Wang et al. BMC Systems Biology 2013, 7(Suppl 6):S11
http://www.biomedcentral.com/1752-0509/7/S6/S11

BMC Systems Biology

RESEARCH Open Access

Shrinkage regression-based methods for microarray missing value imputation

Hsiuying Wang[1], Chia-Chun Chiu[2], Yi-Ching Wu[1], Wei-Sheng Wu[2*]

From 24th International Conference on Genome Informatics (GIW 2013)
Singapore, Singapore. 16-18 December 2013

Abstract

Background: Missing values commonly occur in the microarray data, which usually contain more than 5% missing values with up to 90% of genes affected. Inaccurate missing value estimation results in reducing the power of downstream microarray data analyses. Many types of methods have been developed to estimate missing values. Among them, the regression-based methods are very popular and have been shown to perform better than the other types of methods in many testing microarray datasets.

Results: To further improve the performances of the regression-based methods, we propose shrinkage regression-based methods. Our methods take the advantage of the correlation structure in the microarray data and select similar genes for the target gene by Pearson correlation coefficients. Besides, our methods incorporate the least squares principle, utilize a shrinkage estimation approach to adjust the coefficients of the regression model, and then use the new coefficients to estimate missing values. Simulation results show that the proposed methods provide more accurate missing value estimation in six testing microarray datasets than the existing regression-based methods do.

Conclusions: Imputation of missing values is a very important aspect of microarray data analyses because most of the downstream analyses require a complete dataset. Therefore, exploring accurate and efficient methods for estimating missing values has become an essential issue. Since our proposed shrinkage regression-based methods can provide accurate missing value estimation, they are competitive alternatives to the existing regression-based methods.

Background

Nowadays microarray technique has become an important and useful tool in functional genomics research. This high throughput technique allows the characterization of the gene expression of the whole genome by measuring the relative transcript levels of thousands of genes in various experimental conditions or time points [1]. Microarray data analyses have been widely used to investigate various biological processes such as the cell cycle process [2-8] and the stress response [9,10].

Although the microarray technology has been developed for more than a decade, typical microarray data still contain more than 5% missing values with up to 90% of genes affected [11]. Missing values could be generated by various reasons, including technological failures, administrative error, insufficient resolution, image corruption, dust or scratches on the slide [12]. As many downstream analysis methods (such as gene clustering, disease classification and gene network reconstruction) require complete datasets, missing value estimation becomes an important pre-processing step in the microarray data analysis [11-13].

The missing values in the microarray dataset are traditionally estimated by repeating the microarray experiments or simply replacing the missing values with zero or the row average (the average expression over the experimental conditions). Because these approaches are either time-consuming or leading to serious estimation errors, more

* Correspondence: wessonwu@mail.ncku.edu.tw
[2]Department of Electrical Engineering, National Cheng Kung University, No.1 University Road, 701 Tainan, Taiwan
Full list of author information is available at the end of the article

advanced missing value imputation methods are needed to solve the missing value problems. In 2001, Troyanskaya *et al.* published the first two missing value imputation algorithms based on the k-nearest neighbors (kNNimpute) and the singular value decomposition (SVDimpute) [12]. Since then, a lot of missing value imputation methods have been proposed such as *Bayesian* principal component analysis (BPCA) [14], Gaussian mixture clustering imputation (GMCimpute) [11], conditional ordered list imputation [15], random-forest-based imputation [16] and so on.

Among the existing missing value imputation methods, the regression-based methods are very popular and contain many algorithms, including least squares imputation (LSimpute) [17], local least squares imputation (LLSimpute) [18], sequential local least squares imputation (SLLSimpute) [19], and iterated local least squares imputation (ILLSimpute) [13]. LSimpute estimates the missing values in the target gene by using a weighted average of the k estimates from the k most similar genes. Each estimate is attained by constructing a single regression model of the target gene by a similar gene. LLSimpute represents the target gene as a linear combination of k similar genes by a multiple regression model and uses the regression coefficients to estimate the missing values. SLLSimpute modifies the LLSimpute by estimating the missing values sequentially from the gene containing the fewest missing values and partially utilizing these estimated values. ILLSimpute modifies the LLSimpute by not choosing the similar genes with a fixed number k but defining the similar genes as the genes whose distances from the target gene are less than a distance threshold and then runs LLSimpute iteratively.

In this study, we focus on the regression-based methods because these methods have been shown to have better performances than the other existing methods in many testing microarray datasets [20,21]. To further improve the performance of the regression-based methods, we propose shrinkage regression-based methods which use a shrinkage estimator to replace the least square estimator for the estimation of the regression coefficients in the regression model. The shrinkage estimator such as the James-Stein estimator has been shown to dominate the least square estimator in many statistical models [22,23]. By adopting our new regression coefficients in the regression-based methods, we showed that an improvement on missing value estimation in six testing microarray datasets could be achieved.

Methods

In this study, we propose using the well-known shrinkage estimation approach to improve three existing regression-based methods (LLSimpute [18], SLLSimpute [19], and ILLSimpute [13]) for missing value estimation. We call our proposed methods the shrinkage regression-based methods (see Figure 1). In the following subsections, we first introduce the shrinkage estimation approach and then describe the proposed shrinkage LLSimpute, shrinkage SLLSimpute, and shrinkage ILLSimpute.

Shrinkage estimation approach

One of the shrinkage estimators, the James-Stein estimator, for the normal distribution is introduced here. Suppose that $Y_1, Y_2, ..., Y_k$ are independent normal random variables and these k random variables all have a common known variance, but their means are unknown and different. Let $Y_i \sim N(\theta_i, \sigma^2)$ and $\mathbf{Y} = (Y_1, ..., Y_k)$. Then we have $\mathbf{Y} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ and $\mathbf{I}$ is a $k \times k$ identity matrix. Let $\mathbf{d}(\mathbf{Y}) = (d_1(\mathbf{Y}), ..., d_k(\mathbf{Y}))$ be an estimator of $\boldsymbol{\theta}$. Under the squared error loss function

$$L(\theta - \mathbf{d}(\mathbf{Y})) = \sum_{i=1}^{k} (\theta_i - d_i(\mathbf{Y}))^2 = ||\theta - \mathbf{d}(\mathbf{Y})||^2, \quad (1)$$

we are interested in finding estimators of $\boldsymbol{\theta}$ such that the mean squared error $E_{\mathbf{Y}}[L(\boldsymbol{\theta}, \mathbf{d}(\mathbf{Y}))]$ is minimized. An intuitive estimator of $\boldsymbol{\theta}$ is $\mathbf{Y}$ (i.e. $\hat{\theta}_i = Y_i,\ i = 1, \ldots, k$). However, Stein [22] showed that when $k \geq 3$, there exists other estimators with smaller mean squared error than the intuitive estimator $\mathbf{Y}$. For $k \geq 3$, under the squared error loss, the intuitive estimator $\mathbf{Y}$ is dominated by the estimator

$$\hat{\theta}^{JS} = \left(1 - \frac{k-2}{S_Y^2}\right)\mathbf{Y}, \quad (2)$$

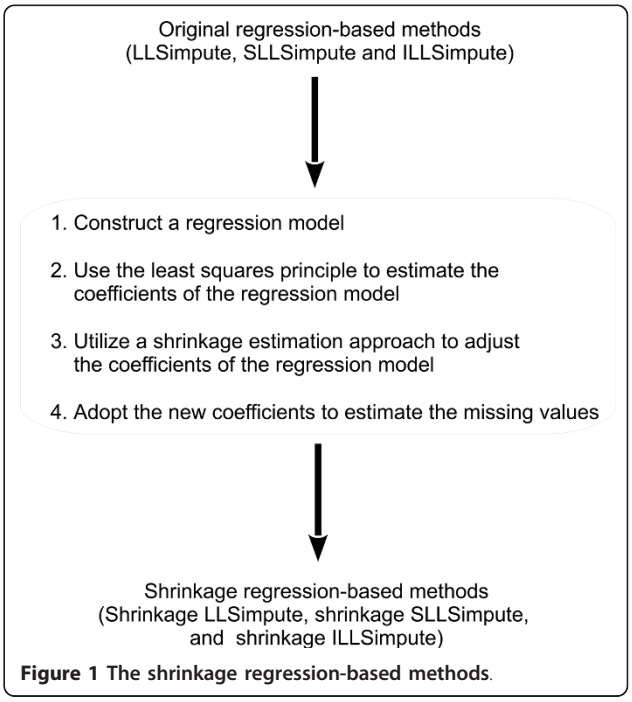

Figure 1 The shrinkage regression-based methods.

