

Microarray Missing Data Imputation based on a Set Theoretic Framework and Biological Constraints

Author

Gan, Xiangchao, Liew, Alan Wee-Chung, Yan, Hong

Published

2006

Conference Title

18TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, VOL 3, PROCEEDINGS

DOI

[10.1109/ICPR.2006.796](https://doi.org/10.1109/ICPR.2006.796)

Rights statement

© 2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Downloaded from

<http://hdl.handle.net/10072/24404>

Griffith Research Online

<https://research-repository.griffith.edu.au>

Microarray Missing Data Imputation based on a Set Theoretic Framework and Biological Constraints

Xiangchao Gan¹, Alan Wee-Chung Liew² and Hong Yan^{1,3}

¹*Department of Electronic Engineering*

City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

²*Department of Computer Science and Engineering*

The Chinese University of Hong Kong, Shatin, Hong Kong

³*School of Electrical and Information Engineering*

University of Sydney, NSW 2006, Australia

Abstract

Gene expressions measured using microarrays usually suffer from the missing value problem. Existing missing value imputation algorithms have some limitations. For example, some algorithms have good performance only when strong local correlation exists in data while some provide the best estimate when data is dominated by a global structure. In addition, these algorithms do not take into account many biological constraints in the imputation procedure. In this paper, we propose a set theoretic framework for missing data imputation. We design our algorithm by taking into consideration the biological characteristic of the data and exploit the local correlation and the global correlation structure adaptively. Experiments show that our algorithm can achieve a significant reduction of error compared with existing methods.

1. Introduction

DNA Microarray has been widely used in numerous biological disciplines, such as the investigation of drug action and cancer prognosis. However, microarray data often contain missing values with measurements for many genes affected. Missing values occur due to various reasons, including hybridization failures, artifacts on the microarray, insufficient resolution and image noise.

There are several simple methods to deal with missing values [1]. Although all these algorithms have shown good performance to deal with missing values when the required condition is satisfied, they also have their limitations. KNNimpute performs better on non-time series data or noisy time series data, while SVDimpute works well on time series data with low

noise level and with strong global correlation structure. BPCA is suitable when a global structure is dominant in data. Nevertheless, these algorithms do not consider biological constraints related to the microarray experiments.

In this paper, we propose a new missing value imputation algorithm that has a superior performance to existing algorithms. This is an extension of our earlier work [2] and the novelty of our current method in this paper lies in two aspects. First, our algorithm can exploit the local and global correlation structures in microarray data adaptively. Second, we make explicitly use synchronization loss, a biological phenomenon in microarray experiment, as a constraint in the imputation process.

2. Biological property for missing values imputation

In a microarray experiment, synchronization is achieved by first arresting cells at a specific biological life point and then releasing cells from the arrest so that all cells are at the same point when the experiment begins [3]. However, even if cells are synchronized perfectly at the beginning of the experiment, they do not remain synchronized forever [4]. For example, yeast cells seem to remain relatively synchronized for two cycles while wild type human cells lose their synchronization very early or halfway through the first cycle depending on the arresting method. This causes the peak expression value to be lower in the second cycle and the lowest expression value to be higher for most cycling genes. A typical gene expression profile with synchronization loss is given in Fig. 1. Due to this phenomenon, we find that the average signal power in successive cycles decreases significantly. Table 1

shows our statistical results of four datasets in Spellman et al.'s experiment.

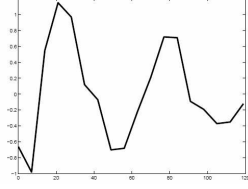


Fig. 1. The gene expression profile of Smc3 in Spellman et al.'s experiment. The synchronization loss is significant.

Dataset	No. of sampling points	No. of complete genes	Average energy, 1st cycle	Average energy, 2nd cycle
CDC28	17	1383	352.78	288.73
CDC15	24	4381	846.28	834.57
Alpha factor	18	4489	474.58	306.24
Elutriation	14	5766	898.48	435.32

Table 1. The statistical result of synchronization loss for 4 datasets in Spellman et al.'s experiment. Note that since the signal of Elutriation is available only for one cycle, we compare the average signal energy for first-half cycle and the second-half cycle.

