Online principal component analysis based onperturbation method

Chunjie Wei & 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*

Email address: wjzhenhua@126.com

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

Abstract: Principal component analysis (PCA) is widely used in data processing and dimensionality reduction. Its online version is useful in many modern applications where the data are large or constantly updated. We introduce an online PCA (OPCA) method based on perturbation matrix updating. The OPCA method based on the perturbation method utilizes the interlaced property of the eigenvalues of the covariance matrix under the rank-1 correction to recursively update and sort the eigenvalues, eigenvectors. Numerical example shows that the OPCA method reduces the computational complexity and can detect faults in time when faults exist.

Keywords: principal component analysis; perturbation matrix; rank-one update; online monitoring.

1 Introduction

Principal component analysis (PCA) is one of the most common methods to reduce dimension, which is simple and feasible. However, in the modern big data environment, the speed of data updating is rapidly, and the demand for time and space is also greatly increased. The continuous updating of data requires the continuous updating of output results. The PCA method has not the ability of time-varying tracking, so we need to consider the online form of PCA-Online PCA (OPCA). Li, et al. (2018) proposed the OPCA based on random approximation. However, these algorithms have slow convergence speed and their performance largely depend on selection of step size.

Many scholars use recursive PCA to monitor data in real time, see also Elshenawy et al. (2010), Chenet al. (2011) and Mitz et al. (2019). Li et al. (2000) proved the convergence and convergence rate of the recursive algorithm. This paper introduces an OPCA based on perturbation method; see also Hegde et al. (2006). In perturbation method, PCA updating is reduced to finding the root of rational function by using the staggered property of the eigenvalues of a covariance matrix under rank-1 correction. Interestingly, this recursive method is accurate and produces the same results as offline PCA. Since the perturbation method is updated based on the matrix with rank-1, the computational complexity is significantly reduced and the storage is greatly saved.

The paper is organized as follows. In Section 2, we describe the perturbation method. The experiment analysis resultsillustrate the advantages of our approach for real data set in Section 3 while conclusions are presented in Section 4.

2 Perturbation method

Let $\mathbf{x}_i$ be p-dimensional random vectors, have mean vector $\bar{\mathbf{x}}_d = \sum_{i=1}^{d} \mathbf{x}_i / d$and sample covariance matrix $S_d = U_d \Lambda_d U_d^T$,where $U_d = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p)$ and$\Lambda_d = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$.

When the new sample $\mathbf{x}_{d+1}$ comes, the updated mean vector and sample covariance matrix arewritten as

$$\bar{\mathbf{x}}_{d+1}=\frac{1}{d+1}(d\bar{\mathbf{x}}_d+\mathbf{x}_{d+1}), \tag{2.1}$$

$$S_d=\frac{d}{d+1}S_d+\frac{1}{d+1}(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})^T. \tag{2.2}$$

By using U_d,Λ_d,$U_{d+1}$and Λ_{d+1}, equation (2.2)can be rewritten as

$$U_{d+1}((d+1)\Lambda_{d+1})U_{d+1}^T=dU_d\Lambda_d U_d^T+(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})^T. \tag{2.3}$$

Define $Q_{d+1}=U_d^T(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})$ as a rank-1 matrix, and substitute equation (2.3) to be expressed as

$$U_{d+1}((d+1)\Lambda_{d+1})U_{d+1}^T=U_d[d\Lambda_d+Q_{d+1}Q_{d+1}^T]U_d^T. \tag{2.4}$$

We decompose matrix$d\Lambda_d+Q_{d+1}Q_{d+1}^T$into$V_{d+1}D_{d+1}V_{d+1}^T$, where $V_{d+1}$is a orthogonal eigenvector matrix andD_{d+1} is diagonal matrix, andequation (2.4) becomes

$$U_{d+1}((d+1)\Lambda_{d+1})U_{d+1}^T=U_dV_{d+1}D_{d+1}V_{d+1}^TU_d^T. \tag{2.5}$$

