Journal of Statistics: Advances in Theory and Applications
Volume 17, Number 1, 2017, Pages 21-32
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100121760

MISSING DATA FILLING BASED ON PRINCIPAL COMPONENT ANALYSIS AND SAVITZKY-GOLAY DENOISING METHOD

XIANGYU WANG

Department of Mathematics
Jinan University
Guangzhou, 510632
P. R. China
e-mail: WangXY926@163.com

Abstract

In this paper, a new missing data filling method, SG-PCA filling algorithm, is provided. This algorithm is based on principal component analysis and Savitzky-Golay denoising method. For given incomplete data, we apply the PCA filling algorithm and the Savitzky-Golay denoising method alternatively to approximate the missing values. As an example, a filling experiment is performed by using the Breast Cancer data set in the University of California Irvine (UCI). The results show that the SG-PCA filling algorithm is more effective in filling accuracy.

1. Introduction

Data mining is an important topic both in computer science and statistics and is widely used in industry, business, healthcare, etc. If the data we are dealing with is not complete, the difficulty of data mining will increase, and

2010 Mathematics Subject Classification: 62-07, 62Pxx.
Keywords and phrases: Savitzky-Golay denoising (SG), principal component analysis (PCA), missing data filling.
Received January 18, 2017

then the efficiency of decision-making will be affected. So how to fill the missing data has been paid more and more attention by many researchers.

Deletion of missing data, may discard the hidden value in the data, resulting in waste of data (Litter and Rubin [1]; Han and Kamber [2]). In general, we use a filling algorithm for the missing data. There are many statistical methods to fill missing data, such as half-minimize algorithm (Xia and Psychogis [3]), k-nearest neighbour algorithm (Pan and Chen [4]; Wang [11]), singular value decomposition algorithm (Alter et al. [5]; Wang [11]; Oba and Sato [8]), Bayesian principal component analysis algorithm (Bishop [6]; Oba and Sato [8]), and multiple imputation algorithm (Rubin [7]).

Principal component analysis (PCA) is a classical method of data dimensionality reduction, and it can be used to fill the missing data (Litter and Rubin [1]).

However, the noise of the data is inevitable (Pan and Chen [4]). In this paper, we will use the combination of Savitzky-Golay denoising and principal component analysis to fill the missing data under the condition of random missing data.

In Section 2, we introduce the SG-PCA filling algorithm, which is based on the principal component model and Savitzky-Golay denoising model. We provide a numerical experiment in Section 3 to compare SG-PCA filling algorithm with some other filling methods.

2. Algorithm Models

2.1. Principal component analysis filling algorithm

Principal component analysis (PCA) is a nonparametric method to transform complex data from high dimension to low dimension and extract the main information of data. The principal component analysis filling algorithm uses incomplete data to find the principal components, and uses the obtained principal components to reconstruct the missing values.

Definition 2.1 (Singular value decomposition (SVD) [9])**.**

Suppose $M \in R^{m\times n}$, there is a decomposition

$$M = U\sum V^T, \tag{1}$$

where $U \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, $\sum \in \mathbb{R}^{m\times n}$ is a diagonal matrix and the diagonal elements are nonnegative, $V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and V^T is its transpose. This decomposition is called the singular value decomposition of matrix *M*.

For the sample data *X*, the principal component of matrix *X* is the columns of matrix *V*, which is based on the definition and the concept of principal components.

We present the filling algorithm as follows:

PCA Filling Algorithm

Task: For the given incomplete data $Y \in \mathbb{R}^{m\times n}$, fill the missing values in *Y*.

Parameters: We set the threshold ε_0 and the cumulative rate $\alpha_0 \in (0, 1)$.

Initialization: Let $\varepsilon = \varepsilon_0$. Use the incomplete mean of $Y(\overline{Y} = (\overline{y}_1, \overline{y}_2, \cdots, \overline{y}_n))$ to fill the missing values in *Y*, denote the filled matrix by Y_1.

Main iteration:

Step 1. Apply singular value decomposition on the matrix Y_1, obtain $U, \sum, V$.

Step 2. Find the least k such that the cumulative rate of the first k principal components reaches α_0.

