A New Data Imputing Algorithm

Ahmed Sobhy Sherif , Hany Harb and Sherif Zaky[3]

**[1] Department of Computer Science, Faculty of Computers & Informatics,
Suez Canal University, Ismailia, 41511, Egypt**

**[2] Department of Computers & Eng., Faculty of Engineering,
Al Azhar University, Cairo, 11651, Egypt.**

**[3] Department of Mathematics, Faculty of Science,
Suez Canal University, Ismailia,41511, Egypt.**

Abstract

DNA microarray analysis has become the most widely used functional genomics approach in the bioinformatics field. Microarray gene expression data often contains missing values due to various reasons. Clustering gene expression data algorithms requires having complete information. This means that there shouldn't be any missing values. In this paper, a clustering method is proposed, called "Clustering Local Least Square Imputation method (ClustLLsimpute)", to estimate the missing values. In ClustLLsimpute, a complete dataset is obtained by removing each row with missing values. K clusters and their centroids are obtained by applying a non-parametric clustering technique on the complete dataset. Similar genes to the target gene (with missing values) are chosen as the smallest Euclidian distance to the centroids of each cluster. The target gene is represented as a linear combination of similar genes. Undertaken experiments proved that this algorithm is more accurate than the other algorithms, which have been introduced in the literature.

Keywords: *Missing Values, Imputation, Microarray, Regression***.**

1. Introduction

In the last decade, molecular biologists have been using DNA microarrays as a tool for analyzing information in gene expression data. During the laboratory process, some spots on the array may be missing due to various factors e.g. insufficient resolution, image corruption, or simply due to dust or scratches on the slide. Repeating the experiments is often very costly or time consuming. As a result, molecular biologists, statisticians, and computer scientist have made attempts to recover the missing gene expressions by some ad-hoc and systematic methods.

Microarray gene expression data have been formulated as gene expression matrix *E* with *m* rows, which correspond to genes, and *n* columns, which correspond to experiments. Many analysis methods, such as Principal component analysis, singular value decomposition or clustering analysis, require complete matrices. Missing $\log_2$ transformed data are often replaced by zeros [1] or, less often, by an average expression over the row, or 'row average'. This approach is not optimal, since these methods do not take into consideration the correlation structure of the data. Thus, many analysis techniques, as well as other analysis methods such as hierarchical clustering, k-means clustering, and self-organizing maps, may benefit from using more accurately estimated missing values.

There is not a lot of work in the literature that deals with missing value estimation for microarray data, but much work has been devoted to similar problems in other fields.

The question has been studied in contexts of non-response issues in sample surveys and missing data in experiments [11]. Common methods include filling in least squares estimates, iterative analysis of variance methods [3] randomized inference methods, and likelihood-based approaches [21]. An algorithm similar to the nearest neighbors was used to handle missing values in CART-like algorithms [8]. Most commonly applied statistical techniques for dealing with missing data are model-based approaches. Local least squares imputation as k-nearest neighbor imputation (KNNimpute) [2] and an estimation method based on Bayesian principal component analysis (BPCA) have been introduced [5].

In this paper, a local least squares imputation is proposed, where a target gene that has missing values is represented as a linear combination of similar genes. A k-means clustering algorithm has been used to cluster the complete microarray matrix. Rather than using all available genes in the data, only the genes with high similarity with the target gene are used in the proposed method which has the smallest Euclidian distance between the target gene and the centeroid of each cluster. The rest of the paper is organized as follows:

Section 2 includes a description of a mathematical model of local least squares imputation based on regression model. Section 3, discusses the proposed k-means algorithm which is used in the clustering process. Section 4, introduces the proposed PCA as a solution for the initial number of clusters parameter and the initial centeroid for each of the clusters. Section 5, explains the proposed novel imputing algorithm based on the previous solutions. The results of numerical experiments are given in Section 6. Section 7 concludes the paper.

