2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51.21

An Accurate Incremental Principal Component Analysis method with capacity of update and downdate

Wang Li, Chen Shuo and Wu Chengdong

School of Information Science & Engineering, Northeastern University, Shenyang, China

Abstract—Principal Component Analysis is a popular and powerful method in many machine learning task, The traditional PCA is implemented in batch mode, which means the much lower efficiency, especially for the task which training dataset is updated or downdated frequently, so it is reasonable to develop the incremental version of PCA. However, most of the existing incremental PCA is based on approximation with high estimation error, or lack of the downdate function. In this paper, a new accurate IPCA algorithm (AIPCA) which can provide both update and downdate capacity with higher accuracy because of direct accurate algebraic derivation is proposed based on the matrix additive modification. Experimental analysis is also given for evaluating the time cost and calculation accuracy of the AIPCA, the result demonstrates that the proposed method has high calculation accuracy and acceptable time consuming.

Keywords-Incremental PCA; update; downdate; SVD; Accurate

1. Introduction

Principal Component Analysis (PCA) is one of the most useful techniques in multivariate analysis, and is employed by many applications of pattern recognition and signal processing in last decades [1] [2] [3] [4]. However, the original PCA is performed in batch mode, which means all training data need to be included in the stage when calculation is carried on. When some new observations are incorporated in training dataset or some are removed out, the dataset's projection result of PCA has to be recalculated directly from all training data, This is a extremely time-consuming process. Furthermore, if the data are obtained from rank-r manifold which embedded in dimension-d observation space ($r \ll d$), only the form r principal components (PCs) are valuable for revealing the data pattern, this suggested that a new PCA algorithm can only focus on the form r PCs calculation in this condition to achieve low time computational complexity but not all d PCs which is implemented in the original PCA. To overcome the disadvantage of time-consuming of the original PCA when update or downdate training dataset, some incremental versions are developed by researchers.

The new PCs generated by the existing incremental PCA (IPCA) algorithms [5] [6] [7] are the estimation of batch mode PCA. The candid covariance-free IPCA (CCIPCA) [5] is one superior method of the recent works, it gives a simple form and with higher estimation accuracy and faster computing speed. But the CCIPCA's estimation accuracy is limited by the number of samples, or in other words, the number of iterations. The PCs of CCIPCA will converge to batch mode result when new data are added in training set continually, however it does not guarantee an acceptable estimation error upper bound between the two adjacent iterations. In some learning task, training set is often changed frequently, so the unguaranteed error may led to unreasonable PCs which may cause failure of the task. In viewing of this limitation, a new IPCA called SVDU-IPCA [9] based on a singular value decomposition (SVD) updating algorithm [10] is proposed. The SVDU-IPCA have been mathematically proved that the estimation error is bounded, but it only provide the updating function, and cannot deal with the downdating situation when some data are removed from

E-mail address: wl1986_ren_ren@163.com

training set. Besides this deficiency, Matthew Brand [11] also points out that the SVD updating algorithm employed by SVDU-IPCA requires a full SVD of a dense matrix on each update, which reduces the performance. So there are strong reasons to design a new accurate IPCA algorithm which can calculate new PCs incrementally through direct accurate algebraic derivation.

In this paper, we focus on the solution of a kind of PCA learning issue that is very common in the field of image or video pattern recognition, in which sample dimension d is usually larger than training samples number q. A new accurate IPCA algorithm (AIPCA) which can provide both update and downdate capacity is proposed in this paper based on the matrix additive modification presented in [11]. Experimental analysis is also given for evaluating the time cost and calculation accuracy of the AIPCA.

2. Batch updating/downdating of SVD

Incremental version of SVD has been studied in several literatures [10] [11] [12] [13], A derivation of batch updating/downdating algorithm of SVD proposed by Matthew Brand in [11] is employed in this paper for the theoretical basis of AIPCA. The capacity of update/downdate is obtained through matrix additive modification.

