

Research Article

Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares

Fuxi Shi,¹ Dan Zhang,² Jun Chen,¹ and Hamid Reza Karimi³

¹ College of Mechanical and Electronic Engineering, Northwest A&F University, No. 22 Xinnong Road, Yangling, Xi'an, Shaanxi 712100, China

² School of Electronics and Information Engineering, Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an, Shaanxi 710049, China

³ Department of Engineering, Faculty of Technology and Science, University of Agder, Service Box 509, 4898 Grimstad, Norway

Correspondence should be addressed to Jun Chen; cmee.chen@gmail.com

Received 1 March 2013; Revised 11 March 2013; Accepted 13 March 2013

Academic Editor: Rongni Yang

Copyright © 2013 Fuxi Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Missing values are prevalent in microarray data, they course negative influence on downstream microarray analyses, and thus they should be estimated from known values. We propose a BPCA-iLLS method, which is an integration of two commonly used missing value estimation methods—Bayesian principal component analysis (BPCA) and local least squares (LLS). The inferior row-average procedure in LLS is replaced with BPCA, and the least squares method is put into an iterative framework. Comparative result shows that the proposed method has obtained the highest estimation accuracy across all missing rates on different types of testing datasets.

1. Introduction

Data generated from DNA microarray data is useful for various biological applications; the data is in the form a large matrices. Generally, a row in a matrix represents a gene, and a column represents an experimental condition. But as large matrices, the data often suffer from missing values due to technical reasons such as spotting problems and background noise [1]. However, downstream analyses always need full matrices as input; thus these missing values should be estimated from existing values. Various methods to estimate missing values in microarray data have been proposed in the past decades. Generally, methods to estimate missing values can be divided into four categories [2]: (i) global based methods, (ii) local based methods, (iii) hybrid methods, and (iv) knowledge-based methods. Singular value decomposition (SVD) [3] and Bayesian principal component analysis (BPCA) [4] are two major global based approaches. SVD estimates the missing value j in gene i by first regressing this gene against K eigengenes and use the coefficients of the regression to reconstruct j from a linear combination of the

k eigengenes. BPCA estimates the target gene (i.e., a gene that contains missing values) by a linear combination of K principal axis vectors, where the parameters are identified by a Bayesian estimation method. Local based category includes some classical and newly proposed methods. The most well-studied local based method is local least squares (LLS) [5]. LLS uses a multiple regression model to estimate the missing values from K nearest neighbor genes of the target gene. Most recently proposed local methods are based on LLS, including iterated Local Least Squares (iLLS), weighted local least squares (wLLS) and iterative bicluster-based least squares (bi-ILS). Hybrid methods aim to capture both global and local correlations in the data. LinCmb [6] and EMDI [7] are two typical hybrid methods which estimate the missing values by a combination of other estimation methods from global approaches and local approaches. In the knowledge-based category, domain biological knowledge or external information is integrated into the estimation process.

Among all kinds of microarray missing value estimation methods, BPCA and local least squares (LLS) are two most widely used approaches. The former is based on the global

structure of the matrix, and the latter is based on local similarity of the matrix. According to a survey [8] about different microarray missing value estimation methods, BPCA performs better than LLS on datasets with lower complexity, whereas due to another survey [9], LLS is superior than BPCA in the presence of data with dominant local similarity structures. This phenomenon inspires us to integrate the two methods, with the hope of improving the estimation accuracy and robustness. The idea of iterated local least squares again inspired us to put the integrated method into an iterative framework, which will further improve the estimation accuracy. We will give a brief review of BPCA and LLS in Section 2, the new method will be described in Section 3, comparative test of the proposed method with LLS and BPCA will be given Section 4, and a conclusion is drawn in Section 5.

