The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC，pp. 280–287

Weighted Local Least Squares Imputation Method for Missing Value Estimation

Wai-Ki Ching[1,*] Kwai-Wa Cheng[2,†] Li-Min Li[1,‡]
Nam-Kiu Tsing[1,§] Alice S. Wong[3,¶]

[1] Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Hong Kong
[2] M.D. Anderson Cancer Center, The University of Texas
[3] Department of Zoology, The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract Missing values often exist in the data of gene expression microarray experiments. A number of methods such as the Row Average (RA) method, KNNimpute algorithm and SVDimpute algorithm have been proposed to estimate the missing values. Recently, Kim et al. proposed a Local Least Squares Imputation (LLSI) method for estimating the missing values. In this paper, we propose a Weighted Local Least Square Imputation (WLLSI) method for missing values estimation. WLLSI allows training on the weighting and therefore can take advantage of both the LLSI method and the RA method. Numerical results on both synthetic data and real microarray data are given to demonstrate the effectiveness of our proposed method. The imputation methods are then applied to a breast cancer dataset.

Keywords Missing values; microarray data; row average method; local least squares imputation method; weighted local least squares imputation method.

1 Introduction

Microarray data analysis is a successful method in genomic research. Many clustering techniques and classification methods for analyzing the microarray data such as Support Vector Machines (SVMs) [15], Principal Component Analysis (PCA) [3, 16], Singular Value Decomposition (SVD) [2] require a complete data set. However, very often gene expression data sets contain missing values, due to various reasons such as insufficient resolution, image corruption, dust or scratches on the slides or experimental errors [9]. Therefore the treatment of missing values is an important step in the preprocessing of the data. It is expensive and also time consuming to repeat the experiment. Therefore a number of imputation methods have been developed for estimating the missing values. For example, the SVD based method

[*] Corresponding Author. Email:wching@hkusua.hku.hk
[†] Email:kwcheng@mdanderson.com
[‡] Email:liminli@hkusua.hku.hk
[§] Email:nktsing@hku.hk
[¶] Email:awong1@hku.hk

(SVDimpute) and the weighted k-Nearest Neighbors Imputation (KNNimpute) have been introduced by Troyanskaya et al. [14].

The KNN-based method actually chooses genes with expression profiles similar to the gene having missing value. Suppose that Gene 1 has one missing value in Experiment 1, this method will find k other genes having no missing values in Experiment 1 and their expressions are most similar (in the sense of Pearson correlation coefficient) to Gene 1 in the rest of the experiments. Then a weighted average of values in Experiment 1 from the k selected genes is used as an estimate for the missing value in Gene 1, see for instance [14]. In the SVDimpute method, SVD is employed to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the data set. These patterns are referred to as eigengenes [1]. The k most significant eigengenes are then identified by sorting the eigengenes based on their corresponding singular values. The missing values can then be obtained by first regressing the corresponding gene against the k eigengenes and then using the coefficients of the regression to reconstruct the missing value from a linear combination of the k eigengenes [14]. It is known that KNNimpute method gives better results on noisy time series data and SVDimpute method performs well on time series data with low noise levels. Nevertheless, estimating unknown values in a given data set has many potential applications in the other fields such as survey sampling [6].

The remainder of the paper is organized as follows. In Section two, we briefly describe the Local Least Squares Imputation (LLSI) method proposed by Kim et al. [4]. In Section three, we present our proposed Weighted Local Least Squares Imputation (WLLSI) method. In Section four, numerical results on both synthetic data and real gene expression microarray data are given to demonstrate the effectiveness of our proposed method when compared with LLSI method and the RA method. We then apply the imputation methods to a breast cancer dataset and interesting results are obtained. Finally, concluding remarks are given in Section five to address further research issues.

2 Local Least Square Imputation Method

In this section, we briefly describe the LLSI method proposed by Kim et al. [4]. We will use the matrix $G \in R^{m\times n}$ to denote a gene expression data matrix with m genes and n experiments. Very often, m is much bigger than n, i.e, $n << m$ and we assume this in our discussion. We adopt the notations in [4]. In the matrix G, a row $\mathbf{g_i}^T \in R^{1\times n}$ represents the expressions of the ith gene in the n experiments. For simplicity of discussion, we assume that there is a missing value in the first experiment of the first gene, i.e., $G(1,1) = \mathbf{g}_1(1) = \alpha$. There are two steps in the LLSI method. The first step is to choose k genes by the L_2-norm or by Pearson correlation coefficients [11]. The second step is to conduct a regression analysis and estimation. To recover a missing value in the first location $\mathbf{g}_1(1)$ of $\mathbf{g}_1$ in the matrix $G \in R^{m\times n}$, the k-nearest neighbor gene vectors for $\mathbf{g}_1$, $\mathbf{g}_{s_i}^T \in R^{1\times n}, \quad 1 \leq i \leq k$ are found for LLSimpute based on the L_2-norm. Here the first component of each gene is ignored as $\mathbf{g}_1(1)$ is missing.

