Identification of Peculiar Data by Using Restoration Method Based on Principal Component Analysis

Manabu YUASA, Nawo YAMAMOTO*, Masafumi UMETANI†
RIST, Kinki University, Higashi-Osaka
577-8502, JAPAN
and
Harinder P. SINGH
Department of Physics & Astrophysics,
University of Delhi,
New Delhi 110021, INDIA

(Received 20 December, 2006)

Abstract

We have developed the restoration method for missing data based on Principal Component Analysis in the previous issues (Yuasa et al. 2005; 2006). From another point of view, this method is able to be regarded as a tool to distinguish a peculiar data from the other most of the data which can be classified normally. We show some examples in the study of classification of the stellar spectra.

Key words: Peculiar data, Restoration of data, Principal Component Analysis

1 Introduction

Recently, a large number of huge database have been constructed in the field of observational astronomy. In these observational database, the small partial lack of the data is common. If we execute an analysis of the data statistically, we should adopt the data as much as possible regarding not only physical quantities but also the number of observed stars in order to produce the increased statistical accuracy. From this point of view, we have developed the restoration method for missing data based on generalized Principal Component Analysis (Unno, Yuasa 1992) in the previous paper (Yuasa et al. 2005; 2006). The method has been applied to the reconstruction of dynamical systems, determination of distances of 183 mass- losing super giants and also to the preliminary study of the restoration of missing data in the spectral data successfully (Yuasa et al. 1999; Unno, Yuasa 2000; Singh, Yuasa et al. 2006).In the present paper, we use this restoration method for missing data as a tool to distinguish peculiar data from the other most of data which can be classified normally.

*Present address: Software Research Associates, Inc.
†Present address: Department of Astronomical Science, The Graduate University for Advanced Studies

2 Restoration Method

The method of restoration which has successfully used in previous issues (Yuasa et al. 2005; 2006; Singh, Yuasa et al. 2006) is adopted from Unno and Yuasa (1992). Here we describe the method briefly for the set of data used in this paper. As mentioned in the next section, we have 20 flux values at 1 Å interval in each case (case(a): 4000Å$\sim$ 4019Å, case(b): 4077Å$\sim$ 4096Å, case(c): 4281 Å$\sim$ 4300 Å) of the range for 4000 $\sim$ 4300 Å for 300 stars.

For the $i-th$ star, let $F_j^{(i)}$ be the $j-th$ observed flux value, where $j = 1, \cdots, 20$ and $i = 1, \cdots, 300$. Using the method of the restoration (Unno and Yuasa 1992) which gives the most probable adjusted values to the missing data, we have examined the restoration of the data.

Embedding these data $F_j^{(i)}$ in the 20-dimensional space, we eliminate one of the data, for example $F_1^{(s)}$.The normalized data $f_j^{(i)}$ $(j = 1, \cdots, 20; i = 1, \cdots, 300)$ for applying PCA is introduced by

$$f_j^{(i)} = \frac{F_j^{(i)} - < F_j >}{\sigma_j}, \qquad (1)$$

where $< F_j >$ and σ_j represent the mean value and the standard deviation of the quantity F_j respectively. If we introduce the weight $w_j^{(i)}$ for $F_j^{(i)}$ and the other weight $v_j^{(i)} = 1 - w_j^{(i)}$ for the virtual added data $x_j^{(i)}$, the elimination of the data $F_1^{(s)}$ means that $w_1^{(s)} = 0$ and all other weights $w_j^{(i)}$ except for $w_1^{(s)}$ are equal to 1.

In this simple case, the virtual added data $x_1^{(s)}$ becomes the restored data of $F_1^{(s)}$ (Yuasa et al. 2005). The value of $x_1^{(s)}$ is given by the solution of the following separated simultaneous algebraic equations:

$$\Big(\sum_{l=1}^{20} \frac{\mu_{l1}^2}{\lambda_l}\Big) x_1^{(s)} + \sum_{l=1}^{20} \frac{\mu_{l1}}{\lambda_l} \Big(\sum_{k=2}^{n} \mu_{lk} f_k^{(s)}\Big) = 0$$
$$(s = 1, \cdots, 300), \qquad (2)$$

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.

