Two-Way Factor Analysis for Missing Value Estimation of Matrix Data

Kazuhiro Sodebayashi, Shigeyuki Oba
Graduate School of Information Science,
Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

Shin Ishii
Graduate School of Informatics,
Kyoto University,
Gokasho, Uji, Kyoto 611-0011, Japan

Abstract

Missing value estimation is usually an important preprocess for analyzing gene expression matrices because subsequent statistical analyses and machine learning algorithms usually require complete data matrices. In this article, we propose a novel probabilistic model, two-way factor analysis (2FA), which assumes heterogeneous noise variances which are specific to both samples and features simultaneously. We applied this model to missing value estimation tasks of synthetic and real gene expression data and compared the performance with those of conventional models such as probabilistic principal component analysis (PPCA) and factor analysis (FA). The 2FA model showed superior estimation performance to those by other models.

Keywords
Missing value estimation, gene expression analysis, linear-Gaussian latent variable model, low-rank matrix approximation.

1 Introduction

Biological experiments often include missing observations due to troubles in the measurement process, low qualities of samples or many other reasons. In bioinformatics, such missing values should be imputed in advance of the subsequent data analyses because many analysis methods based on statistics and machine learning algorithms, such as clustering, classification and dimension reduction, require complete-data matrices. In particular, DNA microarray, which is a high-throughput measurement technology used in a wide range of biological area, could include considerable number of missing entries possibly due to injury and dirt on arrays. Various estimation methods are proposed in order to achieve high accuracy of the missing value estimation [1, 2, 3].

When dealing with matrix data, there have been matrix factorization techniques, such as singular value decomposition (SVD), weighted low-rank matrix factorization [7], probabilistic principal component analysis (PPCA) [4], and factor analysis (FA) [5]. PPCA and FA are based on linear-Gaussian latent variable models with different noise models, namely, PPCA assumes that each component in an observed matrix includes i.i.d. noise with an identical variance, and FA assumes that each sample vector in an observed matrix includes i.i.d. noise whose variance depends on each gene [6]. The assumptions in PPCA and FA are insufficient, however, when considering a microarray dataset which may include both bad samples (i.e., certain experiments whose measurement qualities are worse than the others) and bad features (i.e., certain genes whose probe qualities are worse than the others) simultaneously.

In this study, we consider to integrate two kinds of noises; one is specific to each sample and the other to each feature, which no longer allows i.i.d. assumption, and therefore is called a two-way noise situation. In this study, we first show a generalization of linear-Gaussian latent variable models to handle weighted low-rank matrix approximation and to obtain simpler and more sophisticated treatment. Then, we propose a probabilistic model to deal with the two-way noise situation, called 2FA, and its estimation method are derived based on the maximum-likelihood framework. We apply the proposed model and some conventional models to synthetic and real datasets and show the advantage of the proposed model in terms of accuracies of the missing value estimation.

2 Method

Low-Rank Matrix Approximation Low-rank matrix approximation is a numerical method to obtain compact representation of a matrix, by obtaining a factored form which minimizes a pre-determined cost function. This method is robust against observation noise and missing values, and is efficiently used for noise filtering and missing value imputation. Srebro &

Jaakkola [7] proposed a weighted low-rank matrix approximation (WLRA) which minimizes the weighted Frobenius norm given by

$$J(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}\left(Y_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk}\right)^2, \quad (1)$$

where $\mathbf{U} \in \mathbb{R}^{M\times K}$ and $\mathbf{V} \in \mathbb{R}^{N\times K}$ are factorization matrices of the target matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$, w_{ij} is a weight that represents the approximation error of each element. These authors also performed missing value estimation by setting zero weight on missing elements in the matrix.

A Linear-Gaussian latent variable model is represented as a special case of WLRA. The observed variable $\mathbf{y} \in \mathbb{R}^M$ is obtained by applying a linear transformation to a latent variable $\mathbf{z} \in \mathbb{R}^K (K < M)$ with an additional Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^M$:

$$\mathbf{y} = \mathbf{U}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad (2)$$

where $\mathbf{U} \in \mathbb{R}^{M\times K}$ is a factor loading matrix, $\boldsymbol{\mu} \in \mathbb{R}^M$ is a mean parameter vector, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ is a Gaussian noise vector. PPCA and FA are probabilistic generative models which assume an isotropic covariance matrix $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$ and a diagonal covariance matrix $\boldsymbol{\Psi} = diag(\sigma_1^2, \cdots, \sigma_M^2)$, respectively.

