ORIGINAL RESEARCH

# Use of biclustering for missing value imputation in gene expression data

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

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

**Correspondence:** N. F. Law. Address: Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. Email: ennflaw@polyu.edu.hk.

**Received:** May 30, 2012 **Accepted:** July 12, 2012 **Online Published:** February 21, 2013
**DOI:** 10.5430/air.v2n2p96 **URL:** http://dx.doi.org/10.5430/air.v2n2p96

## Abstract

DNA microarray data always contains missing values. As subsequent analysis such as biclustering can only be applied on complete data, these missing values have to be imputed before any biclusters can be detected. Existing imputation methods exploit coherence among expression values in the microarray data. In view that biclustering attempts to find correlated expression values within the data, we propose to combine the missing value imputation and biclustering into a single framework in which the two processes are performed iteratively. In this way, the missing value imputation can improve bicluster analysis and the coherence in detected biclusters can be exploited for better missing value estimation. Experiments have been conducted on artificial datasets and real datasets to verify the effectiveness of the proposed algorithm in reducing estimation errors of missing values.

## Key words

Missing value imputation, Biclustering, Gene expression data analysis, Biclusters detection

## 1 Introduction

The gene expression data shows the expression values of ten thousands of genes under hundreds of experimental conditions [1]. The data is useful for various applications such as cellular processes analysis, gene functions prediction and diseases diagnoses [2, 3]. However, some values in the gene expression data are missing due to image corruption, dust or scratches on the slides or experimental errors. As many subsequent analysis tools work on complete datasets only, recovery of missing values is necessary. A straightforward approach is to repeat the experiment; but this might not be feasible because of economic reasons or sometimes limitations of samples. Thus, computational-based missing values imputation becomes necessary and crucial.

Early approaches in missing value imputation are simply to replace the missing values with zeros or row averages. Later, methods that explore coherence inside the gene expression data were developed. There are mainly two ways to explore the coherence information, namely the global and the local approaches [4]. The global approaches assume a global covariance structure in all genes [5-10] while the local approaches exploit local correlations existing in subsets of genes for estimation [11, 12]. Hybrid approaches that combine local and global information have also been proposed [13]. Besides, external information such as gene ontology [14], external datasets [15] or histone acetylation information [16] can be exploited for missing value imputation.

Recently, a multi-stage approach to clustering and missing value imputation was proposed [17]. The rationale behind is that both clustering and missing value imputation explore coherence inside the gene expression data. The combination of missing value imputation and clustering was found to improve the accuracy of missing values imputation [18]. In reality, related genes often co-express under certain conditions only [19]. Biclustering that groups genes and conditions simultaneously should be performed to characterize coherence inside the gene expression data. Hence, biclustering, instead of clustering, should be combined with missing value imputation. In this article, we present a framework that combines biclustering and missing value imputation. A model-based imputation is developed for missing value imputation inside biclusters. In this way, coherence found inside biclusters can be used to improve missing value imputation. As a result, accuracy of missing value imputation and quality of identified biclusters can both be enhanced by the proposed framework.

This paper is organized as follows. In Section 2, reviews on biclustering techniques and missing value imputation method are given. Section 3 presents the proposed framework which incorporates biclustering in missing value imputation. Section 4 shows and discusses experimental results of the proposed algorithm on artificial datasets and real datasets. Finally, we draw a conclusion in Section 5.

# 2 Background

There are two main components of our proposed iterative algorithm, namely biclustering and missing value imputation. In the imputation process, an existing imputation method known as LLSimpute is employed to estimate some of the missing values. In this section, biclustering and LLSimpute are reviewed so as to provide background knowledge of our algorithm.

## 2.1 Biclustering

Data from microarray experiments is frequently given as a large matrix showing expression levels of genes (rows) under different experimental conditions (columns) [1, 3]. Biclustering tries to identify homogeneous patterns known as biclusters in the gene expression data. Biclusters consist of a subset of rows that show related expression patterns across a subset of columns [19]. One useful bicluster model is the additive model [20-25]. Denote $b_{ij}$ as an expression value in a bicluster at position $(i, j)$, the additive model can be described as,

$$b_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{1}$$

where $\mu$ is a constant, $\alpha_i$ is a row dependent factor, $\beta_j$ is a column dependent factor and $\varepsilon_{ij}$ is the error term. As the identification of all biclusters is a NP-hard problem [19, 26], an efficient biclustering algorithm called "BiVisu" that has polynomial-time complexity [22] is adopted in our proposed algorithm to reduce the computational time. It is freely available from [27].

## 2.2 LLSimpute

The local least squares imputation (LLSimpute) [12] 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 data based on Pearson correlation or Euclidean distance. Then the target gene is regressed on these *k* similar genes at the conditions which do not have missing values in the target gene. This process is similar to that used in KNNimpute [9]. However, unlike KNNimpute which computes the regression coefficients as the gene similarity, LLSimpute adopts least square approach and finds the coefficients using pseudo-inverse. The computed regression coefficients are substituted into the regression model to estimate the missing values in the target gene from the values of the similar genes at corresponding conditions. In the work of Kim et al. [12], an automatic selection method is proposed to determine the parameter *k*. In particular, some non-missing values are artificially set to be missing values. Estimation errors on the artificial missing values are evaluated at several different values of *k*. The selected value is the one that gives the lowest estimation error.