2. Local Least Squares Imputation

A matrix $G \in \mathbb{R}^{m\times n}$ denotes a gene expression data matrix with m genes and n experiments, and assume$m \gg n$. In the matrix G , a row $g_i^T \in \mathbb{R}^{1\times n}$ represents expressions of the ith gene for n experiments. In order to recover the total of q missing values in any locations of a target gene g, the k-nearest neighbor genes of g,

$$g_{s_i}^T \in \mathbb{R}^{1\times n}, 1 \le i \le k,$$

are found. In this process of finding the similar genes, the q components of each gene at the q locations of missing values in g are ignored. Then, based on these k-nearest neighbor genes, a matrix $A \in \mathbb{R}^{k\times(n-q)}$, a matrix $B \in \mathbb{R}^{k\times q}$, and a vector $w \in \mathbb{R}^{(n-q)\times 1}$ are formed. The ith row vector a_i^T of the matrix A consists of the ith nearest neighbor genes $g_{s_i}^T \in \mathbb{R}^{1\times n}$, $1 \le i \le k$, with their elements at the q missing locations of missing values of g excluded. Each column vector of the matrix B consists of the values of the jth location of the missing values ($1 \le j \le q$) of the k vectors $g_{s_i}^T$. The elements of the vector w are the $n - q$ elements of the gene vector g whose missing items are deleted. After the matrices A and B and a vector w are formed, the least squares problem is formulated as

$$\min_x \|A^T x - w\|_2, \qquad (1)$$

Then, the vector $u = \left(\alpha_1\alpha_2\cdots\alpha_q\right)^T$ of q missing values can be estimated as

$$u = B^T x = B^T (A^T)^\dagger w\,, \qquad (2)$$

where $(A^T)^\dagger$ is the pseudo inverse of A^T.

For example, assume that the target gene g has two missing values in the 1st and the 10th positions among total 10 experiments. If the missing value is to be estimated by the k similar genes, each element of the matrix A and B, and a vector w are constructed as

$$\begin{pmatrix} g_1^T \\ g_2^T \\ \vdots \\ g_{s_k}^T \end{pmatrix} = \begin{pmatrix} \alpha_1 & w_1 & w_2 & \dots & w_s & \alpha_2 \\ B_{1,1} & A_{1,1} & A_{1,2} & \dots & A_{1,s} & B_{1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{k,1} & A_{k,1} & A_{k,2} & \dots & A_{k,s} & B_{k,2} \end{pmatrix},$$

where α_1 and α_2 are the missing values and $g_{s_1}^T, \dots, g_{s_k}^T$ are the k genes that are most similar to g. The known elements of w can be represented by

$$w \cong x_1 a_1 + x_2 a_2 + \cdots + x_k a_k,$$

where x_i are the coefficients of the linear combination, found from the least squares formulation (1). And, the missing values in g can be estimated by

$$\alpha_1 = B_{1,1}x_1 + B_{2,1}x_2 + \cdots + B_{k,1}x_k,$$
$$\alpha_2 = B_{1,2}x_1 + B_{2,2}x_2 + \cdots + B_{k,2}x_k,$$

where α_1 and α_2 are the first and the second missing values in the target gene. For estimating missing values of each gene, we need to build the matrices A and B and a vector w, and solve the least squares problem of Eq. (1).

3. K-Means Clustering

K-means [7] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids; one for each cluster. These centroids should be placed in an accurate way because different locations cause different results. As a result, the best choice is to place them as much as far away as possible from each other. The next step is to take each point belonging to a given data set and associate it with the nearest centroid. When no point is pending, the first step is completed and an early group page is done. At this point we need to re-calculate k new centroids as new centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2, \qquad (3)$$

where $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and c_j the cluster centre , is an indicator of the distance of the n data points from their respective cluster centers.
The algorithm is composed of the following steps:

1- Place *K* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2- Assign each object to the group that has the closest centroid.
3- When all objects have been assigned, recalculate the positions of the *K* centroids.
4- Repeat Steps 2 and 3 until the centroids are longer moving. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