It is obvious to get the recursive update form of the eigenvector matrix and the diagonal matrix as

$$U_{d+1}=U_dV_{d+1},\Lambda_{d+1}=D_{d+1}/(d+1). \tag{2.6}$$

Consider perturbation analysis of matrix $d\Lambda_d+Q_{d+1}Q_{d+1}^T$. When d is large, $d\Lambda_d+Q_{d+1}Q_{d+1}^T$approximates a diagonal matrix; that is, D_{d+1}will be close to $d\Lambda_d$, and V_{d+1}will be close to identity matrix$I$.Therefore, $Q_{d+1}Q_{d+1}^T$ (Q_{d+1} is the rank-1 matrix) is regarded as perturbation term of diagonal matrix $d\Lambda_d$.Using the approximations:$D_{d+1}=d\Lambda_d+P_\Lambda$ and $V_{d+1}=I+P_V$, where P_Λ and P_V arecorresponding perturbation matrices.The matrix $V_{d+1}D_{d+1}V_{d+1}^T$ can be expressed as

$$V_{d+1}D_{d+1}V_{d+1}^T=(I+P_V)(d\Lambda_d+P_\Lambda)(I+P_V)^T$$

$$=d\Lambda_d+P_\Lambda+D_{d+1}P_V^T+P_VD_{d+1}+dP_V\Lambda_dP_V^T+P_VP_\Lambda P_V^T.$$

If $dP_V\Lambda_dP_V^T$ and $P_VP_\Lambda P_V^T$can be ignored, the matrix $Q_{d+1}Q_{d+1}^T$ is expressed as

$$Q_{d+1}Q_{d+1}^T=P_\Lambda+D_{d+1}P_V^T+P_VD_{d+1}. \tag{2.7}$$

BecauseV_{d+1} is orthonormal, which satisfies $V_{d+1}V_{d+1}^T=I$. Supposing$P_VP_V^T\approx 0$, we get$P_V=-P_V^T$.SinceP_Λ andD_{d+1} are diagonal matrices, thesolution for perturbation matrices are as follows

$$P_\Lambda(i,i)=q_i^2,$$

$$\left.\begin{array}{l} P_V(i,j)=\frac{q_iq_j}{\lambda_j+q_j^2-\lambda_i-q_i^2},i\neq j \\ P_V(i,i)=0 \end{array}\right\}, \tag{2.8}$$

whereq_i is the i-th element of Q_{d+1},λ_i and λ_j are the diagonal elements of the eigen diagonal matrix $d\Lambda_d$.These perturbation matrices P_Λ and P_V can be calculated. Therefore, V_{d+1} and D_{d+1} can be obtained from $D_{d+1}=d\Lambda_d+P_\Lambda$ and $V_{d+1}=I+P_V$. According to equation (2.6), U_{d+1} and Λ_{d+1} have been updated.

Generally, we assign the same weighting coefficient to all samples, called $\beta_{d+1}=1/(d+1)$ as forgetting factor, where $\beta_{d+1}\in(0,1)$. The corresponding covariance matrix is expressed as

$$S_d=(1-\beta_{d+1})S_d+\beta_{d+1}(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})(\mathbf{x}_{d+1}-\bar{\mathbf{x}}_{d+1})^T. \tag{2.9}$$

According to the above derivation process, we have $D_{d+1} = (1-\beta_{d+1})\Lambda_d + \beta_{d+1}P_\Lambda$ and $V_{d+1} = I + P_V$. The elements of matrix P_Λ remain unchanged, and the elements of P_V arerewritten as

$$\left.\begin{array}{c} P_V(i,j) = \frac{\beta_{d+1}q_i q_j}{\lambda_j+\beta_{d+1}q_j^2-\lambda_i-\beta_{d+1}q_i^2}, i \neq j \\ P_V(i,i) = 0 \end{array}\right\}. \quad (2.10)$$

In this way, we complete the recursive process of sample covariance matrix.