# 3 Use of biclustering in missing value imputation

Traditionally, missing entries are imputed prior to and independent of biclusters detection. As missing values imputation and biclustering influence each other, we propose to have a joint approach of missing value imputation and biclusters detection. There are two parts in our joint framework. In the first part, missing values are estimated using an existing imputation method, LLSimpute. Then biclusters are detected using the BiVisu software. In the second part, coherence inside the detected biclusters is used to impute the missing values that fall inside the detected biclusters. In this section, a least square framework for estimating missing values inside a bicluster is first discussed. Then, the proposed iterative framework for biclusters detection and missing values imputation is presented.

## 3.1 Estimation of bicluster model parameters

Eq. (1) shows the additive bicluster model. An approach for the model parameters estimation is to formulate it as a constrained optimization problem as

$$\sum_i \sum_j \left(b_{ij} - \mu - \alpha_i - \beta_j\right)^2 \text{ subject to } \sum_i \alpha_i = 0 \text{ and } \sum_j \beta_j = 0 \qquad (2)$$

The sums of $\{\alpha_i\}$ and $\{\beta_j\}$ are set to zero to ensure a unique additive-related bicluster. Equivalently, the problem can be reformulated using a linear regression model, i.e.,

$$b_{ij} = \alpha_1 p_{ij}^1 + \alpha_2 p_{ij}^2 + \cdots + \alpha_m p_{ij}^m + \beta_1 q_{ij}^1 + \beta_2 q_{ij}^2 + \cdots + \beta_n q_{ij}^n + \mu + \varepsilon_{ij} \qquad (3)$$

where *m* and *n* are the number of rows and columns in the bicluster, $\{p_{ij}^{i'}\}$ and $\{q_{ij}^{j'}\}$ are indicator variables such that

$$p_{ij}^{i'} = \begin{cases} 1, & i = i' \\ 0, & i \neq i' \end{cases} \text{ and } q_{ij}^{j'} = \begin{cases} 1, & j = j' \\ 0, & j \neq j' \end{cases} \qquad (4)$$

The advantage of considering the regression model in eq. (4) instead of eq. (2) is that methods developed for linear models such as ANOVA [28] can be used. The two constraints in eq. (2) lead to

$$\alpha_m = -\sum_{i=1}^{m-1} \alpha_i \text{ and } \beta_n = -\sum_{j=1}^{n-1} \beta_j \qquad (5)$$

Using eq. (5), eq. (3) can be rewritten as,

$$b_{ij} = \alpha_1 x_{ij}^1 + \alpha_2 x_{ij}^2 + \cdots + \alpha_{m-1} x_{ij}^{m-1} + \beta_1 y_{ij}^1 + \beta_2 y_{ij}^2 + \cdots + \beta_{n-1} y_{ij}^{n-1} + \mu + \varepsilon_{ij} \qquad (6)$$

where $x_{ij}^{i'} = p_{ij}^{i'} - p_{ij}^m$, $i'$= 1, 2, …, *m*-1 and $y_{ij}^{j'} = q_{ij}^{j'} - q_{ij}^n$, $j'$= 1, 2, …, *n*-1. Eq. (6) can be rewritten in a matrix form as,

$$\mathbf{b} = \mathbf{A}\mathbf{w} + \boldsymbol{\varepsilon} \qquad (7)$$

where **b** is a *mn*×1 vector obtained by concatenating columns of the bicluster one by one, $\mathbf{w} = (\alpha_1, \cdots, \alpha_{m-1}, \beta_1, \cdots, \beta_{n-1}, \mu)^{\mathrm{T}}$ is a (*m*+*n*–1)×1 vector of the model parameters, **A** is a *mn*×(*m*+*n*-1) data matrix of variables $\{x_{ij}^{i'}\}$ and $\{y_{ij}^{j'}\}$, and ε is a *mn*×1 vector of noise terms . The least square solution of eq. (7) is given by,

$$\hat{\mathbf{w}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \qquad (8)$$

By using eq. (8), bicluster model parameters can be obtained for all the detected biclusters.

## 3.2 Outlier elimination

Section 3.1 presents least square estimation of an additive bicluster model using all data in the bicluster. In practice, outliers may appear in the bicluster and reduce estimation accuracy. Thus, covariance ratio statistics [28] is adopted to eliminate the outliers. The covariance ratio of a datum at position $t = (i_0, j_0)$ is defined as,

$$CVR_t = \left|(\mathbf{A}_t^{\mathrm{T}}\mathbf{A}_t)^{-1}\hat{\sigma}_t^2\right| / \left|(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\hat{\sigma}^2\right| \tag{9}$$

where $\hat{\sigma}^2$ is the estimated variance calculated from all data, $\hat{\sigma}_t^2$and $\mathbf{A}_t$ are the estimated variance and variable matrix excluding the datum at *t*. 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 data, the removal of the datum at *t* can reduce the variance significantly [28]. Thus, the datum at *t* is omitted in bicluster model estimation. After removing these outlier data, two matrices: a $(mn$-$r)\times 1$ vector $\tilde{\mathbf{b}}$ and a $(mn$-$r)\times(m$+$n$-$1)$ matrix $\widetilde{\mathbf{A}}$, where *r* is the number of the removed data, are formed. The least square estimation in eq. (8) is performed through replacing $\mathbf{b}$ by $\tilde{\mathbf{b}}$ and $\mathbf{A}$ by $\widetilde{\mathbf{A}}$. In order to avoid performing an inverse of a singular or nearly singular matrix, the determinant of $\widetilde{\mathbf{A}}^{\mathrm{T}}\widetilde{\mathbf{A}}$ is calculated. If the determinant is less than a small positive value, the estimation is not performed for that particular bicluster.