Suppose that the intrinsic dimension of sample data is r, let data matrix $\mathbf{Y} \in \mathbb{R}^{p\times q}$ have economy SVD $\mathbf{Y} = \mathbf{U}_y \mathbf{S}_y \mathbf{V}_y^{\mathrm{T}}$ with $\mathbf{S}_y \in \mathbb{R}^{r\times r}$, so the update question can be translated into searching a incremental SVD solution of $\begin{bmatrix}\mathbf{Y} & \mathbf{B}\end{bmatrix} = \mathbf{U}_u \mathbf{S}_u \mathbf{V}_u^{\mathrm{T}}$, where $\mathbf{S}_u \in \mathbb{R}^{r\times r}$, $\mathbf{B} \in \mathbb{R}^{p\times l}$. And the downdate question can be translated into finding a incremental SVD solution of $\mathbf{Y}' = \mathbf{U}_d \mathbf{S}_d \mathbf{V}_d^{\mathrm{T}}$, where $\mathbf{Y}'$ is a submatrix of $\mathbf{Y} = \begin{bmatrix}\mathbf{Y}' & \mathbf{D}\end{bmatrix}$, $\mathbf{S}_d \in \mathbb{R}^{r\times r}$, $\mathbf{D} \in \mathbb{R}^{p\times l}$. The procedure of the SVD update/downdate algorithm [11] is described as follows:

2.1. Update

1) Obtain the QR decomposition

$$\mathbf{Q}_1 \mathbf{R}_1 = \left(\mathbf{I} - \mathbf{U}_y \mathbf{U}_y^{\mathrm{T}}\right)\mathbf{B} \tag{1}$$

$$\mathbf{Q}_2 \mathbf{R}_2 = \left(\mathbf{I} - \begin{bmatrix}\mathbf{V}_y \\ \mathbf{0}_{l\times r}\end{bmatrix}\begin{bmatrix}\mathbf{V}_y & \mathbf{0}_{l\times r}\end{bmatrix}\right)\begin{bmatrix}\mathbf{0}_{(q-l)\times l} \\ \mathbf{I}_{l\times l}\end{bmatrix} \tag{2}$$

2) Suppose the rank of $\mathbf{R}_1$ and $\mathbf{R}_2$ are r_1 and r_2, obtain the matrixes $\mathbf{Q}_1'$ and $\mathbf{Q}_2'$ which are formed by the former r_1 and r_2 columns of $\mathbf{Q}_1$ and $\mathbf{Q}_2$ respectively, $\mathbf{R}_1'$ and $\mathbf{R}_2'$ which are formed by the former r_1 and r_2 rows of $\mathbf{R}_1$ and $\mathbf{R}_2$ respectively.

3) Obtain the smaller sparse matrix

$$\begin{aligned}\mathbf{K} = &\begin{bmatrix}\mathbf{U}_y^{\mathrm{T}}\mathbf{B} \\ \mathbf{R}_1'\end{bmatrix}\begin{bmatrix}\mathbf{0}_{l\times(q-l)} & \mathbf{I}_{l\times l}\end{bmatrix}\begin{bmatrix}\mathbf{V}_y \\ \mathbf{0}_{l\times r}\end{bmatrix} \quad \mathbf{R}_2'^{\mathrm{T}}\Bigg] \\ &+\begin{bmatrix}\mathbf{S}_y & \mathbf{0}_{r\times r_2} \\ \mathbf{0}_{r_1\times r} & \mathbf{0}_{r_1\times r_2}\end{bmatrix}\end{aligned} \tag{3}$$

4) Obtain the SVD decomposition

$$\mathbf{K} = \mathbf{U}_k \mathbf{S}_k \mathbf{V}_k^{\mathrm{T}},\ \mathrm{S}_k \in \mathbb{R}^{r\times r} \tag{4}$$