The model mentioned above can also be represented by a matrix, in which an observation matrix $\mathbf{Y} \in \mathbb{R}^{M\times N}$ whose columns are $\{\mathbf{y}_j\}$, $j = 1, \ldots, N$, is given by

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^{\mathrm{T}} + \mathbf{E}, \quad (3)$$

where $\mathbf{V} \in \mathbb{R}^{N\times K}$ is a coefficient matrix. $\mathbf{E}$ is a noise matrix whose components are Gaussian noises $E_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$. Rows of the coefficient matrix $\mathbf{V}$ correspond to $\{\mathbf{z}_j\}$, $j = 1, \ldots, N$, and the latent variables are usually generated by Gaussian distribution in a standard formulation. Consequently, the likelihood function of the data matrix $\mathbf{Y}$ is given by

$$\begin{aligned}\mathcal{L} &= \ln p\left(\mathbf{Y}|\mathbf{U},\mathbf{V},\{\sigma_{ij}^2\}\right) + \ln p(\mathbf{V}) \\ &= -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}\left\{\ln 2\pi\sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2}\left(Y_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk}\right)^2\right\} \\ &\quad -\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{K}\left\{\ln 2\pi + V_{jk}^2\right\}.\end{aligned}$$

Comparing to the cost function (1) of the weighted low-rank matrix approximation, it is obvious that the estimated matrix $\mathbf{U}\mathbf{V}^{\mathrm{T}}$ by the maximum likelihood and that by the WLRA are equivalent if the WLRA weights are set at the inverse variance of each component.

Missing Value Estimation with 2FA Model Based on the above-introduced view-point of the low-rank matrix approximation, we propose a novel probabilistic model which assumes that the noise levels in elements of the matrix depend on both samples and features simultaneously, while the i.i.d. assumption of observation vectors no longer holds. Namely, we assume Gaussian noise whose variance σ_{ij}^2 is a sum of sample-wise variance σ_{ri}^2 and feature-wise variance σ_{cj}^2, called two-way noise;

$$p(E_{ij}) = \mathcal{N}(0, \sigma_{ij}^2), \qquad (\sigma_{ij}^2 = \sigma_{ri}^2 + \sigma_{cj}^2). \quad (4)$$

We call the WLRA model incorporating the two-way noise model above, the two-way factor analysis (2FA).

The likelihood of parameters based on observed data $\mathbf{Y}_o$ including missing values is then given by

$$\begin{aligned}L &= \ln p\left(\mathbf{Y}_o|\mathbf{U},\mathbf{V},\{\sigma_{ij}^2\}\right) + \ln p(\mathbf{V}) \\ &= -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N} w_{ij}\left\{\ln \sigma_{ij}^2 + \frac{1}{\sigma_{ij}^2}\left(Y_{ij} - \sum_{k=1}^{K} U_{ik}V_{jk}\right)^2\right\} \\ &\quad -\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{K} V_{jk}^2 + \text{const.}, \quad (5)\end{aligned}$$

where $w_{ij} = 0$ if the observation Y_{ij} is missing or $w_{ij} = 1$ otherwise. Note that we can regard the negative likelihood as the weighted norm of the low-rank matrix approximation.

To perform matrix factorization and estimate the parameters of the two-way noise model, we applied a conjugate gradient procedure, as Srebro & Jaakkola [7] recommended for WLRA. As Maeda & Ishii [8] pointed out, convergence of the EM algorithm can be slow if there are strong correlations between parameters because the conventional EM updates are done in a coordinate-descent manner.