where $S_Y^2 = \sum_{i=1}^{k} Y_i^2$ [23]. The estimator in (2) is called the James-Stein estimator in the literature [23]. With the form in (2), the James-Stein estimator of θ_i is

$$\hat{\theta}_i^{JS} = \left(1 - \frac{k-2}{S_Y^2}\right) Y_i. \quad (3)$$

It is worth noting that the estimator of θ_i in (3) depends on not only the random variable Y_i, but also the other variables Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_k because of the term S_Y^2. On the contrary, the intuitive estimator $\hat{\theta}_i = Y_i$ does not use the other variables Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_k but only uses Y_i to estimate θ_i. It has been shown that estimators using other variables' information provide more accurate estimation for θ than the intuitive estimator does [22]. In fact, except for the estimator in (3), the estimators of the form

$$\hat{\theta}_i^{JS} = \left(1 - \frac{c}{S_Y^2}\right) Y_i \quad (4)$$

all have uniformly smaller mean squared error than the intuitive estimator Y_i, for $k \geq 3$ and $0 < c < 2(k-2)$. Among all the estimators of the form in (4), the estimator in (3) has the minimized mean squared error. The shrinkage estimation approach has also been shown to have good performance in interval estimation [24,25]. Based on the James-Stein estimator in (3), we developed shrinkage regression-based imputation methods.

Notations

In a typical microarray data matrix, the rows are the genes under investigation and the columns are the experimental conditions or time points. The microarray data matrix is obtained by performing a series of experiments on the same set of genes. We use $\mathbf{G} \in \mathbb{R}^{m \times n}$ to represent a microarray data matrix with m genes and n experiments, and assume $m \gg n$ which is true for microarray data. In the matrix $\mathbf{G}$, a row $\mathrm{g}_i^T \in \mathbb{R}^{1 \times n}$ represents the expressions of the ith gene in n experiments:

$$\mathrm{G} = \begin{pmatrix} \mathrm{g}_1^T \\ \vdots \\ \mathrm{g}_m^T \end{pmatrix} \in \mathbb{R}^{m \times n} \quad (5)$$

where g_i^T denotes the transpose of a column vector $\mathbf{g_i}$. If there is a missing value in the lth position of the ith gene, we denote it as α, i.e. $G_{i,l} = g_{il} = \alpha$.

Shrinkage local least squares imputation (Shrinkage LLSimpute)

In the LLSimpute method [18], a target gene with missing values is represented as a linear combination of k similar genes. Rather than using all genes in the dataset, only k genes with high similarity to the target gene are used. The procedure of selecting k similar genes is as follows. Suppose that the target gene is the first gene and has a missing value α in the first position, i.e. $\alpha = g_{11}$ in the matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$. The Pearson correlation coefficient is used to find the k similar genes. These k similar genes are called the k-nearest neighbor genes, which have the k largest absolute values of the Pearson correlation coefficients. The Pearson correlation coefficient r_{1j} between the target gene and the jth gene is defined as

$$r_{1j} = \frac{1}{n-2} \sum_{t=2}^{n} \left(\frac{g_{1t} - \bar{g}_1}{\sigma_1}\right)\left(\frac{g_{jt} - \bar{g}_j}{\sigma_j}\right) \quad (6)$$

where $\bar{g}_j$ and σ_j denote the average and the sample standard deviation of the vector $(g_{j2}, ..., g_{jn})$. When computing the correlation coefficients, g_{j1} is not used because it corresponds to the position of the missing value in the target gene. Based on these selected k-nearest neighbor genes, a matrix $\mathbf{A} \in \mathbb{R}^{k \times (n-1)}$ and two vectors $\mathbf{b} \in \mathbb{R}^{k \times 1}$ and $\mathrm{w} \in \mathbb{R}^{(n-1) \times 1}$ can be formed as follows

$$\begin{pmatrix} \mathrm{g}_1^T \\ \mathrm{g}_{S_1}^T \\ \vdots \\ \mathrm{g}_{S_k}^T \end{pmatrix} = \begin{pmatrix} \alpha & \mathrm{w}^T \\ \mathrm{b} & \mathrm{A} \end{pmatrix} = \begin{pmatrix} \alpha & w_1 & w_2 & & w_{n-1} \\ b_1 & A_{1,1} & A_{1,2} & & A_{1,n-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ b_k & A_{k,1} & A_{k,2} & & A_{k,n-1} \end{pmatrix}$$

where $\alpha$is the missing value in the target gene $\boldsymbol{g_1}$ and $\mathbf{g_{s_1}}, \cdots, \mathbf{g_{s_k}}$ are the k-nearest nieghbor genes of the target gene g_1. Each row of matrix $\mathbf{A}$ consists of the last $n-1$ elements of one k-nearest neighbor gene $\mathbf{g_{s_i}}$, $1 \leq i \leq k$. The elements of the vector $\mathbf{b}$ comprise of the first elements of all these k-nearest neighbor genes and the elements of the vector $\mathbf{w}$ are the last $n-1$ elements of the target gene $\mathbf{g}_1$. With the matrix $\mathbf{A}$, and the vectors $\mathbf{b}$ and $\mathbf{w}$, the least squares problem is formulated in LLSimpute as

$$\min_{\mathrm{x}} ||\mathbf{A}^T\mathrm{x} - \mathbf{w}||_2. \quad (7)$$

Solving the above problem, the least square regression coefficients $\hat{\mathbf{x}} \in \mathbb{R}^{k \times 1}$ are acquired as

$$\hat{\mathrm{x}} \triangleq (\hat{x}_1, \hat{x}_2, \ \ldots, \ \hat{x}_k)^T = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}\mathrm{w}. \quad (8)$$

In the LLSimpute, the missing value is then estimated by

$$\alpha = \mathbf{b}^T\hat{\mathrm{x}} = \hat{x}_1 b_1 + \hat{x}_2 b_2 + \ldots + \hat{x}_k b_k. \quad (9)$$

In this study, we want to improve the performance of LLSimpute by adjusting the regression coefficients in (8). Our shrinkage LLSimpute associates the LLSimpute method with the shrinkage estimator to impute the missing values. Our method replaces the regression coefficient

estimators $\hat{\mathbf{x}}$ in (8) by the shrinkage estimator, and then use the new estimator to estimate the missing value α in (9). However, we found that applying the existing shrinkage estimator in (3) did not always improve the performance of LLSimpute. Therefore, we tested different forms of the shrinkage coefficient estimators and conceived a feasible coefficient estimator to improve the LLSimpute method. We proposed using the shrinkage regression coefficients

$$\hat{x}_i^{JS} = \left(1 - \frac{(k-2)\sigma^2}{\tilde{n}S^2}\right)\hat{x}_i \tag{10}$$

to replace the conventional coefficients in (8), where σ^2 is the variance of the coefficients $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k)$, S is the norm of the coefficients (i.e. $S^2 = \sum_{i=1}^{k} \hat{x}_i^2$), k is the row number of the matrix $\mathbf{A}$ and $\tilde{n}$ is the column number of the matrix $\mathbf{A}$, which equals n - 1 in this case. Finally, the missing value is estimated as

$$\alpha = \mathbf{b}^T\hat{\mathbf{x}}^{JS} = \hat{x}_1^{JS}b_1 + \hat{x}_2^{JS}b_2 + \ldots + \hat{x}_k^{JS}b_k \tag{11}$$

where $\hat{\mathbf{x}}^{JS} = (\hat{x}_1^{JS}, \ldots, \hat{x}_k^{JS})^T$.

