Supplementation of Adjusted Values to the Imperfect Data

Manabu YUASA, Masafumi UMETANI, Nawo YAMAMOTO
RIST, Kinki University,Higashi-Osaka
577-8502, JAPAN
and
M.K.DAS
Institute of Informatics & Communication,
University of Delhi South Campus,
Benitojuarez Road, New Delhi 110021, INDIA

(Received 22 December, 2004)

Abstract

It happens frequently that the observational data is not complete but missing partly by various reasons. A preliminary study for supplementing adjusted values to such imperfect data based on Principal Component Analysis (PCA) is executed.
IRAS 3 colors of mass-losing stars and their expanding velocity on the ground based observations are adopted for the experiment. One of these 4 data is eliminated for each star and the adjusted value is restored. The original data and the restored one are compared and the distribution of the restored errors is studied.

Key words: Adjusted value, Imperfect data, Principal Component Analysis

1 Introduction

If we analize, using PCA(Unno and Yuasa 2000), the dynamical system which is given by a large number of observational data constituted by the many physical quantities with respert to many stars, we should adopt the data as much as possible concerning not only physical quantities but also the number of stars to increase the accuracy statistically. From this point of view, imperfect data should be included in the analysis if the imperfectness is small and the adjusted values are supplemented by the valid method.
In this paper, the preliminary study for the supplementation of adjusted values to the imperfect data using the method based on PCA(Unno and Yuasa 1992) is executed.
We have examined the data concerning the mass losing-stars (Yuasa et al 1999). IRAS 3 colors and the expanding velocity which is observed on the ground are adopted as four physical quantities and the number of stars is 121. Only one physical quantity is eliminated for each star and the adjusted value for it, is restored by using all the remained data ($121 \times 4 - 1 = 483$ data) based on PCA. Finally the eliminated original values and the restored values are compared.The restored error, the absolute value of the difference between the original data and the restored (adjusted) data, is rather small for the region where the absolute value of the normalized data is smaller than 1.

2 Method of Restoration

The method of restoration which is adopted in this study is written in the paper Unno and Yuasa(Unno and Yuasa 1992). Here, we sum-

marize the method. Let $Q_j^{(i)}$ and $w_j^{(i)}$ be the j-th observed physical quantity of the i-th star and its weight respectively, where $j = 1, \cdots, n$ and $i = 1, \cdots, N$. If a quantity is not observed the weight of it is equal to zero.
The normalized data $q_j^{(i)}$ for applying PCA are defined by

$$q_j^{(i)} = \frac{\{Q_j^{(i)} - \langle Q_j \rangle\}}{\sigma_j}, \tag{1}$$

where $\langle Q_j \rangle$ and σ_j are mean value and the standard deviation of Q_j respectively and they are given by

$$\langle Q_j \rangle = \frac{\sum_{i=1}^{N} w_j^{(i)} Q_j^{(i)}}{\sum_{i=1}^{N} w_j^{(i)}} \quad and$$
$$\sigma_j^2 = \frac{\sum_{i=1}^{N} w_j^{(i)} \{Q_j^{(i)} - \langle Q_j \rangle\}^2}{\sum_{i=1}^{N} w_j^{(i)}}. \tag{2}$$

Now, we introduce virtual added data $x_j^{(i)}$ and their weight $v_j^{(i)}$ for each observed quantity $q_j^{(i)}$ of each star as follows:

$$v_j^{(i)} = 1 - w_j^{(i)} \quad , \tag{3}$$
$$\sum_{i=1}^{N} v_j^{(i)} x_j^{(i)} = 0 \quad , \quad \sum_{i=1}^{N} v_j^{(i)} x_j^{(i)2} = \sum_{i=1}^{N} v_j^{(i)}. \tag{4}$$