By changing the columns of original data $F_1^{(i)}$ and $F_2^{(i)}$, we can compute the restored value $x_2^{(s)}$for supplementing the eliminated value $f_2^{(s)}$. In the same manner, we can get restored values for any eliminated one data.

3 Identification of Peculiar Data

We have adopted the data from the Indo-US coudé feed stellar spectral library (CFLIB) by Valdes et al. (2004). The library contains spectra of 1273 stars in the spectral region 3460 to 9464 Å at a high resolution of 1 Å and a wide range of spectral types. In this study a set of spectra of 300 stars is selected in the wavelength region 4000 $\sim$ 4300 Å from the CFLIB.

We show the identification of peculiar data in the following three cases as examples.

Case (a) uses a flux region of 20 Å starting from 4000 Å and the 20 principal components to reconstruct the fluxes at 4000 Å for all the 300 stars.

Case (b) uses a flux region of 20 Å starting from 4077 Å and the 20 principal components to reconstruct the fluxes at 4077 Å for all the 300 stars.

Case (c) uses a flux region of 20 Å starting from 4281 Å and the 20 principal components to reconstruct the fluxes at 4281 Å for all the 300 stars.

The restoration error, namely the difference between the restored data and the original data is shown in Fig.1 $\sim$ Fig.3 for the case of the elimination and the restoration of $f_{4000}^{(s)}$, $f_{4077}^{(s)}$ and $f_{4281}^{(s)}$ $(s = 1, \cdots, 300)$ respectively. In each Figure the horizontal axis is the difference between the normalized original variable $f_j^{(s)}$ (the mean value is 0 and the standard deviation is 1) and the restored value $x_j^{(s)}$ and the vertical axis represents the frequency distribution of the corresponding data. These Figures show the restoration error is small and we can conclude the restoration is successfully performed.

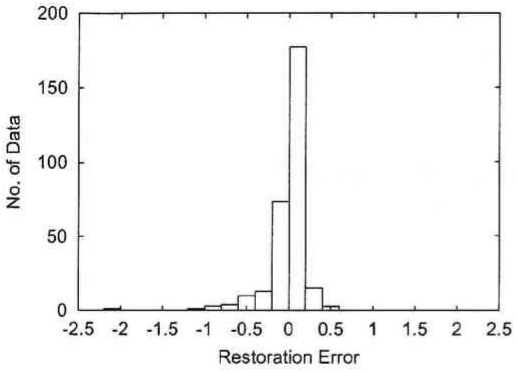

Fig.1. The frequency distribution of restored data is shown against the restoration error, i.e. $f^{(s)}_{4000}$(eliminated value)-$x^{(s)}_{4000}$(restored value) for $s = 1, \cdots, 300$.

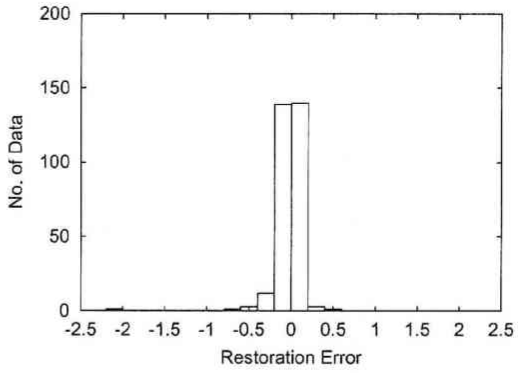

Fig.2. The frequency distribution of restored data is shown against the restoration error, i.e. $f^{(s)}_{4077}$(eliminated value)-$x^{(s)}_{4077}$(restored value) for $s = 1, \cdots, 300$.

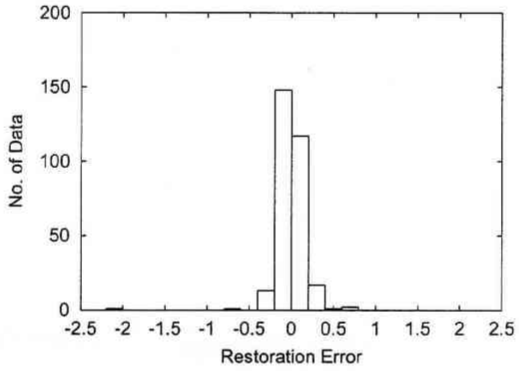

Fig.3. The frequency distribution of restored data is shown against the restoration error, i.e. $f^{(s)}_{4281}$(eliminated value)-$x^{(s)}_{4281}$(restored value) for $s = 1, \cdots, 300$.