Step 3. The p-th missing attribute (column) of the i-th sample (row) in Y is filled by

$$y_i = \overline{y_i} + \mathrm{ent}_{ip}(U\sum\nolimits_k V^T), \tag{2}$$

where $\sum_k$ is the diagonal matrix obtained by setting all diagonal elements of $\sum$ to be zero except the first k elements. Denote the filled matrix by Y_2.

Step 4. Compute $\varepsilon = \dfrac{\|Y_1 - Y_2\|_2}{\|Y_1\|_2}$, if $\varepsilon < \varepsilon_0$, stop. Otherwise, let $Y_1 = Y_2$, return to Step 1.

Output: The filled matrix Y_2 and k: the number of principal components we selected.

2.2. SG algorithm

Savitzky-Golay denoising method was proposed by Savitzky and Golay [10] in 1964, and has been widely used. It is a kind of filtering method based on local polynomial least squares fitting in time domain.

The main idea of Savitzky-Golay algorithm model is as follows:

Set a set of data $x(i)(i = -m, \cdots, \cdots, m)$, constructing a polynomial of order n:

$$p(i) = \sum_{k=0}^{n} a_k i^k. \tag{3}$$

The residual is:

$$\varepsilon_n = \sum_{i=-M}^{M} (p(i) - x(i))^2 = \sum_{i=-M}^{M} (\sum_{k=0}^{n} a_k i^k - x(i))^2. \tag{4}$$

By Fermat's rule, the above residual takes the smallest value when all its partial derivatives about a_k are zero, that is,

$$\frac{\partial \varepsilon_n}{\partial a_t} = \sum_{i=-M}^{M} 2i^t(p(i) - x(i)) = \sum_{i=-M}^{M} 2i^t(\sum_{k=0}^{n} a_k i^k - x(i)) = 0, \tag{5}$$

In the other words,

$$\sum_{k=0}^{n}(\sum_{i=-M}^{M} i^{t+k}) \cdot a_k = \sum_{i=-M}^{M} i^t x(i), \quad t = 0, \cdots, n. \tag{6}$$

Let $A = (a_{it})$, $a_{it} = i^t$, $-M \le i \le M$, $0 \le t \le n$, $B = A^T A$, then

$$b_{tk} = \sum_{i=-M}^{M} a_{ti} a_{ik} = \sum_{i=-M}^{M} i^{t+k} = b_{kt}. \tag{7}$$

So

$$Ba = A^T Aa = A^T x, \tag{8}$$

$$a = (A^T A)^{-1} A^T x = Hx, \tag{9}$$

where $H := (A^T A)^{-1} A^T$ is the desired convolution factor.

2.3. PCA algorithm based on Savitzky-Golay denoising

The missing data filling algorithm based on principal component analysis and Savitzky-Golay denoising method is:

SG-PCA Filling Algorithm

Task: For the given incomplete data $Y \in \mathbb{R}^{m \times n}$, fill the missing values in Y.

Parameters: Threshold $\varepsilon_0 > 0$.

Initialization: Let $\varepsilon = \varepsilon_0$. Use the incomplete mean of Y to fill the missing values in Y, denote the filled matrix by Y_1.

Main iteration:

Step 1. Apply the SG algorithm on each column of Y_1, denote the denoised matrix by Y_2.

Step 2. Apply the PCA filling algorithm on Y_2 to obtain the filled matrix Y_3, here we regard Y_2 as an incomplete matrix which has the missing values on the same positions as Y.

Step 3. Compute $\varepsilon = \frac{\|Y_1 - Y_3\|_2}{\|Y_1\|_2}$.

Step 4. If $\varepsilon < \varepsilon_0$, stop. Otherwise, let $Y_1 = Y_3$, return to Step 1.

Output: The filled matrix Y_3.

3. Numerical Results

In order to verify the effectiveness of the SG-PCA filling algorithm, we apply this algorithm, BPCA algorithm and HM algorithm on breast cancer data set in UCI [13], then compare the results.

