

Imputation of Missing Values in DNA Microarray Gene Expression Data

Hyunsoo Kim
Dept. of Computer Science
University of Minnesota
Minneapolis, MN 55455
hskim@cs.umn.edu

Gene H. Golub
Computer Science Department
Stanford University
Stanford, CA 94305
golub@stanford.edu

Haesun Park
Dept. of Computer Science
University of Minnesota
Minneapolis, MN 55455
hspark@cs.umn.edu

Abstract

Most multivariate statistical methods for gene expression data require a complete matrix of gene array values. In this paper, a imputation method based on least squares formulation is proposed to estimate missing values. It exploits local similarity structures in the data as well as least squares optimization process. The proposed local least squares imputation method (LLSimpute) represents a target gene that has missing values as a linear combination of similar genes. This algorithm showed better performance than the other imputation methods such as k -nearest neighbor imputation and an imputation method base on Bayesian principal component analysis.

1. Introduction

Gene expression data sets often contain missing values due to various reasons, e.g. insufficient resolution, image corruption, dust or scratches on the slides, or experimental error during the laboratory process. Since it is often very costly or time consuming to repeat the experiment, many algorithms have been developed to recover the missing values [3, 2]. In this paper, a local least squares imputation (LLSimpute) is proposed, where a target gene that has missing values is represented as a linear combination of similar genes. 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. It is compared with k -nearest neighbor imputation (KNNimpute) [3] and an estimation method based on Bayesian principal component analysis (BPCA) [2].

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 $\mathbf{g}_i^T \in \mathbb{R}^{1 \times n}$ represents expressions

of the i th gene for n experiments. In order to recover the total of q missing values in any locations of a target gene \mathbf{g} , the k -nearest neighbor genes of \mathbf{g} ,

$$\mathbf{g}_{s_i}^T \in \mathbb{R}^{1 \times n}, \quad 1 \leq i \leq k,$$

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

$$\min_{\mathbf{x}} \|A^T \mathbf{x} - \mathbf{w}\|_2. \quad (1)$$

Then, the vector $\mathbf{u} = (\alpha_1 \ \alpha_2 \ \dots \ \alpha_q)^T$ of q missing values can be estimated as

$$\mathbf{u} = B^T \mathbf{x} = B^T (A^T)^\dagger \mathbf{w}, \quad (2)$$

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

For example, assume that the target gene \mathbf{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 \mathbf{w} are constructed as

$$\begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \alpha_1 & \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_8 & \alpha_2 \\ B_{1,1} & A_{1,1} & A_{1,2} & \dots & A_{1,8} & B_{1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{k,1} & A_{k,1} & A_{k,2} & \dots & A_{k,8} & B_{k,2} \end{pmatrix},$$

where α_1 and α_2 are the missing values and $\mathbf{g}_{s_1}^T, \dots, \mathbf{g}_{s_k}^T$ are the k genes that are most similar to \mathbf{g} . The known ele-

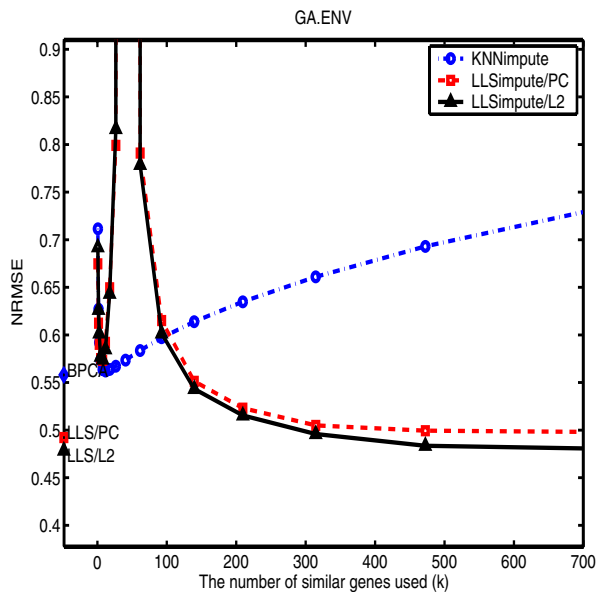


Figure 1. Comparison of the normalized RMS errors of different methods and effect of the number of genes for estimating missing values on GA.ENV data set. The results of the methods that do not depend on the number of genes are shown on the y-axis.

ments of w can be represented by

$$w \simeq x_1 a_1 + x_2 a_2 + \dots + x_k a_k,$$

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

$$\begin{aligned} \alpha_1 &= B_{1,1}x_1 + B_{2,1}x_2 + \dots + B_{k,1}x_k, \\ \alpha_2 &= B_{1,2}x_1 + B_{2,2}x_2 + \dots + B_{k,2}x_k, \end{aligned}$$

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 Eqn. (1).

3. Results and Discussion

The data set we used was from a study of response to environmental changes in yeast [1]. It contains 6361 genes and 156 experiments that have time-series of specific treatments. A complete matrix of 2641 genes and 44 experiments was formed after removing experimental columns that have more than 8% missing values and then selecting gene rows that do not have any missing value (GA.ENV).

Given an initial expression data matrix, 5% of the data elements of the matrix were randomly chosen and regarded as missing values. The performance of the missing value estimation is evaluated by normalized root mean squared error (NRMSE). In KNNimpute, a weighted average of the k -nearest neighbors of a target gene was used as an estimate for each missing value in the target gene. The similarity between two genes is defined by the reciprocal of the Euclidian distance calculated with only non-missing components. Then, a missing entry was estimated as an average weighted by the similarity values. For LLSimpute, the similar genes can be chosen by k -nearest neighbors or k -most coherent genes that have large magnitude of Pearson correlation coefficient. If a method uses k -nearest neighbors, 'L2' is appended to its name, while it is based on k -most coherent genes, a suffix, 'PC', is appended. Nonparametric missing values estimation methods of LLS/L2 and LLS/PC were designed by estimating an optimal k -value only using non-missing parts. This k -value estimating procedure considers some elements of the non-missing parts as artificial missing values, and finds an expected k -value that produces the best estimation ability for the artificial missing values.

In Figure 1, the NRMSE values of imputation methods are presented. LLSimpute outperformed BPCA as well as KNNimpute when k is large. Even though BPCA showed better performances than KNNimpute for all data sets tested in the study of BPCA [2], when genes have dominant local similarity structures, BPCA may be less accurate than KNNimpute [2]. However, LLSimpute takes advantage of the local similarity structures as well as the optimization process by the least squares, which is one of the most important advance of LLSimpute.

4. Conclusion

The local least squares imputation (LLSimpute) is successfully designed for the missing value estimation in microarray gene expression data.

References

- [1] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Mol. Biol. Cell*, 12(10):2987–3003, 2001.
- [2] 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):2088–2096, 2003.
- [3] 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):520–525, 2001.