3 Experiments

We applied our model to estimating low-rank matrix representations of some matrix datasets including missing values and compared its missing value estimation performance with those by some conventional procedures. The SVD imputation applies singular value decomposition to an observed matrix whose missing

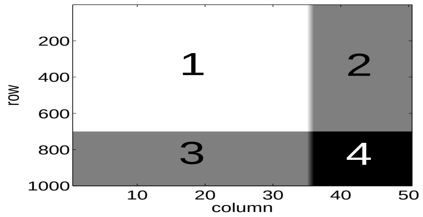

Figure 1: Color map of noise variance levels in an artificial two-way noise matrix **E**. Numbers in the figure are indices of regions. The variance of the gray region is given by a product of the original noise level $\sigma_{\mathbf{X}}^2$ and a noise ratio A. The variance of the black region is twice as large as that of gray regions. The variance of white region is almost zero, $\ln(1 + A\sigma_{\mathbf{X}}^2) \times 10^{-3}$.

values are simply imputed with 0, then matrix is reconstructed by employing a certain number of singular vectors with the largest singular values. PPCA, FA and FAc procedures are almost equivalent to the proposed 2FA procedure except that, the Gaussian noise variances behind the noise matrix are assumed to be same in all elements in the matrix, same for all features (row vectors), and same for all samples (column vectors), respectively.

We prepared a 1000 × 50 synthetic matrix of rank 3 and a 3170 × 22 gene expression matrix taken from the breast cancer database, BRCA [9]. These are the base matrices. We then added a two-way noise matrix **E** to each of the base matrices **X**, so that one-third rows and one-third columns of **E** have a higher variance than the other rows and columns, and thus **E** is constituted by four regions with different noise variances (See illustration of Figure 1). This two-way noise process simulates the existence of bad features and bad samples. The variance levels in the two-way noise are determined as $A\sigma_{\mathbf{X}}^2$, where $\sigma_{\mathbf{X}}^2$ is the variance of the original base matrix and A is an artificially-introduced noise ratio ($A = 0, 1, 3, 5, 10$). Although the BRCA matrix already includes unknown noise, we further added two-way noise to BRCA to simulate the situation we assume.

After that, 10% entries of these matrices were randomly masked as missing. From the observed part of the matrix, the artificially-introduced missing entries were estimated by 2FA and other conventional models, where we set the rank of the estimated matrix at consistently 3. The performance of the missing value prediction was evaluated by normalized root mean squared error (NRMSE) in region 1 (in Figure

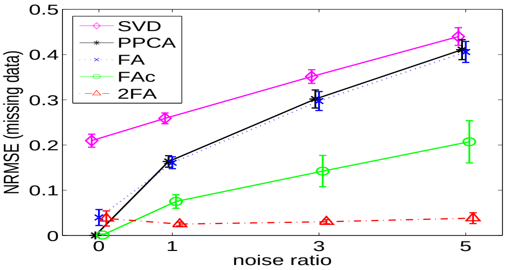

Figure 2: Comparison of the NRMSEs of matrix approximation by SVD, PPCA, FA, FAc and 2FA models on the synthetic matrix added by a two-way noise. The horizontal axis represents the noise ratio. The NRMSEs were examined on region 1 with respect to the artificially-introduced missing values. The markers and the errorbars denote means and standard deviations over ten simulations, respectively.

1):

$$\mathrm{NRMSE} = \sqrt{\mathrm{mean}\left[(X_{\mathrm{answer}} - X_{\mathrm{guess}})^2\right]} \Big/ \mathrm{sd}\left[X_{\mathrm{answer}}\right],$$

where X_{answer} and X_{guess} are sets of true and estimated values, respectively. When 2FA is applied, it is also important to estimate the noise variance appropriately. We assessed the estimated variance on all of the four regions when noise ratio was set at large, $A = 10$. Figures 2 and 3 show the missing value prediction errors for the synthetic matrix of rank 3 and the BRCA data matrix, respectively. In each figure, estimation errors for various noise ratios are shown. We compared performances by 2FA, SVD, PPCA, FA and FAc. In these figures, we see the 2FA exhibited better performance than the others especially when the noise ratio was large.