5) Obtain the SVD decomposition of $\begin{bmatrix}\mathbf{Y} & \mathbf{B}\end{bmatrix}$

$$\begin{aligned}\begin{bmatrix}\mathbf{Y} & \mathbf{B}\end{bmatrix} &= \mathbf{U}_u\mathbf{S}_u\mathbf{V}_u^{\mathrm{T}} \\ &= \left[\begin{bmatrix}\mathbf{U}_y & \mathbf{Q}_1'\end{bmatrix}\mathbf{U}_k\right]\mathbf{S}_k\left[\begin{bmatrix}\begin{bmatrix}\mathbf{V}_y \\ \mathbf{0}_{l\times r}\end{bmatrix} & \mathbf{Q}_2'\end{bmatrix}\mathbf{V}_k\right]\end{aligned} \tag{5}$$

2.2. Downdate

1) Obtain the QR decomposition

$$\mathbf{Q}_1\mathbf{R}_1 = \left(\mathbf{U}_y\mathbf{U}_y^{\mathrm{T}} - \mathbf{I}\right)\mathbf{D} \tag{6}$$

$$\mathbf{Q}_2\mathbf{R}_2 = \left(\mathbf{I} - \mathbf{V}_y\mathbf{V}_y^{\mathrm{T}}\right)\begin{bmatrix}\mathbf{0}_{(q-l)\times l} \\ \mathbf{I}_{l\times l}\end{bmatrix} \tag{7}$$

2) The same as the second step of update procedure, obtain $\mathbf{Q}_1'$, $\mathbf{Q}_2'$, $\mathbf{R}_1'$, $\mathbf{R}_2'$.

3) Obtain the smaller sparse matrix

$$\begin{aligned}\mathbf{K} = &\begin{bmatrix}-\mathbf{U}_y^{\mathrm{T}}\mathbf{D} \\ \mathbf{R}_1'\end{bmatrix}\begin{bmatrix}\begin{bmatrix}\mathbf{0}_{l\times(q-l)} & \mathbf{I}_{l\times l}\end{bmatrix}\mathbf{V}_y & \mathbf{R}_2'^{\mathrm{T}}\end{bmatrix} \\ &+\begin{bmatrix}\mathbf{S}_y & \mathbf{0}_{r\times r_2} \\ \mathbf{0}_{r_1\times r} & \mathbf{0}_{r_1\times r_2}\end{bmatrix}\end{aligned} \tag{8}$$

4) The same as the forth step of update procedure, obtain $\mathbf{U}_k$, $\mathbf{S}_k$ and $\mathbf{V}_k$.

5) Obtain the SVD decomposition $\mathbf{Y}' = \mathbf{U}_d\mathbf{S}_d\mathbf{V}_d^{\mathrm{T}}$, where $\mathbf{U}_d = \left[\begin{bmatrix}\mathbf{U}_y & \mathbf{Q}_1'\end{bmatrix}\mathbf{U}_k\right]$, $\mathbf{S}_d = \mathbf{S}_k$, $\mathbf{V}_d$ is formed by the former $q-l$ rows of $\begin{bmatrix}\mathbf{V}_y & \mathbf{Q}_2'\end{bmatrix}\mathbf{V}_k$.

3. Accurate IPCA (AIPCA)

There are three ways to achieve PCA, which are eigendecomposition of data covariance matrix, SVD of data matrix and SVD of data inner product matrix respectively. AIPCA focus on resolution of high dimension accurate IPCA issue that has the sample dimension d larger than training samples number q, which is very common in the field of image or video pattern recognition, it means the size of data inner product matrix is smaller than the other two. So inspired by SVDU-IPCA, SVD of data inner product matrix is chosen to develop AIPCA.

3.1. Update

Let data matrix $\mathbf{X} = [x_1, x_2, \cdots, x_q]$ has q observations of dimension p. Suppose all data are sampled from a manifold, and the lowest dimension of linear space in which the data manifold can be embedded is r, $\mathbf{A}$ is new data matrix with c observations. The inner product matrix of $\mathbf{X}$ is $\sum_1 = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, so the inner product matrix of new data matrix $\begin{bmatrix}\mathbf{X} & \mathbf{A}\end{bmatrix}$ is $\sum = \begin{bmatrix}\mathbf{X} & \mathbf{A}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{X} & \mathbf{A}\end{bmatrix}$. It is easy to prove that the inner product matrix is real symmetric positive semidefinite , so in rank reserving SVD decomposition of rank r, according to the derivation in [9], $\sum$ can be written as