3. The STF-based Imputation Algorithm

The Set Theoretic Framework (STF) provides a convenient framework to allow multiple pieces of prior information of different nature to be utilized to get an optimal solution. It has been used successfully in signal and image processing [5, 6]. In this method, every known *a priori* property about the original signal is formulated as a corresponding convex set in a Hilbert space. Given m closed convex sets $C_i, i=1,2,\dots,m$, and nonempty intersection $C_0 = \bigcap_{i=1}^m C_i$, the simultaneous projections onto the convex sets

$$a_{n+1} = \sum_{l=1}^m w_l P_l(a_n) \quad (1)$$

will converge to a point in the intersection C_0 for any initial a_0 , where a_n is the estimation of the signal at iteration n , and P_i is the projector onto C_i defined by

$$\| \mathbf{x} - P_i(\mathbf{x}) \| = \min_{\mathbf{g} \in C_i} \| \mathbf{x} - \mathbf{g} \| \quad (2)$$

Another useful feature of the POCS algorithm is its adaptivity in finding a good solution. This can be explained as follow. Suppose we have correlation information between genes and between samples, and this two pieces of information are modeled as two convex sets C_u and C_v , respectively. In one dataset, the first piece of information may be more reliable than the second. In another dataset, it may be the opposite. This

situation is depicted in Fig. 2. When the information is more reliable, the corresponding convex set will be smaller in range. Since POCS always converge to the intersection, the final solution will always be dominated by the smaller set, while still satisfying the constraint imposed by the less reliable set. In this manner, a good solution that makes trade-off between different prior information can be obtained.

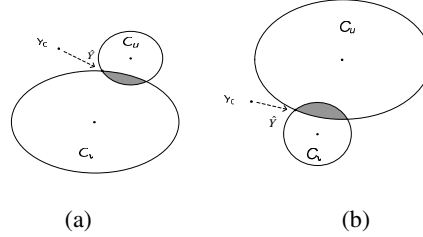


Fig. 2. POCS adaptivity: (a) In a data dominated by gene-wise correlation, the final solution is dominated by C_u , and (b) in a data dominated by array-wise correlation, the final solution is dominated by C_v .

3.1 Capturing gene-wise correlation

In gene expression data, genes that have close biological functions would express similarly. To capture this localized gene-wise correlation in the gene expression data, we construct a convex set based on local least square regression as in [7] as follows. First, we select the K -most correlated genes in Y whose expression profile vectors are similar to gene i except the j -th component and with the j -th component available. Then we estimate the missing value in the target gene using each reference gene based on the single regression model. If we denote the expression profile vector of the target gene as \mathbf{y} , for a reference gene \mathbf{x} , we have

$$\mathbf{y} = \alpha + \beta \mathbf{x} + \mathbf{e} \quad (3)$$

where \mathbf{e} represents random noise. The parameters α and β are estimated by

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \quad \text{and} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad (4)$$

where $S_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$, and $S_{xx} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$.

Thus the least squares estimate of a variable f given a variable t can be written as $\hat{f} = \bar{y} + \frac{S_{xy}}{S_{xx}}(t - \bar{x})$ and the

variance of the residual error is given by

$$\tau = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \quad (5)$$

Since we have K reference genes, we can obtain K estimates \hat{f}_l with variance of estimation error as τ_l ($l = 1, 2, \dots, K$) for a missing value. A weighted average of the estimates and the corresponding variance of the estimation error can be computed.

Defining the positions in matrix Y of all missing values as a set I , we can get an estimate $\hat{B}(l), l \in I$ for every missing value using the above regression. Considering the possible estimation error in this method, we obtain a set

$$C_u = \{Y : \hat{B}(l) - \tau\sqrt{p} \leq Y(l) \leq \hat{B}(l) + \tau\sqrt{p}, l \in I\} \quad (6)$$

The projection P_u onto convex set C_u is then given by

$$P_u(Y(l)) = \begin{cases} \hat{B}(l) - \tau\sqrt{p} & \text{for } Y(l) < \hat{B}(l) - \tau\sqrt{p} \\ \hat{B}(l) + \tau\sqrt{p} & \text{for } Y(l) > \hat{B}(l) + \tau\sqrt{p} \\ Y(l) & \text{otherwise} \end{cases} \quad (7)$$

3.2 Capturing array-wise correlation

We use the PCA approach to capture the global array-wise variation. Assume we have a complete data matrix with no missing value. PCA represents the variation of each array vector \mathbf{y} as a linear combination of principle axis vector \mathbf{u}_l ($0 < l < K$)

$$\mathbf{y} = \sum_{l=1}^K x_l \mathbf{u}_l + \boldsymbol{\varepsilon} \quad (8)$$