Figure 4 shows the estimated noise variance level in each region, and Table 1 shows true variance in each region. The variance was estimated well in each region. The good estimation performance of the sample-wise or feature(gene)-wise variances implies that we could reject bad samples or bad features by applying a certain threshold to the estimated variances.

4 Conclusions

In this study, we first reformulated weighted low-rank matrix approximation as a probabilistic model, and gave a unified viewpoint of linear-Gaussian latent variable models. Based on this framework, we proposed the 2FA model which assumes noise process depending of both features and samples. For the

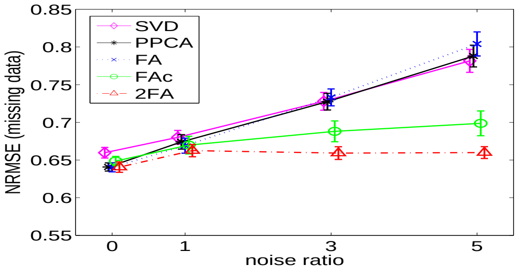

Figure 3: Comparison of the NRMSEs of matrix approximation by SVD, PPCA, FA, FAc and 2FA models on the gene expression matrix (BRCA) added by an artificial two-way noise. The horizontal axis represents the noise ratio. The NRMSEs were examined on region 1 with respect to the artificially-introduced missing values. The markers and the errorbars denote means and standard deviations over ten simulations, respectively.

synthetic data and gene expression data, 2FA model showed superior performance of missing value estimation to the other models especially when there is a large noise ratio between the high and low variance parts in the two-way noise matrix. Since the 2FA could also estimate the variances of the two-way noise with high accuracy, such estimation could be a guide to reject bad samples or bad features in a data matrix, along with performing missing value estimation.

For practical applications, we will incorporate a model selection framework into the 2FA model to select the appropriate rank, for example, by using automatic relevance determination. In addition, we will improve the convergence speed and the computational cost of the estimation algorithm. These are our near-future works.

Acknowledgments

This work was supported by a Grant-in-Aid for Young Scientists 19710172 from the MEXT.

References

[1] S. Oba, MA. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii, "A Bayesian missing value estimation method for gene expression profile data", *Bioinformatics,* **19**(16):2088-2096, 2003.

[2] G. Sanguinetti, M. Milo, M. Rattray, ND. Lawrence, "Accounting for probe-level noise in principal component analysis of microarray data", *Bioinformatics,* **21**(19):3748-3754, 2005.

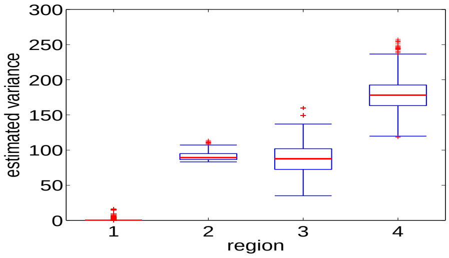

Figure 4: Box and whisker plot summarizes the estimated noise variance level on each region by 2FA. The bottom border of a box is lower quartile, the middle border is median, and the top border is upper quartile. The distance between the lower quartile and upper quartile is the interquartile range (IQR). The wisker range is 1.5 × IQR, and the points denoted by '+' are values out of the range.

Table 1: True variance corresponding to each region ($B = 10$)

region	1	2	3	4
true variance	0.01	91.67	91.67	183.34

[3] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, "Missing value estimation methods for DNA microarrays", *Bioinformatics,* **17**, 520-525, 2001.

[4] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis", *J. Roy. Statist. Soc. B,* **6**(3), 611-622, 1999.

[5] W. A. Kamakura and M. Wedel, "Factor analysis and missing data", *Journal of Marketing Research,* pp. 490-498, 2000.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning* , Springer, 2006.

[7] N. Srebro and T. Jaakkola, "Weighted low-rank approximations", In *Proceedings of ICML,* pp. 720-727, 2003.

[8] S. Maeda and S. Ishii, "Convergence analysis of the EM algorithm and Joint minimization of free energy", In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing,* pp.318-323, 2007.

[9] D. Hedenfalk *et al.*, "Gene-expression profiles in hereditary breast cancer", *N. Engl. J. Med.,* **344**(8):539-48, 2001.