It was found that the least square estimation with at most one data removed ($r \leq 1$) and the covariance ratio can be expressed in a non-matrix form. In this way, computational complexity can be reduced. The derivation of non-matrix form solutions are provided in next section. However, eq. (8) is still important in calculating the estimates after outlier elimination because the number of outliers can be larger than 1.

## 3.3 Least Square Solutions in a Non-matrix Form

An alternative way to solve the constraint optimization problem (2) is to use the method of Lagrange multipliers. Let $I = \{1, 2, \dots, m\}$ and $J = \{1, 2, \dots, n\}$ be the sets of row and column indices of the bicluster respectively. Furthermore, denote the set of positions in the bicluster as $D = I \times J$. In the method of Lagrange multipliers, the constraint optimization problem is equivalent to minimize the following Lagrange function,

$$\Lambda(\widetilde{\mathbf{w}}, \lambda_1, \lambda_2) = \sum_{(i,j)\in D}\left(b_{ij} - \mu - \alpha_i - \beta_j\right)^2 + \lambda_1 \sum_{i\in I}\alpha_i + \lambda_2 \sum_{j\in J}\beta_j \tag{10}$$

where $\widetilde{\mathbf{w}} = (\alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_n, \mu)^{\mathrm{T}}$ is a vector of parameters. By setting the partial derivatives of $\Lambda(\widetilde{\mathbf{w}}, \lambda_1, \lambda_2)$ to zeros, we obtain the following set of linear equations

$$\begin{cases} \dfrac{\partial\Lambda}{\partial\alpha_i} = 0,\ i = 1, 2, \dots, m \\ \dfrac{\partial\Lambda}{\partial\beta_j} = 0,\ j = 1, 2, \dots, n \\ \dfrac{\partial\Lambda}{\partial\lambda_u} = 0, \qquad u = 1, 2 \end{cases} \tag{11}$$

By solving the above set of linear equations, the model parameters can be estimated as

$$\begin{cases} \hat{\mu} = \dfrac{1}{mn}s \\ \hat{\alpha}_i = \dfrac{1}{n}r_i - \hat{\mu},\ i = 1, 2, \dots, m \\ \hat{\beta}_j = \dfrac{1}{m}c_j - \hat{\mu},\ j = 1, 2, \dots, n \end{cases} \tag{12}$$

where $s = \sum_{(i,j)\in D} b_{ij}$, $r_i = \sum_{j\in J} b_{ij}$ and $c_j = \sum_{i\in I} b_{ij}$. The variance of the model can be estimated as

$$\hat{\sigma}^2 = \frac{1}{mn-m-n+1}\sum_{(i,j)\in D}\left(b_{ij} - \hat{b}_{ij}\right)^2 \tag{13}$$

where $\hat{b}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$. Compared with eq.(8), eq. (12) provides solution in a non-matrix form. From eq. (12), it is obvious that the least square solution always exists. The use of eq. (12) instead of eq. (8) can avoid matrix inversion which can be computationally expensive and instable.

In the outlier elimination, the parameter estimation is performed with one of the data removed. Let $t = (i_0, j_0)$ be the position of the removed datum. Define $I' = I\backslash\{i_0\}, J' = J\backslash\{j_0\}$ and $D' = D\backslash\{t\}$. The corresponding Lagrange function is given by

$$\Lambda'(\widetilde{\mathbf{w}}, \lambda_1, \lambda_2) = \sum_{(i,j)\in D'}\left(b_{ij} - \mu - \alpha_i - \beta_j\right)^2 + \lambda_1 \sum_{i\in I}\alpha_i + \lambda_2 \sum_{j\in J}\beta_j \tag{14}$$

Following the previous approach, solutions in a non-matrix form can be proved to be

$$\begin{cases} \hat{\mu}' = \frac{1}{(m-1)(n-1)}\left(\frac{mn-m-n}{mn}s' + \frac{1}{n}r_{i_0}{}' + \frac{1}{m}c_{j_0}{}'\right) \\ \hat{\alpha}_i' = \begin{cases} \frac{1}{n}r_i{}' - \hat{\mu}',\ i \in I' \\ (m-1)\hat{\mu}' + \frac{1}{n}(r_i{}' - s'),\ i = i_0 \end{cases} \\ \hat{\beta}_j{}' = \begin{cases} \frac{1}{m}c_j{}' - \hat{\mu}',\ j \in J' \\ (n-1)(\hat{\mu}' + \hat{\alpha}_{i_0}{}') - r_{i_0}{}',\ j = j_0 \end{cases} \end{cases} \tag{15}$$

where $s' = \sum_{(i,j)\in D'} b_{ij}$, $r_i{}' = \begin{cases} \sum_{j\in J} b_{ij}, i \in I' \\ \sum_{j\in J'} b_{ij}, i = i_0 \end{cases}$ and $c_j{}' = \begin{cases} \sum_{i\in I} b_{ij}, j \in J' \\ \sum_{i\in I'} b_{ij}, j = j_0 \end{cases}$. The variance estimate can be obtained as

$$\hat{\sigma}'^2 = \frac{1}{mn-m-n}\sum_{(i,j)\in D'}\left(b_{ij} - \hat{b}_{ij}{}'\right)^2 \tag{16}$$

where $\hat{b}_{ij}{}' = \hat{\mu}' + \hat{\alpha}_i{}' + \hat{\beta}_j{}'$. Eq.(15) implies that the least square solutions always exist even one of the data is ignored provided that $m > 1$ and $n > 1$.

