Iterative Bicluster-based Least Square Framework for Missing Values Estimation

K.O. Cheng[a,*], N.F. Law[b] and W.C. Siu[c]

Center for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.

E-mail: {encko[a], ennflaw[b], enwcsiu[c]}@polyu.edu.hk

* Corresponding author. Tel.: 852-2766 6201; fax: 852-2362 8439

ABSTRACT

DNA microarray experiment inevitably generates gene expression data with missing values. An important and necessary pre-processing step is thus to impute these missing values. Existing imputation methods exploit gene correlation among all experimental conditions for estimating the missing values. However, related genes coexpress in subsets of experimental conditions only. In this paper, we propose to use biclusters which contain similar genes under subset of conditions for characterizing the gene similarity and then estimating the missing values. To further improve the accuracy in missing value estimation, an iterative framework is developed with a stopping criterion on minimizing uncertainty. Extensive experiments have been conducted on artificial datasets, real microarray datasets as well as one non-microarray dataset. Our proposed biclusters-based approach

is able to reduce errors in missing value estimation.

Keywords: missing value imputation, biclustering, iterative estimation, gene expression analysis

1. INTRODUCTION

The DNA microarray technology [1] allows acquisition of gene expression data from ten thousands of genes under hundreds of experimental conditions. The data is useful for various applications such as cellular processes analysis, gene functions prediction and diseases diagnoses [2-5]. However, the gene expression data is often incomplete in that some values are lost because of image corruption, dust or scratches on the slides and experimental errors. As many subsequent analysis tools work on complete datasets only, recovery of missing values is necessary [6, 7]. A straightforward approach is to repeat the experiment; but this might not be feasible because of economic reasons or sometimes limitations of samples. Thus, computation based estimation becomes necessary and crucial.

Early approaches to deal with missing entries are simply to replace them with zeros or row averages. Later, coherence inside the gene expression data is used for their estimation. There are mainly two ways to explore the coherence information, namely the global and the local approaches [8]. The global approaches assume a global covariance structure in all genes [9, 10] while the local approaches exploit correlations among certain genes only [11-14].

For both local and global approaches, a measure of gene similarity is critical for finding the coherence structure. Often, the gene similarity is measured based on the similarity of the expression profiles across all experimental conditions [15]. In reality, genes are co-expressed under certain conditions only [16-20]. Hence the gene similarity should be measured by considering only those related experimental conditions, rather than all the conditions. In this article, we incorporate this biclustering idea into the framework of local least square imputation (LLSimpute) [12] for missing value estimation. In particular, genes and conditions are grouped alternately based on a weighted distance and correlation respectively. A regression model is then used for least square based missing value estimation. To further improve the selection of coherent genes and correlated conditions, an iterative framework is developed. A stopping criterion that minimizes the uncertainty in estimation is introduced to improve the convergence of the proposed algorithm.

This paper is organized as follows. In Section 2, the LLSimpute for missing value estimation is reviewed. Section 3 then presents the proposed algorithm. In Section 4, the proposed algorithm is evaluated on artificial datasets, real microarray datasets as well as a non-microarray dataset. Besides, issues such as convergence and parameters sensitivity are also addressed. Finally, a conclusion is drawn in Section 5.

2. Review – Local Least Square Imputation

Data from Microarray experiments is frequently given as a large matrix showing expression levels of genes (rows) under different experimental conditions (columns) [1]. It is estimated that the data can contain 10% missing values and in some datasets, up to 90% of genes have at least one missing values. The local least squares imputation (LLSimpute) [12] is a popular state-of-the-art method that explores local coherence information in the gene expression data for missing value estimation. For each target gene which contains at least one missing value, *k* most similar genes are selected based on either Pearson correlation or Euclidean distance. Then the missing values are estimated under a least square framework. Let $\mathbf{E}$ be the expression matrix consisting of *m* genes and *n* conditions. Denote the target gene with *p* missing values as $\mathbf{g}_t^T \in R^{1\times n}$. Without loss of generality, assume that all the missing values are located in the first *p* conditions. Hence,

$$\mathbf{g}_t^T = \begin{pmatrix} \boldsymbol{\alpha}^T & \mathbf{w}^T \end{pmatrix} \tag{1}$$

where $\boldsymbol{\alpha}^T \in R^{1\times p}$ is a 1×*p* vector containing the *p* missing values in the target gene and $\mathbf{w}^T \in R^{1\times(n-p)}$ is a 1×(*n*-*p*) vector containing the non-missing values. To estimate $\boldsymbol{\alpha}^T$, a regression model in the form of $\boldsymbol{\alpha}^T = \mathbf{w}^T\mathbf{Y}$ is adopted where the matrix $\mathbf{Y}$ contains the regression coefficients. The regression coefficients are obtained from the *k* similar genes under a least square framework. In particular, columns of the *k* similar genes are re-arranged in a manner similar to $\mathbf{g}_t^T$ as follows,

$$\begin{pmatrix} \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \mathbf{B} & \mathbf{A} \end{pmatrix} \tag{2}$$

where $\mathbf{g}_{s_i}{}^T \in R^{1\times n}$ for i = 1, 2, ..., k denotes the k similar genes, $\mathbf{B} \in R^{k\times p}$ and $\mathbf{A} \in R^{k\times(n-p)}$ denote respectively the expression values in the first p conditions and remaining (n-p) conditions of the similar genes. The regression coefficients are obtained from these k similar genes by minimizing the following equation,

$$\arg\min_{\mathbf{Y}} \left\| \mathbf{AY} - \mathbf{B} \right\|_2 \tag{3}$$

The closed form solution to Eq.(3) can be written as

$$\hat{\mathbf{Y}} = \mathbf{A}^{+}\mathbf{B} \tag{4}$$

where $\mathbf{A}^{+}$ is the pseudoinverse of $\mathbf{A}$. Hence, the missing values in the target gene can be approximated as,

$$\hat{\boldsymbol{\alpha}}^T = \mathbf{w}^T\hat{\mathbf{Y}} = \mathbf{w}^T\mathbf{A}^{+}\mathbf{B} \tag{5}$$

In LLSimpute, k is the only parameter. A heuristic approach for its estimation has been proposed in [12]. First, some of the non-missing values in the expression matrix are considered to be missing. Then, the value of k is obtained by minimizing the normalized root mean square error (NRMSE) that is defined as,

$$\text{NRMSE} = \left(\sqrt{\sum_{(i,j)\in S} (\alpha_{ij} - \hat{\alpha}_{ij})^2 \,/\, |S|} \right) \Big/ \sigma \tag{6}$$

where α_{ij} is the actual value in the data matrix at position (i, j), $\hat{\alpha}_{ij}$ is the corresponding estimated value, S is a set of missing entries, $|S|$ is the cardinality of the set S and σ is the standard deviation of the actual values at positions in S.

3. Proposed Algorithm for Missing Value Estimation

In LLSimpute, a group of genes that is similar to the target gene is identified so that the group can be used to estimate the missing entries in the target gene. The gene similarity is measured by considering the similarity of the expression profiles across all experimental conditions. However, studies have found that gene profiles are similar under some experimental conditions only. For instances, it was found in a yeast expression dataset that genes express more coherently in a subset of conditions than in the whole set of conditions [16]. Comparing with traditional clustering approaches, simultaneously clustering in both genes and experimental conditions (i.e., biclustering) is more effective in identifing patterns with similar gene functions [18]. Biclustering also performs better sample classification than clustering [21].

Further evidence can be obtained by studying a set of similar genes in a yeast expression matrix (called Ronen dataset as described in Section 4) [22]. The dataset has 5342 genes and 26 experimental conditions. The experimental conditions are glucose pulse (2g/l) from 10min to 240min and glucose pulse (0.2g/l) from 2min to 150min. Three experimental conditions considered are “glucose pulse (2g/l) on galactose chemostat at 10 min” (condition A), “glucose pulse (2g/l) on galactose chemostat at 15 min” (condition B) and “glucose pulse (2g/l) on galactose chemostat at 180 min” (condition C). Fig. 1 shows the expression levels of a set of similar genes in these three conditions. It can be observed that the responses in conditions A and B are highly correlated but the

responses in condition C appear to be uncorrelated with the other two conditions. In fact, the correlation value between conditions A and B is 0.7860 while that between conditions A and C is 0.0390. Hence, the set of genes are similar only under conditions A and B, but not in condition C. The finding is consistent with the nature of experimental conditions that the measurements for conditions A and B were taken with only a few minutes apart but that for condition C was taken almost 3 hours later.