The LLSimpute based on the Pearson correlation coefficient takes the advantage of the coherent genes. When there is a missing value in the first location of $\mathbf{g}_1$, the Pearson correlation coefficient between two genes (both with first entries removed) is computed. We remark that the components of $\mathbf{g}_1$ corresponding to the missing values are not considered in computing the coefficients. All Pearson correlation coefficients between $\mathbf{g}_1$ and the other genes are computed. To recover a missing value in the first location of the gene $\mathbf{g}_1$, $G(1,1) = \mathbf{g}_1(1) = \alpha$, the k genes $\mathbf{g}_{s_1}, \ldots, \mathbf{g}_{s_k}$ with the largest Pearson correlation coefficients in magnitude are found.

Now if the missing value is to be estimated by the k similar genes, the matrix A, and vectors $\mathbf{b}$ and $\mathbf{w}$ can be constructed as follows:

$$\begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_{n-1} \\ \mathbf{b}_1 & A_{1,1} & A_{1,2} & \cdots & A_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_k & A_{k,1} & A_{k,2} & \cdots & A_{k,n-1} \end{pmatrix}$$

where α is the missing value and $\mathbf{g}_{s_1}^T, \ldots, \mathbf{g}_{s_k}^T$ are genes similar to $\mathbf{g}_1^T$. We then solve the following minimization problem:

$$\min_x ||A^T \mathbf{x} - \mathbf{w}||_2 \tag{1}$$

to get $\mathbf{x}$. It is well known that the least squares solution $\mathbf{x}$ to Problem (1) is given by $\mathbf{x} = (\mathbf{A^T})^\dagger \mathbf{w}$ where $A^\dagger$ is the pseudo-inverse of A. Since we assume that $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{n-1}) \approx \sum_{i=1}^k \mathbf{x}_i (A_{i,1}, A_{i,2}, \ldots, A_{i,n-1})$, then the missing value α can be estimated as follows: $\alpha = \sum_{i=1}^k \mathbf{x}_i \mathbf{b_i} = \mathbf{b}^T \mathbf{x} = \mathbf{b}^T (A^T)^\dagger \mathbf{w}$. Finally we remark that the method can be easily extended to the case of multiple missing values as discussed in [4].

3 Weighted Local Least Squares Imputation

In this section, we present our Weighted Local Least Squares Imputation (WLLSI) method for missing value estimation. We assume that there are p missing values $\mathbf{a} \in R^{p \times 1}$ in the first gene of the data. The motivation of our proposed WLLSI method is the following. We observe that the RA method can be efficient when the gene expression data follows certain probability distribution, while LLSI method can be efficient and outperforms the other methods [4] when the rows of the data are strongly correlated. Thus it is natural to consider a method which can take advantage of the two methods. Before proposing our method, we first give some assumptions. In the following, we assume that there are p missing entries in the first p positions of the gene $\mathbf{g_1^T}$ in the matrix $G \in R^{m \times n}$. The k genes most similar to $\mathbf{g_1}$ in the sense of Pearson correlation coefficient or L_2-norm are $\mathbf{g}_{s_i}^T \in R^{1 \times n}, 1 \leq i \leq k$. Then we construct the matrices A and B,and the vectors $\mathbf{a}$ and $\mathbf{w}$ as follows:

$$\begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \left(\begin{array}{c|c} \mathbf{a}^T & \mathbf{w^T} \\ \hline B & A \end{array} \right) = \left(\begin{array}{ccc|ccc} \alpha_1 & \cdots & \alpha_p & \mathbf{w}_1 & \cdots & \mathbf{w}_{n-p} \\ \hline B_{1,1} & \cdots & B_{1,p} & A_{1,1} & \cdots & A_{1,n-p} \\ \vdots & & \vdots & \vdots & & \vdots \\ B_{k,1} & \cdots & B_{k,p} & A_{k,1} & \cdots & A_{k,n-p} \end{array} \right).$$

Here A is a $k \times (n-p)$ matrix, B is a $k \times p$ matrix, $\mathbf{a}$ is a $p \times 1$ column vector and $\mathbf{w}$ is an $(n-p) \times 1$ column vector.