2. Brief Review of BPCA and LLS

2.1. Bayesian Principal Component Analysis. Bayesian methods have been widely used in many fields such as face recognition and decision making [10–13], and it also has successful application in microarray missing value estimation. Bayesian principal component analysis (BPCA) represents the D -dimensional microarray expression vectors \mathbf{Y} as a linear combination of K ($K < D$) principal axis vectors w_l ($1 \leq l \leq K$):

$$y = \sum_{l=1}^K x_l w_l + \varepsilon, \quad (1)$$

where the coefficient x_l is called a factor score and ε denotes the residual error. The principal axis vectors are obtained by computing the eigenvalues and eigenvectors of the covariance matrix of the dataset \mathbf{Y} . As there are missing values in the original matrix \mathbf{Y} , the principal axis vectors are separated into two parts as $\mathbf{W} = (\mathbf{W}^{\text{obs}}, \mathbf{W}^{\text{miss}})$, corresponding to the observed part and missing part, respectively. Factor scores $x = (x_1, x_2, \dots, x_k)$ are obtained by minimizing the residual error of the observed part:

$$\text{err} = \|\mathbf{y}^{\text{obs}} - \mathbf{W}^{\text{obs}} \mathbf{x}\|^2. \quad (2)$$

Equation (2) is a least squares problem LLS which can be solved easily in BPCA. By using the factor scores x and \mathbf{W}^{miss} , the missing part of the dataset is estimated as

$$\mathbf{y}^{\text{miss}} = \mathbf{W}^{\text{miss}} \mathbf{x}. \quad (3)$$

In BPCA, the factor scores x and the residual error ε in (1) are assumed to obey normal distributions; BPCA utilizes a probabilistic PCA (PPCA) model [14] to estimate parameters in the normal distribution. The parameter \mathbf{W} , along with another two parameters μ and τ in the normal distribution, forms a parameter set $\theta = \{\mathbf{W}, \mu, \tau\}$. BPCA introduces a Bayesian estimation method for the PPCA model, where the posterior distributions of θ and \mathbf{Y}^{miss} are estimated by a variational Bayes algorithm [15] simultaneously.

2.2. Local Least Squares. Local least squares (LLS) uses the linear correlation of the target gene and its k nearest neighbors to recover unknown entries in the target gene. To explain how LLS works, we take an $m \times n$ microarray matrix as an example. Assuming that gene y has p missing values, take g_1 and its k nearest neighbors $g_{s1}, g_{s2}, \dots, g_{sk}$ as a column vector, where in finding the nearest neighbors, the measurement can be l_2 -norm distance or Pearson's correlation; then, rewrite the vector as (4):

$$\begin{pmatrix} g_1 \\ g_{s1} \\ g_{s2} \\ \vdots \\ g_{sk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{w}^T \\ \mathbf{B} & \mathbf{A} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p & w_1 & w_2 & \cdots & w_{n-p} \\ B_{1,1} & B_{1,2} & \cdots & B_{1,p} & A_{1,1} & A_{1,2} & \cdots & A_{1,n-p} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,p} & A_{2,1} & A_{2,2} & \cdots & A_{2,n-p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{k,1} & B_{k,2} & \cdots & B_{k,p} & A_{k,1} & A_{k,2} & \cdots & A_{k,n-p} \end{pmatrix}. \quad (4)$$

In (4), $\boldsymbol{\alpha}$ is the vector of unknown entries of the target gene and \mathbf{w}^T is the vector of known entries of the target gene. \mathbf{B} and \mathbf{A} are the k neighbors' corresponding columns with $\boldsymbol{\alpha}$ and \mathbf{w}^T , respectively. A linear coefficient vector \mathbf{X} is established as a least squares problem with \mathbf{A}^T and \mathbf{w} :

$$\min_{\mathbf{X}} \|\mathbf{A}^T \mathbf{X} - \mathbf{w}\|. \quad (5)$$

Then the unknown entries of the target gene can be reconstructed by a linear combination of \mathbf{B}^T and \mathbf{X} :

$$\boldsymbol{\alpha}^T = (a_1 \cdots a_p)^T = \mathbf{B}^T \mathbf{X} = \mathbf{B}^T (\mathbf{A}^T)^\dagger \mathbf{w}, \quad (6)$$

where $(\mathbf{A}^T)^\dagger$ is the pseudoinverse of \mathbf{A}^T . Repeat the procedure for all rows that have missing values and the full matrix can be recovered.