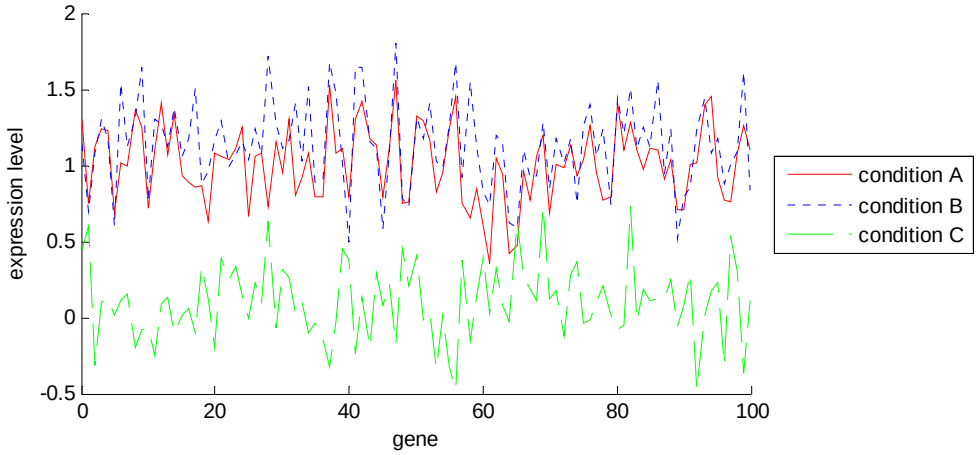


Fig.1. Gene expression levels of a set of similar genes in three selected experimental conditions in the Ronen dataset. The number of similar genes is 101. The three experimental conditions are "glucose pulse (2g/l) on galactose chemostat at 10 min" (condition A), "glucose pulse (2g/l) on galactose chemostat at 15 min" (condition B) and "glucose pulse (2g/l) on galactose chemostat at 180 min" (condition C).

Fig. 1 provides evidence for the assumption that genes are coexpressed under some conditions only. If one considers the coherence across the entire experimental conditions, local coherence might not be captured correctly which in turn affects the accuracy in the missing value estimation. Biclusters which are coherent clusters consisting of correlated genes under some experimental conditions

should thus be used for characterizing the local coherence information. In this part, we incorporate the biclustering idea in the LLSimpute so that gene similarity is measured within the correlated conditions only.

3.1. Bicluster-based Least Square Framework

Similar to the LLSimpute, a set of *k* similar genes is first identified using the Euclidean distance. From these *k* similar genes, coherence information among different conditions is estimated. Note that condition *i* and condition *j* can have very different correlation with other conditions. Thus, correlation among different conditions for each missing value in the target gene should be estimated separately. Hence, we have,

$$\mathbf{R} = \mathbf{B}^T \mathbf{A} \tag{7}$$

Using $\mathbf{R}$, the set of *k* similar genes for the *j*th missing value of the target gene is reselected from the expression matrix by considering a weighted Euclidean distance. In particular, the similarity between the target gene $\mathbf{g}_t$ and the other gene $\mathbf{g}_s$ is calculated for the *j*th missing value as,

$$d_j(\mathbf{g}_t, \mathbf{g}_s) = \sqrt{\sum_{v=p+1}^{n} r_j(v-p)^2 [g_t(v) - g_s(v)]^2} \Big/ \sqrt{\sum_{v=1}^{n-p} r_j(v)^2} \tag{8}$$

where $r_j(v)$ is the (*j*, *v*)th element of $\mathbf{R}$ and $g(v)$ is the *v*th element of the vector $\mathbf{g}$. Using Eq.(8), coherence among genes in some related experimental conditions are considered for selecting the *k* similar genes. Then, in estimating the *j*th missing value of the target gene, the "uncorrelated"

conditions are removed in the least square framework. Let $r_{j,\max} = \max_{v \in \{1,\ldots,n-p\}} |r_j(v)|$, the conditions are said to be related if $|r_j(v)| \geq T_0 r_{j,\max}$ where T_0 is a pre-set threshold. After removing all the "uncorrelated" conditions, the target gene can be written as,

$$\mathbf{g}_t^T = \begin{pmatrix} \alpha_j & \mathbf{w}_j^T \end{pmatrix} \tag{9}$$

where α_j is the *j*th missing value in the target gene, and $\mathbf{w}_j^T$ denotes the non-missing values from the columns that are correlated to the *j*th column in the target gene. Columns of the *k* similar genes can be re-arranged and truncated in the same manner as $\mathbf{g}_t^T$. Hence, we have,

$$\begin{pmatrix} \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \mathbf{b}_j & \mathbf{A}_j \end{pmatrix} \tag{10}$$

where $\mathbf{b}_j$ is the *j*th column of the data and $\mathbf{A}_j$ is a matrix consisting of the correlated columns of the similar genes. The regression coefficients are obtained from these *k* similar genes using a least square framework as,

$$\arg\min_{\mathbf{y}} \left\| \mathbf{A}_j \mathbf{y} - \mathbf{b}_j \right\|_2 \tag{11}$$

Thus, the *j*th missing value can be estimated as,

$$\alpha_j = \mathbf{w}_j^T \hat{\mathbf{y}}_j = \mathbf{w}_j^T \mathbf{A}_j^+ \mathbf{b}_j \tag{12}$$

The above estimation can be repeated until all the *p* missing values in the target gene are obtained. Our proposed bicluster-based missing value estimation approach has two parameters, the number of similar genes *k* and the threshold for the correlation between columns T_0. These two parameters are determined automatically by employing a heuristic approach using simulated missing values as in

[12]. The simulated missing values are non-missing values which are randomly selected to be missing values. Since the actual values of the simulated missing values are known, the estimation error can be calculated. The values of k and T_0 are determined as those giving the minimum NRMSE for the simulated missing values. In order to minimize the computational cost, k is estimated first and then used to determine T_0.

3.2. Iterative Application of the Proposed Bicluster-based Imputation

An iterative version of the LLSimpute called ILLSimpute was developed in [23]. It aims to improve the selection of k similar genes and thus the estimation of missing values based on estimates from the previous iteration. Experimental results have demonstrated that the ILLSimpute can achieve an improvement of more than 10% in NRMSE at a 10% missing rate as compared to the LLSimpute. Following the same idea, the proposed bicluster-based imputation is iteratively applied to refine the selection of k similar genes and the correlated conditions so as to improve the estimates of missing values. In the iterative framework, one of the important considerations is the convergence rate. The study of ILLSimpute suggests that direct modification for iteration results in slow convergence and deviation from the optimal estimation error [23]. In order to improve the convergence, we use the concept of the uncertainty to update the estimates. In particular, the estimates in the current iteration will replace that in the previous iteration only if the uncertainty is

decreased. The uncertainty δ is calculated as the half width of the prediction interval [24] at a significant level of α, i.e.,

$$\delta = t_{\alpha/2,m'-n'}\sqrt{(\mathbf{w}_j{}^T(\mathbf{A}_j{}^T\mathbf{A}_j)^{-1}\mathbf{w}_j + 1)\hat{\sigma}^2} \tag{13}$$

where $t_{\alpha/2,m'-n'}$ is the *t*-value of the Student's *t* distribution with *m*'-*n*' degree of freedom, *m*' is the number of rows of $\mathbf{A}_j$, *n*' is the number of columns of $\mathbf{A}_j$ and $\hat{\sigma}^2$ is the unbiased estimator of noise variance in the regression model, i.e.

$$\hat{\sigma}^2 = \left(\mathbf{b}_j - \mathbf{A}_j\hat{\mathbf{y}}_j\right)^T\left(\mathbf{b}_j - \mathbf{A}_j\hat{\mathbf{y}}_j\right)\Big/(m'-n') \,. \tag{14}$$

A small δ implies a small deviation of the estimate from its actual value. Using Eq.(14), a reliable approximation can be achieved in the statistical sense. Furthermore, the values of uncertainty form a non-increasing sequence consisting of non-negative values. This implies that the update of estimates tends to vanish and hence the estimated values would converge. In the implementation, the iterative process is terminated when the average change in estimates is insignificant. In order to impose a further control on the number of iterations, the maximum number of iterations can also be set.

4. EXPERIMENTAL RESULTS

Our proposed iterative bicluster-based least square estimation is evaluated on two artificial datasets, and five real datasets. In the first artificial dataset, there are twenty data matrices of size 360×30

which were initially generated with uniformly distributed random values in the range of -10 and 10. Then, twelve 30×12 biclusters were implanted into each of the data matrices without row overlap. In total, there is 40% gene expression data covered by the biclusters. Finally, 30dB noise was added. The second artificial dataset was generated in a similar way except that the size of each bicluster is 30×18 so that the percentage of bicluster-region in these data matrices is 60%. Since the bicluster information is known, the artificial datasets allow a systematic study of the proposed algorithm.