To implement our idea above, we combine the RA method and the LLSI method by the following objective function:

$$\min_{\mathbf{y} \in R^{(n-p)\times p}} \left\{ \lambda \|\mathbf{y}^T\mathbf{w} - \frac{\mathbf{e}\mathbf{f}^T\mathbf{w}}{n-p}\|_2^2 + (1-\lambda)\|A\mathbf{y} - B\|_2^2 \right\}, \tag{2}$$

where $\mathbf{e} = (1,1,\ldots,1)^T$ is a p-dimensional column vector and $\mathbf{f} = (1,1,\ldots,1)^T$ is the $(n-p)$-dimensional column vector of all ones. The parameter $\lambda \in [0,1]$ is the weighting factor which can be obtained by training on the known data. The first term of the function (2) corresponds to the objective function (to be minimized) of the RA method and the second term corresponds to the objective function of the LLSI method (compare (1) and (2)). For simplicity, we define the column vector $\mathbf{c} = \frac{\mathbf{e}\mathbf{f}^T\mathbf{w}}{n-p} \in R^{p\times 1}$, then the problem can be formulated as follows:

$$\min_{\mathbf{y}} \left\{ \lambda \|\mathbf{y}^T\mathbf{w} - \mathbf{c}\|_2^2 + (1-\lambda)\|A\mathbf{y} - B\|_2^2 \right\}. \tag{3}$$

To solve the minimization problem (3), we begin with the following proposition.

Proposition 1.
If $\tilde{A} = \lambda \mathbf{w}\mathbf{w}^T + (1-\lambda)A^TA$ is positive definite, then the optimal solution $\mathbf{y}^*$ of the above minimization problem (3) is given by $\mathbf{y}^* = \tilde{A}^{\dagger}\tilde{B}$ where $A^{\dagger}$ is the pseudo-inverse of A and $\tilde{B} = \lambda \mathbf{w}\mathbf{c}^T + (1-\lambda)A^TB$.

Proof. We note that

$$\begin{aligned} f(\mathbf{y}) &= \lambda \|\mathbf{y}^T\mathbf{w} - \mathbf{c}\|_2^2 + (1-\lambda)\|A\mathbf{y} - B\|_2^2 \\ &= \lambda(\mathbf{w}^T\mathbf{y}\mathbf{y}^T\mathbf{w} - 2\mathbf{c}^T\mathbf{y}^T\mathbf{w} + \mathbf{c}^T\mathbf{c}) + (1-\lambda)(\mathbf{y}^TA^TA\mathbf{y} - 2\mathbf{y}^TA^TB \\ &\quad +B^TB). \end{aligned}$$

Then we can easily get

$$\begin{aligned} \tfrac{1}{2}\nabla_{\mathbf{y}} f &= \lambda(\mathbf{w}\mathbf{w}^T\mathbf{y} - \mathbf{w}\mathbf{c}^T) + (1-\lambda)(A^TA\mathbf{y} - A^TB) \\ &= (\lambda \mathbf{w}\mathbf{w}^T + (1-\lambda)A^TA)\mathbf{y} - (\lambda \mathbf{w}\mathbf{c}^T + (1-\lambda)A^TB) = \tilde{A}\mathbf{y} - \tilde{B}. \end{aligned}$$

Since $\tilde{A}$ is positive definite, the optimal solution of the original problem is given by $\mathbf{y}^* = \tilde{A}^{\dagger}\tilde{B}$. □

We remark that if A is a matrix with full column rank, then A^TA is positive definite. The matrix $\tilde{A}$ is a rank one perturbation of the positive definite matrix $(1-\lambda)A^TA$. We note that for $\mathbf{z} \neq \mathbf{0}$ and $\lambda \in [0,1)$, we have

$$\mathbf{z}^T\tilde{A}\mathbf{z} = \lambda(\mathbf{z}^T\mathbf{w})^2 + (1-\lambda)\mathbf{z}^TA^TA\mathbf{z} \geq (1-\lambda)\mathbf{z}^TA^TA\mathbf{z} > 0.$$

Thus $\tilde{A}$ is positive definite for $\lambda \in [0,1)$. If we fix the parameter λ, then we can obtain the optimal solution of the minimization problem. Then the p missing values

$\mathbf{a} \in R^{p\times 1}$ in the Gene $\mathbf{g}_1$ can be estimated by $\mathbf{a} = \mathbf{y}^{*T}\mathbf{w}$. If there are missing values in other genes, we can estimate them one by one.