Breast cancer set was created by Matjaz Zwitter and Milan Soklic (Doctor), oncology institute, University Medical Center, Ljubljana, Yugoslavia, Tan and Jeff Schlimmer donation. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

3.1. Model evaluation index

When data is missing, researchers try to find the data that is the same with reality, which is almost impossible. Therefore, how to find an accurate filling algorithm for missing fill is the key, that is, the missing estimate obtained by the missing fill algorithm is as close as possible to the original value of the missing data. However, there is no single criterion on how to evaluate the quality of the algorithm. In this paper, we use (Wang [11]; Li [12])

$$corrate = \left[1 - \frac{\sum_{t=1}^{q} \left| x^{(t)} - x^{(\tilde{t})} \right|}{q}\right] \times 100\%, \tag{10}$$

where $x^{(t)}$ is the original data, $x^{(\tilde{t})}$ is the filled data, and q is the total number of the missing data,

$$|x^{(t)} - x^{(\tilde{t})}| = \begin{cases} 0, & x^{(t)} = x^{(\tilde{t})}, \\ 1, & x^{(t)} \neq x^{(\tilde{t})}. \end{cases} \tag{11}$$

3.2. Empirical analysis

Figure 1 and Figure 2 are based on the breast cancer data set in UCI, under the conditions of 10% missing rate and 20% missing rate, analyze the effect of the k value on the filling effect. Figure 1, when $k = 10$, the cumulative contribution rate of the principal component has reached 95%, and the accuracy is high, however, Figure 2, although $k = 10$, the cumulative contribution rate of the principal component has reached 95%, but the accuracy is relatively small.

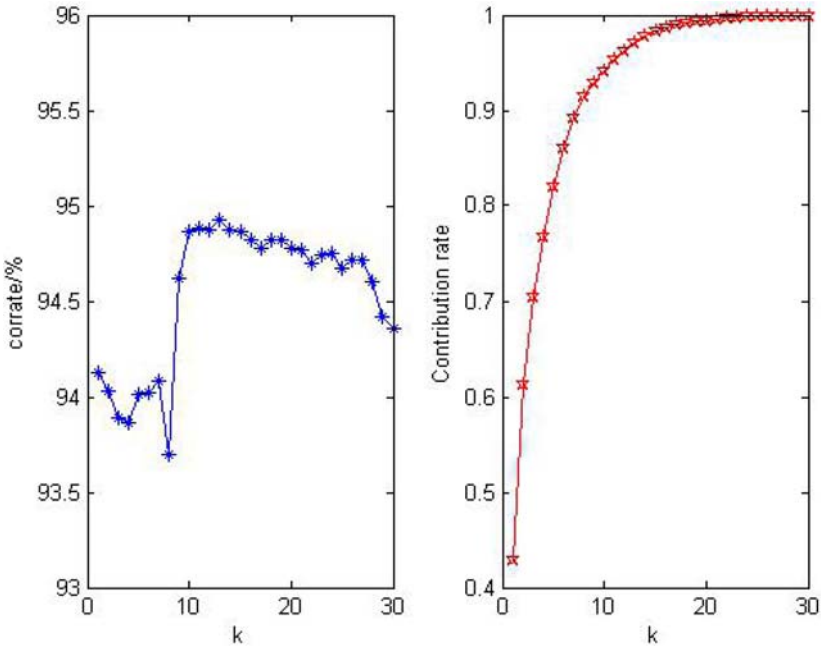

Figure 1. Filling map with 10% missing rate and the cumulative contribution rate of principal components.