In addition to the estimates of parameters, the covariance ratio of the datum at $t$ in eq. (9) can be re-expressed in a non-matrix form [28] as below

$$CVR_t = \frac{1}{1-h_{tt}}(\hat{\sigma}'/\hat{\sigma})^{2k^*} \tag{17}$$

where $k^* = m + n - 1$ and $h_{tt}$ is the diagonal element of the hat matrix $\mathrm{H} = \mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}}$ which corresponds to the datum at $t$. $h_{tt}$ can be shown to be

$$h_{tt} = \frac{m+n-1}{mn} \tag{18}$$

Unlike the least square solution, eq.(17) requires either ($m > 2$ and $n > 3$) or ($m > 3$ and $n > 2$) so as to have a meaningful estimation of variances and non-zero value of 1-$h_{tt}$. Furthermore, it is possible that $\hat{\sigma}$ is equal to or close to zero. To ensure validity of eq.(17), the outlier estimation is performed only when $\hat{\sigma}$ is above a certain small threshold.

## 3.4 The iterative imputation-biclustering algorithm

Since missing value imputation and biclustering are interrelated, a joint approach for these two processes is proposed. In particular, missing value imputation and biclusters detection are applied iteratively. The procedure of the proposed iterative algorithm is summarized as follows,

1) Initialization: the LLSimpute, is applied to impute all missing values in the gene expression data.

2) Biclusters detection using BiVisu.

3) For all detected biclusters, the bicluster model parameters are estimated using least square approach. Then the missing values inside the currently detected biclusters are updated as follows:

   - For missing entries that are also inside the biclusters detected in previous iteration, re-estimation is not performed.

   - For other missing entries, a partial update is performed using sum of the new estimates obtained from the bicluster model with weight $\rho$ and the old estimates with weight 1- $\rho$ , where $\rho$ is equal to one in the first iteration and is between 0 and 1 for subsequent iterations.

4) If the mean absolute difference (MAD) in missing value imputation falls below a pre-set threshold or the maximum number of iterations is reached, output the current estimates and terminates the algorithm.

5) For missing entries that do not fall into the detected biclusters in the current iteration, LLSimpute in step 1 is performed.

   - If the missing entries belong to biclusters detected in previous iteration, the missing entries are updated partially as the sum of the new estimates with weight $\eta$ and the old estimates with weight 1- $\eta$ ;

   - Otherwise, they are replaced by the new estimates directly.

6) Go back to Step 2 for biclusters detection and then missing values imputation.

In steps 1 and 5, LLSimpute [29] is adopted because LLSimpute is comparable to other popular existing approaches such as BPCA [2, 30, 31] and suitable for datasets with high complexity, e.g. experiments involving multiple exposures. Furthermore, its parameter can be determined automatically as described in Section 2.2. In step 2, additive-related biclusters are detected by BiVisu. In step 3, the bicluster model parameters are estimated for all the detected biclusters as described in Sections 3.1-3.3. In outlier elimination, only missing entries are considered because their values are originally unavailable which implies that there is usually a large error.

In step 4, the iteration is stopped if the MAD in missing value imputation falls below a pre-set threshold or the maximum number of iterations is reached. The purpose of step 5 is to update those missing values not in the currently detected biclusters using LLSimpute. If the missing values fall in biclusters detected in previous iteration, a partial update is performed.

# 4 Experimental results

The proposed iterative framework is evaluated by experiments on two artificial datasets and three real datasets. In the first artificial datasets, 10 datasets of size 360×30 were generated. Each of them was initially assigned with uniformly distributed random values in the range of -10 and 10. Then, 16 additive-related biclusters of size 27×10 were implanted into each of the datasets with various degrees of row/column overlaps. In these datasets, the biclusters cover 40% region. Finally, 30dB noise was added. The second artificial datasets was generated in a similar way except that 18 biclusters of size 30×12 were embedded in each of them. Thus, the percentage of bicluster region in these datasets was 60%. Two of the real datasets are the yeast cell cycle expression dataset yeast_alpha and yeast_elu [32]. The yeast dataset whose data were synchronized using α factor arrest is denoted by yeast_alpha while the one whose data were synchronized by elutriation is referred as yeast_elu. Their sizes are 4489×18 and 5766×14 respectively. The third real dataset gut_cell contains gene expression data of Intestinal epithelial cells [33]. The expression levels are zero-transformed and selected to have at least 3 fold change. Its size is 933×10 after removing two columns with most missing values and rows with missing values.

In the experiments for artificial datasets, *r* % of values were set to be missing randomly inside and outside the biclusters, where *r* = 1, 5, 10, 15, 20. The estimation was repeated five times for each dataset. A similar experiment was performed on each real dataset ten times except that the missing values were distributed over the whole matrix as the 'ground-truth' biclusters are unknown. The estimation error was measured by the normalized root mean square error (NRMSE). Let *S* be a set of positions of missing values in a data matrix. The NRMSE is defined as,

$$NRMSE = \sqrt{\sum_{(i,j)\in S}\left(b_{ij} - \hat{b}_{ij}\right)^2/|S|}/\tilde{\sigma} \tag{19}$$

where $b_{ij}$ is the actual value in the data matrix at position $(i, j)$, $\hat{b}_{ij}$ is the corresponding estimate using imputation, $|S|$ is the cardinality of the set *S* and $\tilde{\sigma}$ is the standard deviation of the actual values at positions in *S*. NRMSE has a low value for an accurate estimation.

## 4.1 Performance on missing value imputation using "true" biclusters information

Biclustering implemented in BiVisu [22] was first applied to identify biclusters in the three real datasets yeast_alpha, yeast_elu and gut_cell. The detected biclusters are regarded as "true" biclusters. Then for each of the real datasets, ten datasets with missing values were generated at various missing rates. After that, all the missing values were initially estimated using LLSimpute. The missing values insides the "true" biclusters were re-estimated using least square method for additive models as described in Sections 3.1-3.3.