Among the real datasets, four are gene expression microarray data and one is non-microarray data. The first two are cell cycle expression datasets of yeast *Saccharomyces cerevisiae* (*S. cerevisiae*), Sp.alpha and Sp.cdc15, which are synchronized using α factor and a cdc15 temperature-sensitive mutant respectively [25]. The third microarray dataset, Ogawa, is a non-time series dataset for the analysis of phosphate accumulation and polyphosphate metabolism in *S. cerevisiae* [26]. The fourth microarray dataset is called Bonen which contains two time series of yeast response to glucose pulses in galactose-limited chemostats [22]. The sizes of the four datasets Sp.alpha, Sp.cdc15, Ogawa and Bonen are 4489×18, 4381×24, 5783×8 and 5342×26 respectively after removing the genes with missing values. These four real datasets were used to verify the performance of the proposed algorithm on microarray data. The last real dataset, Finance [27], is a non- microarray dataset. It contains information about 200 French industries in 5 years duration and has a size of 650×36. Each row represents a sample with 35 variables together with one output. The variables

involve several types of data; balance sheet, income statement and market data while the output is the return of assets. The purpose of this dataset is to study the potential application of the proposed algorithm to non-microarray data.

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 to generate average results. For real datasets, the missing values were distributed over the whole data as the 'ground-truth' biclusters are unknown. The estimation was repeated ten times on the Sp.alpha, Sp.cdc15 and Bonen datasets; forty times on the Ogawa and Finance datasets. Our proposed method is compared with several existing algorithms including LLSimpute, Bayesian principal component analysis (BPCA) [10] and ILLSimpute. Since the convergence of ILLSimpute is poor, the best result among the iterations is selected for comparison. The automatic parameter estimation of ILLSimpute as well as the proposed method was done for the first three iterations only so as to maintain a tradeoff between the computational cost and performance. The accuracy of missing value estimation is evaluated using an average of the NRMSE defined by Eq.(6).

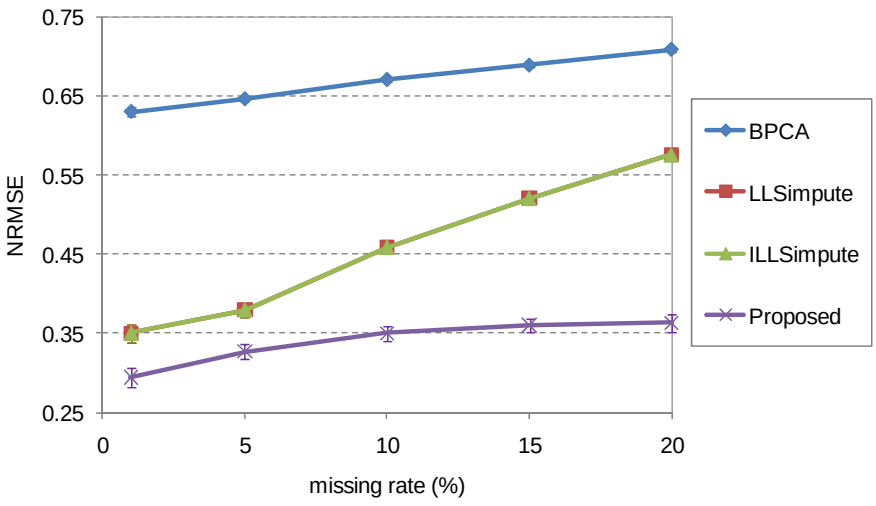
4.1. Artificial Datasets

Fig.2 shows the average NRMSE at different regions (bicluster-region, non-bicluster region and the

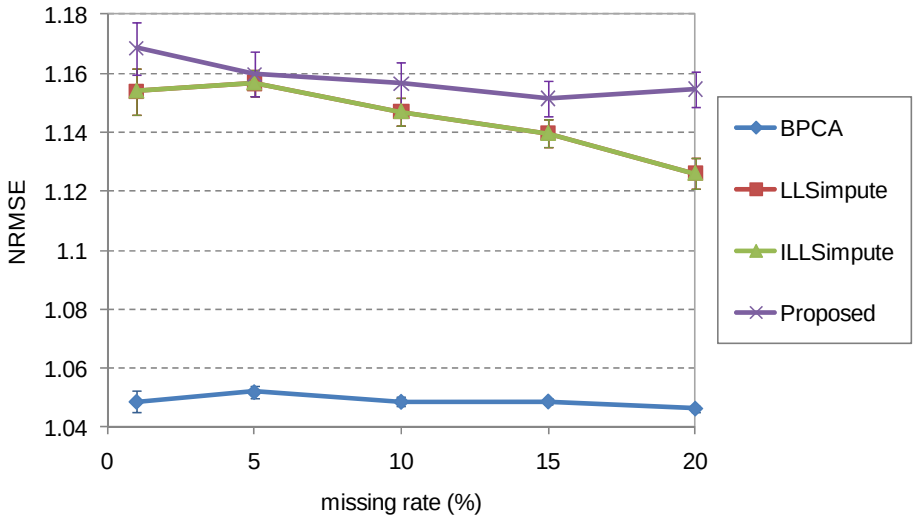
overall expression matrix) for the first artificial dataset with 40% bicluster-region under different missing rates. From Fig.2(a), we can see that our proposed algorithm achieves the best NRMSE at all missing rates in bicluster-region. The improvement over the other algorithms can be attributed to the use of the regression strategy that considers only the related experimental conditions. The improvement is between 13.8% and 53.3% as compared to the other three algorithms. The ILLSimpute attains its minimum NRMSE at the first iteration at all the missing rates. Hence ILLSimpute has essentially the same performance as LLSimpute. This demonstrates that the iterative framework is not effective in characterizing the bicluster information due to the use of clustering. BPCA has the worst NRMSE because it considers the data correlation over the whole matrix instead of the coherent data. In the non-bicluster-region, on the contrary, the BPCA has the best performance and the performance of our proposed algorithm is comparable with that of LLSimpute and ILLSimpute. The main reason is that the data are generated independently so the idea of finding correlated genes and correlated columns is not valid. In fact, such data do not have any significant bicluster pattern. The use of the bicluster models would bias the estimates against the underlying random model. In terms of the overall data matrix as shown in Fig.2(c), our proposed algorithm outperforms LLSimpute and ILLSimpute at the missing rates between 5% and 20%. Comparing with BPCA, the proposed algorithm has higher overall NRMSE because the proportion of the bicluster region is lower than that of the non-bicluster region. The improvement in the

bicluster region cannot compensate for the deterioration in the non-bicluster region. However, when the missing rate increases, the NRMSE of the proposed algorithm get close to that of BPCA.

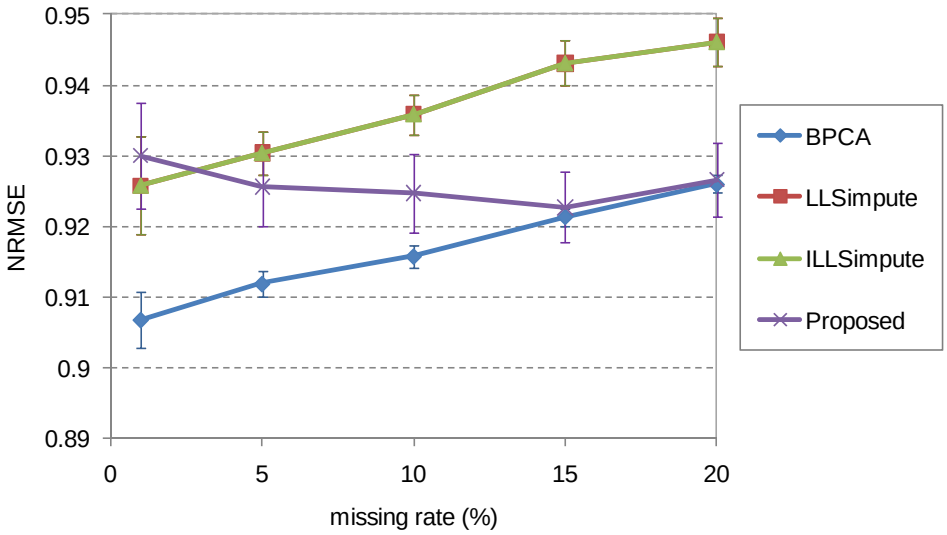
Results for the second artificial dataset with 60% bicluster region are shown in Fig.3. As in the first artificial dataset, the use of the biclustering idea makes our proposed algorithm achieving the lowest NRMSE at all the missing rates in the bicluster region. The reduction in NRMSE is between 12.1% and 42.5%. Although the proposed algorithm is not the best estimation method in the non-bicluster region, its NRMSE is always the lowest when the NRMSE is considered over the whole matrix. The improvement using the proposed algorithm is between 2.0% and 4.5% in the overall NRMSE. Hence, if the local coherent structure becomes more significant as in the second artificial dataset, we can see that our proposed algorithm outperforms the other three algorithms apparently. In next section, experimental results on real datasets are discussed.