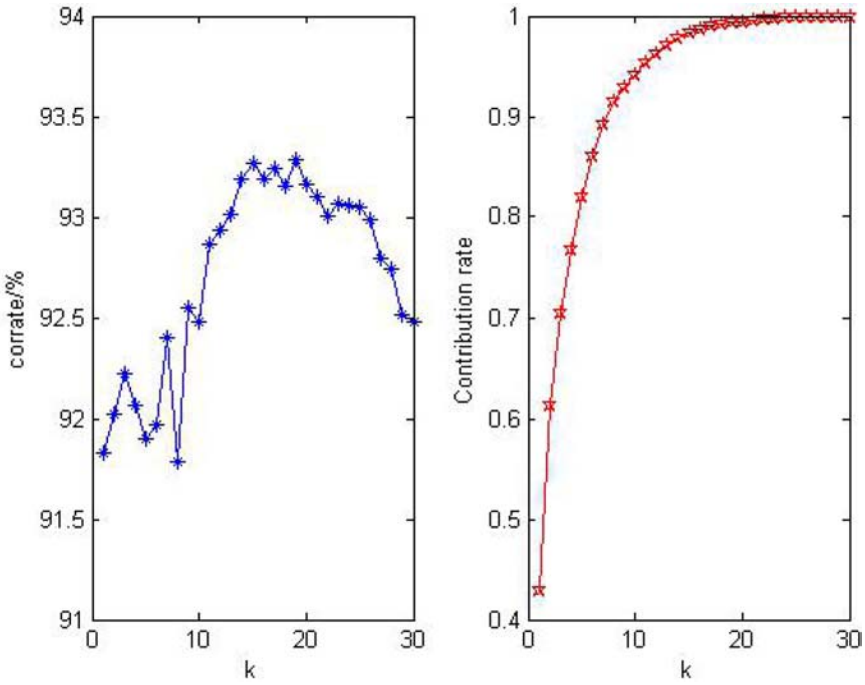

Figure 2. Filling map with 20% missing rate and the cumulative contribution rate of principal components.

Figure 3 is under different missing rate, comparing the accuracy of the PCA filling algorithm and the denoising PCA algorithm, we can figure out that the accuracy of PCA to fill the missing data is significantly less than the PCA algorithm.

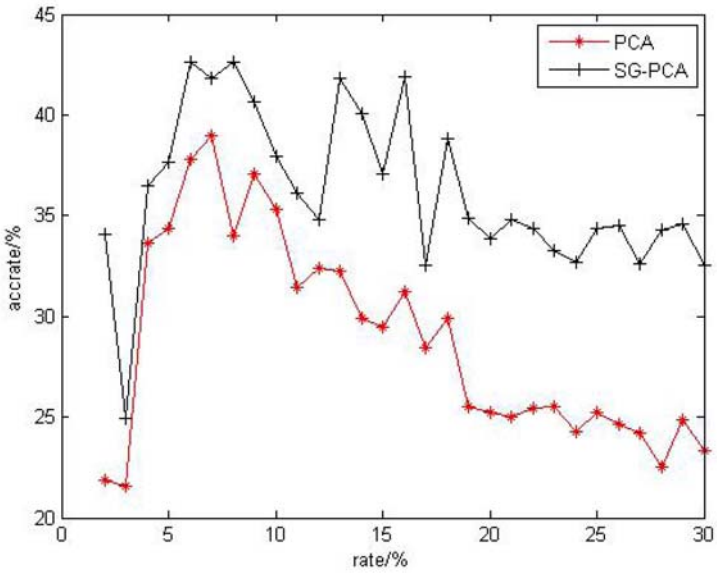

Figure 3. Filling map of PCA and SG-PCA under different missing rate.

Figure 4 shows the filling accuracy of the breast cancer data set through the PCA, BPCA, and HM algorithms at different missing ratios. We can find that although the overall level of PCA algorithm and BPCA be roughly the same, but the filling effect is not obvious.

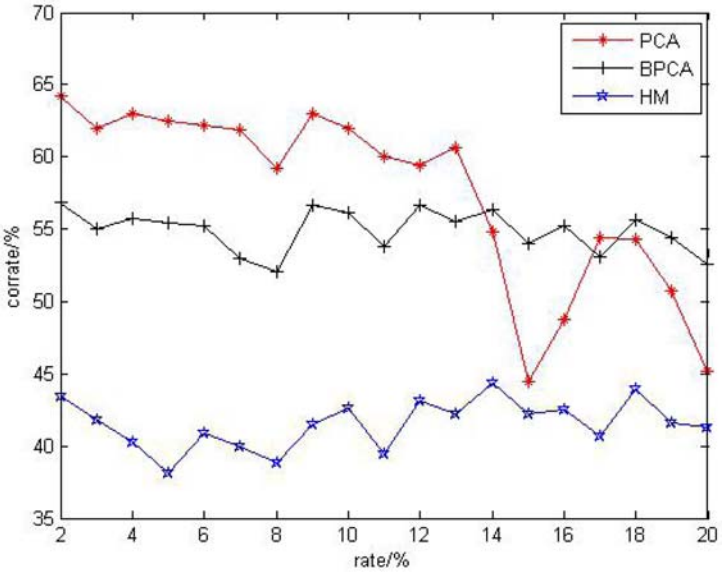

Figure 4. Comparison of PCA and different filling algorithms.