$$\sum = P^{\mathrm{T}}P = \begin{bmatrix}\sum_1 & \sum_2 \\ \sum_2^{\mathrm{T}} & \sum_3\end{bmatrix} = \begin{bmatrix}P_1 & Q_1\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}P_1 & Q_1\end{bmatrix} \tag{9}$$

where $\Sigma_1 = P_1^{\mathrm{T}} P_1 = U_1 \Lambda_1 U_1^{\mathrm{T}}$ is the rank r SVD decomposition of Σ_1, and the matrix $P_1 = \Lambda_1^{1/2} U_1^{\mathrm{T}}$ is the PCA projection result, $\Lambda_1 \in \mathbb{R}^{r \times r}$. Q_1 can be obtained as $Q_1 = \Lambda_1^{-1/2} U_1^{\mathrm{T}} \Sigma_2$ [9].

It can be see that $P_1 = I_{r \times r} \Lambda_1^{1/2} U_1^{\mathrm{T}}$ is rank r SVD of P_1, so using the SVD updating algorithm presented in section 2 to get the decomposition of $\begin{bmatrix} P_1 & Q_1 \end{bmatrix}$ is a straightforward approach, that is

$$\begin{bmatrix} P_1 & Q_1 \end{bmatrix} = U_P S_P V_P^{\mathrm{T}} \tag{10}$$

then we can get

$$\Sigma = \begin{bmatrix} P_1 & Q_1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} P_1 & Q_1 \end{bmatrix} = V_P S_P^{\,2} V_P^{\mathrm{T}} = U \Lambda V^{\mathrm{T}} \tag{11}$$

where $U = V_P$, $\Lambda = S_P^{\,2}$, $V = V_P$.

Finally the PCA is achieved by the incremental SVD of inner product of new data matrix $\begin{bmatrix} \mathbf{X} & \mathbf{A} \end{bmatrix}$.

3.2. Downdate

In downdate procedure we redefine the notation $\mathbf{A}$, let $\mathbf{A} \in \mathbb{R}^{q \times c}$ is formed by the data which needed to be deleted from training sample matrix $\mathbf{X}$, and $\mathbf{X}_1$ is the reserved data, then a new training sample matrix after shifting the columns which will be deleted to the right side is $\hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{A} \end{bmatrix}$. The column shift is a SVD-invariant transform to the data matrix, which means the SVD of $\hat{\mathbf{X}}^{\mathrm{T}} \hat{\mathbf{X}}$ and $\mathbf{X}^{\mathrm{T}} \mathbf{X}$ is equal, so the inner product of $\hat{\mathbf{X}}$ can be re-written in rank reserving SVD as

$$\hat{\mathbf{X}}^{\mathrm{T}} \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1^{\mathrm{T}} \mathbf{X}_1 & \Sigma_2 \\ \Sigma_2^{\mathrm{T}} & \Sigma_3 \end{bmatrix} = P^{\mathrm{T}} P \tag{12}$$

where $U \Lambda V^{\mathrm{T}}$ is the singular value decomposition of $\hat{\mathbf{X}}^{\mathrm{T}} \hat{\mathbf{X}}$, then the SVD of $P = \begin{bmatrix} P_1 & Q_1 \end{bmatrix}$ is $I_{r \times r} \Lambda^{1/2} U^{\mathbf{T}}$, $\Lambda \in \mathbb{R}^{r \times r}$. So similar to update procedure, the SVD-downdating algorithm presented in section 2 can be employed to obtain the SVD of P_1, that is

$$P_1 = U_{P_1} S_{P_1} V_{P_1}^{\mathrm{T}} \tag{13}$$

Then the downdate PCA can be achieved as

$$\begin{aligned} \mathbf{X}_1^{\mathrm{T}} \mathbf{X}_1 &= P_1^{\mathrm{T}} P_1 \\ &= V_{P_1} S_{P_1} U_{P_1}^{\mathrm{T}} U_{P_1} S_{P_1} V_{P_1}^{\mathrm{T}} \\ &= V_{P_1} S_{P_1}^{\,2} V_{P_1}^{\mathrm{T}} \\ &= U_1 \Lambda_1 V_1^{\mathrm{T}} \end{aligned} \tag{14}$$

where $U_1 = V_1 = V_{P_1}$, $\Lambda_1 = S_{P_1}^{\,2}$.