To estimate a proper k value in finding k nearest neighbors, LLS [5] provides a method like this. First, erase a certain number of known entries as missing values. Then, estimate the artificial missing matrix by using different k neighbors by LLS. At last, compare these estimated matrices with the actual matrix; the k value corresponding to the highest accuracy is chosen to be the optimal parameter.

3. BPCA-iLLS

Note that in LLS, in order to find k nearest neighbors and to estimate an optimal k value, a complete matrix is needed. However, in many cases, almost all rows in a microarray matrix contain missing values, which makes the distances between the target gene and other genes unable to be measured. To solve this problem, LLS [5] fills all missing values

in the target gene by the row's average value first. But in our experiment, we found that row-average cannot reflect the real structure of the dataset. Because row-average only uses the information of an individual row, the missing values in a target gene do not only rely on the known values in its own row. In the proposed BPCA-iLLS method, we replace the row-average procedure in LLS with BPCA. The flowchart of the proposed method is shown in Figure 1.

First, the input incomplete matrix is estimated by BPCA, to get a complete matrix. Next, this complete matrix is used as a temporary matrix for a further LLS procedure. In the LLS procedure, the optimal k value is estimated on this temporary matrix, and this k value is used to find matrices \mathbf{A} and \mathbf{B} . Subsequently, the missing values in every target gene are estimated by matrix \mathbf{B} and the coefficient vector \mathbf{X} . LLS is put into an iterative framework in the proposed method; that is, the estimated values by LLS are reused to form the temporary matrix in every iteration, and matrices \mathbf{A} and \mathbf{B} are refined in every iteration. It can be seen from the flowchart that the temporary matrices are different in each iteration. The initial temporary matrix is estimated by BPCA; following that, this matrix turns into the complete matrix that is estimated by LLS in each iteration. It should be mentioned that if the number of complete rows in the original incomplete matrix exceeds a preset threshold (e.g., 400 in LLS [5]), only complete rows are used to form the initial temporary matrix, which will highlight the original information of the matrix. This phenomenon happens only when the missing rates are low (typically below 5%). In most cases, the initial temporary matrices are BPCA-estimated ones in our proposed method. By replacing the row-average procedure in LLS by BPCA, and refining the temporary matrix in each iteration, the proposed method has the advantage over LLS and BPCA to be more robust on all kinds of datasets and has the ability to reduce the estimation error.

4. Comparative Result

4.1. Methods and Evaluation. We compare the proposed BPCA-iLLS method with BPCA and LLS. The only parameter of BPCA (number of principal axis vectors) is set to its default value, and the only parameter of LLS (number of neighbor genes) is learned by its heuristic method. For the proposed method, the number of iterations is a new parameter, and in our experiments, we set this parameter to be 5 because the estimation results do not change much after 5 iterations.

The accuracy is evaluated by normalized root mean square error (NRMSE):

$$\text{NRMSE} = \frac{\sqrt{\sum_{j=1}^N (y_j - \hat{y}_j)^2 / N}}{\sigma_y}, \quad (7)$$

where y_j is the real value, \hat{y}_j is the estimated value, and σ_y is the standard deviation for the N actual values of the missing entries. A smaller NRMSE represents a higher accuracy. The same evaluation criterion was also used in LLS, BPCA, and a survey of different missing value estimation methods [9].

