Regularized PCA for missing data

Haoyue Song & Jian Wang*

Unit: School of Mathematics and Statistics, Shandong University of Technology, Zibo, Shandong, China
&
Unit: School of Mathematics and Statistics, Shandong University of Technology, Zibo, Shandong, China

IJASR 2021
VOLUME 4
ISSUE 3 MAY – JUNE **ISSN: 2581-7876**

Abstract: To solve the overfitting problem in the existing methods for missing data, we introduce a regularized PCA (RPCA) method. We draw a missing data matrix, which obtain missing values and observed values. We make some distribution assumptions on the score matrix, the loading matrix and the error matrix to limit the range of estimation results. In numerical simulation, we fit a missing data set with the RPCA method, the MSE and MAE values are obviously decreased as the increase of the missing ratios, which reflects that the method has advantages on the missing data with high missing ratios.

Keywords: missing data, overfitting, regularized PCA

1. Introduction

Missing data is an important and common problem in data analysis, correct estimate the missing values will improve the utilization of the data and increase the rationality of the conclusion. In the past research, statisticians are committed to finding methods with higher estimation accuracy, and they found serious overfitting problem will occur in many missing values areas, especially for high dimensions and high missing ratios.

A simple way to reduce overfitting is reducing the dimensions by removing missing samples, whileit cause the loss of data information. Another effective method is the regularized method; it sets a rangefor parameter to prevent large changes of parameter, which can effectively avoid the influence of noiseon parameter estimation. Josse and Husson (2012) introduce a method named the regularizedPCA (RPCA) method, theycombined the regularized method and the PCA method to solve the overfitting problem of multivariatemissing data.

The rest structure of the paper is as follows. In Section 2, we introduce missing data matrix and thecalculation process of the RPCA method. In Section 3, we fit a missing data set withmean absolute error (MSE) and mean absolute error (MAE) indicators. In Section 4, we present severaldiscussions.

2. Theregularized PCA method

In this section, we present the RPCA method of missing data. We first define missing data matrix. The $I \times J$ original missing data matrix is $X_0 = (x_{0_{ij}})_{I\times J}$, we assume that the column mean of X_0 is zero.We mark the missing values and the observed values with $x_{0_{ij}}^{\#}$ and $x_{0_{ij}}^{*}$, respectively. The number ofmissing values is n, and the missing ratio value is $M = n/(I \times J)$. We then obtain an updated matrix X written as $X = (x_{ij})_{I\times J}$, where initial interpolate missing values with zero and keep observed valueswith the original values; that is, $x_{ij}^{\#} = 0$ and $x_{ij}^{*} = x_{0_{ij}}^{*}$. We now construct the $J \times J$ covariance matrix S and correlation matrix R, written as respectively

$$S = \frac{1}{n-1} X^{\mathrm{T}} X = (s_{ij})_{J\times J}\,,\ R = (r_{ij})_{J\times J}\,,$$

Where $r_{ij} = r_{ji} = s_{ij} \big/ \sqrt{s_{ii}}\sqrt{s_{jj}}$ for $i, j = 1, \ldots, J$.

Next, we perform singular value decomposition (SVD) on the matrix X as $X = B\Lambda A^{\mathrm{T}}$, where the singular value matrix is $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_J)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J \geq 0$, the left singular matrix B is an $I \times J$ column orthogonal matrix and the right singular matrix A is a $J \times J$ column orthogonal matrix. When perform SVD on the matrix R, we use the superscript (R) to distinguish its decomposition matrices, $A^{(R)} = B^{(R)}$ is named the eigenmatrix on X and $\Lambda^{(R)} = \mathrm{diag}(\lambda_1^{(R)}, \lambda_2^{(R)}, \ldots, \lambda_J^{(R)})$ is named the eigenvalue matrix on X. The cumulative contribution rate is $\zeta_r = \sum_{j=1}^{r} \lambda_j^{(R)} \Big/ \sum_{j=1}^{J} \lambda_j^{(R)}$, and we choose r as the principal components number when ζ_r rexceed 0.7 in real data sets. We use the dominant r singular values and the smaller $J\text{-}r$ singular values to consist two diagonal matrix, written as $\Lambda_r = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $\Lambda_{J\text{-}r} = \mathrm{diag}(\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_J)$, respectively. Corresponding, the r-right singular matrix is $A_r = (\mathbf{a}_{\cdot 1}, \mathbf{a}_{\cdot 2}, \ldots, \mathbf{a}_{\cdot r})$ and the r-left singular matrix is $B_r = (\mathbf{b}_{\cdot 1}, \mathbf{b}_{\cdot 2}, \ldots, \mathbf{b}_{\cdot r})$, respectively.