Shrinkage sequential local least squares imputation (Shrinkage SLLSimpute)

In the LLSimpute, it does not use the information of genes with missing values since the existence of missing values hinders the use of the other observed values of that gene. In the SLLSimpute method, it estimates the missing values sequentially from the gene containing the fewest missing values and partially utilizes these estimated values. The details of SLLSimpute [19] is described as follow. First, the microarray matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$ is divided into two submatrices: a complete matrix $\mathbf{G}_1 \in \mathbb{R}^{m_1 \times n}$ consisting of genes without missing values and an incomplete matrix $\mathbf{G}_2 \in \mathbb{R}^{(m-m_1) \times n}$ consisting of genes with missing values. In the incomplete matrix $\mathbf{G}_2$, the genes are sorted by their missing rates. The first gene has the smallest missing rate and the last gene has the largest missing rate. The missing rate is calculated by

$$r_i = \frac{c_i}{n}, \tag{12}$$

where c_i is the number of missing values in i-th gene. The imputation is executed sequentially from the first gene of $\mathbf{G}_2$. That is, the first gene of $\mathbf{G}_2$ which has the smallest missing rate is selected as the target gene firstly. Then LLSimpute is applied to estimate the missing values in the target gene by finding the k-nearest neighbour genes from the complete matrix $\mathbf{G}_1$ and then using the formula in (9) to estimate the missing values. After filling all the missing values in the target gene, it is moved to $\mathbf{G}_1$. Then the second gene of $\mathbf{G}_2$ is selected as the target gene and repeat the same process again. By moving the genes whose missing values have been imputed to the complete matrix, the previous target genes with imputed values can be utilized for the missing value estimation of the following target gene. However, too many missing values in a gene will result in big estimation error and reusing a gene with too many imputed values will reduce the imputation performance. Therefore, only the genes with missing rates less than a threshold r_0 are reused, where r_0 is set as the average missing rate of all genes containing missing values, i.e.,

$$r_0 = \frac{\sum_{i=1}^{m-m_1} c_i}{(m-m_1) \times n} \tag{13}$$

By a similar argument as for the shrinkage LLSimpute, we apply the shrinkage estimator to SLLSimpute. The shrinkage SLLSimpute adjusts the coefficients of the regression model by the formula in (10) and use the formula in (11) to estimate the missing values.

Shrinkage iterated local least squares imputation (Shrinkage ILLSimpute)

LLSimpute and SLLSimpute methods select k-nearest neighbor genes for a target gene, where k is a fixed number. However, in the ILLSimpute method [13], it does not fix the number of similar genes selected. Alternatively, it defines the similar genes as the genes whose distances to the target genes are less than a distance threshold δ. The rationale of using a distance threshold rather than using a fixed number of similar genes is that some of the k-nearest neighbor genes are already far away from the target gene and are not very similar to the target gene.

The procedure of ILLSimpute is as follows. In the first iteration, missing values of each target gene are filled with the row average. Then a distance threshold δ is used to select the similar genes of each target gene. Finally, LLSimpute method is used to estimate the missing values of each target gene. In the later iteration, ILLSimpute method uses the imputed results from the previous iteration to reselect the similar genes of each target gene (using the same distance threshold) and applies LLSimpute method to re-estimate the missing values.

By a similar argument as for the shrinkage LLSimpute, we apply the shrinkage estimator to ILLSimpute. The shrinkage ILLSimpute adjusts the coefficients of the regression model by the formula in (10) and use the formula in (11) to estimate the missing values.

Results and Discussion

We conducted several experiments to compare the performances of our shrinkage regression-based methods and the original regression-based methods under different scenarios. In the first subsection, we introduce the

Table 1 Benchmark datasets.

Name	Dimension of original datasets	Dimension of reduced complete datasets	Time series data	Ref.
Ogawa	6263 × 8	3069 × 8	N	[26]
BohenSH	2364 × 24	623 × 24	N	[27]
Lymphoma	4096 × 96	854 × 96	N	[28]
Brauer05	6133 × 20	706 × 20	Y	[29]
Shapira04A	4771 × 23	2970 × 23	Y	[30]
Shapira04B	4771 × 14	3340 × 14	Y	[30]

benchmark datasets. In the second subsection, we describe how we measure the performance of various imputation methods. In the following three subsections, we report the comparison results for different number of similar genes used, different missing rates, and different noise levels. Finally, we further compare the performances of our shrinkage regressioni-based methods and three existing non-regression-based methods.

Datasets

Considering the effects of dataset selection and types of microarray experiments on the performance of an imputation method, six representative datasets (three non-time series and three time series) were used in our simulations. They were Ogawa's data from the study of phosphophate accumulation and poly-phosphophate metabolism (denoted as Ogawa, non-time series) [26], Bohen's follicular lymphomas data (denoted as BohenSH, non-time series) [27], the data from a lymphoma study (denoted as Lymphoma, non-time series) [28], the data from Brauer's experiments which studied the physiological response to glucose limitation in batch and steady-state cultures of yeasts (denoted as Brauer05, time series) [29], and Shapira's oxidative stress data (denoted as Shapira04A and Shapira04B, time series) [30]. We divided Shapira's data into two datasets because the authors used one kind of oxidative chemical in the experiment in Shapira04A, but they used another kind of oxidative chemical in the experiment in Shapira04B. The six microarray datasets were used as benchmark datasets in numerical experiments to compare the performances of our shrinkage regression-based methods and the original regression-based methods. Each dataset was processed by deleting the genes with missing values to generate a complete data matrix, and the details of these datasets were listed in Table 1.

The performance measure

A common criterion used to compare the performances of different imputation methods is the normalized root mean squared error (NRMSE) [11-13,17-19]. From a microarray dataset, we can obtain an original data matrix $\mathbf{M}_0$ with m genes and n experiments, and then we can construct a complete matrix $\mathbf{M}_1 \in \mathbb{R}^{m_1 \times n}(m_1 \leq m)$ by deleting the genes with missing values. After the complete data matrix $\mathbf{M}_1$ is established, we randomly select a specific percentage of the elements of $\mathbf{M}_1$ and regard these elements as missing values. Then we estimate the missing values using various imputation methods and compare their performances using NRMSE which is shown below:

$$NRMSE = \frac{\sqrt{mean[(\mathrm{y}_{\mathrm{guess}} - \mathrm{y}_{\mathrm{ans}})^2]}}{std(\mathrm{y}_{\mathrm{ans}})} \tag{14}$$

where $\mathrm{y}_{\mathrm{guess}}$ and $\mathrm{y}_{\mathrm{ans}}$ are vectors whose elements are the estimated values by an imputation method and the known answers for all missing entries, respectively.

Performance comparison for different *k* values

A parameter k, the number of similar genes used, has to be determined before using two regression-based methods (LLSimpute and SLLSimpute). Since the performance of both algorithms is known to be affected by the k value used and different microarray datasets may have different optimal k values [18,19], we tested several possible k values (50, 100, 150, 200, 250 and 300) on six benchmark datasets. Table 2 listed the optimal k values for LLSimpute and SLLSimpute on each of the six benchmark datasets. Another regression-based method (ILLSimpute) does not have the parameter k and therefore was not considered in this numerical experiment.

For each of the six benchmark dataset, we also compared the performances of the proposed shrinkage regression-based methods and the original regression-based methods for several possible k values (50, 100, 150, 200, 250 and 300). In our numerical experiments, missing rate for each benchmark dataset was set to be 5%. Namely, for each dataset, we randomly removed 5% entries of the complete matrix to generate a matrix with missing values, and then estimated the missing values using the shrinkage and the original regression-based methods. The same procedure was run for five independent rounds and the average NRMSE of these five

Table 2 The optimal *k* value for each benchmark dataset.