Equations (4) are the constraints that the mean value of virtual data is zero and the standard deviation of them is 1 for each quantity ($\langle x_j \rangle = 0$ and $\sigma_{x_j} = 1$). Then, the correlation coefficient between the j-th quantity and the k-th quantity is defined by

$$r_{jk} = \frac{1}{N} \sum_{i=1}^{N} (w_j^{(i)} q_j^{(i)} + v_j^{(i)} x_j^{(i)})(w_k^{(i)} q_k^{(i)} + v_k^{(i)} x_k^{(i)}). \tag{5}$$

After the above preparation, most probable values of $x_j^{(i)}$ are given by solving the following n simultaneous linear algebraic equations:

$$\sum_{l=1}^{n} \frac{1}{\lambda_l} \left\{ \mu_{lj}^2 x_j^{(i)} + \sum_{k \neq j} \mu_{lj} \mu_{lk} (w_k^{(i)} q_k^{(i)} + v_k^{(i)} x_k^{(i)}) \right\}$$
$$= 0 \quad (j = 1, \cdots, n), \tag{6}$$

where λ_l is the l-th eigen value of PCA and μ_{lj} represents the j-th component of the l-th eigen vector of PCA.
Finally the adjusted value for the normalized j-th quantity of the i-th star is given by

$$w_j^{(i)} q_j^{(i)} + v_j^{(i)} x_j^{(i)}. \tag{7}$$

3 Application and Results

To test the above mentioned method, we adopt the data used in the previous paper concerning the mass-losing stars (Yuasa, Unno and Magono 1999). Using flux intensity at 12, 25, 60 and 100 μm by IRAS observation (IRAS Science Team 1986;Beichman et al. 1988;Joint IRAS Science Working Group 1988), we adopt three colors $Q_1^{(i)} = 2.5 \log\{F_{25}^{(i)}/F_{12}^{(i)}\}$, $Q_2^{(i)} = 2.5 \log\{F_{60}^{(i)}/F_{25}^{(i)}\}$, and $Q_3^{(i)} = 2.5 \log\{F_{100}^{(i)}/F_{60}^{(i)}\}$. Besides these three colors,we adopt $Q_4^{(i)} = \log V_e^{(i)}$, where V_e expresses the expanding velocity of the circumstellar gas observed by the emission lines of CO and/or HCN on the ground ($n = 4$). In the previous paper (Yuasa, Unno and Magono 1999), the 183 mass-losing stars were divided into two groups, namely group1 (121 stars) and group2 (62 stars). In this paper the group1 (121 stars) is adopted for the analysis ($N = 121$). We have applied the method of the previous section to the case of only one data, $Q_1^{(s)}$, is elimnated. Namely we have considerd the case $w_1^{(s)} = 0$ and all other weight $w_j^{(i)}$ excepting for $w_1^{(s)}$ are equal to 1. In this simple case the complicated simultaneous linear algebraic equations (6) are reduced to separated equations:

$$\Big(\sum_{l=1}^{n} \frac{\mu_{l1}^2}{\lambda_l}\Big) x_1^{(s)} + \sum_{l=1}^{n} \frac{\mu_{l1}}{\lambda_l} \Big(\sum_{k=2}^{n} \mu_{lk} q_k^{(s)}\Big) = 0$$
$$(s = 1, \cdots, N). \tag{8}$$

We can solve equation (8) easily to get the value $x_1^{(s)}$ for the restored (adjusted) value $q_1^{(s)}$.
By exchanging the original data $Q_1^{(i)}$ and $Q_2^{(i)}$, we can compute the restored (adjusted) value $x_2^{(s)}$ for supplementing the eliminated value $q_2^{(s)}$. In the same manner, we can get the restored (adjusted) values $x_3^{(s)}$ and $x_4^{(s)}$ for the eliminated values $q_3^{(s)}$ and $q_4^{(s)}$.
The results are shown in Fig.1 $\sim$ Fig.4, where the horizontal axis is the original normarized data q_j (mean value is 0 and the standard deviation is 1) and the vertical axis is the absolute value of the

difference between the original data and the restored (adjusted) data, $\mid q_j - x_j \mid$, in each figure. These figures show the difference (restored error) becomes larger as the absolute value of q_j becomes larger.