Next we show in Fig.4 $\sim$ Fig.6, the restoration error of all 300 stars in each case against the original data which is standardized as the mean value and the standard deviation have 0 and 1 respectively.

We can find clearly one peculiar star, outlier, at the left bottom in all three Figures. The peculiar star in each Figure is the same star. This peculiar star is HD31996 and it indeed has no MK spectral class assigned to it in the CFLIB. In the 300 stars which we have adopted in this study, 298 stars are classified by MK spectral class normally. There remain two stars which have no spectral class by the reason of their complicated spectral feature. One of them is HD31996 and it has been able to be identified by our restoration method. Another no spectral class star is HD46687 but its spectrum resembles an M type star. For this star, we have been able to reconstruct the flux within a similar error value as the normal stars and we have not able to identify it as a peculiar data.

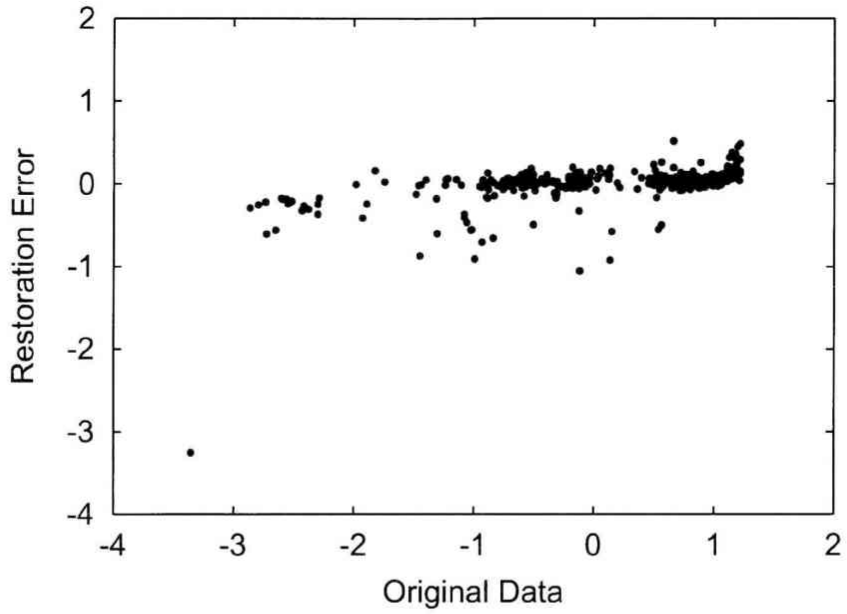

Fig.4. The restoration error ($f^{(s)}_{4000} - x_{4000}(s)$; $s = 1, \cdots, 300$) is plotted against the normalized original data for case (a). Flux value at 4000 Å is restored using 20 principal components between 4000 Å and 4019 Å.