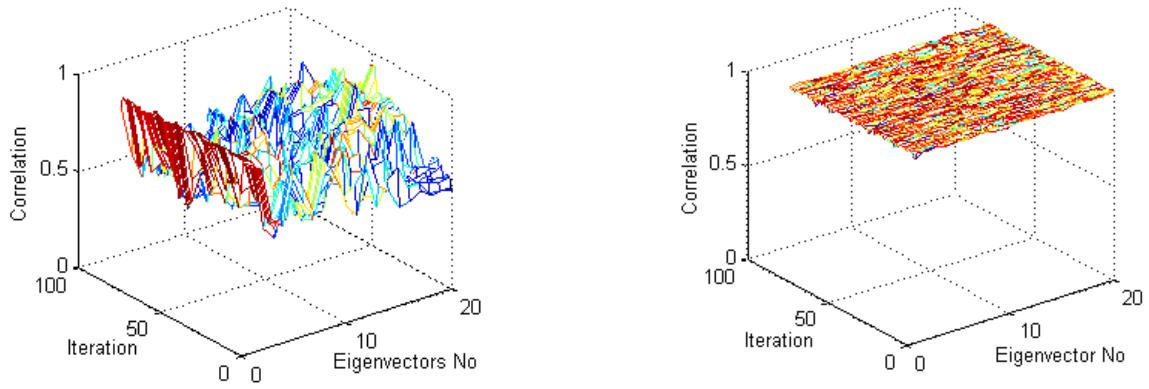
4. Experimental analysis

The time complexity of computing a full SVD is $\mathrm{O}\left(pq \bullet \min\left(p, q \right) \right)$, if other operation for calculating PCA based on SVD are pulsed, the complexity would be more higher. Many powerful algorithm such as CCIPCA has been proposed to overcome this question, however most of them sacrifice accuracy in exchange for efficiency, or just a single direction method like SVDU-IPCA that can only achieve a update task. Suppose there is a training dataset with samples number q, observation dimension p and intrinsic dimension r, for

AIPCA, the focus is kept on solve a subspace learning question which q is smaller than p and r are much less than p in a incremental manner, this is exactly the most conditions which faced by pattern recognition problem. If there are c samples needed to be added in or removed out, the QR decomposition of AIPCA would take $\mathrm{O}\left(p(r+c)^2\right)$ time, the SVD of sparse matrix takes $\mathrm{O}\left((r+c)^3\right)$ time, and the rotation of the subspaces takes $\mathrm{O}\left((p+q)(r+c)^2\right)$ time [11], which can save much time than $\mathrm{O}\left(pq\bullet\min(p,q)\right)$ because of $q<p$ and $r \ll p$.

Experiments have been implemented to evaluate the accuracy and efficiency of AIPCA against CCIPCA and batch mode PCA. We generate a high dimension but low rank dataset which satisfy high dimensional zero mean Gaussian distribution of with small additive noise randomly. The batch mode PCA is realized to obtain the standard subspace of PCA which is used to compare with AIPCA and CCIPCA. The inner product of component on principle axis between standard subspace and the testing algorithm's subspace (AIPCA's or CCIPCA's) is employed to evaluate the correlation which can indicate the accuracy of algorithm.

The correlation represented by inner product of two algorithms is shown as Fig.1. It can be seen that only the correlation of the pair of the first PCs of CCIPCA is near 1, the rest PCs cannot obtain a acceptable accuracy between two adjacent update iteration. However, for AIPCA there are only little error generated by numerical error. The experimental result demonstrates that AIPCA has high accuracy because of the direct accurate algebraic derivation. The time cost of batch mode PCA, AIPCA and CCIPCA when intrinsic of training set is growing is also given as Fig.2, we can see that the efficiency of AIPCA is lower than CCIPCA but beat batch mode PCA, this is acceptable at the premise of ensuring the higher accuracy.



