Probabilistic Principal Component Analysis

Ashutosh Kakadiya, Harsh Patel, Nisarg Tike, Prerak Raja, Riddhesh Sanghvi

School of Engineering and Applied Science
Ahmedabad University

***Abstract*—Principal Component Analysis (PCA) is the most widely used statistical technique for dimensionality reduction. One major drawback of this method is that it is very sensitive to outliers. Also, conventional PCA does not offer a way to combine PCA models. In this report, PCA is formulated within a maximum-likelihood framework, based on a specific form of Gaussian latent variable model. This leads to a well-defined mixture model for probabilistic principal component analysers, whose parameters can be determined using an EM algorithm.**

Keywords: Principal component, PPCA, Estimation Maximization (EM)

I. Introduction

Principal component analysis (PCA) is one of the most popular techniques for processing, compressing and visualising data, but it has some limitations. The covariance matrix needs to be calculated and is not robust with missing data and outliers. Probabilistic Principal Component Analysis addresses limitations of classical PCA. PPCA can be used as a general Gaussian density model in addition to reducing dimensions. Maximum-likelihood estimates can be computed for elements associated with principal components and it captures dominant correlations with few parameters, multiple PCA models can be combined as a probabilistic mixture.

II. Latent Variable Model

Let t be a d-dimensional vector and let x be a q-dimensional vector. Also W is a dxq-matrix which relates t and x. Let μ permit the model to have non-zero mean. ϵ is a gaussian noise modeled as $\epsilon \sim N(0, \Psi)$. Conventionally the latent variable x is taken as $x \sim N(0, I)$. Based on the above given parameters factor analysis model is defined as

$$t = Wx + \mu + \epsilon \qquad (1)$$

t modeled as above can be intuitively visualized as a Gaussian variable such that $t \sim N(\mu, WW^T + \psi)$. PCA linkage to this factor analysis model can be found in [1].

III. Probabilistic PCA Model

The use of isotropic Gaussian model $N(0, \sigma^2 I)$ as ϵ in equation (1) implies that t can be conditionally recreated given probalility distribution of x in t-subspace as,

$$t|x \sim N(Wx + \mu, \sigma^2 I) \qquad (2)$$

In above given conditional model, x is normally distributed with 0 mean and I variance, such that $x \sim N(0, I)$, thus the distribution of t can be represented as

$$t \sim N(\mu, C) \qquad (3)$$

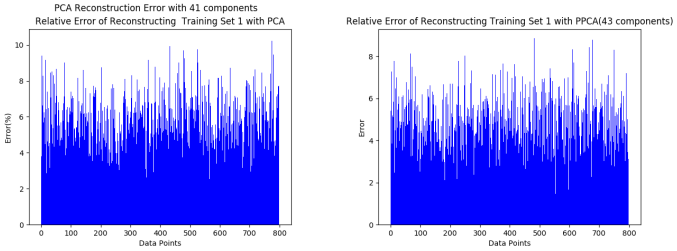

Fig. 1.

IV. Expectation-Maximization

It is an iterative process to estimate parameters consisting of two steps for each iteration.
Expectation (data step): complete all hidden and missing variables (or latent variables) from current set of parameters.
Maximization (likelihood step): Update set of parameters, using Maximum Likelihood Estimator, from complete set of data from previous step.
Likelihood obtained from MLEs guaranteed to improve in successive iterations. Continue iterations until negligible improvement is found in likelihood

V. Advantage of EM algorithm in PPCA model

Convergence: The only stable local extremum is the global maximum at which the true principal subspace is found.
Complexity: Methods that explicitly compute the sample covariance matrix have complexities $O(n^2)$. E-M algorithm does not require computation of sample covariance matrix, O(dnq) Mixtures of probabilistic PCA models can be formulated in a principled way and trained using the EM algorithm. In PPCA-EM approaches, the algorithm finds an approximate principal subspace and the approximate principal component projections regardless of the missing data and then it estimates the missing data for each individual volume.

VI. Results

We generated graphs on PCA PPCA reconstruction of training as well as test data set and successfully using EM and generated precision and covariance matrix of generated data set as shown in fig 1.

References

[1] Tipping, M. E., and C. M. Bishop. Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 61, No.3, 1999, pp. 611622.