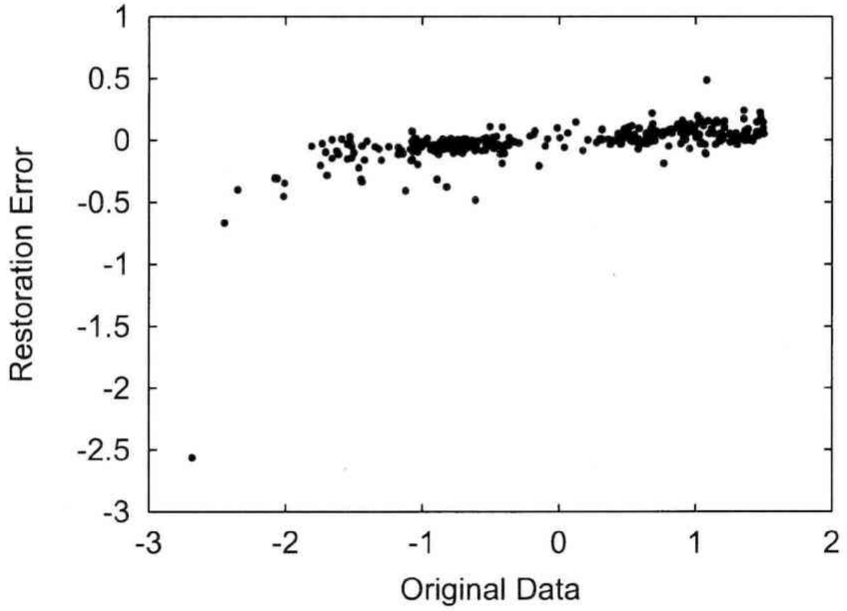

Fig.5. The restoration error ($f^{(s)}_{4077} - x_{4077}(s)$; $s = 1, \cdots, 300$) is plotted against the normalized original data for case (a). Flux value at 4077 Å is restored using 20 principal components between 4077 Å and 4096 Å.

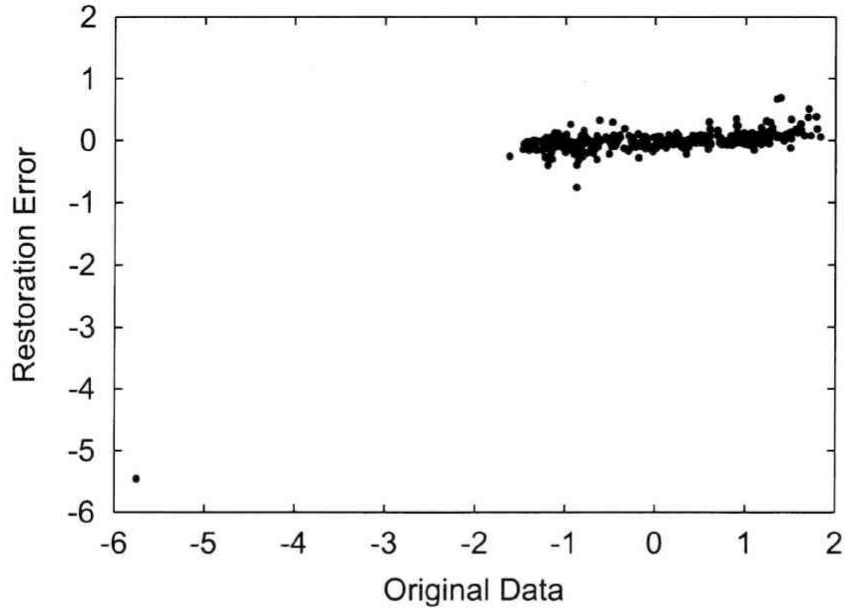

Fig.6. The restoration error ($f^{(s)}_{4281} - x_{4281}(s)$; $s = 1, \cdots, 300$) is plotted against the normalized original data for case (a). Flux value at 4281 Å is restored using 20 principal components between 4281 Å and 4300 Å.

4 Summary

In this study we have directed a spotlight on to the restoration method for missing data, not in the intrinsic meaning for supplementing adjusted values but in the meaning for identification of peculiar data.

The restoration method for missing data is useful not only for supplementing adjusted values to the imperfect observational data, but also for identifying a few of peculiar data included in a large number of normal data.

The restoration method for missing data based on Principal Component Analysis is able to be regarded as a tool to distinguish a peculiar data from the other most of data which can be classified normally.

Acknowledgement

We are grateful to Emeritus Prof. W. Unno of the University of Tokyo for valuable discussions. MY and HPS would like to thank JSPS (Japan Society for Promotion of Science) and DST (Department of Science & Technology, India) for a financial support for exchange visits which made this work possible.

References

[1] Singh H., Yuasa M., Yamamoto N. and Gupta R. 2006, PASJ 58, 177

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

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

[4] Valdes F., Gupta R., Rose J.A., Singh H.P., & Bell D.J. 2004, ApJS, 152,251

[5] Yuasa M., Unno W., & Magono S. 1999, PASJ 51, 197

[6] Yuasa M., Umetani M., Yamamoto N. and Das M.K. 2005, Sci. & Tech. Kinki Univ. No.17,1

[7] Yuasa M., Yamamoto N., Umetani M. and Das M.K. 2006, Sci. & Tech. Kinki Univ. NO.18,1