Algorithms\Datasets	Ogawa	BohenSH	Lymphoma	Brauer05	Shapira04A	Shapira04B
LLS	100	250	300	300	250	200
SLLS	150	300	250	300	250	200

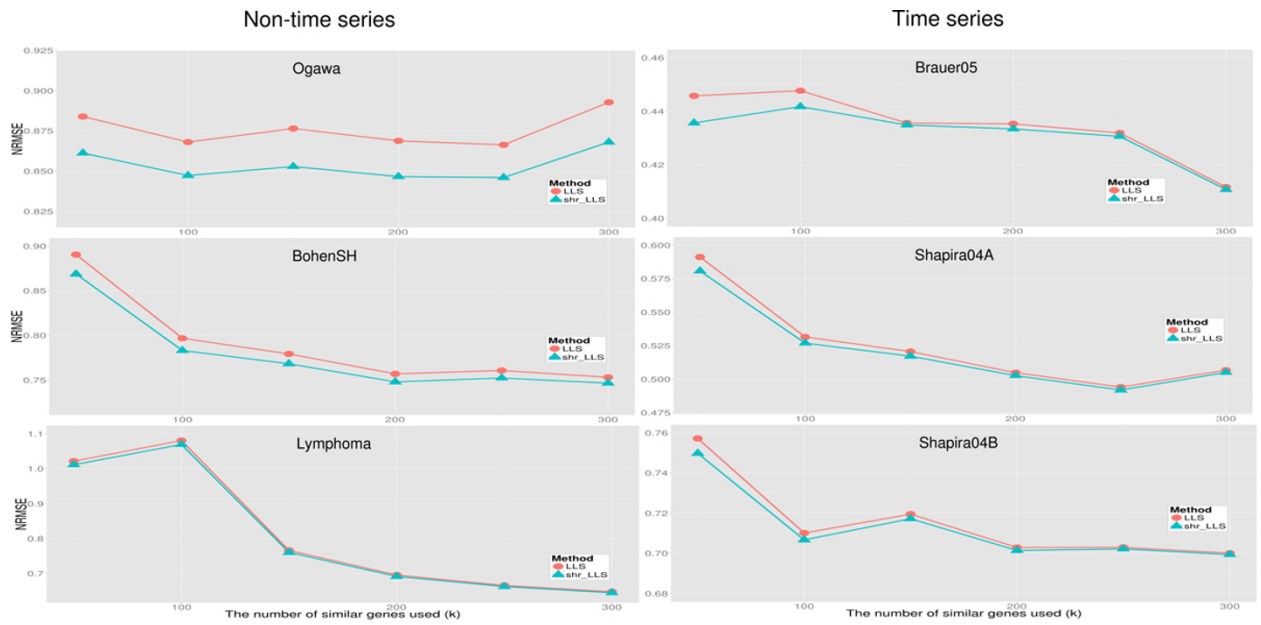

Figure 2 Performance comparison between shrinkage LLS (shr_LLS) and LLS for different *k* values.

simulations was used to compare the performances of different imputation methods.

As shown in Figure 2, the proposed shrinkage LLSimpute outperforms LLSimpute for all *k* values and all benchmark datasets. Similarly, the proposed shrinkage SLLSimpute outperforms SLLSimpute for all *k* values and all benchmark datasets (see Figure 3). The simulation results suggest that utilizing a shrinkage estimation approach to adjust the coefficients of the regression model can improve the performances of the original regression-based methods.

Performance comparison for different missing rates

In real applications, different microarray data may have different missing rates to be imputed. It is informative to know how an imputation method performs for different

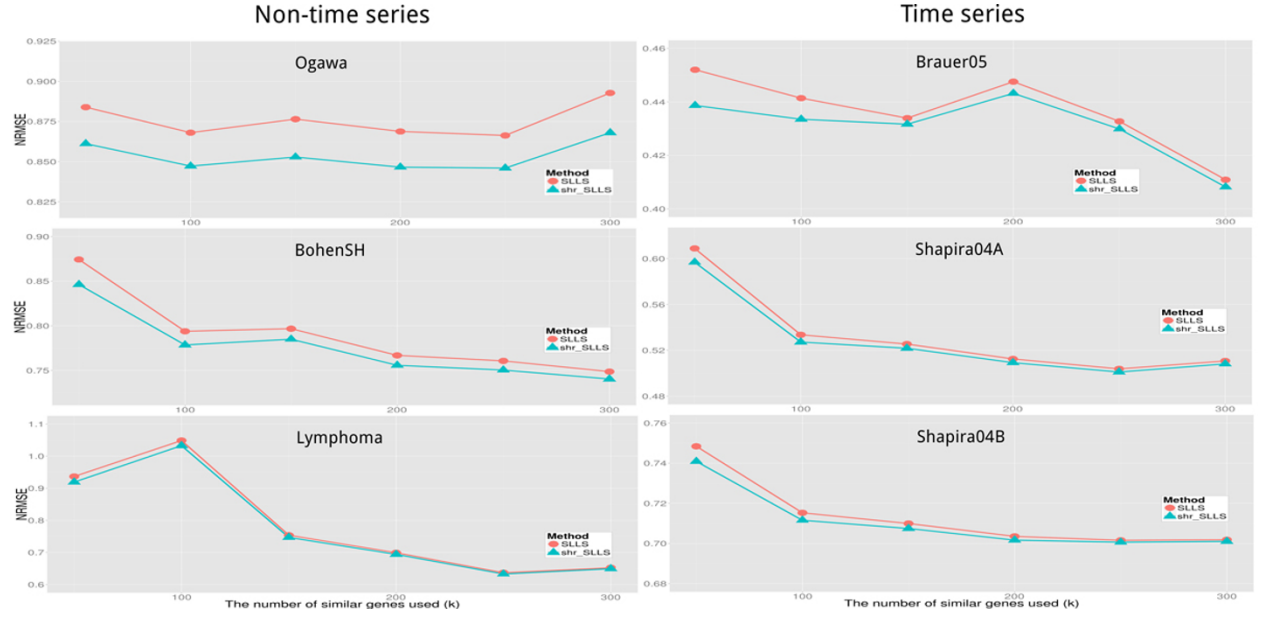

Figure 3 Performance comparison between shrinkage SLLS (shr_SLLS) and SLLS for different *K* values.

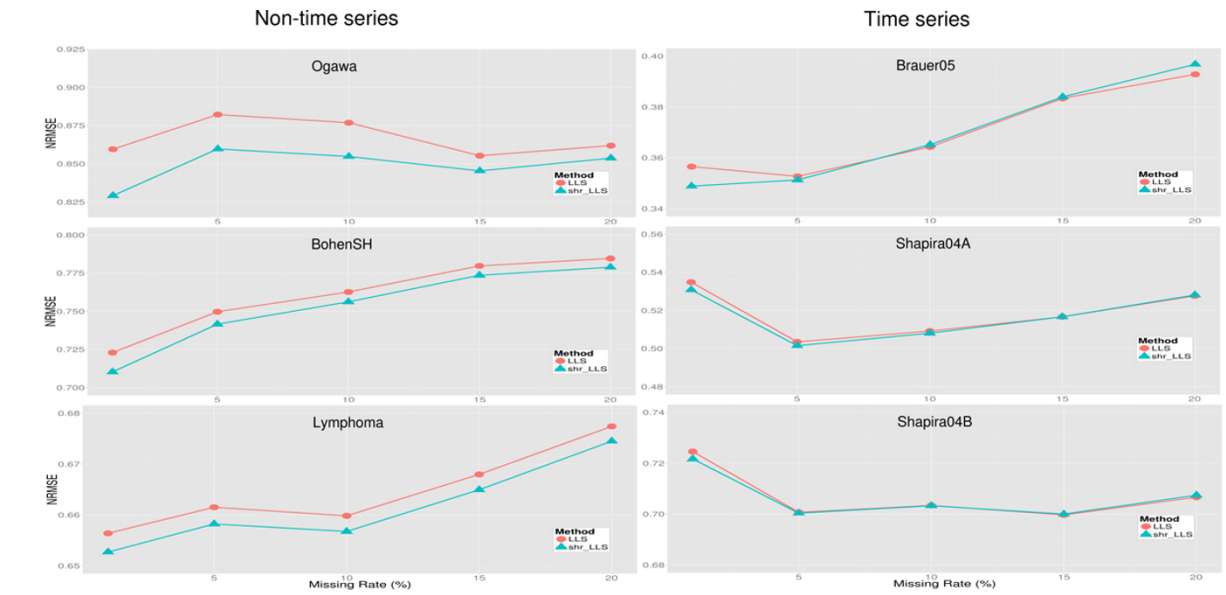

Figure 4 Performance comparison between shrinkage LLS (shr_LLS) and LLS for different missing rates.

missing rates. Therefore, we compared the performances of the shrinkage regression-based methods and the original regression-based methods on the microarray data with different missing rates (1%, 5%, 10%, 15% and 20%). Namely, for each of the six benchmark dataset, we randomly removed x% (x = 1, 5, 10, 15 or 20) entries of the complete matrix to generate a matrix with missing values, and then estimated the missing values using the shrinkage and the original regression-based methods. The same procedure was run for five independent rounds and the average NRMSE of these five simulations was used to compare the performances of different imputation methods. Note that the optimal k value used for each benchmark dataset was listed in Table 2.