This tendency is plausible because our method supplements the restored (adjusted) values by the correlation between the eliminated quantity and the other three quantities, namely by the averaged virtual star.

Therefore, our method to restore the missing data is effective if the stars whose data is partly lacking are similar to an averaged star. In practice, if we consider the region where the absolute value of the normalized data $q_j^{(i)}$ is less than 1 in Fig.1$\sim$Fig.4, the restored error is less than 0.5 for the most of stars.

The frequency of the stars with respect to the restored error is shown in Fig.5$\sim$Fig.8. The distribution of the restored error is similar to the Gaussian distribution in each figure. Somewhat large distribution at the both ends of the horizontal axis is the results of the sum effects which is taken to the stars exceeding the value ± 2.0 with respect to the restored error.

The fact that the restored error distribution is similar to the Gaussian distribution may support the validity of our method. We will apply our method to the another data which consists of more number of the physical quantities and the stars to investigate the detailed conditions.

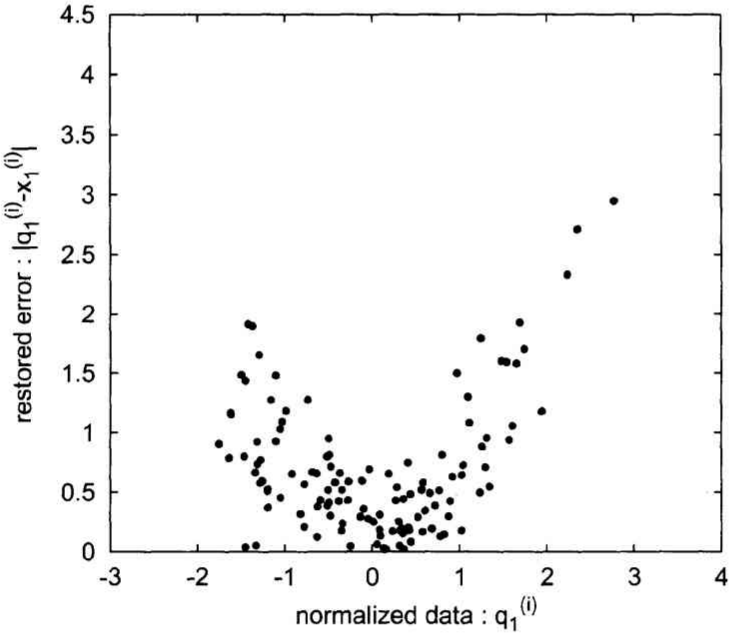

Fig.1. The restored error, the absolute value of the difference between the original data $q_1^{(i)}$ and the adjusted data $x_1^{(i)}$, is shown with respct to the normalized original data $q_1^{(i)}$. The adjusted data $x_1^{(i)}$ is obtained under the elmination of each $q_1^{(i)}$.

