Probabilistic Principal Component Analysis

Hemang Jethva, Jainil Vachhani, Nand Parikh, Palash Hariyani, Parth Shah

School of Engineering and Applied Science
Ahmedabad University

***Abstract*—Principal component analysis(PCA) is a well perceived technique for dimensionality reduction, data analysis, but a conventional PCA is not a probabilistic model. Probabilistic PCA (PPCA) model absorbs the essense of factor analysis for finding the pricipal axes of a set of observed data vectors using maximum-likelihood estimation of the latent variables's parameter. An iterative model for finding principal subspaces using EM algorithm and considering the properties of maximum-likelihood technique is proposed here. Applications of PPCA spans from simply dimensionality reduction to more complex encounters like finding principal axes for an incomplete or corrupted data matrix.**

Keywords: Latent variables, PPCA, Factor Analysis, Estimation Maximization (EM).

I. Factor Analysis

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

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

PCA linkage to factor analysis model can be found in [1].

II. 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,

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

$$t \sim N(\mu, WW^T + \sigma^2 I) \qquad (3)$$

$$S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mu)(t_n - \mu)^T \qquad (4)$$

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 equation(3)

S in equation(4) is the covariance matrix. The relation matrix W and isotropic noise's variance σ^2 are the parameters that are to be found.

A. Closed Form

[1] has mentioned a direct method for finding W and σ^2 as follows,

$$\sigma^2 = \frac{1}{d-q} \sum_{j=q+1}^{j=d} \lambda_j \qquad (4)$$

$$W_{ML} = U_q(\Lambda_q - \sigma^2 I)^{1/2} R \qquad (5)$$

U_q has the eigen vectors of S with corresponding eigen values in Λ and R is just a rotation matrix. Even [1] has acknowledged the computational efficiency of the iterative approach E-M takes which is discussed next.

B. E-M Algorithm

E-Step

$$x_n = (W^T W + \sigma^2 I) W^T (t_n - \mu) \qquad (6)$$

$$x_n x_n^T = \sigma^2 (W^T W + \sigma^2 I) + x_n x_n^T \qquad (7)$$

M-Step

$$W = \sum_{n=1}^{N} (t_n - \mu)(x_n)^T (\sum_{n=1}^{N} x_n x_n^T)^{-1} \qquad (8)$$

$$\sigma^2 = \frac{1}{Nd} \sum_{n=1}^{N} ||t_n - \mu||^2 - 2x_n^T W^T (t_n - \mu) + tr(x_n x_n^T W^T W) \qquad (9)$$

The E-step gives us the parameter estimation of latent variable x and M-step maximizes our parameters W and σ^2

III. Results

Table discusses the computational time using both the approach, where we can see E-M clearly performs better.

PPCA			PPCA-EM	
q	Time		q	Time
25	182.412		25	12.77
50	216.256		50	12.72
100	239.726		100	12.81
200	265.08		200	12.91

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.