Thirdly, we introduce the RPCA method, it can be expressed as

$$X_{\mathrm{RPCA}} = TC^{\mathrm{T}} + E \tag{2.1}$$

Where the matrices T, C and E are the $J \times r$ loading matrix, the $I \times r$ score matrix and the $I \times J$ errormatrix, respectively. We further assume that each row of T is distributed as $N(\mathbf{0}, e_r e_r^{\mathrm{T}})$ and each row of E is distributed as $N(\mathbf{0}, \sigma^2 e_r e_r^{\mathrm{T}})$, where all the elements in the r-vector e_r are 1, see also Bellas, et al. (2013).With $\Lambda_{J\text{-}r}$, the estimator of σ^2 can be expressed as $\hat{\sigma}^2 = \mathrm{trace}(\Lambda_{J-r}^2)/(J-r)$. We then construct the MLEsof the score matrix T and the load matrix C in (2.1), expressed as

$$\hat{T} = I^{\frac{1}{2}} B_r (\Lambda_r^2 - \hat{\sigma}^2 e_r e_r^{\mathrm{T}})^{\frac{1}{2}} \Lambda_r^{-1}, \hat{C} = I^{-\frac{1}{2}} A_r (\Lambda_r^2 - \hat{\sigma}^2 e_r e_r^{\mathrm{T}})^{\frac{1}{2}}, \tag{2.2}$$

see also Loisel and Takane (2019). Then, we substitute $\hat{T}$ and $\hat{C}$ into (2.1), expressed as

$$\hat{X}_{\mathrm{RPCA}} = (\hat{x}_{\mathrm{RPCA}_{ij}})_{I \times J} = \hat{T}\hat{C}^{\mathrm{T}} = B_r (\Lambda_r - \hat{\sigma}^2 \Lambda_r^{-1}) A_r^{\mathrm{T}}.$$

Moreover, we just replace $x_{ij}^{\#}$ with $\hat{x}_{\mathrm{RPCA}_{ij}}^{\#}$ and still keep the original value of $x_{0_{ij}}^{*}$. Thus, the final estimatorof X_0 is $\hat{X} = (\hat{x}_{ij})_{I \times J}$, where $\hat{x}_{ij}^{\#} = \hat{x}_{\mathrm{RPCA}_{ij}}^{\#}$ and $\hat{x}_{ij}^{*} = x_{0_{ij}}^{*}$.

3. Numerical simulation

In this section, we select MSE and MAE to examine the estimation accuracy of the RPCA method, expressed as

$$\mathrm{MSE}(\hat{x}_{ij}^{\#}) = \frac{1}{n} \sum (\hat{x}_{ij}^{\#} - x_{0_{ij}}^{\#})^2, \mathrm{MAE}(\hat{x}_{ij}^{\#}) = \frac{1}{n} \sum | \hat{x}_{ij}^{\#} - x_{0_{ij}}^{\#} |. \tag{3.1}$$

The smaller these two values are, the better the estimation accuracy is.

We select a real data set named Seeds data set, it measures 210 wheat samples in 7 geometrical properties of kernels, see also Charytanowicz, et al. (2010). It is completed, we take the missing ratio values as M = 10%,20%,30%,40% in simulation, estimate the missing values with the RPCA method and with the mean method as comparison. We display the MSE and MAE values in Figure 1. As shown in Panel (a), the MSE values of the RPCA method range from 1.826 to 1.869, those of the mean method range from 1.867 to 1.887. As shown in Panel (b), the MAE values of the RPCA method range from 0.839 to 0.856, those of the mean method range from 0.853 to 0.866. The values of the RPCA method always below those of the mean method, it means that the RPCA method has better estimation effect than the mean method.

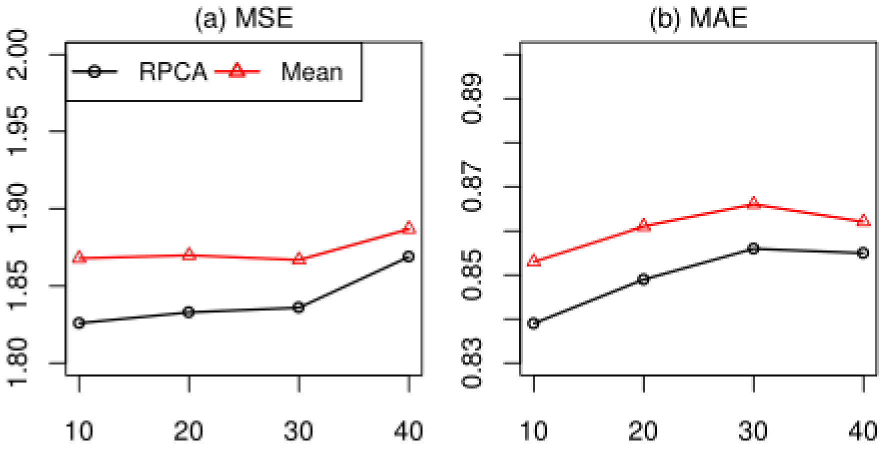

Figure 1: Comparison results of the seeds data set

4. Discussions

The RPCA method corrects the biases of the algorithm. It is precisely because of the advantage of the RPCA method in solving the overfitting problem of missing data, the method is worthy of application in more missing data problems.

In the future, we can research the effect of the RPCA method on the missing data with larger missing ratios. We can further search a criteria as a preliminary preparation to judge the applicability of the RPCA method to missing data sets.

References

1. Bellas, A., Bouveyron, C., Cottrell, M., and Lacaille, J. (2013). Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA. Advances in Data Analysis andClassification, 7(3), 281-300.
2. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Lukasik, S., and Zak, S. (2010).Complete gradient clustering algorithm for features analysis of X-ray images. Information Technologies in Biomedicine, 2, 15-24.
3. Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysismethods. Journal de la Societe Francaise de Statistique. 153(2), 1-21.
4. Josse, J., Chavent, M., Liquet, B., and Husson, F. (2012). Handling missing values with regularizediterative multiple correspondence analysis. Journal of Classification, 29(1), 91-116.
5. Loisel, S. and Takane, Y. (2019). Comparisons among several methods for handling missing data inprincipal component analysis (PCA). Advances in Data Analysis and Classification, 13(2), 495-518.