(a) (b)

Figure 1. The correlation between batch mode PCA and CCIPCA (a), batch mode PCA and AIPCA (b).

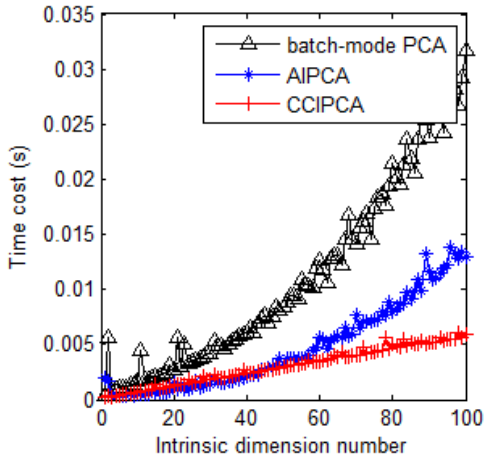


Figure 2. The time cost of batch-mode PCA, AIPCA and CCIPCA

5. Conclusion

A new accurate IPCA algorithm (AIPCA) which can provide both update and downdate capacity is proposed in this paper, we focus on the solution of a kind of PCA learning issue that is very common in the field of image or video pattern recognition, in which sample dimension d is usually larger than training samples number q. based on the matrix additive modification presented. The experiment shows that the proposed method has high calculation accuracy and acceptable efficiency. How to reduce the time cost is the focus of the future research.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 60874103) and the Fundamental Research Funds for the Central Universities (No. N100604018). The authors would like to thank for these support.

7. References

[1] Bakshi, R Bhavik, "Multiscale PCA with Application to Multivariate Statistical Process Monitoring", AIChE Journal, vol. 44, pp. 1596-1610, July 1998.

[2] F Castells, P Laguna, "Principal component analysis in ECG signal processing", Eurasip Journal on Advances in Signal Processing, vol. 2007, 2007.K. Elissa, "Title of paper if known," unpublished.

[3] J Yang, D Zhang and A F Frangi, et al., "Two-dimensional PCA: A new approach to appearance-based face representation and recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, pp. 131-137, January 2004.

[4] L Haiping, K N Plataniotis and A N Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects", IEEE Transactions on Neural Networks, vol. 19, pp. 18-39, January 2008.

[5] D Skocai, A Leonardis, "Weighted and Robust Incremental Method for Subspace Learning", Proceedings Ninth IEEE International Conference on Computer Vision, vol.2, pp. 1494-501, 2003.

[6] W Juyang, Z Yilu and H Wey-Shiuan, "Candid Covariance-Free Incremental Principal Component Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 1034-40, August 2003.

[7] L Yongmin, "On incremental and robust subspace learning", Pattern Recognition, vol. 37, pp. 1509-18, July 2004.

[8] Y Wongsawat, "Fast PCA via UTV Decomposition and Application on EEG Analysis", 2009 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5669-72, 2009.

[9] Z Haitao, Y Pong-Chi and J T Kwok, "A Novel Incremental Principal Component Analysis and Its Application for Face Recognition", IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), vol. 36, pp. 873-86, August 2006.

[10] Z Hongyuan, H D Simon, "On Updating Problems in Latent Semantic Indexing", SIAM Journal of Scientific Computing, vol. 21, pp. 782-791, September 1999.

[11] M Brand, "Fast low-rank modifications of the thin singular value decomposition", Linear Algebra and Its Applications, vol. 415, pp. 20-30, May 2006.

[12] J C Wan, K J Hyoung and L J Gyu, "On Updating the singular value decomposition", 1996 International Conference on Communication Technology Proceedings, vol. 2, pp. 675-678, 1996.

[13] J R Bunch, C P Nielsen, "Updating the Singular Value Decomposition", Numerische Mathematik, vol. 31, pp. 111-129, 1978.