(a)

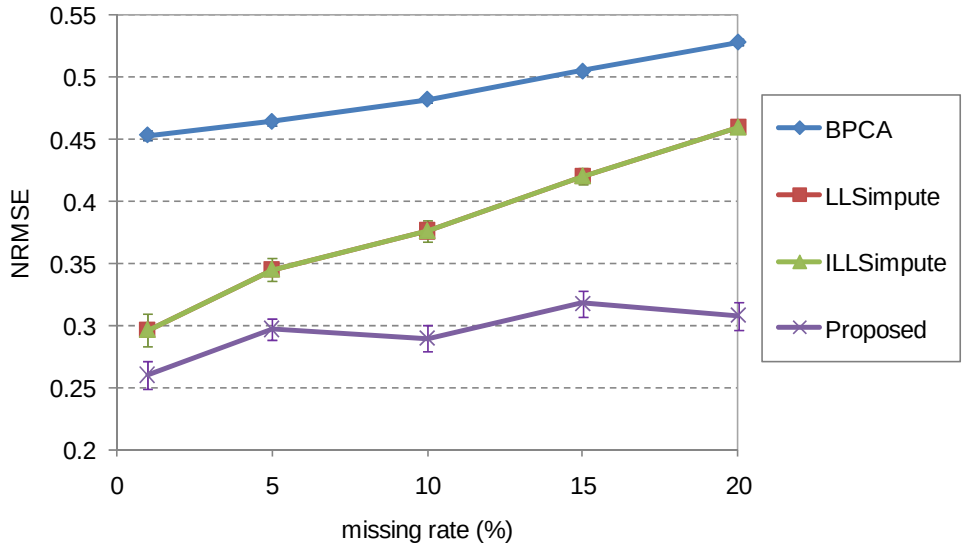


(b)

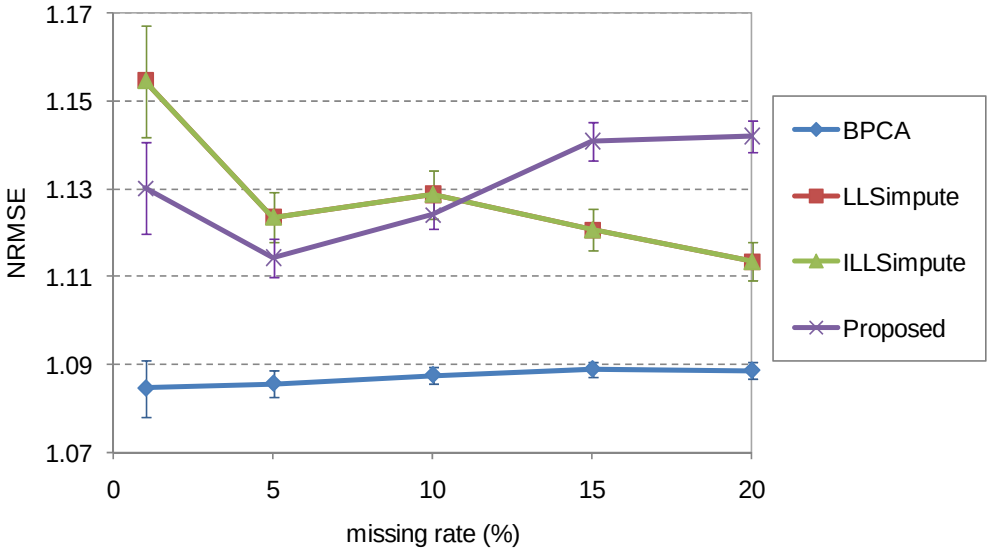


(c)

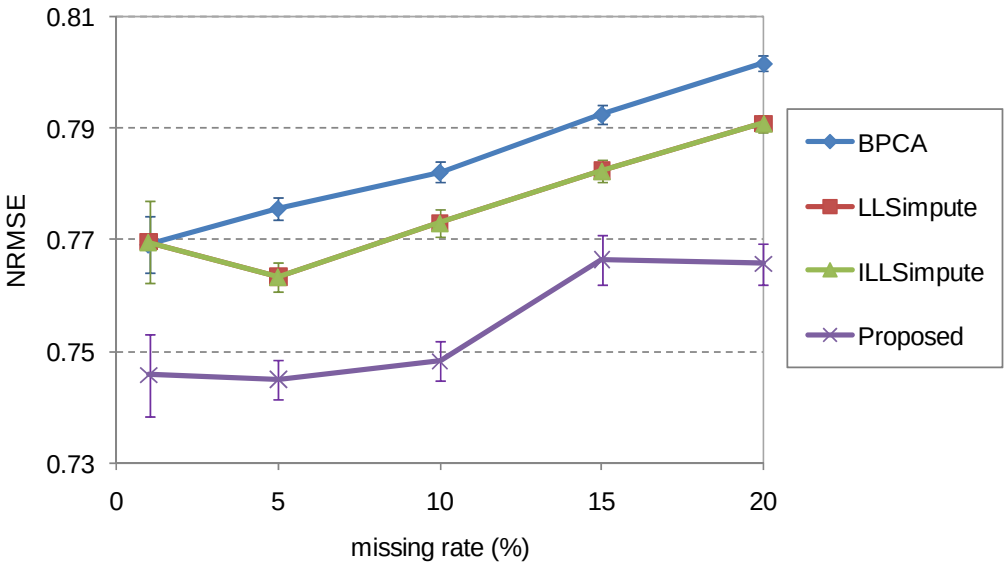
Fig.2. Average NRMSE in (a) the bicluster-region, (b) the non-bicluster-region and (c) the whole data matrix achieved by BPCA, LLSimpute, ILLSimpute and our proposed algorithm in the first artificial dataset (with 40% bicluster-region) for various missing rates. The error bars indicate the standard error of mean in the experiments.


0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
NRMSE
0
5
10
15
20
missing rate (%)
BPCA
LLSimpute
ILLSimpute
Proposed

(a)


1.17
1.15
1.13
1.11
1.09
1.07
NRMSE
0
5
10
15
20
missing rate (%)
BPCA
LLSimpute
ILLSimpute
Proposed

(b)



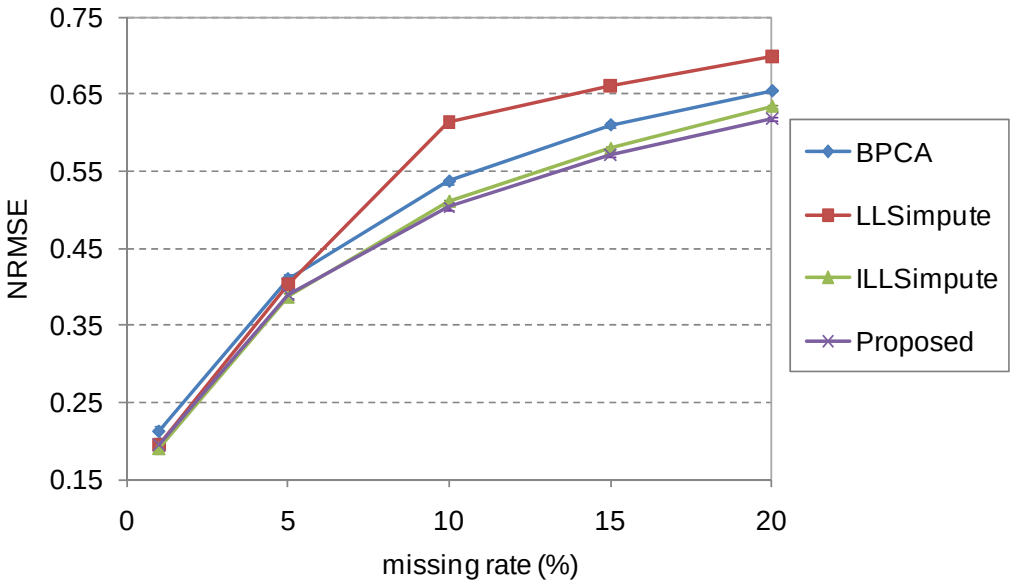
(c)

Fig.3. Average NRMSE in (a) the bicluster-region, (b) the non-bicluster-region and (c) the whole data matrix achieved by BPCA, LLSimpute, ILLSimpute and our proposed algorithm in the second artificial dataset (with 60% bicluster-region) for various missing rates. The error bars indicate the standard error of mean in the experiments.