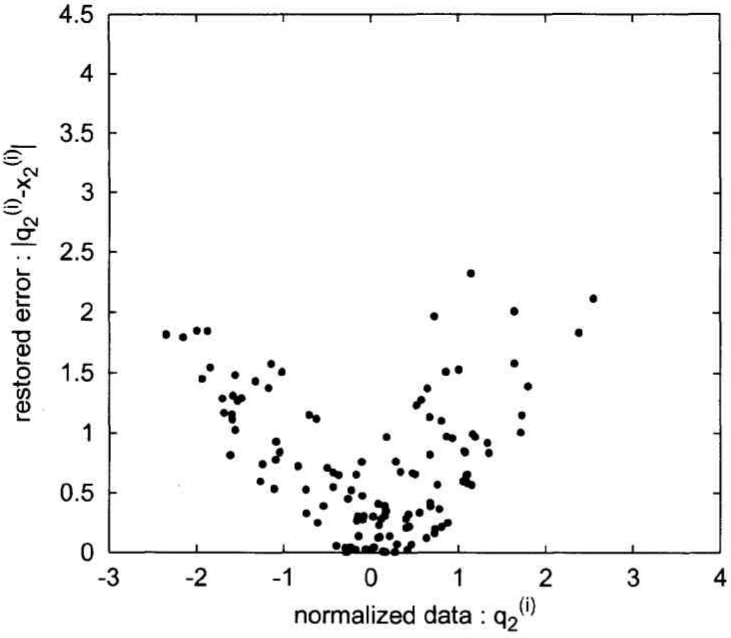

Fig.2. The restored error, the absolute value of the difference between the original data $q_2^{(i)}$ and the adjusted data $x_2^{(i)}$, is shown with respct to the normalized original data $q_2^{(i)}$. The adjusted data $x_2^{(i)}$ is obtained under the elmination of each $q_2^{(i)}$.

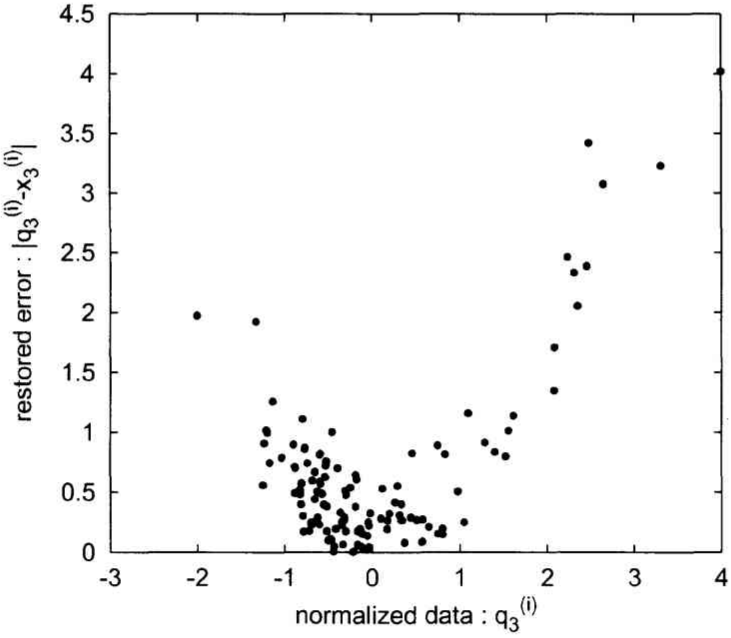

Fig.3. The restored error, the absolute value of the difference between the original data $q_3^{(i)}$ and the adjusted data $x_3^{(i)}$, is shown with respct to the normalized original data $q_3^{(i)}$. The adjusted data $x_3^{(i)}$ is obtained under the elmination of each $q_3^{(i)}$.