The linear coefficients x_l ($0 < l < K$) are called factor scores and $\boldsymbol{\varepsilon}$ denotes the residual error. For each \mathbf{u}_l , there is a corresponding eigenvalue λ_l . For gene expression data, eigenvalue λ_l indicates the relative significance of the l -th eigenarray in term of the fraction of the overall expression they captured. In PCA for gene expression data, only the K ($0 < K < L$) most significant eigenarray are used [8]. The other $L-K$ eigenarray are treated as noise and the signal-to-noise ratio is given by

$$p = \frac{\sum_{k=1}^K \lambda_k^2}{\sum_{k=K+1}^L \lambda_k^2} \quad (9)$$

The estimation error is given by $\boldsymbol{\varepsilon}$. As in Equation (6), when a solution $\tilde{y}_{i,j}$ for a missing value is found using this method, a more reliable estimate is that the missing value lies in an interval $[(1 - \tau\sqrt{p})\tilde{y}, (1 + \tau\sqrt{p})\tilde{y}]$, where τ is a parameter determined statistically. Defining the positions of all missing values in matrix Y as a set I , we can construct a convex set

$$C_v = \{Y : \varepsilon_1 \tilde{A}(l) \leq Y(l) \leq \varepsilon_2 \tilde{A}(l), l \in I\} \quad (10)$$

where the $\tilde{A}(l)$ is the estimated missing value using the eigenarrays, and $\varepsilon_1 = (1 - \tau\sqrt{p})$ and $\varepsilon_2 = (1 + \tau\sqrt{p})$. The projection onto set C_v is then given by

$$P_v(Y(l)) = \begin{cases} \varepsilon_1 \tilde{A}(l) & \text{for } Y(l) < \varepsilon_1 \tilde{A}(l) \\ \varepsilon_2 \tilde{A}(l) & \text{for } Y(l) > \varepsilon_2 \tilde{A}(l) \\ Y(l) & \text{otherwise} \end{cases} \quad (11)$$

3.3 Capturing the phenomenon of synchronization loss

We propose here a series of convex sets to take advantage of the phenomenon of synchronization loss. Define the positions in matrix Y of all missing values belonging to the i -th period as a set I_i , and the positions of all observed values belonging to the i -th period as a set Ω_i . Let function $u(I)$ be the cardinal number of set I .

We get

$$C_i = \left\{ Y : \frac{1}{u(I_i)} \sum_{l \in I_i} Y^2(l) = \varphi_i \right\} \quad (12)$$

with $\varphi_i = \frac{1}{u(\Omega_i)} \sum_{l \in \Omega_i} Y^2(l)$ and $i = 1, \dots, n$ denotes the number of periods considered.

The projection of an arbitrary $Y(i, j)$ onto the set C_i is then given by

$$P_i(Y(l)) = \begin{cases} \frac{Y(l)}{\sqrt{\varphi_i \cdot u(I_i)}} & \text{for } l \in I_i, \text{ and } \sum_{l \in I_i} Y^2(l) \neq 0 \\ Y(l) & \text{otherwise} \end{cases} \quad (13)$$

3.4 Our algorithm

With the convex sets defined, the set theoretic estimation yields the following missing values imputation algorithm:

(1) Select average of the gene expression profile as initial estimation Y_0 .

(2) For $k = 1, 2, \dots$, compute Y_k from

$$Y_k = w_1 P_u(Y_{k-1}) + w_2 P_v(Y_{k-1}) + w_3 P_1 \cdots P_n(Y_{k-1}) \quad \text{where} \\ P_u, P_v, P_1, \dots, P_n \text{ denote the projectors onto the} \\ \text{constraint sets } C_u, C_v, C_1 \cdots C_n, \text{ respectively, and} \\ w_1, w_2 \text{ and } w_3 \text{ are weighting parameters of POCS} \\ \text{with } \sum_{l=1}^3 w_l = 1 \text{ and } w_l > 0 \text{ for all } l = 1, 2, 3. \text{ A}$$

convenient choice is to let w_l 's be equal.

(3) If $Y_k = Y_{k-1}$, exit the iteration, else go to step 2.