4. Principal Component Analysis

Principal component analysis (PCA) is probably the most popular multivariate statistical technique and it is used by almost all scientific disciplines. It is also likely to be the oldest multivariate technique. In fact, its origin can be traced back to Pearson [14] or even Cauchy [7].
The modern instantiation was formalized by Hotelling [9] who also coined the term principal component. PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the important information from the data table and to express this information as a set of new orthogonal variables called principal components. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in maps [11][8].
The data table to be analyzed by PCA comprises *I* observations described by *J* variables and it is represented by the $I \times J$ matrix *X*, whose generic element is $x_{i,j}$. The matrix *X* has rank *L* where $L \leq min\{I, J\}$.
The matrix *X* has the following singular value decomposition [19][2]:

$$X = P\Delta Q^T \tag{4}$$

where *P* (*Principal direction*) is the $I \times L$ matrix of left singular vectors, *Q* (*Principal components*) is the $J \times L$ matrix of right singular vectors, and Δ is the diagonal matrix of singular values. Eq. (4) can also be rewritten as

$$\mathrm{X} = \sum_{\ell=1}^{\mathrm{L}} \delta_\ell \mathrm{p}_\ell \mathrm{q}_\ell^{\mathrm{T}} \tag{5}$$

with ℓ being the rank of *X* and δ_ℓ , p_ℓ and q_ℓ being (respectively) the ℓ^{th} singular value, left and right singular vectors of *X*. This shows that *X* can be reconstituted as a sum of *L* rank one matrices (i.e., the $\delta_\ell \mathrm{p}_\ell \mathrm{q}_\ell^{\mathrm{T}}$ terms). The first of these matrices gives the best reconstitution of *X* by a rank one matrix, the sum of the first two matrices gives the best reconstitution of *X* with a rank two matrix, and so on, and, in general, the sum of the first *M* matrices gives the best reconstitution of *X* with a matrix of rank *M*.
The goals of PCA are to (a) extract the most important information from the data table, (b) compress the size of the data set by keeping only this important information, (c) simplify the description of the data set, and (d) analyze the structure of the observations and the variables.
In order to achieve these goals, PCA computes new variables called principal components which are obtained as linear combinations of the original variables.
The first principal component is required to have the largest possible variance. Therefore, this component will "explain" or "extract" the largest part of the inertia of the data table. The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed likewise. The values of these new variables for the observations are called factor scores, these factors scores can be interpreted geometrically as the projections of the observations onto the principal components.
In PCA, the components are obtained from the singular value decomposition of the data table I. Specially, with $\mathrm{X} = \mathrm{P}\Delta\mathrm{Q}^{\mathrm{T}}$, the $\mathrm{I} \times \mathrm{L}$ matrix of factor scores, denoted *F* is obtained as

$$\mathrm{F} = \mathrm{P}\Delta, \tag{6}$$

The matrix *Q* gives the coefficients of the linear combinations used to compute the factors scores. This matrix can also be interpreted as a projection matrix because multiplying *X* by *Q* gives the values of the projections of the observations on the principal components. This can be shown by combining Eq.(4) and Eq.(6) as

$$\mathrm{F} = \mathrm{P}\Delta = \mathrm{P}\Delta\mathrm{Q}\mathrm{Q}^{\mathrm{T}} = \mathrm{X}\mathrm{Q}, \tag{7}$$

The components can also be represented geometrically by the rotation of the original axes. The matrix *Q* is also called a loading matrix. The matrix *X* can be interpreted as the product of the factors score matrix by the loading matrix as:

$$\mathrm{X} = \mathrm{F}\mathrm{Q}^{\mathrm{T}} \text{ with } \mathrm{F}^{\mathrm{T}}\mathrm{F} = \Delta^2 \text{ and } \mathrm{Q}\mathrm{Q}^{\mathrm{T}} = \mathrm{I}, \tag{8}$$

This decomposition is often called the bilinear decomposition of *X* .

5. The proposed algorithm

In this section, a Local Least Squares imputation that depends on the Clustering model will be introduced. Clustering complete data sets into K clusters with K centroid per each clusters will also be discussed. Target gene that has missing values represented as a linear combination of similar genes. The similar genes are the clusters whose centroids have the smallest Euclidian distance to the target gene.