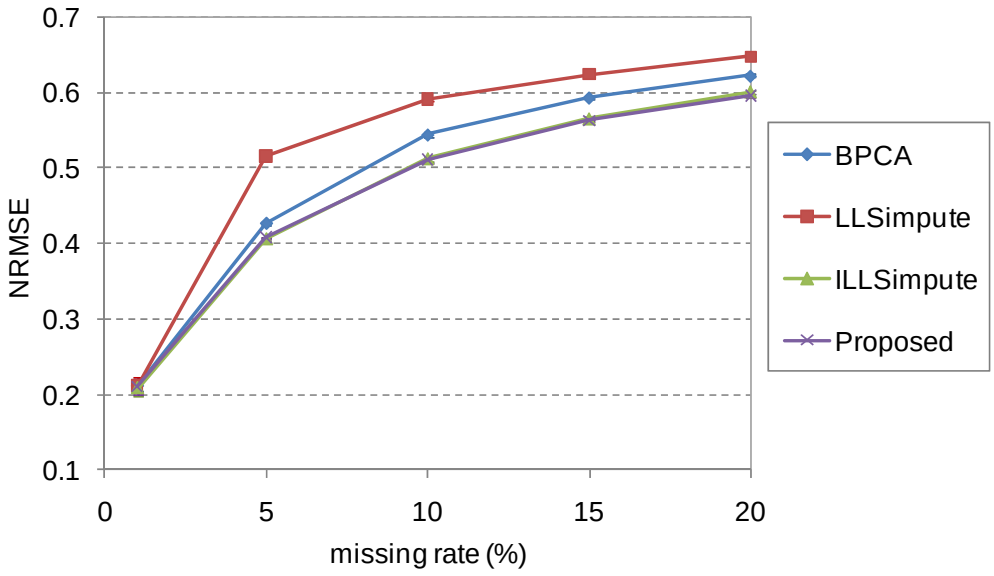
4.2. Real Datasets

The average NRMSE against missing rate for the four real microarray datasets, namely Sp.alpha, Sp.cdc15, Ogama and Ronen is plotted in Fig.4. For the datasets Sp.alpha and Sp.cdc15, the proposed algorithm outperforms BPCA and LLSimpute in all the cases with significant improvement at mid and high missing rates. The overall improvement over BPCA is 6.4% and 4.2% for datasets Sp.alpha and Sp.cdc15 respectively. Compared with LLSimpute, the proposed algorithm reduces the NRMSE by 9.3% and 10.6% for datasets Sp.alpha and Sp.cdc15 respectively. ILLSimpute also demonstrates higher performance than BPCA and LLSimpute. This implies that iterative framework

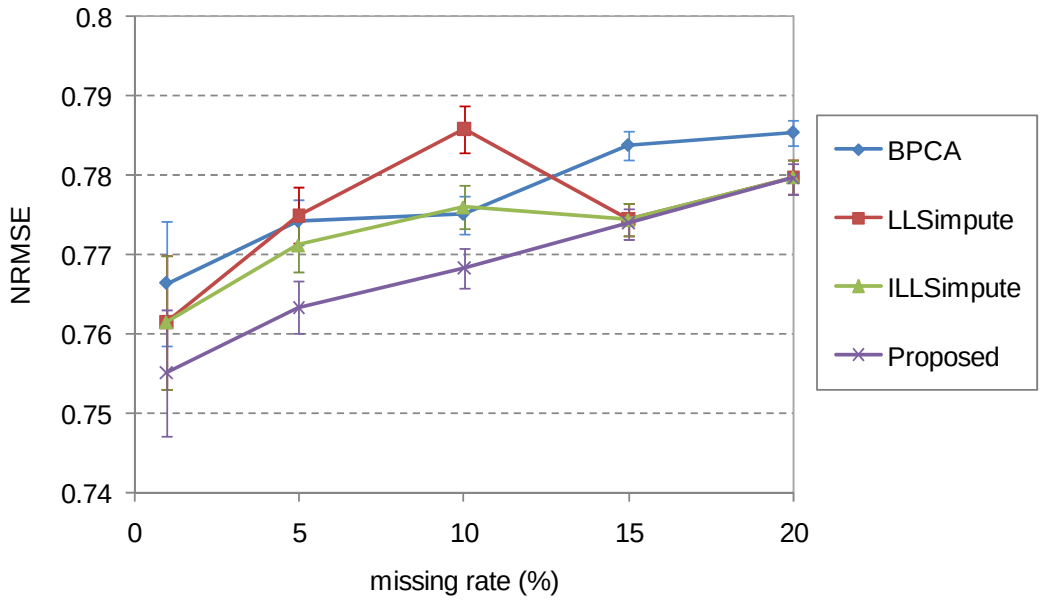
can refine and improve the estimates of missing values. Our proposed algorithm has lower NRMSE than ILLSimpute at mid and high missing rates (10-20%) for the two cell-cycle datasets.



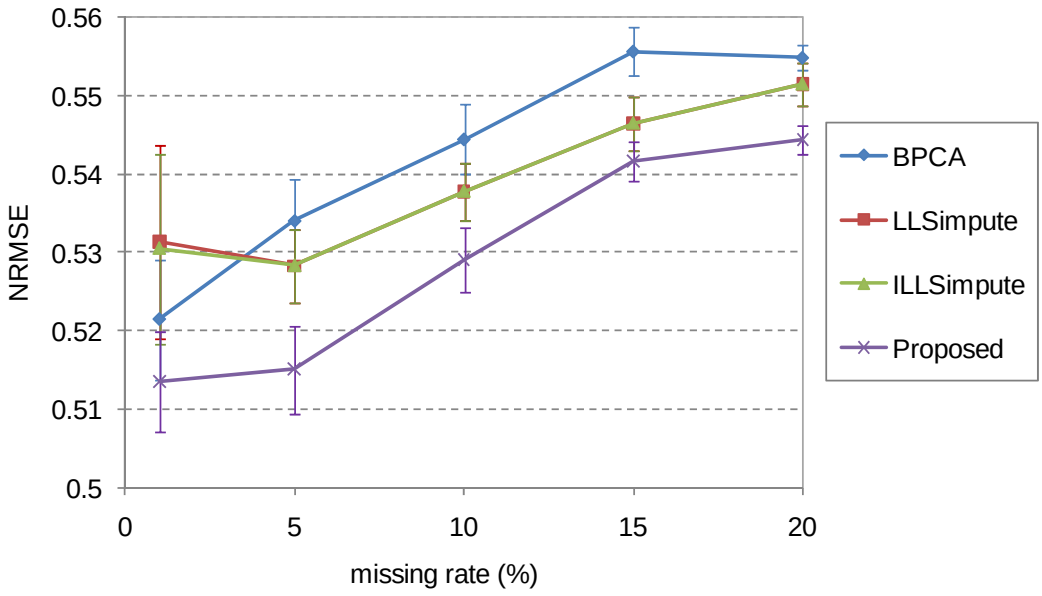
(a)



(b)



(c)



(d)

Fig.4. Average NRMSE of BPCA, LLSimpute, ILLSimpute and the proposed algorithm at different missing rates for microarray datasets (a) Sp.alpha, (b) Sp.cdc15, (c) Ogama and (d) Bonen. The error bars indicate the standard error of mean in the experiments.

For the other two microarray datasets Ogama and Bonen, our proposed algorithm again shows better

performance than all the other algorithms. In Ogama dataset, a large improvement over ILLSimpute (the second best algorithm on average) is found at low to mid missing rates instead of high missing rates. When missing rate increases, there are fewer correlated conditions available for estimation so that the performance of the proposed method becomes close to that of ILLSimpute which also uses local least square estimation but with gene clustering only. In Bonen dataset, ILLSimpute and LLSimpute essentially have the same performance. Thus the iterative framework using clustering cannot improve the missing value estimation. However, our proposed method that uses the biclustering under an iterative framework is able to achieve the lowest NRMSE for all the missing rates.