Before giving a method for choosing the parameter λ, we first introduce the Normalized Root Mean Squared Error (NRMSE). NRMSE is used to evaluate the performance of the estimation methods for missing values, see for instance [10]. The NRMSE is defined as follows:

$$\text{NRMSE} = \frac{\sqrt{mean(\mathbf{a}_{guess} - \mathbf{a}_{ans})}}{std(\mathbf{a}_{ans})}$$

where $\mathbf{a}_{guess}$ and $\mathbf{a}_{ans}$ are vectors containing the estimated values and the true values for all missing entries respectively. The mean and the standard deviation are then calculated over missing entries in the entire matrix. For a given gene expression data with missing values, the smaller the value of NRMSE, the better the method will be.

The model parameter λ remains to be determined. In our experiments, the following heuristic grid search algorithm is used to obtained the best parameter λ^*. We divide the interval $[0,1]$ into N (say $N = 100$) sub-intervals. Then we get $N+1$ values of λ as follows: $\lambda_j = \frac{j}{N}, \quad \text{for} \quad j = 0,1,\ldots,N$. If there are missing values in the Gene g_i, then we pretend one or two existing values in this gene to be missing. Then for different $\lambda_j, j = 0,1,\ldots,N$, we use our WLLSI method to estimate these pretended missing values, thereby calculate the NRMSEs over the entire matrix. Finally we choose the optimal λ^* corresponding to the smallest NRMSE as the best weighting for our model. We then use this model to estimate the missing values. In Proposition 1, we have described our proposed method by formulating the model and giving a feasible solution to it. In the following section we will give some numerical experiments to illustrate the effectiveness of our proposed method.

4 Numerical Results

In this section, we compare our proposed WLLSI method with the Row Average method and LLSI method in both synthetic data and real data. For the synthetic data, it is generated by combining a matrix with entries following the uniform distribution $U(0,1)$, and another matrix having strongly linear dependent rows. The real data set comes from the yeast gene expression data [16]. We then consider a breast cancer dataset. Interesting results are obtained in the clustering analysis when different imputation methods are used in recovering the missing values in the dataset.

For the synthetic data, we generate data in the form of $G = (1-w)P + wQ$ where $w \in [0,1]$. Here P is a 474×15 matrix such that its ith row is given by the (i mod 15)th row of the matrix $M = (I - zz^T) \in R^{15\times 15}$ where $z = (-1/7, -1/6, \ldots, -1, 0, 1, \ldots, 1/6, 1/7)^T$ and Q is a random matrix whose entries follow the uniform distribution $U(0,1)$. Here a mod b is the remainder when a is divided by b. We then randomly pick 0.2% of the entries of G and assume they are missing. Then we apply the Row Average method, the LLSI method and our WLLSI method to estimate these missing entries and the results of their NRMSEs are reported in Tables I. We observe that our WLLSI method is robust and has the best performance in general.

Table 1: NRMSE when the number of missing values is 132 (0.2%)

w	RA Method	LLSI Method	WLLSI Method	λ^*
1.0	1.01	1.03	1.01	0.98
0.9	0.98	1.03	0.99	0.94
0.8	0.98	1.03	0.98	0.94
0.7	0.98	1.03	0.99	0.92
0.6	0.98	1.02	0.98	0.92
0.5	0.98	1.02	0.99	0.89
0.4	0.98	1.01	0.98	0.90
0.3	0.98	1.00	0.98	0.80
0.2	1.00	0.98	0.98	0.45
0.1	1.05	0.75	0.75	0.00
0.0	1.10	0.00	0.00	0.00

Table 2: NRMSE of different methods

Number of Missing Values	65 (0.1%)	33 (0.05%)	8 (0.02%)
RA Method	0.66	0.38	0.52
LLSI Method	0.47	0.40	0.15
WLLSI Method	0.47	0.33	0.08
optimal λ^*	0.00	0.03	0.45

For the real data example, we use a practical data set taken from yeast data set (Yeung and Ruzzo [16]). The raw matrix is available at http://hkumath.hku.hk/∼wkc/yeast.xls. The gene expression data is a 384×17 matrix. We randomly pick 0.1%, 0.05%, 0.02% of the entries of the matrix and assume that they are missing. We then use the three different methods to estimate the missing values. Table II reports the results of NRMSEs, and our WLLSI method is the best.