5.1. Getting number of clusters

Clustering algorithms are unsupervised learning processes i.e. users are usually required to set some parameters for these algorithms. These parameters vary from one algorithm to another, but most clustering algorithms require a parameter that either directly or indirectly specifies the number of clusters. This parameter is typically either k, the number of clusters to return, or some other parameter that indirectly controls the number of clusters to return, such as an error threshold. Setting these parameters requires either detailed prior knowledge of the data, or time-consuming trial and error. The latter case still requires that the user has sufficient domain knowledge to know what a good clustering "looks" like. However, if the data set is very large or is multidimensional, human verification could become difficult. It is necessary to have an algorithm that can efficiently determine a reasonable number of clusters to return from any clustering algorithm. The following proposed algorithm will identify the correct number of clusters to return from a clustering algorithm.
The algorithm is composed of the following steps:

1. $E_c^{q\times m}$will be the complete microarray matrix by removing each gene row with missing value.
2. By Eq. (4) get the eigengene matrix $\Delta^{q\times m}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$.
3. Compute $\rho_t = \frac{\sum_{j=1}^{t}\sigma_j^2}{\sum_{i=1}^{q}\sigma_i^2}, t = 1, \cdots, q$
4. Choose the ℓ eigengenes that contribute ρ_t to be about 70%-75% of the total expression level as the number of clusters K .

5.2. Getting initial centroids of clusters

The k-means algorithm starts by initializing the K cluster centers. Two simple approaches to cluster center initialization are either to select the initial values randomly, or to choose the first K samples of the data points. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters [14]. Therefore, different methods have been proposed in literature [5][8][4]. When random initialization is used, different runs of K-means typically produce different clusters groups and the resulting clusters are often poor. Another problem with the basic K-means algorithm given earlier is that empty clusters can be obtained. This paper proposes that principal components are actually the continuous solution of the cluster membership indicators in the K-means clustering method.
The main basis of PCA-based dimension reduction is that PCA picks up the dimensions with largest variances Eq. (5). Mathematically, this is equivalent to finding the best low rank approximation of the data via the singular value decomposition Eq. (6). As a result, the first component is used as an index indicator to the K initial centroids.

The algorithm is composed of the following steps:

1. $E_c^{q\times m}$will be the complete microarray matrix by removing each gene row with missing value.
2. By Eq. (4) get P, Δ .
3. Compute first component by Eq.(6).
4. Sort first component vector.
5. Let the first K component indexes of E_c be the first K initial centroids.

5.3. Clustering

K-means clustering algorithm, as shown in section 3, has been used as a clustering algorithm for our proposed imputation algorithm. After applying this algorithm, $\boldsymbol{K}$ of disjoint subsets $A^j, j = 1, \cdots, k$ are obtained. Each cluster A^j is identified by its centroid $x^j \in \mathcal{R}^n, j = 1, \cdots, k$.

5.4. Imputing.

K-nearest neighbor method (KNNimpute) does not introduce an optimal and restricted method to find the nearest neighbor. Bayesian Principal Component (BPCA), depends on a probabilistic model. This model requires certain statistical parameters that must be known before. Local Least Squares Imputation (LLSimpute), depends on K coherent genes that have large absolute values of Pearson correlation coefficients. This can be costly in calculations. This paper proposes Local Least Squares imputation. This method represents a target gene that has missing values as a linear combination of similar genes. The similar genes are the cluster whose centroid has the smallest Euclidian distance to the target gene.
The algorithm is composed of the following steps:

1. $X^{m\times n}$ will be the original microarray matrix.
2. $g^{1\times n}$ will be the target gene (with q missing elements).
3. By using algorithm proposed in section 5.1 get K.
4. By using algorithm proposed in section 5.2 get K

centroids.

5. By using algorithm proposed in section 5.3 get K clusters.
6. Get the nearest cluster $C^{r\times n}$ to the target gene.