In the experiments, the parameters of the biclustering algorithm were varied to obtain optimal performance. On the other hand, the number of similar genes *k* is automatically selected as the value which results in the smallest NRMSE of artificial missing values using the selection method reviewed in Section 2.2 [12]. Table 1 provides averages of selected values of *k*. In addition to missing rates and nature of datasets, the selected values usually vary during iterations. The results suggest that the value of *k* should be selected adaptively as in the selection method rather than being set at a fixed value.

**Table 1.** Selected values of *k* at different missing rates

| Dataset | Missing rates | | | | |
|---|---|---|---|---|---|
| | **1%** | **5%** | **10%** | **15%** | **20%** |
| yeast_alpha | 366-380 | 249-377 | 263-431 | 238-694 | 216-740 |
| yeast_elu | 222-284 | 352-681 | 235-735 | 303-800 | 355-1642 |
| gut_cell | 276-355 | 229-450 | 267-331 | 261-350 | 247-319 |

Figures. 1 - 3 show the NRMSE for datasets yeast_alpha, yeast_elu and gut_cell respectively. Results of the local method LLSimpute, the global approach BPCA and the hybrid approach POCSimpute without cyclic loss model are included for comparison. It can be seen that the estimation using "true" biclusters outperforms the three existing algorithms especially at high missing rates. At 20% missing rate, 12% – 26% improvement can be achieved using bicluster information. This observation is related to the change of correlation embedded among expression data. The more missing values are, the less data correlation can be conserved. The improvement of the estimation using “true” biclusters can be attributed to the strong homogeneity existing in biclusters. The results verify that the bicluster information is useful for missing values imputation.

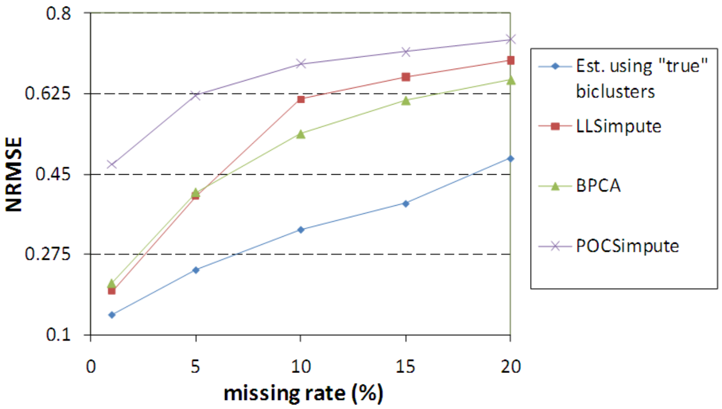


**Figure 1.** NRMSE of imputation using “true” biclusters, LLSimpute, BPCA and POCSimpute on yeast_alpha

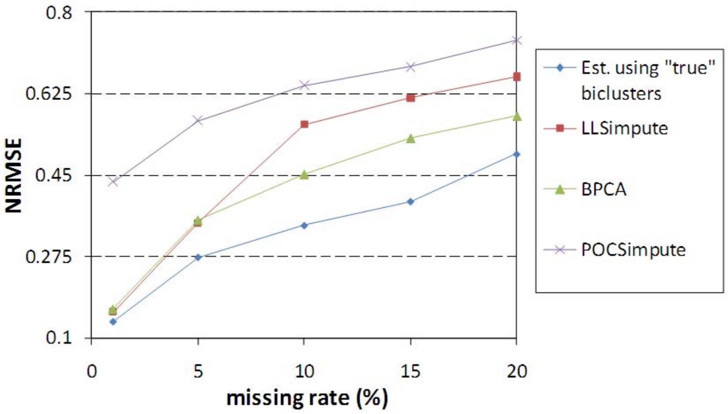


**Figure 2.** NRMSE of imputation using “true” biclusters, LLSimpute, BPCA and POCSimpute on yeast_elu

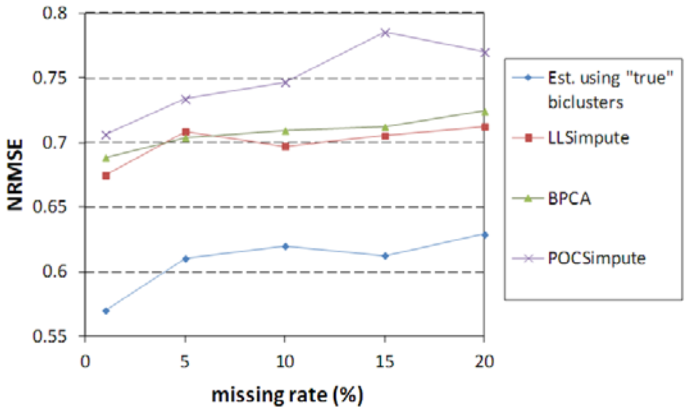


**Figure 3.** NRMSE of imputation using “true” biclusters, LLSimpute, BPCA and POCSimpute on gut_cell

## 4.2 Performance on missing value imputation in the proposed iterative framework

In the experiments for the artificial datasets, biclusters were set to have at least 21 rows and 8 columns. The update weights $\rho$ and $\eta$ were set to 0.7 so that a slightly higher weighting is given to the new estimates than the previous estimates. The parameter $k$ of LLSimpute is determined by the algorithm itself as in the experiment discussed in Section 4.1. Table 2 shows the NRMSE for the artificial datasets with 40% bicluster region. Inside biclusters, the proposed bicluster-based algorithm achieves a smaller NRMSE than LLSimpute over all the missing rates. The improvement in NRMSE decreases from the highest value 20.10% to the lowest value 11.19% when missing rate increases from 1% to 20%. When missing rate increases, the bicluster structure is destroyed so that the estimation accuracy is affected. Table 3 shows the results for the artificial datasets with 60% bicluster region. At the same missing rate, lower NRMSE was found for the datasets with 60% than those with 40% bicluster region. It is because datasets with larger bicluster region exhibit higher coherence. For the missing values outsides biclusters, the proposed algorithm performs a bit poorer because there is no correlation among the data. As the NRMSE outside biclusters is much larger that inside biclusters, the overall improvement appears to be not as high as that inside biclusters. Nevertheless, the results show that the use of coherence information in biclusters can positively affect the accuracy of the missing value imputation.