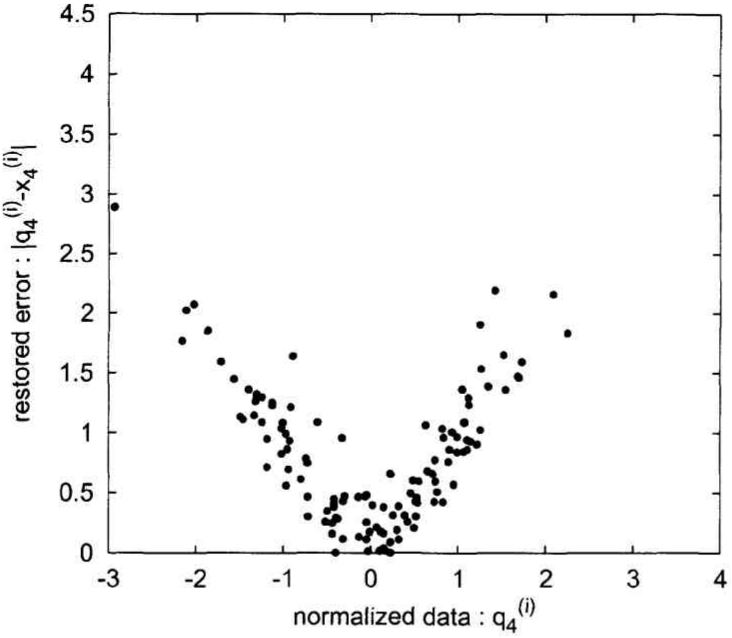

Fig.4. The restored error, the absolute value of the difference between the original data $q_4^{(i)}$ and the adjusted data $x_4^{(i)}$, is shown with respct to the normalized original data $q_4^{(i)}$. The adjusted data $x_4^{(i)}$ is obtained under the elmination of each $q_4^{(i)}$.

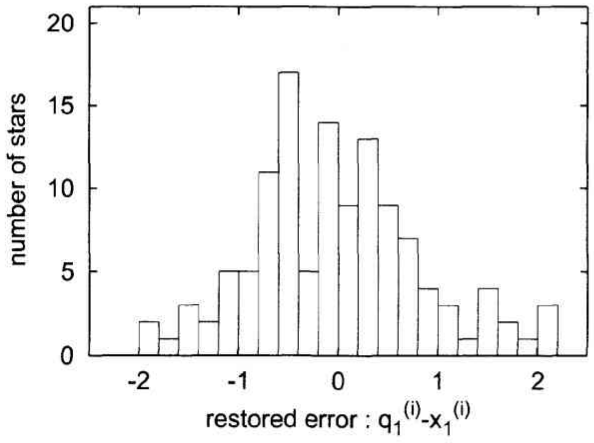

Fig.5. The frequency (number) of stars is shown with respect to the restored error $q_1^{(i)} - x_1^{(i)}$. At the both ends of the horizontal axis, the stars which exceed ± 2.0 are summed up.

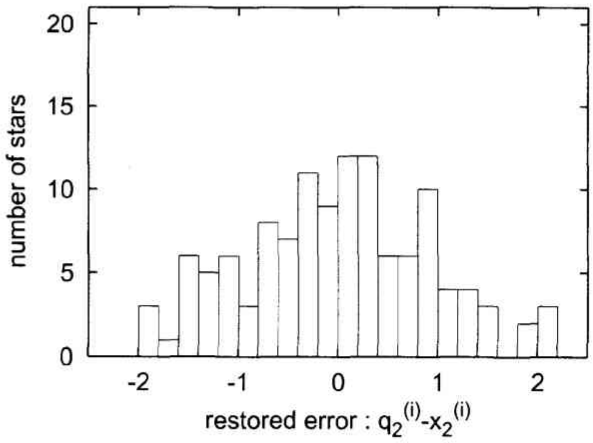

Fig.6. The frequency (number) of stars is shown with respect to the restored error $q_2^{(i)} - x_2^{(i)}$. At the both ends of the horizontal axis, the stars which exceed ± 2.0 are summed up.

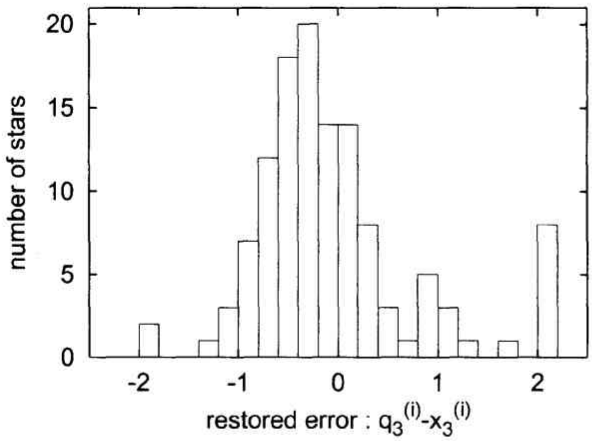

Fig.7. The frequency (number) of stars is shown with respect to the restored error $q_3^{(i)} - x_3^{(i)}$. At the both ends of the horizontal axis, the stars which exceed ± 2.0 are summed up.