We then apply the imputation methods to a published breast cancer gene expression dataset (Sortlie et al., [12]) to recover the missing data. In general, there are more genes being identified after the missing data analysis. Using the original data, we have identified 89 differentially expressed genes between normal and breast cancer sample. With the aid of missing data analysis, 9 additional genes were identified. Among them, 3 genes (PPAP2N, CD01, CDKN1C) are common among all missing value estimation methods, whereas LLSI method or WLLSI method and RA method specifically identify 1 (CCNA2) and 5 (PTPN1, LEPROTL1, CCNF, CCL7 and C21orf45) additional genes, respectively.

We remark that some of the genes that are identified after data preprocessing by using our imputation methods do have pathological significance. In fact,it is well known that cell division cycle is tightly controlled by activation and inactivation of cyclin-dependent kinases (CDKs), which trigger the transition to subsequent phases of the cycle. CDKs are small serine/threonine protein kinases that require association with a cyclin subunit for their activation. CDK inhibitors (CKIs) can prevent cell cycle progression by negatively regulating cyclin-CDK complexes ([8] and [13]). Interestingly, many of the differential expressed genes, such as CCNA2 (cyclin A2), CCNF (cyclin F), and CDKN1C (p57 Kip2), as well as those from the original data (CDC2, CDKN2C, p18 CDK4) are known to regulate cell cycle progression, in particular G1-S and G2-M transitions.

5 Concluding Remarks

In this paper, we proposed the WLLSI method for missing value estimation. The WLLSI method is a combination of the Row Average method and the LLSI method. The method allows the model parameter λ to be trained. Numerical results based on synthetic data and real yeast data show that our method is more effective and robust in general. We remark that WLLSI can be easily extended to the case when we consider the combination of "Column Average method" and LLSI method.

We observe that when the number of missing entries in the matrix are relatively too large (e.g. more than 10%), the information which can be used in the genes with missing values is very limited, so hence making it difficult to obtain a good model parameter λ through training. The success of our WLLSI method also relies on the process of choosing the k most similar genes. Therefore we need a good measurement to measure the similarity of two genes. We will further develop our method so as to cope with the above two difficulties.

References

[1] Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl Acad. Sci. USA, 97, 10101-10106.

[2] Golub, G.H. and van Loan, C.F. (1996) Matrix Computations. 3rd edition. Johns Hopkins University Press, Baltimore, CA.

[3] Jolliffe, I.T. (2002) Principal component analysis. 2nd edition, Springer, New York.

[4] Kim, H., Golub, G.H. and Park, H., (2005) Missing value estimation for DNA microarray gene expression data: local least square imputation. Bioinformatics, 21, 187-197.

[5] Kong, M., Barnes, E.A., Ollendorff, V., and Donoghue, D.J. (2000) Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction. EMBO J. 19, 1378-88.

[6] Longford, N.T. (2005) Missing data and small-area estimation : modern analytical equipment for the survey statistician. Springer, New York.

[7] Ma, Y., and Cress, W.D. (2006) Transcriptional upregulation of p57 (Kip2) by the cyclin-dependent kinase inhibitor BMS-387032 is E2F dependent and serves as a negative feedback loop limiting cytotoxicity. Oncogene (Epub ahead of print).

[8] Morgan D.O. (1997) Cyclin-dependent kinases: engines, clocks, and microprocessors. Annu. Rev. Cell Dev. Biol. 13, 261-91.

[9] Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19, 2088-96.

[10] Ouyang,M. et al. (2004) Gaussian mixture clustering and imputation of microarray data. Bioinformatics, 20, 917-23.

[11] Pearson, K. (1894) Contributions to the mathematical theory of evolution. Phil. Trans. R. Soc. London, 185, 71-110.

[12] Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Eystein Lonning, P., and Borresen-Dale, A.L. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci., 10869-74.

[13] Tannoch, V.J., Hinds, P.W., and Tsai, L.H. (2000) Cell cycle control. Adv. Exp. Med. Biol. 465, 127-40.

[14] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarray. Bioinformatics, 17, 520-25.

[15] Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer-Verlag, New York.

[16] Yeung, K. and Ruzzo, W. (2001). An empirical study on principal component analysis for clustering gene expression data. Bioinformatics, 17, 763-74.