Figure 4 shows that the proposed shrinkage LLSimpute outperforms LLSimpute for all missing rates and all benchmark datasets. Figure 5 shows that the proposed shrinkage SLLSimpute outperforms SLLSimpute for all

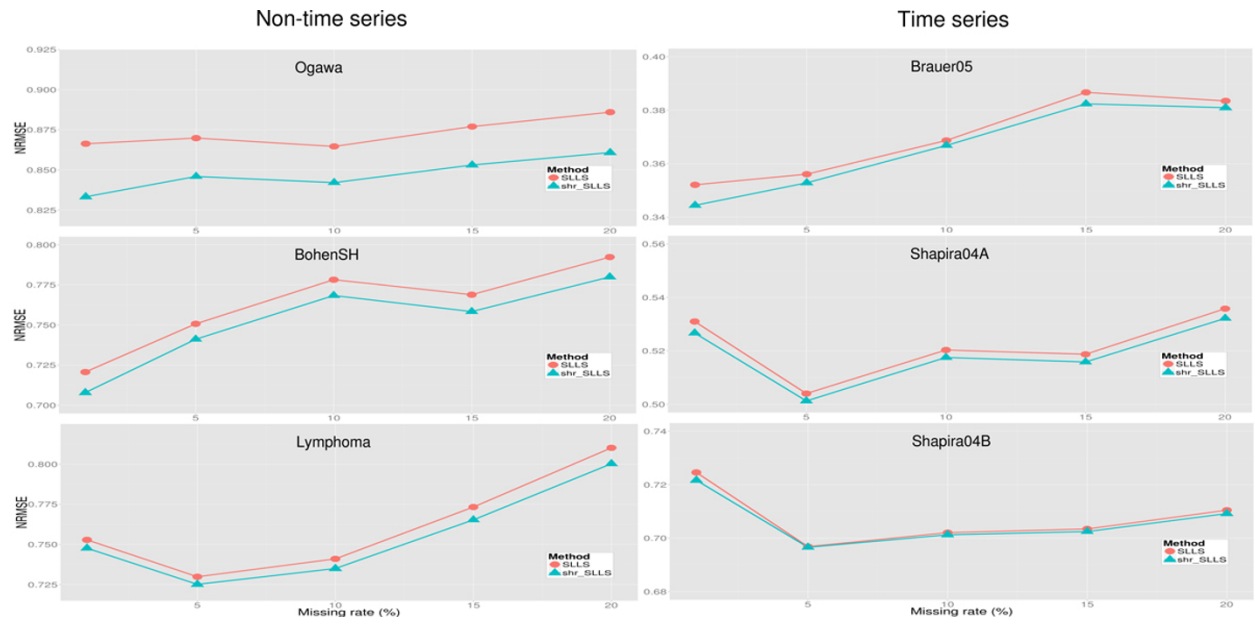

Figure 5 Performance comparison between shrinkage SLLS (shr_SLLS) and SLLS for different missing rates.

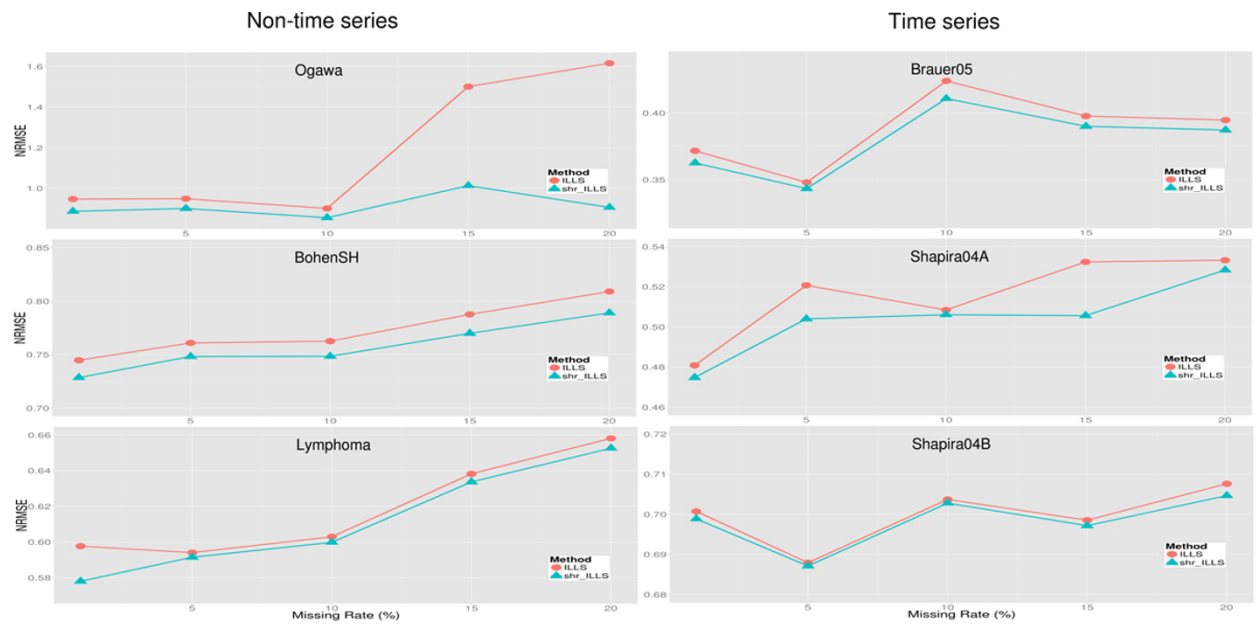

Figure 6 Performance comparison between shrinkage ILLS (shr_ILLS) and ILLS for different missing rates.

missing rates and all benchmark datasets. Figure 6 shows that the proposed shrinkage ILLSimpute outperforms ILLSimpute for all missing rates and all benchmark datasets. The simulation results suggest that utilizing a shrinkage estimation approach to adjust the coefficients of the regression model can improve the performances of the original regression-based methods.

Performance comparison for different noise levels

In real applications, different microarray data may contain different levels of noises. It is informative to know how an imputation method performs for different levels of noises inherent in the microarray data. Therefore, we compared the performances of the shrinkage regression-based methods and the original regression-based methods on the microarray data with different noise levels. For each of the six benchmark dataset, we added Gaussian noises with different levels into the data. The magnitudes of the noises were set in terms of the standard deviations ranging from 0 to 0.25 with a step size 0.05. In our numerical experiments, missing rate for each benchmark dataset was set to be 5% and the optimal k value used for each benchmark dataset was listed in Table 2. Namely, for each dataset (after adding Gaussian noises into the data), we randomly removed 5% entries of the complete matrix to generate a matrix with missing values, and then estimated the missing values using the shrinkage and the original regression-based methods. The same procedure was run for five independent rounds and the average NRMSE of these five simulations was used to compare the performance of different imputation methods.

Figure 7 shows that the proposed shrinkage LLSimpute outperforms LLSimpute for all noise levels and all benchmark datasets. Figure 8 shows that the proposed shrinkage SLLSimpute outperforms SLLSimpute for all noise levels and all benchmark datasets. Figure 9 shows that the proposed shrinkage ILLSimpute outperforms ILLSimpute for all noise levels and all benchmark datasets. The simulation results suggest that utilizing a shrinkage estimation approach to adjust the coefficients of the regression model can improve the performances of the original regression-based methods.

Performance comparison with three existing non-regression-based methods

We have shown that our shrinkage regression-based methods perform better than the existing regression-based methods. Still, it would be interesting to know whether our shrinkage regression-based methods provide more accurate missing value imputation than the existing non-regression-based methods do. Therefore, we compared the performances of our shrinkage regression-based methods and three existing non-regression-based methods (kNNimpute [12], SVDimpute [12], and BPCA [14]) on the six benchmark microarray datasets. As shown in Figures 10, 11, 12, the proposed shrinkage regression-based methods outperform these three existing non-regression-based methods for almost all missing rates and all benchmark datasets. Taken together, our shrinkage regression-based methods are competitive alternatives to the existing methods for microarray missing value imputation.