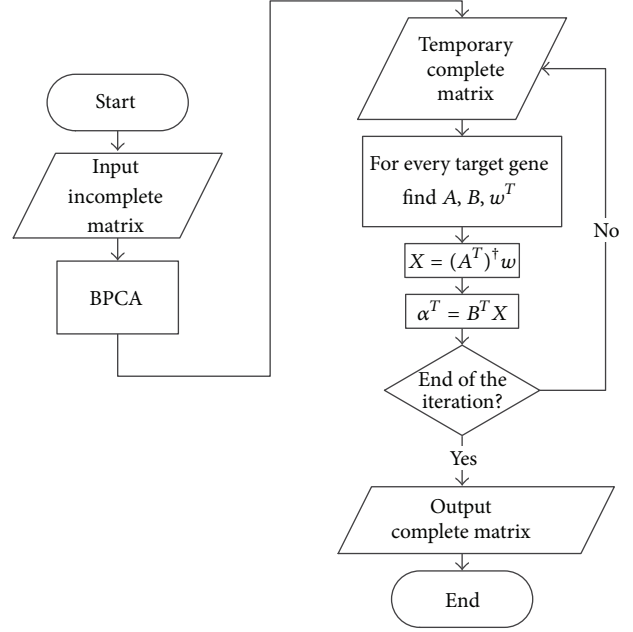


FIGURE 1: Flowchart of BPCA-iLLS.

TABLE 1: Testing datasets.

Dataset	Reference	Original size	Complete size	Type
CDC15_28	[16]	6178 × 41	869 × 41	TS
SP_ALPHA	[16]	6178 × 18	4489 × 18	TS
NCI60	[17]	9706 × 60	2266 × 60	NTS
Yoshi	[18]	6166 × 24	4380 × 24	MIX

4.2. Datasets. Three types of datasets are tested for the proposed method, they are time series data (TS), non-time-series data (NTS), and mixed data (MIX). Table 1 shows details of the testing datasets. Here, CDC15_28 is the same time series data as what was used in survey [9]; SP_ALPHA was also used in [5] to test the performance of LLS. NCI60 and Yoshi come from the non-time-series data and mixed data in survey [9], respectively.

All original datasets contain missing values. To compute the estimation error rates, only complete rows of these datasets are used. A number of entries are randomly removed from the complete part to get artificial missing values in different missing rates. As the real values of these entries are actually known, the error rates can be calculated following (7). The same testing method was also employed in BPCA, LLS, and surveys [2, 8, 9].

4.3. Experimental Result. We estimate different rates of simulated missing values on the abovementioned datasets by three comparative methods: LLS, BPCA, and BPCA-iLLS, and calculate NRMSE following (7). Figures 2(a), 2(b), 2(c), and 2(d) provide the NRMSE across different missing rates for the three comparative methods on datasets CDC15_28, SP_ALPHA, NCI60, and Yoshi, respectively. Every NRMSE is a mean value of five independent experiments.

