주성분 분석을 위한 새로운 EM 알고리듬

안종훈 오종훈
포 항 공 과 대 학 교 물 리 학 과
{jonghune, jhoh}@postech.ac.kr

New EM algorithm for Principal Component Analysis

Jong-hune Ahn and Jong-Hoon Oh
Dept. of Physics, Postech

Abstract

We present an expectation-maximization algorithm for principal component analysis via orthogonalization. The algorithm finds actual principal components, whereas previously proposed EM algorithms can only find principal subspace. New algorithm is simple and more efficient than probabilistic PCA specially in noiseless cases. Conventional PCA needs computation of inverse of the covariance matrices, which makes the algorithm prohibitively expensive when the dimension of data space is large. This EM algorithm is very powerful for high dimensional data when only a few principal components are needed.

1. Introduction

Principal component analysis is a multivariate technique which reduces the dimensionality and decorrelates the output data. The reduced data can best preserve the variance of data, so that it is widely used in data compression, image processing, etc. Scientists who are dealing with high dimensional data with low dimensional structure may need dimensionality reduction ways such as PCA(principal component analysis), and variety of methods for PCA has been introduced in literatures.[7]

Nevertheless, existing methods for PCA have several shortcomings. The classical eigenvalues analysis is quick and exact method for a low dimensional data, but not suitable for very high-dimensional or large data set. Computing a full covariance matrix is very laborious, and specially inefficient when only a few principal components are required. PCA neural networks such as Hebbian type or auto-associative network are not optimally scaled gradient methods, and hold redundancy of computation in spite of biological motivations.[5,6] In fact, optimal gradient recalling makes the additive network model equal to noiseless probabilistic PCA or multiplicative update rule of non-negative matrix factorization. But noiseless probabilistic PCA algorithm has rotational ambiguity, which means that it finds only PCA subspace whose axes are linear combinations of orthogonal principal axes.[1,2] Also, it is time-consuming to compute matrix inversion per each step.

In this paper, we suggest a new EM algorithm for PCA via orthogonalization. The developed method provides two advantages. First, the EM update rules converge to the actual principal components and decorrelated output data as well as principal subspace. Second, the orthogonalization processes cancel with respect to matrix inversion,

so we don't have to do matrix inversion as it was done in noiseless probabilistic PCA algorithm, so that it serves a triple purpose of simple detection of PCA subspace, decorrelation of output data and orthogonal basis.

2. Equivalence for PCA

Dimensionality reduction, variance maximization and a linear orthogonal model are important properties of PCA. The dimensionality reduction is a fundamental property. It is a method for low dimensional treatment of data, so that dimensional reduction is naturally required. Second property is that the reduced output data should best contain variance of data. That is, square error function should be minimized and first axis should be oriented along the direction in which the data has its highest variance. Similarly, subsequent axes are oriented so as to account for as much as possible of the variance in the data, subject to the constraint that they must be orthogonal to preceeding axes. The fact that PCA defines a linear, orthogonal model space gives it favorable computational properties, though it is its main limitation. Then one obtains orthogonal basis and decorrelated data as well as PCA subspace. Therefore PCA may be exactly executed by three independent constraints - least square error, orthogonal basis, decorrelated output data.[7]

$$\varepsilon = ||V - WH||^2 \quad (1)$$

$$W^T W = D_w^2 \quad (2)$$

$$HH^T = D_h^2 \quad (3)$$

(where D is the diagonal norm matrix of W column vectors or H row vectors[8] and W is orthogonal basis and H is decorrelated output data.)

Least square error is a good error measure for finding PCA subspace, but not for decorrelating output data and orienting toward actual principal direction. Orthogonality of the basis doesn't mean decorrelation of output data. Above three constraints assure us of actual principal component. Because of magnitude ambiguity, solutions under the constraints justly form manifolds, not one point in abstract data space.

3. Orthogonalization

With respect to matrix factorization we should find W and H that approximate V(input data) by the product WH.

$$V \approx WH \quad (4)$$

There are two approaches for actual PCA EM algorithm. - noiseless probabilistic PCA and nonnegative matrix factorization[3].