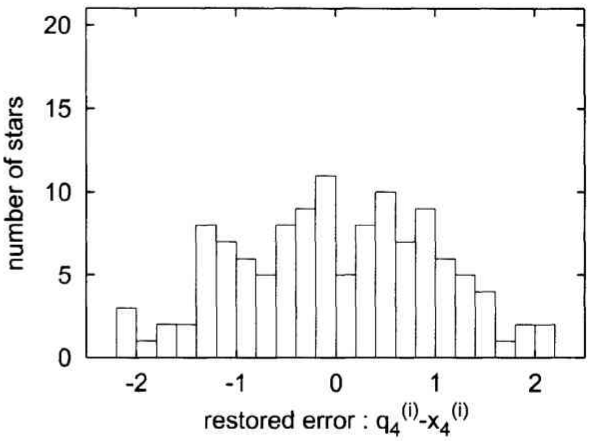

Fig.8. The frequency (number) of stars is shown with respect to the restored error $q_4^{(i)} - x_4^{(i)}$. At the both ends of the horizontal axis, the stars which exceed ± 2.0 are summed up.

4 Discussion

We have analized the data whose star number is 121 and the observed quantity is 4. If we analize the dynamical system whose dimension (the number of the independent simultaneous differential equations which constitute the dynamical system) is d, the embedding space for data should be equal to or larger than d in the case of non-complex system and be equal to or larger than $2d + 1$ in the case of complex dynamical system (Takens 1981). In addition, to determine n principal component axes in the embedding space the number of data N should exceed at least 2^n statistically. Then we have the inequality $N \geq 2^n$,$n \geq d$ in the case of non-complex system and $N \geq 2^n$, $n \geq 2d + 1$ in the case of complex dynamical system.

In our data for mass-losing stars, if the dynamical system can be described by 4 dimensions (mass, age, chemical composition and environment) and the system is not very complex, $n = 4$ and $N = 121$ are sufficient because of $N = 121 > 2^n = 16$. But there is a possibility that the dynamical system of mass-losing star can not be described by 4 dimensions and also it can not be recognized as non-complex system.

Then, if we can adopt more number of the embedding dimension n' (physical quantities) and the star number $N'(N' > 2^{n'})$, the error of the restored data $|q_j - x_j|$ may possibly have the distribution with smaller values than the results in this study. From this point of view other data should be tried to analize in the near future.

Acknowledgement

We are grateful to Emeritus Prof. W. Unno of the University of Tokyo for valuable discussions. M. Yuasa would like to thank JSPS (Japan Society for Promotion of Science) and DST (Department of Science & Technology, India) for a financial support to visit Delhi University, India where a part of this work was completed.

References

[1] Beichman C.A., Neugebauer G., Habing H.J., Clegg P.E., Chester T.J. (ed.) 1988, IRAS Catalogs and Atlases, Explanatory Supplement (NASA, Washington DC)

[2] IRAS Science Team 1986, A&AS 65, 607

[3] Joint IRAS Science Working Group 1988, IRAS Catalogs and Atlases, Point Source Catalog (NASA, Washington DC)

[4] Takens F. 1981, in D.A.Rand and L.-S.Young (eds.), Lecture Note in Mathematics, Springer-Verlag, Berlin, p. 366

[5] Unno W. and Yuasa M. 1992, Ap&SS 189, 271

[6] Unno W. and Yuasa M. 2000, PASJ 52, 127

[7] Yuasa M., Unno W. and Magono S. 1999, PASJ 51, 197