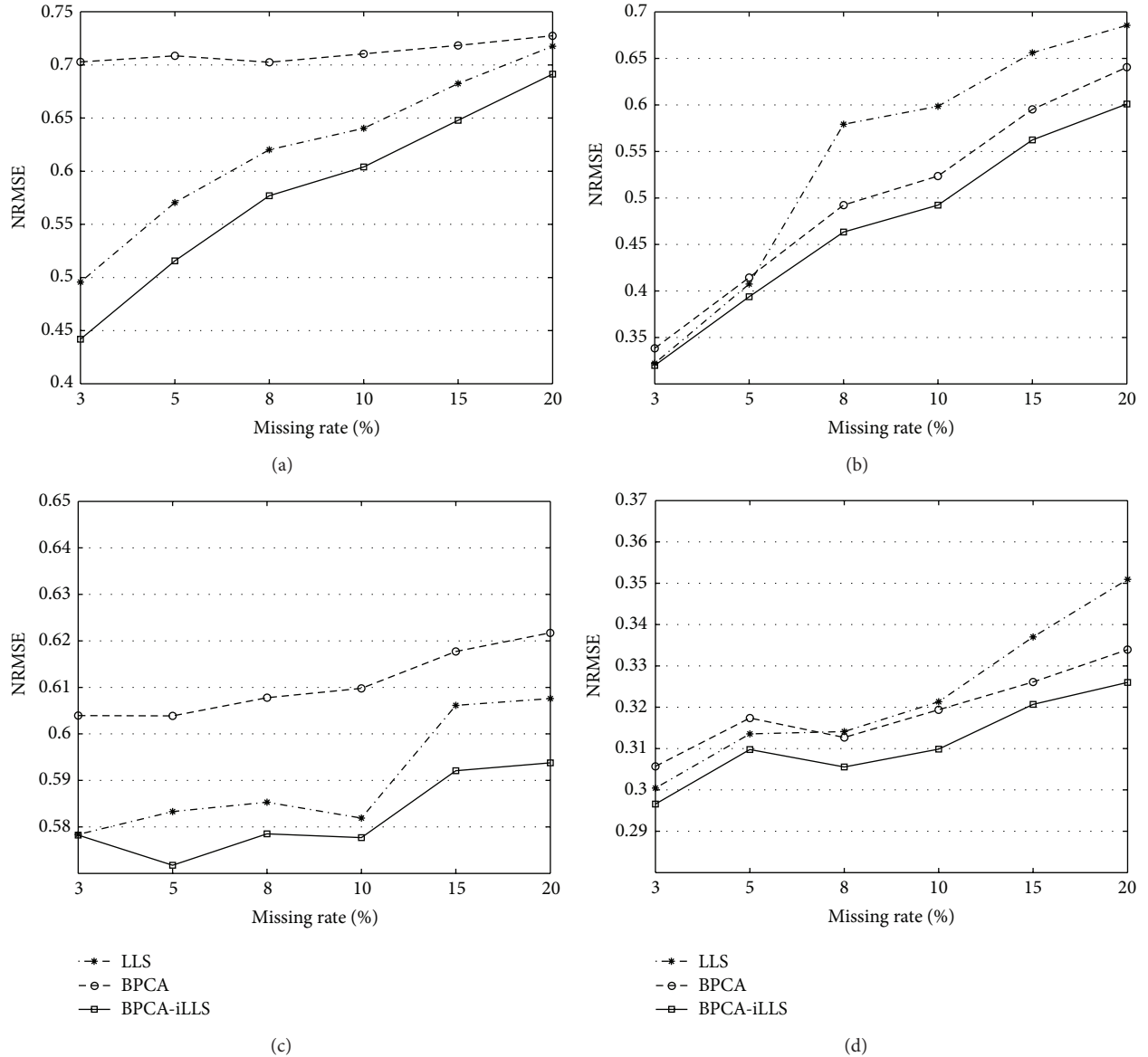


FIGURE 2: NRMSE on the four testing datasets. (a) CDC15_28, (b) SP_ALPHA, (c) NIC60, and (d) Yoshi.

It can be seen from Figure 2 that on all the four testing datasets, BPCA-iLLS obtains the lowest NRMSE across all missing rates. LLS outperforms BPCA on datasets CDC15_28 and NCI60, and BPCA outperforms LLS on dataset SP_ALPHA; this reveals that the two methods are complementary with each other. As an integration of the two methods, BPCA-iLLS shows its robustness on different datasets.

Table 2 shows the computational time of different methods on dataset CDC15_28. The time is obtained from running experiments by Matlab R2011b on an ordinary 64 bit Windows 7 computer with 3.4 GHz quad-core processor and 16 GB internal memory. Intuitively, as an integration of two methods, BPCA-iLLS requires more computational time. It can be seen from Table 2 that the computational time of BPCA-iLLS is indeed longer than that of BPCA and LLS.

TABLE 2: Computational time (seconds) on CDC15_28.