   6.1. From $C^{r\times n}$ get $A^{r\times(n-q)}$ with columns corresponding to complete elements of g.

   6.2. From $C^{r\times n}$ get $B^{r\times q}$ with columns corresponding to missing elements of g.

7. $w^{1\times(n-q)}$ with columns corresponding to complete elements of g .
8. Solve Eq. 2 to get estimated q missing values of .
9. Repeat steps from 2 to 8 until estimation of all missing genes.

6. Results and Discussion

Six microarray datasets were obtained for the purpose of comparison. The first data set was obtained from α-factor block release that was studied for identification of cell-cycle regulated genes in *Saccharomyces cerevisiae*[19]. A complete data matrix of 4304 genes and 18 experiments (ALPHA) that does not have any missing value to assess missing value estimation methods. The second data set of a complete matrix of 4304 genes and 14 experiments (ELU) is based on an elutriation data set [19]. The 4304 genes originally had no missing values in the α-factor block release set and the elutriation data set. The third data set was from 784 cell cycle regulated genes, which were classified by Spellman et al. [19] into five classes, for the same 14 experiments as the second data set. The third data set consists of 2856 genes and 14 experiments (CYC-a). The fourth data set of 242 genes and 14 experiments (CYC-b). The fifth data set is from a study of response to environmental changes in yeast [6]. It contains 5431 genes and 13 experiments that have time-series of specific treatments (ENV). The sixth data set is the cDNA microarray data relevant to human colorectal cancer (CRC)[21]. This data set contains 758 genes and 205 primary CRCs that include 127 non-metastatic primary CRCs, 54 metastatic primary CRCs to the liver and 24 metastatic primary CRCs to distant organs exclusive of the liver, and 12 normal colonic epithelia (CRC).

This is a challenging data set with multiple experiments with no time course relationships. The ALPHA, ELU, and CRC are the same data sets that were used in the study of BPCA [5] and LLsimpute[8].

The performance of the missing value estimation is evaluated by normalized root mean squared error(NRMSE) :

$$NRMSE = \sqrt{mean[(y_{guess} - y_{ans})^2]}/std[y_{ans}], \qquad (9)$$

Where y_{guess} and y_{ans} are vectores whose elements are the estimated values and the known answer values, respectively, for all missing entries.

The similarity between the target genes and the closest centroid is defined by the reciprocation of the Euclidian distance calculated for non-missing components.