**Table 2.** Comparison of NRMSEs using LLSimpute and the proposed bicluster-based algorithm on artificial datasets with 40% bicluster region

| Algorithm | Region | Missing rate | | | | |
|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 15% | 20% |
| LLSimpute | Bicluster | 0.6855 | 0.7019 | 0.7128 | 0.7301 | 0.7446 |
| | Other | 1.0957 | 1.0917 | 1.0792 | 1.0737 | 1.0687 |
| | Overall | 1.0307 | 1.0290 | 1.0197 | 1.0175 | 1.0156 |
| Proposed | Bicluster | 0.5477 | 0.5685 | 0.5783 | 0.6217 | 0.6613 |
| | Other | 1.0998 | 1.0957 | 1.0818 | 1.0804 | 1.0745 |
| | Overall | 1.0191 | 1.0170 | 1.0058 | 1.0097 | 1.0097 |
| Improvement | Bicluster | 20.10% | 19.01% | 18.87% | 14.85% | 11.19% |
| | Other | -0.37% | -0.37% | -0.24% | -0.62% | -0.54% |
| | Overall | 1.13% | 1.17% | 1.36% | 0.77% | 0.58% |

**Table 3.** Comparison of NRMSEs using LLSimpute and the proposed bicluster-based algorithm on artificial datasets with 60% bicluster region

| Algorithm | Region | Missing rate | | | | |
|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 15% | 20% |
| LLSimpute | Bicluster | 0.5074 | 0.5243 | 0.5427 | 0.5683 | 0.5908 |
| | Other | 1.1557 | 1.1319 | 1.1215 | 1.1078 | 1.1000 |
| | Overall | 0.9779 | 0.9669 | 0.9627 | 0.9572 | 0.9564 |
| Proposed | Bicluster | 0.3742 | 0.3907 | 0.4095 | 0.4426 | 0.4816 |
| | Other | 1.1596 | 1.1396 | 1.1361 | 1.1240 | 1.1186 |
| | Overall | 0.9591 | 0.9508 | 0.9512 | 0.9468 | 0.9498 |
| Improvement | Bicluster | 26.25% | 25.48% | 24.54% | 22.12% | 18.48% |
| | Other | -0.34% | -0.68% | -1.30% | -1.46% | -1.69% |
| | Overall | 1.92% | 1.67% | 1.19% | 1.09% | 0.69% |

For the real datasets, biclusters were set to have at least 5 rows and 4 columns in the experiments. The NRMSE for the real datasets yeast_alpha and yeast_elu are provided in Tables 4 and 5 respectively. At low missing rates of 1% and 5%, the improvement of the proposed iterative algorithm over LLSimpute is not obvious. However, the NRMSE obtained by using the proposed algorithm is reduced by 14.03% and 19.67% in yeast_alpha and yeast_elu respectively at 10% missing rate. At high missing rates, the coherence structure becomes ambiguous so stricter models based on biclusters is significant. At

20% missing rates, our improvement over LLSimpute are about 5.85% and 10.58% for yeast_alpha and yeast_elu respectively. For the real dataset gut_cell, the performance of the proposed algorithm is still positive although the improvement is smaller as shown in Table 6. The bicluster-based estimation using “true” biclusters studied in Section 4.1 should impose the upper bound on the performance of our proposed iterative algorithm as it uses the biclusters identified in gene expression data without missing values. The performance of the proposed algorithm is considerably poorer than the estimation using “true” biclusters in the three real datasets. For example at 20% missing rates of the yeast datasets, the NRMSE for LLSimpute, our proposed algorithm using true biclusters and our proposed iterative algorithm are respectively 0.7, 0.5 and 0.6. Hence, the accuracy of biclustering should have a great impact on missing value estimation. The difference also means that there is still a large room for improvement. How to enhance the estimation based on inaccurate biclusters is one of important problems to solve in our future works.

**Table 4.** Comparison of NRMSEs using LLSimpute and the proposed bicluster-based algorithm on the yeast dataset yeast_alpha

| Algorithm | Missing rate | | | | |
|---|---|---|---|---|---|
| | 1% | 5% | 10% | 15% | 20% |
| LLSimpute | 0.1950 | 0.4024 | 0.6129 | 0.6609 | 0.6980 |
| Proposed | 0.1944 | 0.3966 | 0.5269 | 0.6027 | 0.6572 |
| Improvement | 0.31% | 1.44% | 14.03% | 8.81% | 5.85% |

**Table 5.** Comparison of NRMSEs using LLSimpute and the proposed bicluster-based algorithm on the yeast dataset yeast_elu

| Algorithm | Missing rate | | | | |
|---|---|---|---|---|---|
| | 1% | 5% | 10% | 15% | 20% |
| LLSimpute | 0.1559 | 0.3467 | 0.5588 | 0.6153 | 0.6609 |
| Proposed | 0.1559 | 0.3455 | 0.4489 | 0.5393 | 0.5910 |
| Improvement | 0% | 0.35% | 19.67% | 12.35% | 10.58% |