Fig. 5 illustrates the performance on the Finance dataset. The proposed algorithm has lower NRMSE than the other three algorithms at missing rates 5%-20%. The results suggest that even for non-microarray datasets, promising performance can still be achieved if the data correlation fits our assumption. If the data correlation model is not modeled by biclusters, modifications on the proposed algorithm are required to adapt to the appropriate coherent patterns in the dataset. A detailed study of the proposed algorithm on different data correlation model will be investigated in future.

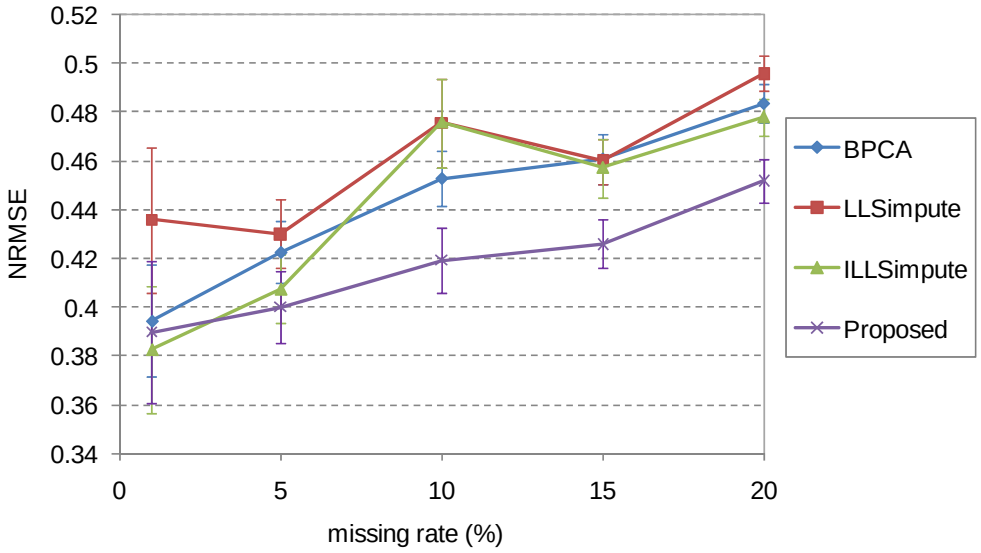
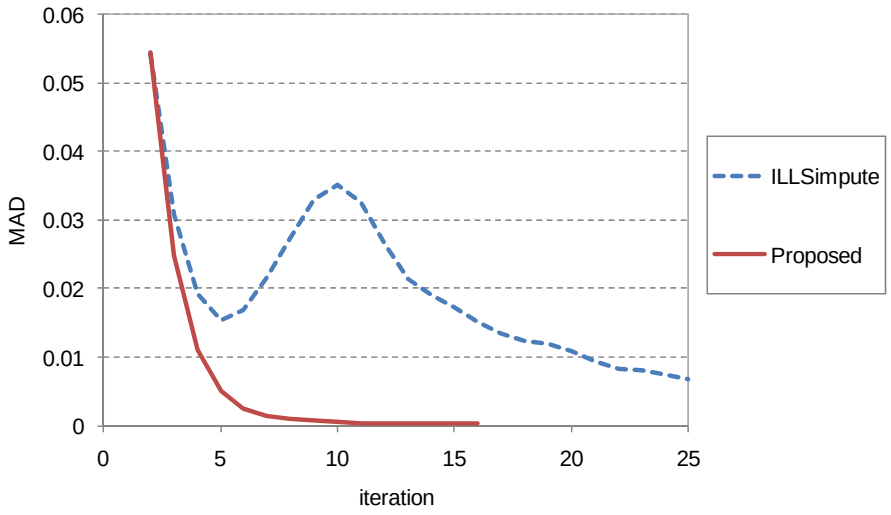


Fig.5. Average NRMSE of BPCA, LLSimpute, ILLSimpute and the proposed algorithm at different missing rates for the Finance dataset. The error bars indicate the standard error of mean in the experiments.

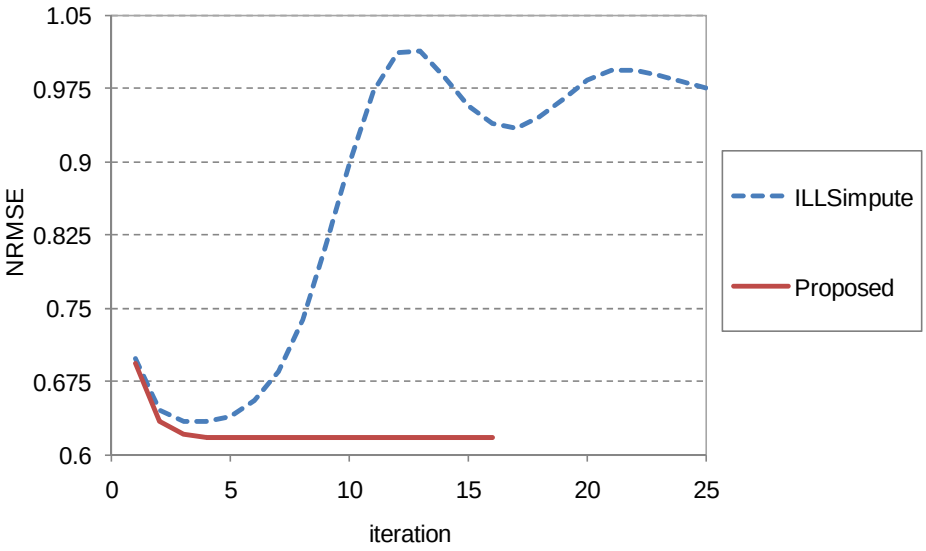
4.3. Convergence Analysis

One of the concerns on iterative algorithms is the convergence, i.e. whether the estimates finally remain unchanged at certain values and whether the values produce the lowest error. As discussed in Section 3.2, the estimates of the proposed algorithm are enforced to converge to certain values which are optimal in the statistical sense so that the prediction error is likely to be the lowest. In order to study the convergence rate, mean absolute difference (MAD) between the current and previous estimates of missing values is calculated in iterations from 2 to 25. Fig.6(a) illustrates the MAD for the real dataset Sp.alpha at 20% missing rate. The MAD of the proposed algorithm is always decreasing and becomes stable within 25 iterations. As ILLSimpute lacks a control on convergence, the MAD tends to take longer time to drop. Furthermore, there is even a crest at around 10 iterations

and the MAD cannot drop below the threshold after 25 iterations. Fig.6(b) shows the NRMSE of the estimates at iterations up to 25 for the dataset Sp.alpha at a missing rate of 20%. In general, the update criterion based on the prediction interval allows the proposed algorithm to improve the estimation accuracy. For ILLSimpute, the NRMSE, however, may increase substantially after it passes the minimum point.



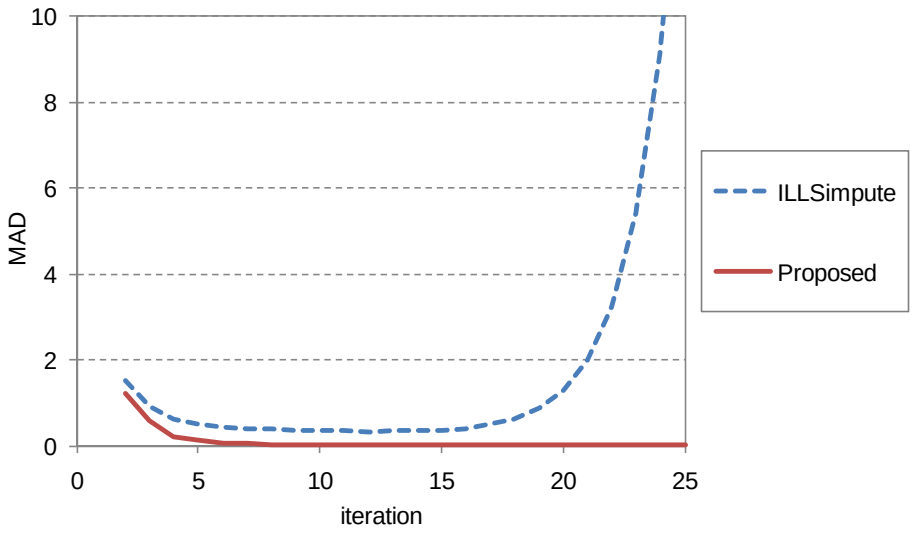
(a)