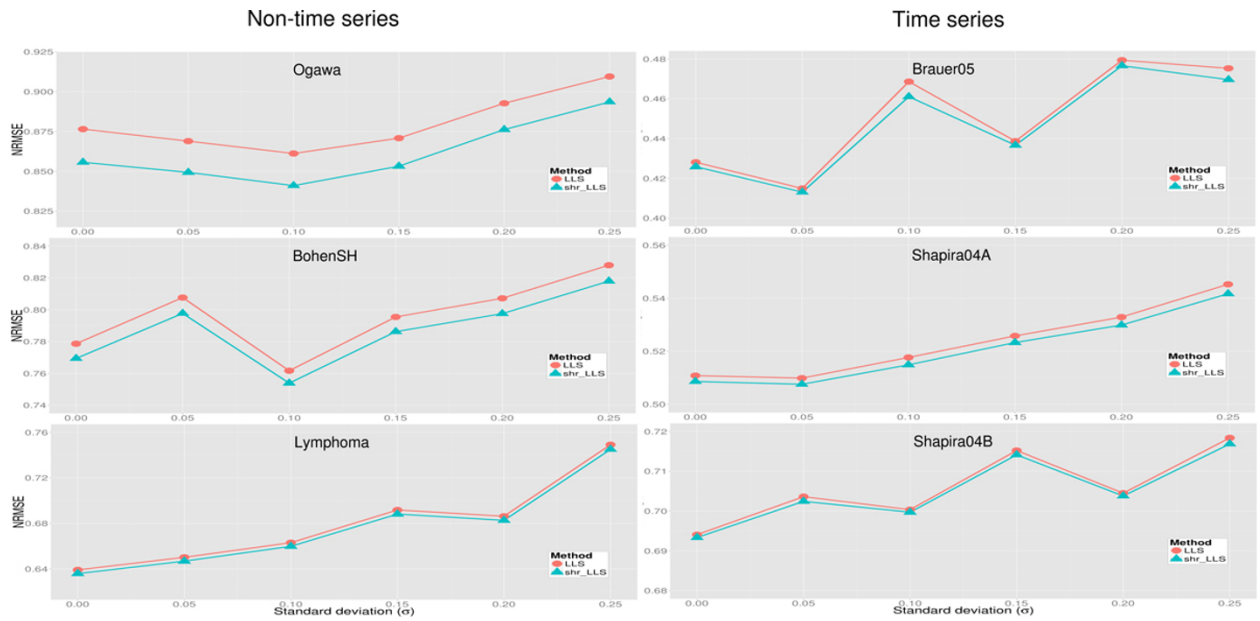

Figure 7 Performance comparison between shrinkage LLS (shr_LLS) and LLS for different noise levels.

Conclusions

Imputation of missing values is a very important aspect of microarray data analyses because most of downstream analyses require a complete dataset. Therefore, exploring accurate and efficient methods for estimating missing values has become an essential issue. In this study, regression-based methods associated with a shrinkage estimation approach are proposed to estimate missing values in the microarray data. Our methods take the advantage of the correlation structure existing in the microarray data and select similar genes for the target gene by Pearson correlation coefficients. Besides, our methods incorporate the least squares principle, utilize a shrinkage estimation approach to adjust the coefficients of the regression model, and apply the new coefficients of the regression model to estimate missing values.

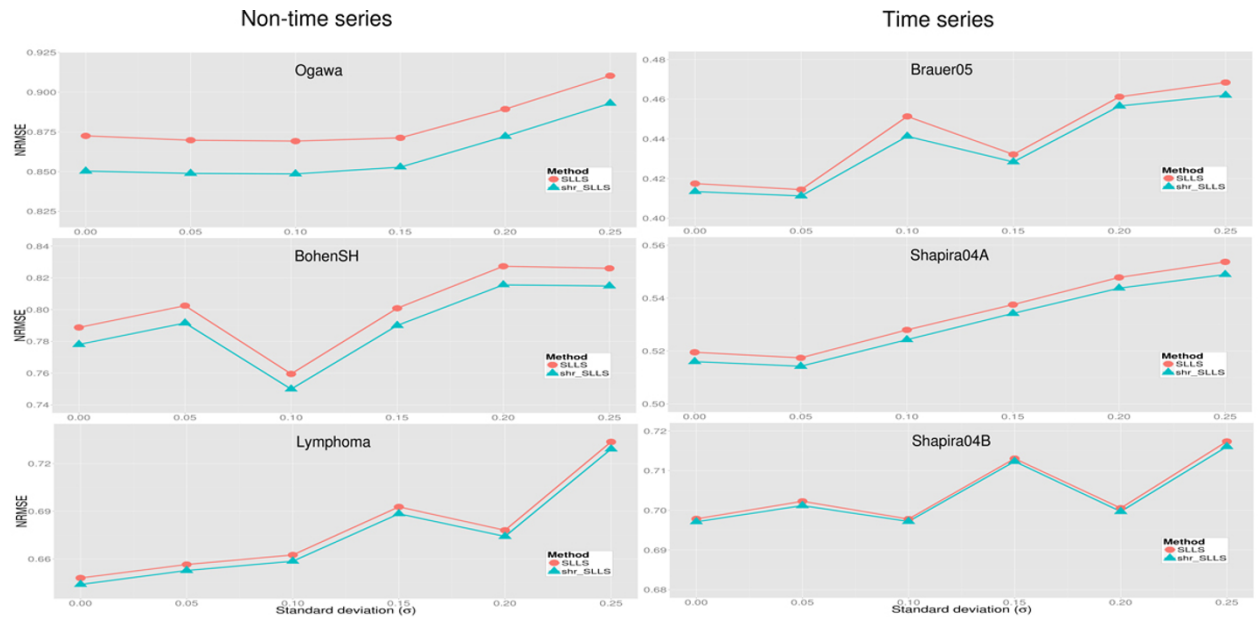

Figure 8 Performance comparison between shrinkage SLLS (shr_SLLS) and SLLS for different noise levels.

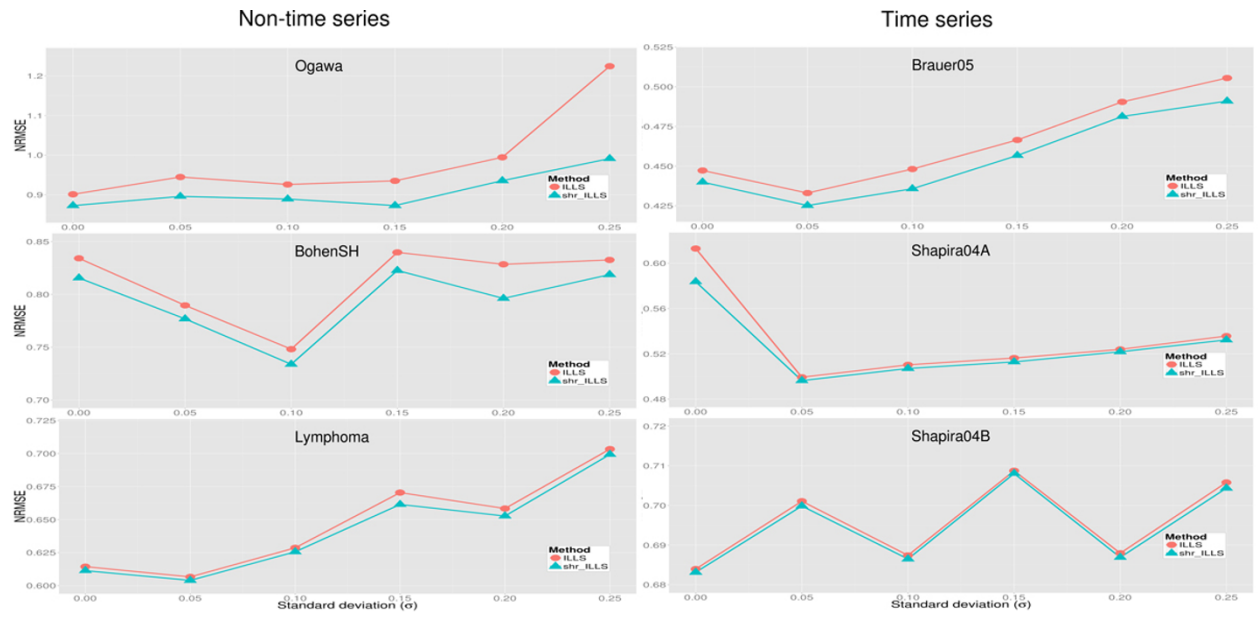

Figure 9 Performance comparison between shrinkage ILLS (shr_ILLS) and ILLS for different noise levels.