Determine the number of principal components m according to the cumulativepercent variance (CPV) $\eta_m = (\sum_{i=1}^{m}\lambda_i)/(\sum_{i=1}^{p}\lambda_i)$. For a real data set, make $\eta_m \geq 85\%$.

3 Experiment analysis

3.1 T2-statistic

When updating the $d+1$ sample, score vector t is expressed as$t_{d+1} = (\mathbf{x}_{d+1}^T U_{d+1}^{(m)})^T = \left(U_{d+1}^{(m)}\right)^T \mathbf{x}_{d+1}$,where $U_{d+1}^{(m)}$ is the first m column of the updated eigenvector matrix.The T2-statistic is as follows

$$T_{d+1}^2 = t_{d+1}^T H^{-1} t_{d+1} = \mathbf{x}_{d+1}^T U_{d+1}^{(m)} \Lambda_{d+1}^{-1} \left(U_{d+1}^{(m)}\right)^T \mathbf{x}_{d+1}, \quad (3.1)$$

where H is the diagonal matrix constituted by the standard deviation of the score vector t. The control limit of T2-statistics is expressed as

$$T_{d+1}^2 \sim \frac{m[(d+1)^2-1]}{(d+1)[(d+1)-m]} F(m, d+1-m). \quad (3.2)$$

When the significance level is 0.05, the control limit can be determined.

3.2 Statistical accuracy

We establish relative error to evaluate estimation accuracy of the method, see also Cardot et. al (2017).Let $W_m = UU^T$ bethe orthogonal projector on this eigenspace. Given a matrix $\widehat{U}$ of estimated eigenvectors suchthat $\widehat{U}^T\widehat{U} = I_m$, we consider the orthogonal projector $W_m = \widehat{U}\widehat{U}^T$and measure the relative error by

$$L(\widehat{W}_m) = \frac{\|\widehat{W}_m - W_m\|_F^2}{\|W_m\|_F^2} = 2(1 - \frac{tr[\widehat{W}_m W_m]}{m}), \quad (3.3)$$

where$\|\cdot\|_F$ denotes the Frobenius norm.

3.3 Numerical application

In this section, we select the wine data set to test the performance of the OPCA method. The wine data set from the UCI database includes 13 different substances in three wines and 178 sample data. For the data set, we take $d = n/2$ as the training sample, and the remaining $1/2$ samples for online learning. According to CPV criterion, we choose the number of principal components m=6 for experiment.

Figure 1 shows T2-statistic values at each moment, and the solid red line is the 95% control limit. It can be seen from panel (a), there is no over-limitphenomenon in the updating process of wine data set. In order to test the ability of the method in monitoring faults, we replace the 120-th sample with outliers.As shown in panel (b), there are obvious faults in detection of 120-th sample, indicating that the OPCA method based on the perturbation method is very reliable.