Noiseless probabilistic PCA

Noiseless probabilistic PCA is a fast EM algorithm for finding PCA subspace from high dimensional data space. It can also be proven in the framework of deterministic model, where H or W of each step is determined to best minimize the square error from the given W or H.[4] The EM algorithm shows the fastest convergence to PCA subspace. However, because of rotational ambiguity, the column vectors of converged W are not actual principle components, but linear combinations of them. Maximum likelihood for probabilistic PCA ensures only least square error : More operations, which should not break fast monotonic convergence to least square error, are needed. Fortunately, WH is invariant under matrix transformation and inverse transformation including orthogonalization, and one can naturally apply the orthogonality to each EM-step. It is not amazing that this orthogonality removes calculations of matrix inversion and simplify the update rule.

Noiseless probabilistic PCA algorithm has the update rule with

$$H^{new} = (W^T W)^{-1} W^T V \quad (5)$$

$$W^{new} = VH^T (HH^T)^{-1} \quad (6)$$

This iterative algorithm gives monotonic convergent result of PCA subspace. Expectation step can be transformed by an arbitrary matrix A which orthogonalizes the given matrix W. So the E step may be written as

$$W^{new} = WA \quad (7)$$

$$H^{new} = A^{-1}(W^T W)^{-1} W^T V$$

$$= (W^{new\,T} W^{new})^{-1} W^{new\,T} V \quad (8)$$

Obviously, matrix product WH is invariant under the transformation A and inverse transformation A^{-1}. W is orthogonalized about column vectors, but H is not yet about row vectors. However, keeping

step with monotonic convergence to least square error the H is gradually converged to an orthogonal matrix. The matrix inversion $\{(W^{new})^T W^{new}\}^{-1}$ in H^{new} update rule is diagonal matrix, and E step rule may be rewritten as simpler form without calculation of matrix inversion.

$$W^{new} = F(W) \tag{9}$$

$$H^{new} = \frac{W^{new\,T} V}{W^{new\,T} W^{new}\, 1_H} \tag{10}$$

Similarly, in the M step,

$$H^{new} = F(H) \tag{11}$$

$$W^{new} = \frac{V H^{new\,T}}{1_W\, H^{new} H^{new\,T}} \tag{12}$$

where F is orthogonalizing function[9] and 1_A is ones matrix whose size is the same to matrix A. The matrix division is changed from matrix inversion to componentwise division.

Least square error formulation of NMF

There are two update rules for non-negative matrix factorization(NMF). One is to minimize generalized Kullback-Leibler divergence which is related to parts based learning, the other is to minimize square error function. The latter update rule minimizes the square error monotonically like EM algorithm with non-negativity. However the non-negativity constraint is also free under orthogonalization. The result is the same as above algorithm.(Eq. 9~12)

The multiplicative update rule for least square error is

$$H^{new} = H \odot \frac{W^T V}{W^T W\, H} \tag{13}$$

$$W^{new} = W \odot \frac{V\, W^T}{W\, H\, H^T} \tag{14}$$

($\odot$ is elementwise product and matrix division is elementwise division)

The convergent property of the rule is the same as noiseless probabilistic PCA algorithm. It is a good compromise between speed and ease of implementation for square error minimization. However numerical division restricts the application to positive matrix. The constraint is removed through orthogonalization. Because of orthogonality of W and H, W^TW and HH^T are diagonal, so that the update rules is reduced to Eq. 9~12. Therefore the same update rule is obtained.

References and Notes

[1] Michael E. Tipping, Christopher M. Bishop Probabilistic Principal Component Analysis, Technical Report NCRG/97/010, Neural Computing Research Group, Aston University, September 1997.

[2] Sam Roweis, EM algorithms for PCA and SPCA, NIPS'97

[3] Daniel D. Lee, H. Sebastian Seung, Algorithms for Non-negative Matrix Factorization, NIPS'00

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Royal. Stat. Soc.(1977) 39, 1-38.

[5] Erkki Oja, Principal components, Minor components, and Linear neural networks, Neural Networks.(1992) 5, 927-935

[6] P. Baldi and K. Hornik, Neural networks and Principal component analysis : learning from examples without local minima, Neural Networks.(1989) 2, 53-58.

[7] Jolliffe, IT(1986), Principal Component Analysis, New York : Springer-Verlag.

[8] $d_{ij} = \delta_{ij} \sqrt{W_i^T W_j}$ or $d_{ij} = \delta_{ij} \sqrt{H_i H_j^T}$

[9] F is a continuous mapping which satisfies

$$W' = F(W)$$

if W is orthogonal, $W' = W$

else $W' \neq W$, W' is orthogonal.