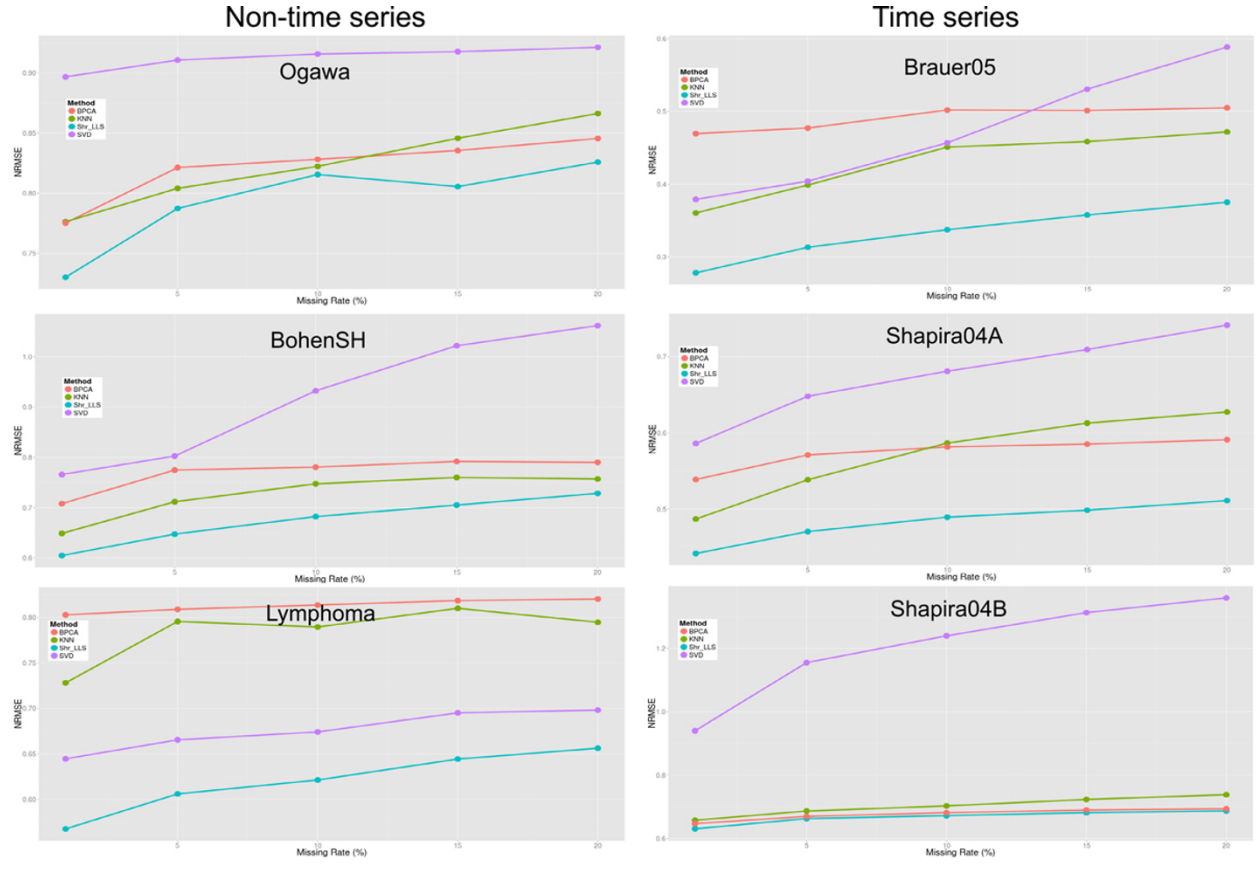

Figure 10 Performance comparison between shrinkage LLS (shr_LLS) and three non-regression-based methods for different missing rates.

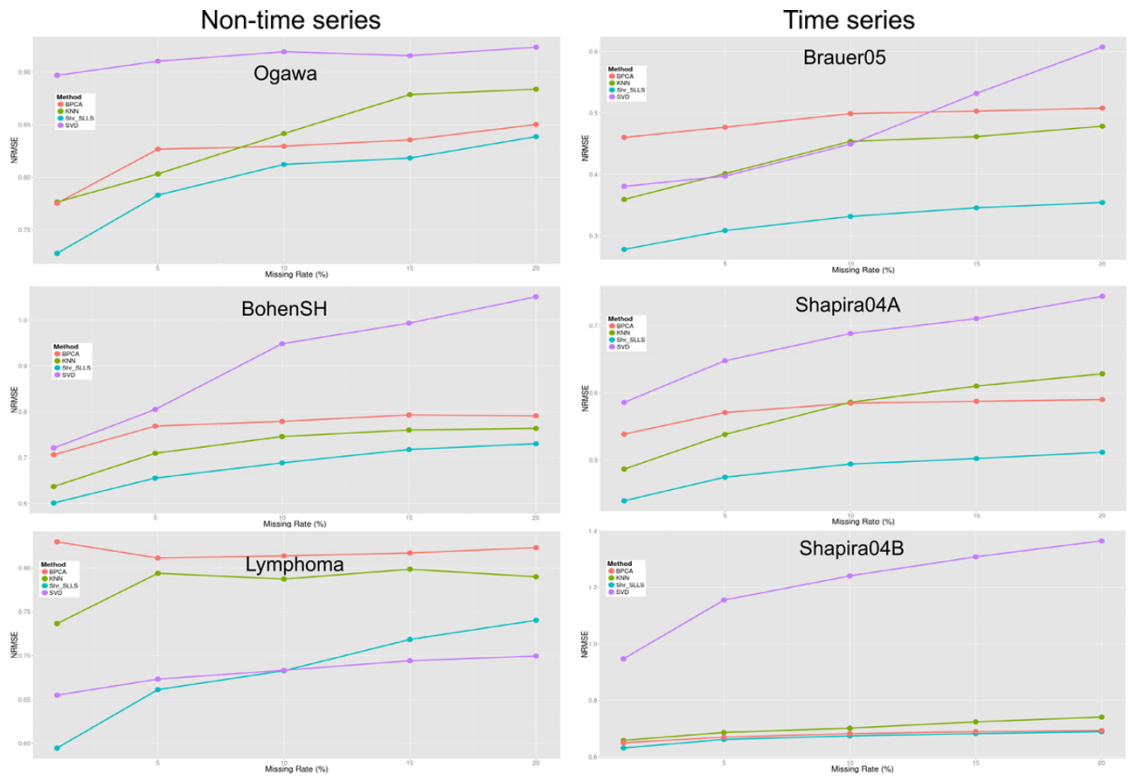

Figure 11 Performance comparison between shrinkage SLLS (shr_SLLS) and three non-regression-based methods for different missing rates.