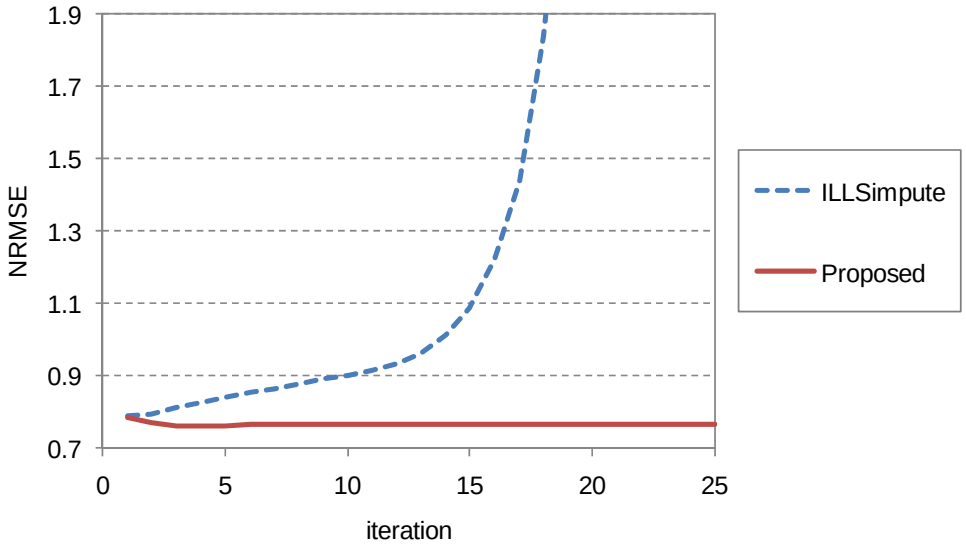
(b)

Fig.6(a) The MAD of estimates between two consecutive iterations and (b) average NRMSE against iterations using ILLSimpute and the proposed algorithm in the experiments on the real microarray dataset Sp.alpha at a missing rate of 20%.

In addition to the real dataset, the convergence in artificial datasets is also studied. Fig.7(a) and (b) show the MAD and NRMSE in the second artificial datasets (with 60% bicluster-region) at the missing rate of 20% respectively. As in the experiments on the real dataset Sp.alpha, the estimates by the proposed algorithm converge in both MAD and NRMSE. On the other hand, the convergence problem of ILLSimpute becomes more serious in the artificial dataset that there is an increasing trend for both MAD and NRMSE after 25 iterations. This further confirms the significance of the proposed convergence control in the iterative framework, especially at high missing rates.



(a)



(b)

Fig.7. (a) The MAD of estimates between two consecutive iterations and (b) average NRMSE against iterations using ILLSimpute and the proposed algorithm in the experiments on the second artificial dataset (with 60% bicluster-region) at a missing rate of 20%.

4.4. Parameters Analysis

As discussed in Section 3.2, our proposed algorithms have two parameters: the number of similar genes k and the threshold for the correlation between columns T_0. Although our proposed algorithm has an automatic strategy to select these parameters, it is important to study the sensitivity of the proposed algorithm to these parameters. In the following, the automatic parameter selection strategy is not used so that k and T_0 can be set manually. Experiments were conducted on the real microarray dataset Ronen at a missing rate of 10% with manually selected k and T_0. Fig.8 shows the average NRMSE obtained with different values of parameters. T_0 is set to be between 0 and 0.8 while k is between 1 and 4096. It can be seen that k cannot be set to be too small. For k to be

larger than 64, the NRMSE did not have a large variation for various T_0. The minimum NRMSE is 0.5286 found at $k = 512$ and $T_0 = 1\times10^{-4}$. Using our automatic selection strategy, the NRMSE is 0.5290 which is slightly larger than the minimum error. Since the difference in NRMSE between the optimal parameter values and the selected parameter values is small, the proposed algorithm together with the automatic parameters selection strategy is practical for missing value estimation.

k \ T_0	0	0.0001	0.001	0.01	0.1	0.2	0.4	0.6	0.8
1	0.703	0.7031	0.7032	0.705	0.7173	0.7332	0.7591	0.7731	0.8318
2	0.6445	0.6444	0.6445	0.6452	0.8247	4.3059	7.9625	38.2395	37.7268
4	0.6187	0.6187	0.6191	0.6189	6.842	9.6936	31.9775	42.5806	37.8898
8	0.6351	0.6343	0.6346	0.6353	5.7825	24.3952	324.3478	69.636	6.2199
16	0.8518	0.8644	0.9156	6.1367	456.6069	150.5135	11.442	1.1264	0.6114
32	0.8409	0.8436	0.8405	0.8232	0.7315	0.672	0.6227	0.6078	0.6054
64	0.6076	0.6089	0.6077	0.6038	0.5885	0.5828	0.5933	0.6014	0.6104
128	0.55	0.5504	0.5494	0.5499	0.5478	0.55	0.5781	0.6017	0.6198
256	0.5316	0.5317	0.5316	0.5319	0.5337	0.5397	0.567	0.604	0.6305
512	0.5288	0.5286	0.5289	0.5289	0.5332	0.54	0.5631	0.5991	0.6375
1024	0.5316	0.5317	0.5316	0.5318	0.5368	0.5439	0.5643	0.5932	0.6396
2048	0.5368	0.5368	0.5368	0.5371	0.5419	0.5498	0.5714	0.5919	0.6347
4096	0.5482	0.5482	0.5481	0.5484	0.5536	0.5634	0.58	0.5963	0.6307

Fig.8. Average NRMSE for the proposed algorithm applied on the Ronen dataset at a missing rate 10% using different values of parameters k and T_0. The minimum NRMSE 0.5286 (highlighted) is achieved at $k = 512$ and $T_0 = 0.0001$.

5. CONCLUSIONS

Existing state-of-the-art missing value algorithms always measure the gene similarity by considering expression profiles in all experimental conditions. As genes are correlated under some experimental conditions only, a bicluster-based least square algorithm is proposed for estimating

missing values in gene expression data. In our algorithm, biclusters which consist of a subset of genes that is similar in a subset of conditions are identified by performing clustering on genes and conditions alternately. By applying a regression model to the found biclusters, least square estimation can be performed without the influence of unrelated genes and conditions. In addition to the use of biclusters concept, the estimation is iterated so as to refine the selection of similar genes/conditions which in turn improves the accuracy of the missing value estimation. One of the main concerns in an iterative algorithm is convergence. The convergence problem is solved by requiring the uncertainty to be decreased with respect to the iterations. Unlike the existing iterative approach, ILLSimpute, the proposed convergence control can guarantee the algorithm to converge.

Experiments on two artificial datasets, and four real microarray datasets are conducted to study the performance of the proposed algorithm on gene expression data. For the artificial datasets, normalized root mean squared error (NRMSE) is calculated in bicluster-region, non-bicluster-region and over the whole data matrix. Experimental results show that the proposed algorithm has prominent improvement in the bicluster-region compared with BPCA, LLSimpute and ILLSimpute. For the real microarray datasets, only NRMSE over the whole matrix was studied as the ground truth biclusters are not available. The overall performance of the proposed algorithm generally outperforms the three existing algorithms. Since ILLSimpute also adopts an iterative approach, its

performance is the closest to the proposed algorithm among the three existing algorithms. However, it cannot fully exploit the data correlation due to the use of clustering. Furthermore, ILLSimpute suffers from the convergence problems. Experimental results on artificial and real datasets show that ILLSimpute did not guarantee to converge. However, owing to the use of the prediction intervals, the proposed algorithm can always converge. In addition, an experiment on a financial dataset was conducted to evaluate the performance of the proposed algorithm on non-microarray data. The result is promising as our proposed algorithm still outperforms the other three algorithms. Hence, our proposed algorithm is applicable to other datasets as long as the data correlation model fits with our assumption. In the future, we will extend our algorithm so that it can be applied to other data correlation model.

ACKNOWLEDGEMENTS

This work is supported by the Centre for Signal Processing, the Hong Kong Polytechnic University. K.O. Cheng thanks the Hong Kong Polytechnic University for the support he receives under its postdoctoral fellowship scheme.

REFERENCES

[1] D.J. Lockhart, E.A. Winzeler, Genomics, gene expression and DNA arrays, Nature 405 (2000) 827 – 836.

[2] A.W.C. Liew, H. Yan, M.S. Yang, Pattern recognition techniques for the emerging field of bioinformatics: a review, Pattern Recognition 38 (11) (2005) 2055 – 2073.

[3] T.C. Lin, R.S. Liu, C.Y. Chen, Y.T. Chao, S.Y. Chen, Pattern classification in DNA microarray data of multiple tumor types, Pattern Recognition 39 (12) (2006) 2426 – 2438.