4 Experiment results

In this section, we apply our method to two microarray data sets. The first one is the study of yeast cell-cycle from Spellman et al. (<http://cellcycle-www.stanford.edu>) [3]. It contains expression profiles of 6178 genes under different experimental conditions, i.e., *cdc15*, and *cdc28*, alpha factor and elutriation experiments. In addition, one of the time-series data

sets contained less apparent noise than the other. Another time series data set is from [8], which we denote as fkh1_fkh2.

The missing value estimation techniques were tested by randomly removing data values and then computing the estimation error. In the experiments, between 1-15% of the values is removed from each dataset. The normalized RMS (NRMS) error is calculated. We compare the normalized RMS error of our algorithm with the KNNimpute, SVDimpute and LSImpute algorithms. Due to limited space of the paper, we only provide an experiment with 15% missing values in Fig. 3.

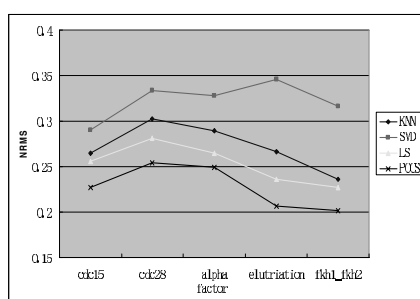


Fig. 3. Performance comparison for 15% missing values.

We now analyze the performance of the imputation algorithms as a function of gene expression level. The box and whisker plots of the estimation errors relative to the true expression values (log2 ratio) in range are provided in Fig. 4. The test data are from Elutriation with 4489 genes and 18 arrays and 1% missing values. From the plots, we can find some interesting properties of our estimation. When the magnitude of true values is small, the performances of three methods are close. However our algorithm has lower error median and spread when true expression values are medium or large (log2 ratio between 0.68 to 1.5). In a microarray experiment, the expression ratios of those genes with medium or high expression level are considered to be more reliable and hence taken with greater faith. If those values are missing due to experimental artifacts or contaminations, we would like them to be more reliably imputed. Our algorithm provides better estimate for these medium or large values.

5. Conclusion

In this paper, we have proposed a set theoretic approach based on POCS which we call POCSimpute for the problem of microarray missing value estimation. POCSimpute can adaptively find an optimal solution regardless of whether the global or local correlation

structure is dominant in the target data. Furthermore, it can conveniently make use of biological constraints to get a better estimate. Experiments show that our algorithm can achieve a significant reduction of error compared existing algorithms.

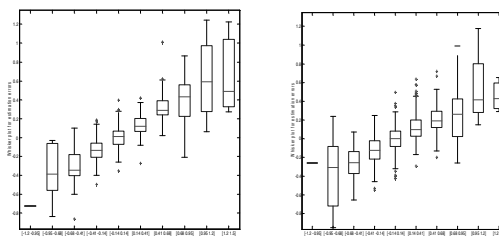


Fig. 4. Box and whisker plots of the estimation errors relative to the true values in range. (left) KNNimpute, and (bottom) POCSimpute.

Acknowledgement: This work is supported by a CityU interdisciplinary grant (project 9010003) and strategic research grant (project 7001706).

REFERENCES

- [1] O. Troyanskaya, et al. "Missing values estimation methods for DNA microarrays," *Bioinformatics*, **17**, 520-525, 2001.
- [2] X. C. Gan, A. W. C. Liew and Yan, H., "Missing value estimation for microarray data based on projection onto convex sets method," *Int'l Conf. Pat. Rec. (ICPR2004)*, Cambridge, UK, III:782-785, 2004..
- [3] P. T. Spellman, P.T. et al., "Comprehensive Identification identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell.*, **9**, 3273-3297, 1998.
- [4] Z. Bar-Joseph, et al., "Deconvolving cell cycle expression data with complementary information," *Bioinformatics*, **20**, i23-i30, 2004.
- [5] H. Stark and Y. Yang, *Vector Space Projections*, John Wiley & Sons, New York, 1998.
- [6] X. C. Gan, A. W. C. Liew and H. Yan, "Blocking artifact reduction in compressed images based on edge-adaptive quadrangle meshes," *J. Visual Communications & Representation*, **14**(4):492-507, 2003.
- [7] Bø, T.H. Dysvik, B. and Jonassen, I. LSImpute: accurate estimation of missing values in microarray data with least squares method. *Nucleic Acids Res.*, **32**, e34, 2004
- [8] DeRisi, J.L., Iyer, V.R. and Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686, 1997.