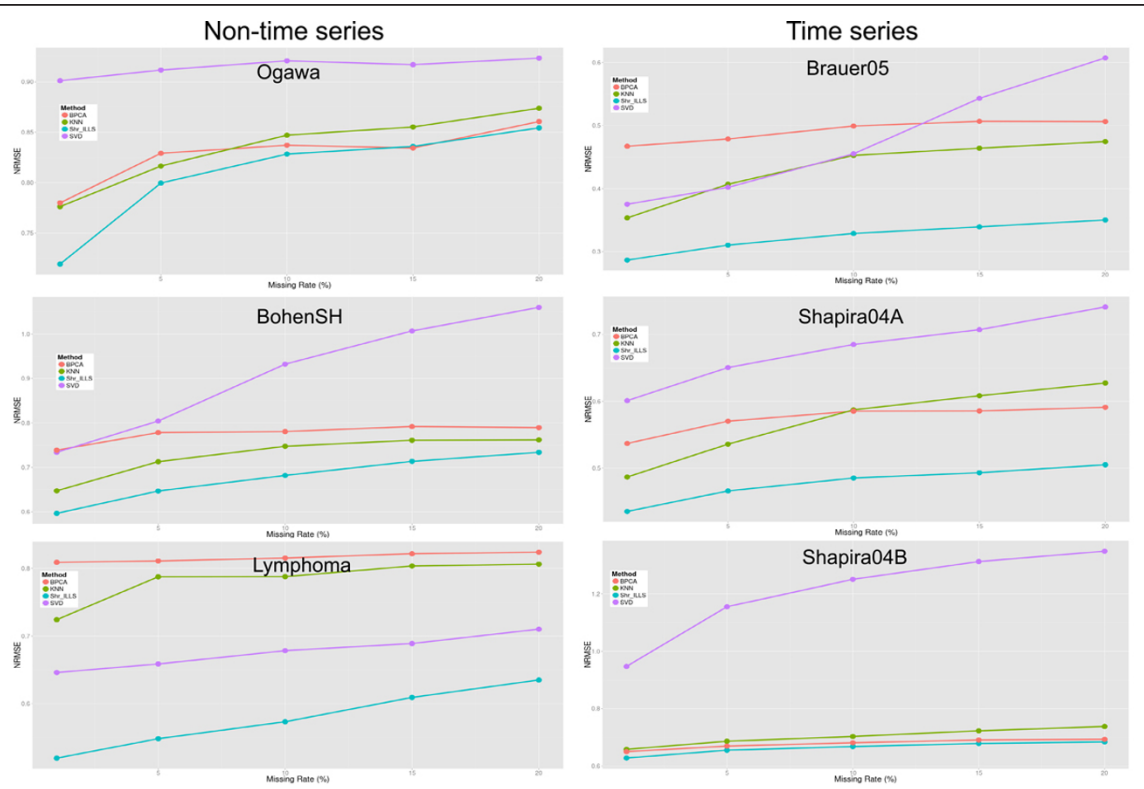

Figure 12 Performance comparison between shrinkage ILLS (shr_ILLS) and three non-regression-based methods for different missing rates.

Simulation results show that the proposed shrinkage regression-based methods provide more accurate missing value estimation for various types of datasets than the original regression-based methods do. Since our proposed methods can be applied to modify any kind of regression-based methods and can provide accurate missing value estimation, they are competitive alternatives to the existing regression-based methods.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
WSW conceived the research topic and provided essential guidance. HW developed the alogirthm. CCC did all the simulations. HW, CCC, YCW, and WSW wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements
This study was supported by the National Cheng Kung University and Taiwan National Science Council NSC 99-2628-B-006-015-MY3 and NSC 101-2118-M-009-006-MY2.

Declarations
The full funding for the publication fee came from Taiwan National Science Council and College of Electrical Engineering and Computer Science, National Cheng Kung University.
This article has been published as part of *BMC Systems Biology* Volume 7 Supplement 6, 2013: Selected articles from the 24th International Conference on Genome Informatics (GIW2013). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/7/S6.

Authors' details
[1]Institute of Statistics, National Chiao Tung University, 1001 University Road, 300 Hsinchu, Taiwan (R. O. C). [2]Department of Electrical Engineering, National Cheng Kung University, No.1 University Road, 701 Tainan, Taiwan.

Published: 13 December 2013

References

1. Schena M, Shalon D, Davis R, Brown P: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467-470.
2. Wu W, Li W, Chen B: **Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle.** *BMC Bioinformatics* 2006, **7**:421.
3. Rowicka M, Kudlicki A, Tu B, Otwinowski Z: **High-resolution timing of cell cycle-regulated gene expression.** *Proc Natl Acad Sci USA* 2007, **104**:16892-16897.
4. Wu W, Li W, Chen B: **Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data.** *BMC Bioinformatics* 2007, **8**:188.
5. Futschik M, Herzel H: **Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis.** *Bioinformatics* 2008, **24**:1063-1069.
6. Wu W, Li W: **Systematic identification of yeast cell cycle transcription factors using multiple data sources.** *BMC Bioinformatics* 2008, **9**:522.
7. Siegal-Gaskins D, Ash J, Crosson S: **Model-based deconvolution of cell cycle time-series data reveals gene expression details at high resolution.** *PLoS Comput Biol* 2009, **5**:e1000460.
8. Wang H, Wang Y, Wu W: **Yeast cell cycle transcription factors identification by variable selection criteria.** *Gene* 2011, **485**:172-176.
9. Gasch A, Spellman P, Kao C, Carmel-Harel O, Eisen M, Storz G, Botstein D, Brown P: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11**:4241-4257.
10. Wu W, Li W: **Identifying gene regulatory modules of heat shock response in yeast.** *BMC Genomics* 2008, **9**:439.
11. Ouyang M, Welsh W, Georgopoulos P: **Gaussian mixture clustering and imputation of microarray data.** *Bioinformatics* 2004, **20**:917-923.
12. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17**:520-525.
13. Cai Z, Heydari M, Lin G: **Iterated local least squares microarray missing value imputation.** *J Bioinform Comput Biol* 2006, **4**:935-957.
14. Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S: **A Bayesian missing value estimation method for gene expression profile data.** *Bioinformatics* 2003, **19**:2088-2096.
15. Yu T, Peng H, Sun W: **Incorporating nonlinear relationships in microarray missing value imputation.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**:723-731.
16. Stekhoven D, Bühlmann P: **MissForest-non-parametric missing value imputation for mixed-type data.** *Bioinformatics* 2012, **28**:112-118.
17. Bø T, Dysvik B, Jonassen I: **LSimpute: accurate estimation of missing values in microarray data with least squares methods.** *Nucleic Acids Res* 2004, **32**:e34.
18. Kim H, Golub G, Park H: **Missing value estimation for DNA microarray gene expression data: local least squares imputation.** *Bioinformatics* 2005, **21**:187-198.
19. Zhang X, Song X, Wang H, Zhang H: **Sequential local least squares imputation estimating missing value of microarray data.** *Comput Biol Med* 2008, **38**:1112-1120.
20. Celton M, Malpertuy A, Lelandais G, de Brevern A: **Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments.** *BMC Genomics* 2010, **11**:15.
21. Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G: **Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes.** *BMC Bioinformatics* 2008, **9**:12.
22. Stein C: **Inadmissibility of the usual estimator for the mean of a multivariate normal distribution.** *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1956, **1**:197-206.
23. James W, Stein C: **Estimation with quadratic loss.** *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1961, **1**:361-379.
24. Wang H: **Brown's paradox in the estimated confidence approach.** *The Annals of Statistics* 1999, **27**:610-626.
25. Wang H: **Improved confidence estimators for the multivariate normal confidence set.** *Statistica Sinica* 2000, **10**:659-664.
26. Ogawa N, DeRisi J, Brown P: **New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis.** *Molecular Biology of the Cell* 2000, **11**:4309-4321.
27. Bohen S, Troyanskaya O, Alter O, Warnke R, Botstein D, Brown P, Levy R: **Variation in gene expression patterns in follicular lymphoma and the response to rituximab.** *Proc Natl Acad Sci USA* 2003, **100**:1926-1930.
28. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, Rosenwald A, Boldrick J, Sabet H, Tran T, Yu X, Powell J, Yang L, Marti G, Moore T, Hudson J, Lu L, Lewis D, Tibshirani R, Sherlock G, Chan W, Greiner T, Weisenburger D, Armitage J, Warnke R, Levy R, Wilson W, Grever M, Byrd J, Botstein D, Brown P, Staudt L: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503-511.
29. Brauer M, Saldanha A, Dolinski K, Botstein D: **Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures.** *Mol Biol Cell* 2005, **16**:2503-2517.
30. Shapira M, Segal E, Botstein D: **Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress.** *Mol Biol Cell* 2004, **15**:5659-5669.

doi:10.1186/1752-0509-7-S6-S11
Cite this article as: Wang *et al.*: Shrinkage regression-based methods for microarray missing value imputation. *BMC Systems Biology* 2013 **7**(Suppl 6):S11.