[4] H. Liu, L. Liu, H. Zhang, Ensemble gene selection for cancer classification, Pattern Recognition 43 (8) (2010) 2763-2772.

[5] H.-Q. Wang, H.-S. Wong, D.-S. Huang, J. Shu, Extracting gene regulation information for cancer classification, Pattern Recognition 40 (12) (2007) 3379-3392.

[6] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognition 41 (12) (2008) 3692-3705.

[7] Y. Sun, U. Braga-Neto, E.R. Dougherty, 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 (2009), doi:10.1155/2009/504069.

[8] A.W.C. Liew, N.-F. Law, H. Yan, Missing value imputation for gene expression data: computational techniques to recover missing data from available information, Briefings in Bioinformatics Advance Access published December 14, 2010, doi: 10.1093/bib/bbq080.

[9] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520 – 525.

[10] O. Oba, M.A. Sato, I. Takemasa, M. Monden, K.I. Matsubara, S. Ishii, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics 19 (16) (2003) 2088 – 2096.

[11] T.H. Bo, B. Dysvik, I. Jonassen, LSimpute: accurate estimation of missing values in microarray data with least squares methods, Nucleic Acids Research 32 (3): e34 (2004).

[12] H. Kim, G.H. Golub, H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics 21 (2) (2005) 187 – 198.

[13] S. Friedland, A. Niknejad, M. Kaveh, H. Zare, An algorithm for missing value estimation for DNA

microarray data, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol.2, May 2006, pp.1092 – 1095.

[14] C.C. Liu, D.Q. Dai, H. Yan, The theoretic framework of local weighted approximation for microarray missing value estimation, Pattern Recognition 43 (8) (2010) 2993 – 3002.

[15] A.W.C. Liew, N.F. Law, H. Yan, Cluster analysis of gene expression data, in: Juan Ramon Rabunal Dopico, Julian Dorado, Alejandro Pazos (Eds.), Encyclopedia of Artificial Intelligence, Information Science Reference, July 2008, pp.289-296.

[16] Y. Cheng, G.M. Church, Biclustering of expression data, in: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 2000, pp.93 – 103.

[17] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE Transactions on Computational Biology and Bioinformatics 1 (1) (2004) 24 – 45.

[18] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, A systematic comparsion and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (9) (2006) 1122 – 1129.

[19] K.O. Cheng, N.F. Law, W.C. Siu, A.W.C. Liew, Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization, BMC Bioinformatics, 9:210 (2008).

[20] K.O. Cheng, N.F. Law, W.C. Siu, T.H. Lau, BiVisu: software tool for bicluster detection and visualization, Bioinformatics 23 (17) (2007) 2342-2344.

[21] A. Tanay, R. Sharan, R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioninformatics 18(suppl_1) (2002) S136-S144.

[22] M. Ronen, D. Botstein, Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source, Proceedings of the National Academy of Sciences of the United States of America 103(2) (2006) 389-394.

[23] Z. Cai, M. Heydari, G. Lin, Iterated local least squares microarray missing value imputation, Journal of Bioinformatics and Computational Biology 4 (5) (2006) 935-957.

[24] B.L. Bowerman, R.T. O'Connell, Linear statistical models: an applied approach, second ed., Duxbury, USA, 1990.

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

[26] N. Ogawa, J. DeRisi, P.O. Brown, New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis, Molecular Biology of the Cell 11 (12) (2000) 4309-4321.

[27] Environmental and Industrial Machine Learning Group, Department of Information and Computer Science, Aalto University
http://research.ics.tkk.fi/eiml/datasets.shtml

VITAE

Kin-On Cheng received the B.Eng. degree and M.Phil. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology, 2001 and 2003 respectively. In 2008, he received the Ph.D. degree from the electronic and information engineering department in The Hong Kong Polytechnic University. Currently, he is working as a postdoctoral fellow in the same university. His research interests include signal processing, image retrieval and bioinformatics.

Ngai-Fong Law received the B.Eng. degree with first class Honours from the University of Auckland, New Zealand, in 1993 and a Ph.D. degree from the University of Tasmania, Australia, in 1997, both

in Electrical and Electronic Engineering. She is currently an assistant professor in Electronic and Information Engineering Department, The Hong Kong Polytechnic University, Hong Kong. Her research interests include signal and image processing, wavelet transform, image enhancement and compression. Recently she has also extended her study into a new area on bioinformatics, working on gene expression and DNA sequence analysis.

Wan-Chi Siu received an Associateship from The Hong Kong Polytechnic University (formerly called the Hong Kong Polytechnic), a M.Phil. degree from The Chinese University of Hong Kong, and a Ph.D. degree from the Imperial College of Science, Technology, and Medicine, London, U.K. in 1975, 1977, and 1984, respectively. He was with The Chinese University of Hong Kong between 1975 and 1980. He then joined The Hong Kong Polytechnic University as a Lecturer in 1980 and became Chair Professor in 1992. He was the Head of Department of Electronic and Information Engineering and subsequently became Dean of the Engineering Faculty between 1994 and 2002. He is now the Director of the Centre for Multimedia Signal Processing at the same university. He has published over 360 research papers in DSP, transforms, fast algorithms, video coding and pattern recognition. He is a Member of the Editorial Board of the *Journal of VLSI Signal Processing Systems* for signal, image and video technology and the *EURASIP Journal on Applied Signal Processing*.

Dr. Siu was a Guest Editor of a Special Issue of the IEEE Transactions on Circuits and Systems II,

published in May 1998, and was an Associate Editor of the same journal from 1995 to 1997. He has been the general chair or the technical program chair of a number of international conferences. In particular, he was the Technical Program Chair of the IEEE International Symposium on Circuits and Systems (ISCAS'97) and the General Chair of the International Symposium on Intelligent Multimedia, Video, and Speech Processing (ISIMP'2001), which were held in Hong Kong in June 1997 and May 2001, respectively. He was the General Chair of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2003), which was held in Hong Kong. Between 1991 and 1995, he was a member of the Physical Sciences and Engineering Panel of the Research Grants Council (RGC), Hong Kong Government, and in 1994, he chaired the first Engineering and Information Technology Panel to assess the research quality of 19 Cost Centers (departments) from all universities in Hong Kong. He is a Chartered Engineer and a Fellow of both the IEE and the HKIE.

FIGURE LEGENDS

Fig.1. Gene expression levels of a set of similar genes in three selected experimental conditions in the Ronen dataset. The number of similar genes is 101. The three experimental conditions are "glucose pulse (2g/l) on galactose chemostat at 10 min" (condition A), "glucose pulse (2g/l) on galactose chemostat at 15 min" (condition B) and "glucose pulse (2g/l) on galactose chemostat at 180 min" (condition C).

Fig.2. Average NRMSE in (a) the bicluster-region, (b) the non-bicluster-region and (c) the whole data matrix achieved by BPCA, LLSimpute, ILLSimpute and our proposed algorithm in the first artificial dataset (with 40% bicluster-region) for various missing rates. The error bars indicate the standard error of mean in the experiments.

Fig.3. Average NRMSE in (a) the bicluster-region, (b) the non-bicluster-region and (c) the whole data matrix achieved by BPCA, LLSimpute, ILLSimpute and our proposed algorithm in the second artificial dataset (with 60% bicluster-region) for various missing rates. The error bars indicate the standard error of mean in the experiments.

Fig.4. Average NRMSE of BPCA, LLSimpute, ILLSimpute and the proposed algorithm at different missing rates for microarray datasets (a) Sp.alpha, (b) Sp.cdc15, (c) Ogama and (d) Bonen. The error bars indicate the standard error of mean in the experiments.

Fig.5. Average NRMSE of BPCA, LLSimpute, ILLSimpute and the proposed algorithm at different

missing rates for the Finance dataset. The error bars indicate the standard error of mean in the experiments.

Fig.6(a) The MAD of estimates between two consecutive iterations and (b) average NRMSE against iterations using ILLSimpute and the proposed algorithm in the experiments on the real microarray dataset Sp.alpha at a missing rate of 20%.

Fig.7. (a) The MAD of estimates between two consecutive iterations and (b) average NRMSE against iterations using ILLSimpute and the proposed algorithm in the experiments on the second artificial dataset (with 60% bicluster-region) at a missing rate of 20%.

Fig.8. Average NRMSE for the proposed algorithm applied on the Ronen dataset at a missing rate 10% using different values of parameters k and T_0. The minimum NRMSE 0.5286 (highlighted) is achieved at $k = 512$ and $T_0 = 0.0001$.