Probability principle component analysis

Michael E. Tipping and Christopher M. Bishop,

Microsoft Research, Cambridge, UK

Abstract—Principle Component Analysis is a technique for data analysis and processing, but it's not based on probability model. Probabilistic principal component analysis (PPCA) is a method to estimate the principal axes when any data vector has one or more missing values.

Index Terms—Density estimation, EM Algorithm, Gaussian mixtures, Maximum likelihood, PCA, probability Model.

I. Introduction

PCA is popular technique for dimensionality reduction. The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space. For a set of observed d-dimensional data vectors $\{t_n\}$, the q principle axes w_j, $j\epsilon\{1, 2,\ldots,q\}$ are those orthonormal axes onto which the variance is maximum. The Advantage of PCA is that it minimizes the squared reconstruction error.

II. What is PPCA?

Probabilistic principal component analysis (PPCA) is a method to estimate the principal axes when any data vector has one or more missing values. PPCA is based on an isotropic error model. It seeks to relate a d-dimensional observation vector t to a corresponding q-dimensional vector of latent (or unobserved) variable x, which is normal with mean zero and covariance I. The Relationship is given by:

$$t = Wx + \mu + \epsilon$$

where t is d-dimensional observation vector, x is q-dimensional vector of latent variables, and ϵ is the isotropic error term. ϵ is Gaussian with mean zero and covariance of σ^2, where σ^2 is the residual variance.

Here, q needs to be smaller than the rank for the residual variance to be greater than 0 ($\sigma^2>0$). Standard principal component analysis, where the residual variance is zero, is the limiting case of PPCA. The observed variables, t, are conditionally independent given the values of the latent variables, x. So, the latent variables explain the correlations between the observation variables and the error explains the variability unique to a particular t_i. The d-by-q matrix W relates the latent and observation variables, and the vector μ permits the model to have a nonzero mean. PPCA assumes that the values are missing at random through the data set. This means that whether a data value is missing or not does not depend on the latent variable given the observed data values.

III. Why EM is BETTER?

It's general Algorithm for missing data problem. It's Iterative Method for finding maximum likelihood. EM algorithm estimates Maximum Likelihood for missing data at each iteration. Maximum Likelihood PCA is computationally heavy for high dimensional data.

IV. EM Algorithm for PPCA

In EM approach for maximizing the likelihood of PPCA, we consider the latent variable $\{x_n\}$ to be "missing data" and the complete data are comprises of latent and observation variables, for the complete-data log-likelihood will be,

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

Where,

$$p(t_n, x_n) = (2\pi\,\sigma^2)^{\frac{-d}{2}} \exp\left(-\frac{||tn - Wxn - \mu||^2}{2\,\sigma^2}\right)(2\pi)^{\frac{-q}{2}} \exp\left(-\frac{||xn||^2}{2}\right)$$

In E step, we take expectation of L_c with respect to the distribution,

$$\langle \mathcal{L}_C \rangle = -\sum_{n=1}^{N} \left\{ \frac{d}{2}\ln(\sigma^2) + \frac{1}{2}\,\mathrm{tr}(\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle) + \frac{1}{2\sigma^2}(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}) - \frac{1}{\sigma^2}\langle \mathbf{x}_n \rangle^{\mathrm{T}} \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\,\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{W}\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle) \right\},$$

And,

$$\langle \mathbf{x}_n \rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}),$$

$$\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle = \sigma^2 \mathbf{M}^{-1} + \langle \mathbf{x}_n \rangle \langle \mathbf{x}_n \rangle^{\mathrm{T}}$$

Here, $M = W^T W + \sigma^2 I$.

In M Step, (L_c) is Maximized with respect to W and σ^2 giving new parameter estimates,

$$\tilde{\mathbf{W}} = \left\{ \sum_{n=1}^{N} (\mathbf{t}_n - \boldsymbol{\mu})\langle \mathbf{x}_n \rangle^{\mathrm{T}} \right\} \left(\sum_{n=1}^{N} \langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \right)^{-1}$$

$$\tilde{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \{ \|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - 2\langle \mathbf{x}_n \rangle^{\mathrm{T}} \tilde{\mathbf{W}}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}) + \mathrm{tr}(\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{W}}) \}$$

To maximize the likelihood, $\langle x_n \rangle$ and $\langle x_n x_n{}^T \rangle$ are calculated and then by substituting values of $\langle x_n \rangle$ and $\langle x_n x_n{}^T \rangle$, we will get new W and new σ^2. These four equations are iterated in sequence until the algorithm is judged to have converged.