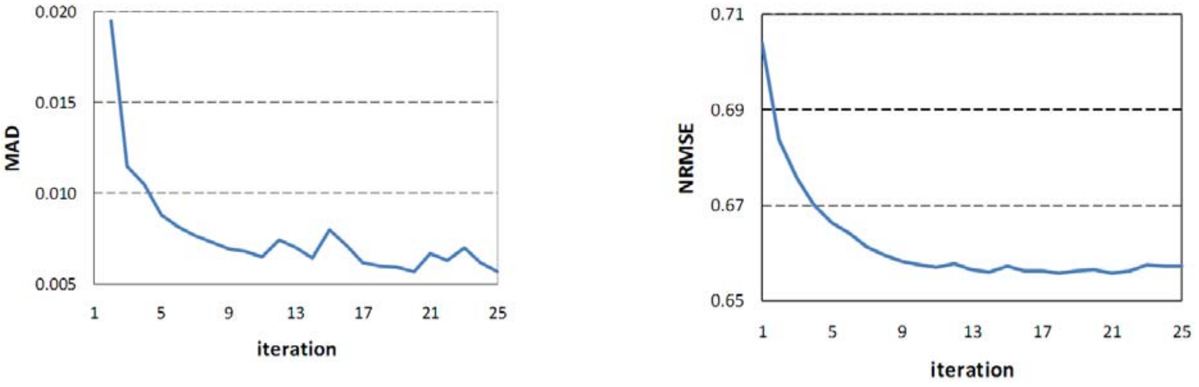
**Table 6.** Comparison of NRMSEs using LLSimpute and the proposed bicluster-based algorithm on the real dataset gut_cell

| Algorithm | Missing rate | | | | |
|---|---|---|---|---|---|
| | 1% | 5% | 10% | 15% | 20% |
| LLSimpute | 0.6746 | 0.7086 | 0.6972 | 0.7053 | 0.7125 |
| Proposed | 0.6688 | 0.7017 | 0.6968 | 0.7030 | 0.7114 |
| Improvement | 0.86% | 0.97% | 0.06% | 0.33% | 0.15% |

As the proposed algorithm iteratively applies missing value imputation and biclustering, the computational complexity should be higher than non-iterative imputation algorithms such as LLSimpute. However, with the use of non-matrix form solutions in the outlier elimination as described in Section 3.3, skip of re-estimation in step 3, part 1 and step 5 and the use of fast biclustering algorithm (BiVisu), we can compromise the estimation accuracy with computational complexity in the real situation. The prototypes of the implementation in Matlab show that the processing time of the proposed algorithm with 25 iterations is about 12 times higher than that of LLSimpute in the dataset yeast_alpha at 20% missing rate. The ratio of increase in computation time is even lower than the number of iterations.

## 4.3. Convergence analysis of the proposed iterative framework

To study the convergence of the proposed iterative approach in missing value imputation and biclustering, mean absolute difference (MAD) between the consecutive estimates of missing values in the first 25 iterations on the real dataset yeast_alpha at 20% missing rate is plotted in Figure 4(a). Although the MAD does not become zero within 25 iterations, it decreases to a small value. Figure 4(b) shows the corresponding plot of NRMSE. It shows that the average estimation error decreases as well. This justifies the use of the MAD as one of termination criteria in our algorithm. Furthermore, the experimental results also suggest that the proposed algorithm can terminate at about 10 iterations with no significant change in average error.



(a) (b)

**Figure 4.** Plots of (a) MAD and (b) NRMSE between consecutive estimates using the proposed bicluster-based algorithm on yeast_alpha at 20% missing rate

# 5 Conclusions

In this paper, a joint approach for missing value imputation and biclustering has been presented so that the two constituent processes can interact constructively and improve each others' results. In particular, our algorithm iterates between missing value imputation and biclustering. At first, the missing values are initialized using LLSimpute. Then a biclustering algorithm implemented in BiVisu is applied to find biclusters. The coherence inside the detected biclusters is utilized to estimate the missing values inside biclusters through additive-models. In order to improve the estimation outside biclusters, LLSimpute is re-applied to estimate missing values outside the biclusters.

Experiments have been conducted on artificial datasets and gene expression data of yeast and intestinal epithelial cells. We found that our iterative approach achieves a smaller estimation error than the standalone missing value imputation method LLSimpute. This shows that the coherence inside biclusters is able to improve the accuracy of the missing values imputation. We have also studied the convergence of the algorithm in terms of change in estimates between consecutive iterations and overall estimation error. Both measures decrease with the increase in iteration on average. Furthermore, the change in estimation error is small after 10 iterations. On the other hand, it is found that there is a gap between imputation using estimated biclusters and "true" biclusters. Hence our future work is to improve our algorithm to approach the performance using the "true" biclusters. In addition, we will investigate the use of other bicluster models such as multiplicative models in the iterative framework.

## Acknowledgement

This work was supported by the Center for Signal Processing, Department of Electronic and Information Engineering, the Hong Kong Polytechnic University (Projects A-PJ24 and G-U876).