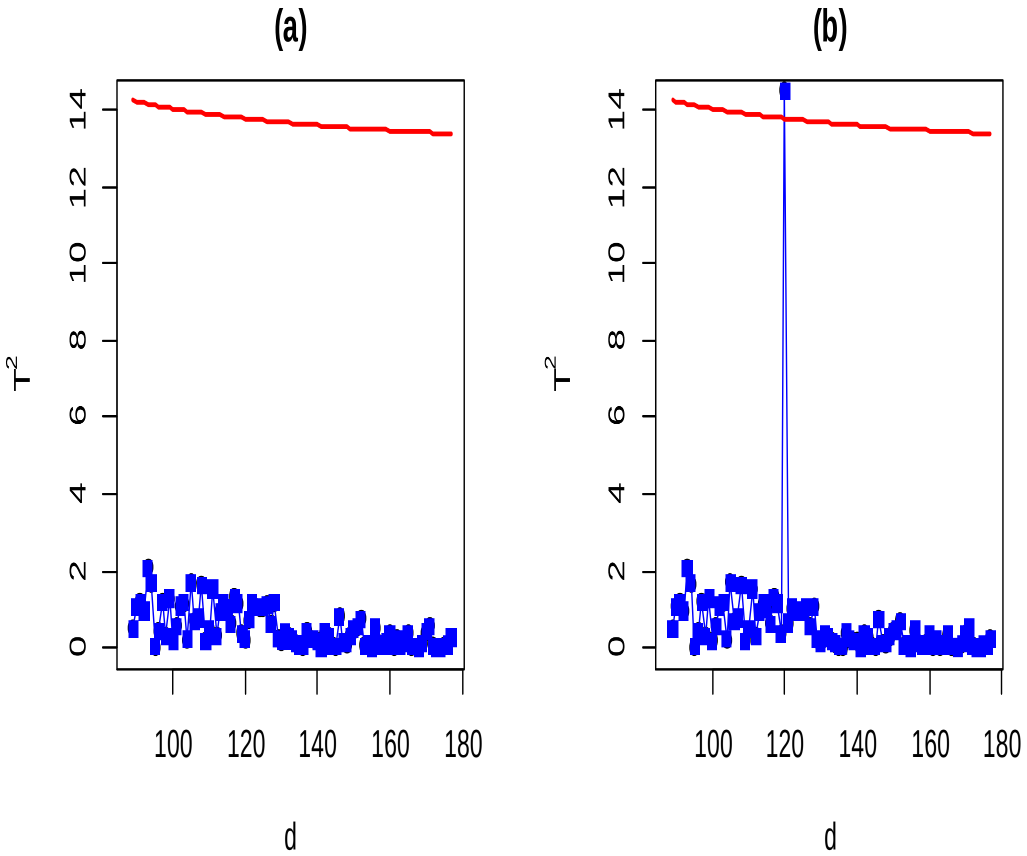

Figure 1 The change of T²-statistic in wine data set

We calculate the relative error $L(\widehat{W}_m) = 0.003293414$, and the running time is 0.667038 seconds. It shows that the OPCA method has high accuracy and fast running time.

4 Conclusions

Compared with the offline PCA method, the OPCA method based on the perturbation matrix updating uses a perturbation matrix to update the data set, and adapts to the time-varying characteristics of the data set.The online update not only makes full use of the information of each sample, but also saves a lot of storage.The results of numerical application show that the method can detect faults in time,improve the online monitoring effect of time-varying process, and improve the estimation accuracy.

References

1. Chen, H. F., Fang, H. T., and Zhang, L. L. (2011). Recursive estimation for ordered eigenvectors of symmetric matrix with observation noise. Journal of Mathematical Analysis and Applications, 382(2), 822-842.
2. Cardot, H. and Degras, D. (2017). Online principal component analysis in high dimension: which algorithm to choose?. International Statal Review, 86(1), 29-50.
3. Elshenawy, L. M., Yin, S., Naik, A. S., and Ding, S. X. (2010). Efficient recursive principal component analysis algorithms for process monitoring. Industrial & Engineering Chemistry Research, 49(1), 252-259.
4. Hegde, A., Principe, J. C., Erdogmus, D., Ozertem, U., Rao, Y. N., andPeddaneni, H. (2006). Perturbation-based eigenvector updates for on-line principal components analysis and canonical correlation analysis. Journal of Vlsi Signal Processing Systems for Signal Image & Video Technology, 45(1/2), 85-95.
5. Li, C. J., Wang, M., Liu, H., and Zhang, T. (2018). Near-optimal stochastic approximation for online principal component estimation. Mathematical Programming, 167(1), 75–97.

6. Li, W., Yue, H. H., Valle-Cervantes, S.,and Qin, S. J. (2000). Recursive pca for adaptive process monitoring. Journal of Process Control, 10(5), 471-486.
7. Mitz, R. andShkolnisky, Y. (2019). Roipca: an online pca algorithm based on rank-one updates. arXiv:1911.11049.