Probabilistic Principal Component Analysis with Expectation Maximization

Group-3
Shreyas Patel[1401025], Amee Bhuva[1401009], Akhil Vavadia[1401095], Manini Patel[164404], Malav Vora[1401062]
School of Engineering and Applied Science, Ahmedabad University

***Abstract*—Principal Component Analysis is a global technique for data analysis. But it does not deal properly with outlying data observations which effects analysis badly, so another PCA model was introduced which demonstrates how the principal components of observed data vectors may be determined through maximum-likelihood estimation of parameters in a latent variable model closely related to factor analysis. This PCA model is known as Probabilistic PCA(PPCA). EM (Expectation Maximization) Algorithm will help for finding principal components through iteratively maximizing the likelihood function.**

***Keywords*—Principal Component Analysis, Probabilistic model, EM algorithm, Latent variable, Factor analysis, Maximum-likelihood, Gaussian mixtures.**

I. Introduction

In classical PCA, as we all known given observed data vectors matrix t_{N*M}, one can find eigen values and eigen vectors of sample co-variance matrix of t. From that consider only q eigen vectors and construct $W = (w_1, w_2, ...w_q)$. We can find principal components of observed vectors ($x_n = W^T(t_n - \mu)$) and this principal component projection minimizes the squared reconstruction error. But observed data vectors may have corrupted data entries and missing entries as well, so modeling with PCA may lead to bad analysis, so probabilistic modeling of observed data vectors is required which leads to idea of Probabilistic PCA. PPCA can be utilized as a general Gaussian density model.

II. Factor Analysis

Factor analysis is based on formal model predicting observed variables from theoretical latent factors(x). where linear relationship is given as

$$t = Wx + \mu + \epsilon \tag{1}$$

Where latent variable x is model with zero mean and unit variance x$\sim \mathcal{N}(0, I)$, additionally specifying the error or noise, model to likewise Gaussian $\epsilon \sim \mathcal{N}(0, \Psi)$ and related observations $t \sim \mathcal{N}(\mu, WW^T + \epsilon)$, So latent variable x explains correlation between observation variables while ϵ_i represents variability unique to a particular t_i.

III. Probabilistic PCA Model

The marginal distribution for the observed data t is obtained as below where the observation co-variance model is specified as $C = WW^T + \sigma^2 I$. Related log-likelihood (ℓ) is

$$t \sim \mathcal{N}(\mu, C)$$

$$\ell = \frac{N}{2} dln(2\pi) + ln(C) + tr(C^{-1}S)$$

Where μ is mean of data and S is sample co-variance matrix of observed data vectors t_n and we want to find W and σ^2 by iterative maximization of log-likelihood and at the end we will find latent variables x from observed data through following equation.

$$x|t \sim \mathcal{N}(M^{-1}W^T(t - \mu), \sigma^2 I)$$

Where M = $W^T W + \sigma^2 I$ and note that M is q*q matrix whereas C is N*N matrix.

IV. Properties of Maximum Likelihood Estimators

Further calculation of maximizing log-likelihood ℓ will lead to stationary points of log-likelihood, and solution to that is given as $W = U(\Lambda - \sigma^2 I)^{1/2} R$, where U is a matrix of eigen vectors of S, Λ is diagonal matrix of eigen value of S and R is arbitrary rotation matrix. It may be shown that for W = W_{ML}, the maximum likelihood estimator for σ^2 is given by lost variance, averaged over lost dimensions.

V. An EM Algorithm for Probabilistic PCA

MLE PCA requires heavy-computation for high dimensional data sets and as it does not deal properly with missing data. EM is employed which deals with corrupted data at every iteration of maximizing likelihood. In approach with maximizing likelihood for PPCA, we consider the latent variables x_n as missing data and the complete data to comprise the observations together with these latent variables and related complete log-likelihood is given as below

$$\ell_c = \sum_{n=1}^{N} ln(p(t_n, x_n))$$

And maximization of above equation, after sufficient calculation, iterative computation of following equations will lead to maximized likelihood for PPCA.

$$< x_n >= M^{-1}W^T(t_n - \mu) \tag{2}$$

$$< x_n x_n^T >= \sigma^2 M^{-1} + < x_n >< x_n >^T \tag{3}$$

$$\tilde{W} = \Big[\sum_n (t_n - \mu) < x_n >^T \Big] \Big[\sum_n < x_n x_n^T > \Big]^{-1} \tag{4}$$