Figure 5 comparing the algorithm with BPCA and HM filling, we can see that the algorithm is 10% higher than BPCA and 20% HM.

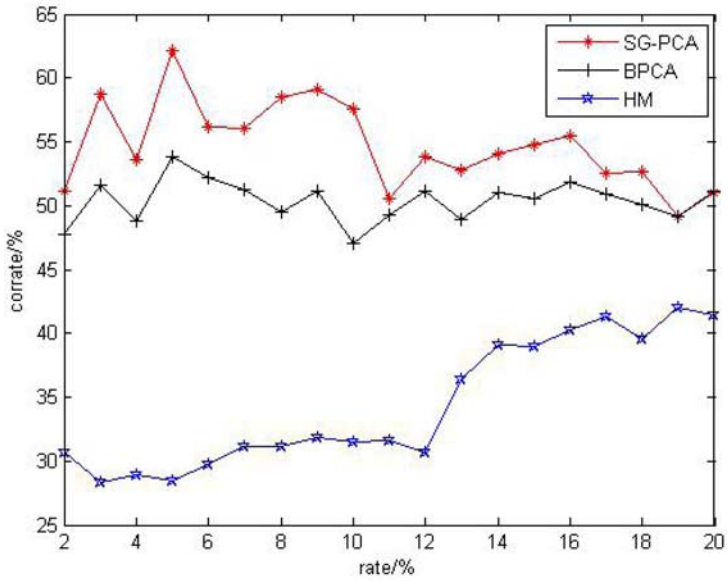

Figure 5. Comparison of SG-PCA and different filling algorithms.

4. Conclusion

In this paper, we propose a missing data filling method based on principal component analysis and Savitzky-Golay denoising method. The advantages of this method are: (a) Savitzky-Golay algorithm is a widely used denoising method; (b) Principal component analysis (PCA) is a non parametric method for extracting data information. According to the experimental results, we can find that the accuracy of this algorithm is higher than that of BPCA and HM algorithm.

Further research work: It is important to choose k in PCA filling algorithm, so it is necessary to study how to find the best k for different missing rates.

References

[1] R. J. A. Litter and D. B. Rubin, Statistical Analysis with Missing Data, John Wiley and Sons, Hoboken, NJ, 1987.

[2] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2nd Edition, San Diego, USA, Academic Press, 2006.

[3] J. G. Xia and N. Psychogios, Metabo analyst: A web server for metabolomic data analysis and interpretation, Nucleic Acids Research 37 (2009), 652-660.

[4] Z. M. Pan and Y. L. Chen, Local outlier detection algorithm based on shared anti k-nearest neighbors, Computer Simulation 30(2) (2010), 269-273.

[5] O. Alter, P. O. Brown and D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Natl. Acad. Sci. USA, 27 (2000).

[6] C. M. Bishop, Variational principal components, In IEE Conference Publication on Artificial Neural Networks (1999), 509-514.

[7] D. B. Rubin, Multiple Imputations for Nonresponse in Surveys, New York, 1987.

[8] S. Oba and M. A. Sato, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics 19(16) (2003), 2088-2096.

[9] W. B. Guo and M. S. Wei, Singular Value Decomposition and its Application in Generalized Inverse Theory, Beijing, Science Press, 2008.

[10] A. Savitzky and M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Analytical Chemistry (36) (1964), 1627-1639.

[11] F. M. Wang, Research and Application of Data Missing Filling Algorithm in Data Preprocessing, Guangdong University of Technology, 2010.

[12] H. Li, Statistical Learning Method, Beijing, Tsinghua University Press, 2012.

[13] http://archive.ics.uci.edu/ml/datasets.html

■