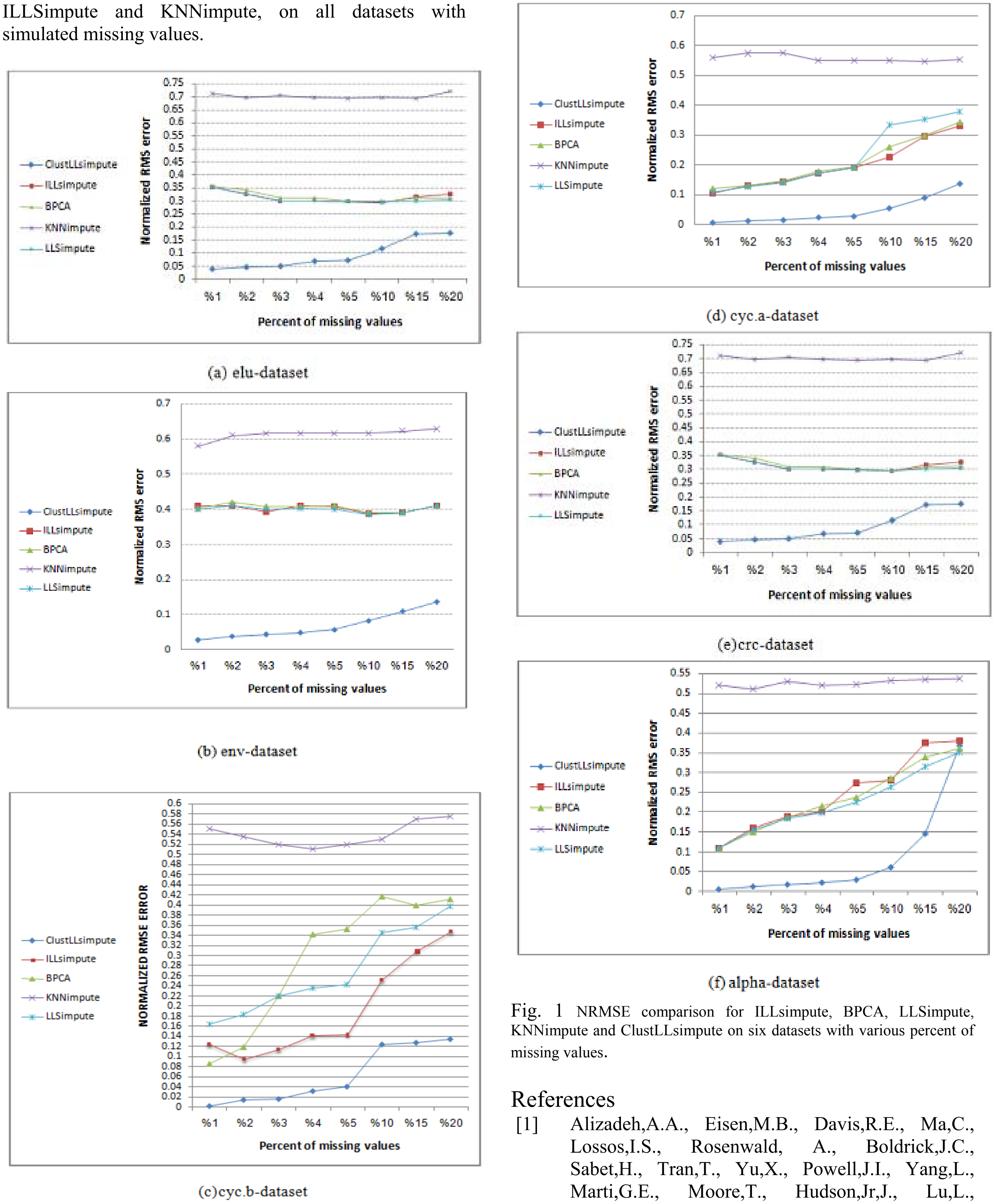
6.1. Experimental results

In the experiments, we randomly removed some percentage, i.e. missing rate, of expression level to create missing values (between 1 and 20% of the data were deleted). Each method was then used to recover the introduced missing values for each data set, and the estimated values were compared to those in the original data set.

From the plots of NRMSE values (Fig. 1) achieved by all five methods on six datasets, we can see that KNNimpute method always performs the worst and ClustLLsimpute always performs the best. For all the other three methods, they perform equally well on env-dataset and crc-dataset but ClustLLsimpute performs better than the other three. In fact, from Figures 1(b) and 1(e), it is hard to tell which one of them performs better than the other three except ClustLLsimpute which is outperform. All other three methods again perform equally well on elu-, cyc-a-, and alpha-datasets when the missing rate is small, i.e. less than 5% (cf. Figures 1(a), 1(d), and 3(f)) and also ClustLLsimpute is outperform all of them. However, the performances differ when the missing rate is large. Our method ClustLLsimpute performs very close to other three as in Figures 1(f) with 20% rate, though still a little better. From these results, it is deduced that the ClustLLsimpute method performs better than both BPCA and LLSimpute, the two most recent imputation methods [3][25].

7. Conclusions

This paper proposes a novel version of Local Least Squares Imputation (ClustLLsimpute) method to estimate the missing values in microarray data. In ClustLLsimpute, the complete dataset is clustered by using a novel clustering k-nearest clustering method to obtain k-clusters and its centroids. The number of nearest neighbors for every target gene is automatically determined as the cluster with the nearest centroid, rather than pre specified in most existing imputation methods. The experimental results on six real microarray datasets show that ClustLLsimpute outperforms the four most well known recent imputation methods BPCA, LLSimpute, , ILLSimpute and KNNimpute, on all datasets with simulated missing values.



(a) elu-dataset

(b) env-dataset

(c)cyc.b-dataset

(d) cyc.a-dataset

(e)crc-dataset

(f) alpha-dataset

Fig. 1 NRMSE comparison for ILLsimpute, BPCA, LLSimpute, KNNimpute and ClustLLsimpute on six datasets with various percent of missing values.