Missing rate	BPCA	LLS	BPCA-iLLS
3%	20.90	12.78	41.04
5%	27.86	12.43	47.90
8%	31.33	11.99	49.76
10%	27.23	11.74	46.85
15%	25.14	10.65	37.16
20%	22.25	9.57	34.29

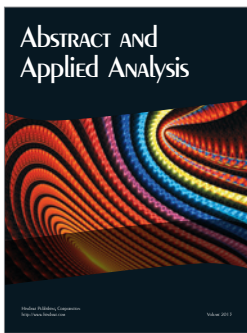
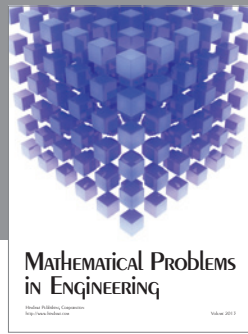
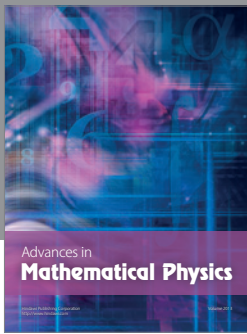
However the increment of time is within a limited scope. Considering its estimation accuracy, the increment of computational time is acceptable.

5. Conclusion

Microarray missing value estimation is an important procedure in biology experiments. As two widely used missing value estimation methods, Bayesian principal component analysis (BPCA) and local least squares (LLS) take advantage of the matrix's global structure and local structure, respectively; these two methods are complementary with each other. The proposed BPCA-iLLS method is an integration of BPCA and LLS, which fully exploits the global structure and local structure of the microarray matrix simultaneously, and the iterative scheme also helps to reduce the estimation error. Experimental results show that BPCA-iLLS has obtained the lowest normalized root mean square error (NRMSE) across all missing rates on all the testing datasets within an acceptable computational time. The performance of BPCA-iLLS also reveals the effectiveness of the integration of both global and local correlations of the microarray data, and such integration is one possible future direction of this field.

References

- [1] R. Jörnsten, H. Y. Wang, W. J. Welsh, and M. Ouyang, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, no. 22, pp. 4155–4161, 2005.
- [2] A. W. Liew, N. F. Law, and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," *Brief Bioinform*, vol. 12, no. 5, pp. 498–513, 2011.
- [3] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [4] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. I. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.
- [5] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2005.
- [6] R. Jörnsten, H. Y. Wang, W. J. Welsh, and M. Ouyang, "DNA microarray data imputation and significance analysis of differential expression," *Bioinformatics*, vol. 21, no. 22, pp. 4155–4161, 2005.
- [7] X. Y. Pan, Y. Tian, Y. Huang, and H. B. Shen, "Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach," *Genomics*, vol. 97, no. 5, pp. 257–264, 2011.
- [8] G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J. Lotz, and G. C. Tseng, "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes," *BMC Bioinformatics*, vol. 9, p. 12, 2008.
- [9] L. P. Brás and J. C. Menezes, "Dealing with gene expression missing data," *IEE Systems Biology*, vol. 153, no. 3, pp. 105–119, 2006.
- [10] M. R. Daliri and M. Saraf, "A Bayesian framework for face recognition," *International Journal of Innovative Computing, Information and Control*, vol. 8, pp. 4591–4603, 2012.
- [11] H. T. T. Nguyen, H. N. Luong, and C. W. Ahn, "An entropy approach to evaluation relaxation for Bayesian optimization algorithm," *International Journal of Innovative Computing, Information and Control*, vol. 8, pp. 6371–6388, 2012.
- [12] M. Hsieh, "A Bayesian approach in making mastery decisions: comparison of two loss functions," *International Journal of Innovative Computing, Information and Control*, vol. 8, pp. 7427–7435, 2012.
- [13] Z. F. Eren-Dogu and C. C. Celikoglu, "Information security risk assessment: Bayesian prioritization for AHP group decision making," *International Journal of Innovative Computing, Information and Control*, vol. 8, pp. 8001–8018, 2012.
- [14] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.
- [15] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI '99)*, pp. 21–30, 1999.
- [16] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.
- [17] U. Scherf, D. T. Ross, M. Waltham et al., "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, no. 3, pp. 236–244, 2000.
- [18] H. Yoshimoto, K. Saltsman, A. P. Gasch et al., "Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*," *Journal of Biological Chemistry*, vol. 277, no. 34, pp. 31079–31088, 2002.



Hindawi

Submit your manuscripts at
<http://www.hindawi.com>

