2003.

[6] S. Friedland, A. Niknejad, M. Kaveh, H. Zare, "An algorithm for missing value estimation for DNA microarray data," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.2, pp.II-1092-1095, 14-19 May 2006.

[7] S.C. Madeira and A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE Transactions on Computational Biology and Bioinformatics*, vol.1, no.1, pp.24-45, Jan-Mar 2004.

[8] A.W.C. Liew, N.F. Law and H. Yan, "Recent Patents on Biclustering Algorithms for Gene Expression Data Analysis", *Recent Patents on DNA and Gene Sequence*, Vol. 5, No. 2, 117-125, 2011.

[9] Y. Cheng and G.M. Church, "Biclustering of expression data," *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology*, pp.93-103, 2000

[10] K.O. Cheng, N.F. Law, W.C. Siu and A.W. Liew, "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization," *BMC Bioinformatics*, vol.9: 210, doi:10.1186/1471-2105-9-210, April 2008.

[11] B.L. Bowerman and R.T. O'Connell, *Linear statistical models: an applied approach*, Duxbury, 2nd edition, USA, 1990.

[12] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol.9, pp.3273 – 3297, Dec. 1998.