# References

[1] Lockhart, D.J. and Winzeler, E.A. Genomics, gene expression and DNA arrays. Nature. 2000; 405: 827 – 836. http://dx.doi.org/10.1038/35015701

[2] Brock, G.N. and et al. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics. 2008; 9: 12. http://dx.doi.org/10.1186/1471-2105-9-12

[3] Kurella, M. and et al. DNA microarray analysis of complex biologic processes. Journal of the America Society of Nephrology. 2001; 12: 1072 – 1078

[4] Liew, A.W.C., Law, N.F. and Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics. 2011; 12(5): 498 – 513. http://dx.doi.org/10.1093/bib/bbq080

[5] Friedland, S., Niknejad, A. and Chihara, L. A simultaneous reconstruction of missing data in DNA microarrays. Linear Algebra and its Applications. 2006; 416 (1): 8–28. http://dx.doi.org/10.1016/j.laa.2005.05.009

[6] Liu, C.-C., Dai, D.-Q. and Yan, H. The theoretic framework of local weighted approaximation for microarray missing value estimation. Pattern Recognition. 2010; 43(8): 2993-3002. http://dx.doi.org/10.1016/j.patcog.2010.02.006

[7] Oba, S. and et al. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003; 19 (16): 2088 – 2096. http://dx.doi.org/10.1093/bioinformatics/btg287

[8] Scholz, M. and et al. Non-linear PCA: a missing data approach. Bioinformatics. 2005; 21 (20): 3887 – 3895. http://dx.doi.org/10.1093/bioinformatics/bti634

[9] Troyanskaya, O. and et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17 (6): 520 – 525. http://dx.doi.org/10.1093/bioinformatics/17.6.520

[10] Wang, X. and et al. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. BMC Bioinformatics. 2006; 7:32. http://dx.doi.org/10.1186/1471-2105/7/32

[11] Bo, T.H., Dysvik, B. and Jonassen, I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research. 2004; 32(3): e34. http://dx.doi.org/10.1093/nar/gnh026

[12] Kim, H., Golub, G.H. and Park, H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics. 2005; 21 (2): 187–198. http://dx.doi.org/10.1093/bioinformatics/bth499

[13] Gan, X., Liew, A.W.C. and Yan, H. Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucleic Acids Research. 2006; 34 (5): 1608-1619. http://dx.doi.org/10.1093/nar/gkl047

[14] Yuikkala, J. and et al. Improving missing value estimation in microarray data with gene ontology. Bioinformatics. 2006; 22 (5): 566-572. http://dx.doi.org/10.1093/bioinformatics/btk019

[15] Hu, J. and et al. Integrative missing value estimation for microarray data. BMC Bioinformatics. 2006; 7:449. http://dx.doi.org/10.1186/1471-2105-7-449

[16] Xiang, Q. and et al. Missing value imputation for microarray gene expression data using histone acetylation information. BMC Bioinformatics. 2008; 9:252. http://dx.doi.org/10.1186/1471-2105-9-252

[17] Wong, D.S.V., Wong, F.K. and Wood, G.R. A multi-stage approach to clustering and imputation of gene expression profiles. Bioinformatics. 2007; 23(8): 998-1005. http://dx.doi.org/10.1093/bioinformatics/btm053

[18] Friedland, S. and et al. An algorithm for missing value estimation for DNA microarray data. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2006; 2: II-1092-1095. http://dx.doi.org/10.1109/ICASSP.2006.1660537

[19] Madeira, S.C. and Oliveira, A.L. Biclustering algorithms for biological data analysis: a survey. IEEE transactions on computational biology and bioinformatics. 2004; 1 (1): 24–45. http://dx.doi.org/10.1109/TCBB.2004.2

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

[21] Cheng, K.O. and et al. BiVisu: software tool for bicluster detection and visualization. Bioinformatics. 2007; 23: 2342 – 2344. http://dx.doi.org/10.1093/bioinformatics/btm338

[22] Cheng, K.O. and et al. Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. BMC Bioinformatics. 2008; 9: 210. http://dx.doi.org/10.1186/1471-2105-9-210

[23] Gan, X., Liew, A.W.C. and Yan, H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. BMC Bioinformatics. 2008; 9:209. http://dx.doi.org/10.1186/1471-2105-9-209

[24] Liu, X. and Wang, L. Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics. 2007; 23 (1): 50 – 56. http://dx.doi.org/10.1093/bioinformatics/btl560

[25] Yoon, S. and et al. Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2005; 2 (4): 339–354. http://dx.doi.org/10.1109/TCBB.2005.55

[26] Prelic, A. and et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006; 22 (9): 1122-1129. http://dx.doi.org/10.1093/bioinformatics/btl060

[27] Law, N.F. 2008. Homepage for BiVisu. Available from: http://www.eie.polyu.edu.hk/~nflaw/Biclustering/index.html.

[28] Bowerman, B.L. and O'Connell, R.T. Linear statistical models: an applied approach. 1990.

[29] Park, H. Missing value estimation software in MATLAB. Available from: http://www.cc.gatech.edu/~hpark/softwareMVE.html.

[30] Sun, Y., Braga-Neto, U. and Dougherty, E.R. Impact of missing value imputation on classification for DNA microarray gene expression data – a model-based study. EURASIP Journal on Bioinformatics and Systems Biology. 2009. http://dx.doi.org/10.1155/2009/504069.

[31] Tuikkala, J. Missing value imputation improves clustering and interpretation of gene expression microarray data. BMC Bioinformatics. 2008; 9:202. http://dx.doi.org/10.1186/1471-2105/9/202

[32] Spellman, P.T. and et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell. 1998; 9(12): 3273-3297.

[33] Baldwin, David N. and et al. A gene-expression program reflecting the innate immune response of cultured intestinal epithelial cells to infection by Listeria monocytogenes. Genome Biology. 2002; 4: R2. http://dx.doi.org/10.1186/gb-2002-4-1-r2