References

[1] Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald, A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X., Powell,J.I., Yang,L., Marti,G.E., Moore,T., Hudson,Jr,J., Lu,L., Lewis,D.B., Tibshirani,R., Sherlock,G., Chan,W.C., Greiner,T.C., Weisenburger, D.D.,

Armitage,J.O., Warnke,R. and Staudt,L.M., et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". Nature, 403, pp. 503–511, 2000.

[2] Abdi, H., and Valentin, D. "Multiple correspondence analysis". In N.J. Salkind (Ed), Encyclopedia of measurement and statistics, pp. 651-657, 2007. Thousand Oaks, CA: Sage.

[3] Cai Z, Heydari M and Lin G. "Iterated local least squares microarray missing value imputation". J Bioinform Comput Biol,4(5):pp. 935-57, 2006.

[4] Ding, C. and He, X. "K-means clustering via principal component analysis". Proceedings of the 21st International Conference on Machine Learning, Banff, Canada,2004.

[5] G. Babu and M. Murty. "A near optimal initial seed value selection in kmeans algorithm using a genetic algorithm". Pattern Recognition Letters Vol. 14, pp. 763-769, 1993.

[6] Gasch, A. P., Huang, M., Metzner, S., Botstein, D., Elledge, S. J. and Brown, P. O. "Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p". Mol. Biol. Cell, 12 (10), pp. 2987–3003, 2001.

[7] Grattan-Guinness, I. "The rainbow of mathematics". New York: Norton, 1997.

[8] H. Kim, G. H. Golub, and H. Park. "Missing value estimation for DNA microarray gene expression data: Local least squares imputation". Bioinformatics, 20:pp. 1–12, 2004.

[9] Hotelling, H. "Analysis of a complex of statistical variables into principal components". Journal of Educational Psychology, 25, pp. 417-441, 1933.

[10] J. B. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1, pp. 281-297, 1967.

[11] Jolliffe, I.T. "Principal component analysis". New York: Springer, 2002.

[12] Little,R.J.A. and Rubin,D.B. "Statistical Analysis with Missing Data,". Wiley, New York,1987.

[13] Loh,W. and Vanichsetakul,N. "Tree-structured classification via generalized discriminant analysis'. J. Am. Stat. Assoc., 83, pp. 715–725, 1988.

[14] M. Ismail and M. Kame. "Multidimensional data clustering utilization hybrid search strategies". Pattern Recognition Vol. 22 (1), pp. 75-89, 1989.

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

[16] Pearson, K. "On lines and planes of closest fit to systems of points in space". Philosophical Magazine, 6, pp. 559-572, 1901.

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

[18] Saporta, G, and Niang, N. "Principal component analysis: application to statistical process control". In G. Govaert (Ed), Data analysis. pp. 1-23, 2009. London: Wiley.

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

[20] Takane, Y. "Relationships among various kinds of eigenvalue and singular value decompositions". In Yanai, H., Okada, A., Shigemasu, K., Kano, Y., and Meulman, J. (Eds.), New developments in psychometrics , 45-56,2002. Tokyo:Springer Verlag.

[21] Takemasa, I., Higuchi, H., Yamamoto, H., Sekimoto, M., Tomita, N., Nakamori, S., Matoba, R., Monden, M. and Matsubara, K. "Construction of preferential cDNA microarray specialized for human colorectal carcinoma: molecular sketch of colorectal cancer. Biochem". Biophys. Res. Commun., 285, pp. 1244–1249, 2001.

[22] Wilkinson,G.N. "Estimation of missing values for the analysis of incomplete data". Biometrics, 14, pp. 257–286, 1958.

[23] Y. Linde, A. Buzo and R. Gray. "An algorithm for vector quantizer design. IEEE trans". Comm. Vol. 28 (1), pp. 84-95,1980.

[24] Yates,Y. "The analysis of replicated experiments when the field results are incomplete". Emp. J. Exp. Agric., 1, pp. 129–142, 1933.