Probabilistic Principal Component Analysis

Chirag Mehta
ai20btech11006@iith.ac.in

November 14, 2023

Abstract

Principal Component Analysis (PCA) is a widely used method for analyzing and processing data, although it lacks a foundation in a probability model. In this article, we look at the utilization of maximum-likelihood estimation in a latent variable model closely linked to factor analysis to identify the principal axes of observed data vectors. We explore the characteristics of the corresponding likelihood function and highlight the benefits offered by this probabilistic approach through illustrative examples.

1 Introduction

Probabilistic Principal Component Analysis (PCA) stands at the intersection of dimensionality reduction and probabilistic modeling. While traditional PCA excels at identifying key patterns in high-dimensional data, it falls short in explicitly addressing uncertainties within the observed information. In response, Probabilistic PCA introduces a probabilistic framework to the classic PCA, offering a more nuanced analysis by accounting for inherent uncertainties. This article navigates the fundamentals of Probabilistic PCA, spotlighting its capacity to model data variability in a probabilistic manner. By delving into its applications and insights, we explore how this method enhances our understanding of complex datasets and provides a robust foundation for decision-making in the face of inherent uncertainties.

2 Mathematical Background

- Definition A random vector is said to be *k-variate* normally distributed if every linear combination of its k components has a univariate normal distribution.
- Affine Transformation

$$X \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

Then

$$AX + b \sim \mathcal{N}\left(A\mu + b, A\Sigma A^T\right) \tag{2}$$

- Marginalization If

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \tag{3}$$

Then

$$X_1 \sim \mathcal{N}\left(\mu_1, \Sigma_{11}\right) \tag{4}$$

- Conditioning If

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right) \tag{5}$$

Then

$$X_1|X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) \tag{6}$$

3 The Probability Approach

Probabilistic PCA is defined in the following manner

$$x = Wz + \epsilon \tag{7}$$

$$where \quad z \sim \mathcal{N}(0, I) \tag{8}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{9}$$

Probabilistic PCA can be used for majorly two tasks, that are

1. $z|x$: Dimensionality Reduction
2. $x|z$: Data Generation

The next question is "how do we obtain the distributions for each of the above formulations?" The latter is straightforward to obtain.

$$x|z \sim \mathcal{N}\left(Wz, \sigma^2 I\right) \tag{10}$$

Next we will find the joint distribution of (x, z). We do that in the following way

$$p(x,z) = p(x|z)p(z) \tag{11}$$

$$p(x,z) \propto \exp\left(\frac{-1}{2\sigma^2}(x - Wz)^T(x - Wz)\right)\exp\left(\frac{-1}{2}z^T z\right) \tag{12}$$

$$\propto \exp\left(\frac{-1}{2}\begin{bmatrix}x^T & z^T\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma^2}I & \frac{-1}{\sigma^2}W \\ \frac{-1}{\sigma^2}W^T & \frac{1}{\sigma^2}W^TW + I\end{bmatrix}\begin{bmatrix}x \\ z\end{bmatrix}\right) \tag{13}$$

Using this, we get

$$p(v) \propto \exp\left(\frac{-1}{2}v^T\Sigma^{-1}v\right) \tag{14}$$

where $v = \begin{pmatrix}x \\ z\end{pmatrix}$, $\Sigma^{-1} = \begin{bmatrix}\frac{1}{\sigma^2}I & \frac{-1}{\sigma^2}W \\ \frac{-1}{\sigma^2}W^T & \frac{1}{\sigma^2}W^TW + I\end{bmatrix}$
The above likelihood is that of a gaussian

$$\therefore \begin{pmatrix}x \\ z\end{pmatrix} \sim \mathcal{N}(0, \Sigma) \tag{15}$$

Next, we will find the distribution for $z|x$
From Bayes Rule

$$p(z|x) = \frac{p(x|z)}{p(x)}p(z) \tag{16}$$

Using the conditioning property of gaussian distribution, we get

$$z|x \sim \mathcal{N}\left(M^{-1}W^Tx, \sigma^2M^{-1}\right), \quad where\, M = W^TW + \sigma^2I \tag{17}$$

Now that we have obtained the distributions for both $x|z$ and $z|x$, we want to obtain the matrix W. We form an optimization problem to achieve that. We want to maximize the likelihood of sampling the training data from the distribution of x. The distribution for x can be obtained easily using marginalization.

$$x \sim \mathcal{N}\left(0, WW^T + \sigma^2I\right) \tag{18}$$

We form the optimization problem that is to maximize the log-likelihood for the training dataset.

$$\mathcal{L} = -\frac{N}{2}\left\{d\ln(2\pi) + \ln|C| + tr(C^{-1}S)\right\} \tag{19}$$

where $S = \frac{1}{N}\sum_{k=1}^{N}x_kx_k^T$

$$\frac{\partial \mathcal{L}}{\partial W} = N\left(C^{-1}SC^{-1}W - C^{-1}W\right) \tag{20}$$

$$SC^{-1}W = W \tag{21}$$

Using SVD

$$SC^{-1}ULV^T = ULV^T \tag{22}$$

multiply by V on both sides

$$SC^{-1}UL = UL \tag{23}$$

After some simplification, we obtain

$$SUL = U\left(L^2 + \sigma^2 I\right)L \tag{24}$$

For $l_j \neq 0$

$$Su_j = \left(\sigma^2 + l_j^2\right)u_j \tag{25}$$

Therefore, each column of U must be an eigenvector of S, with corresponding eigenvalue $\lambda_j = \sigma^2 + l_j^2$. So,

$$l_j = (\lambda_j - \sigma^2)^{1/2} \tag{5}$$

Therefore,

$$W = U\left(K_n - \sigma^2 I\right)^{1/2} R \tag{26}$$

where K_n is a $n \times n$ diagonal matrix

$$k_{jj} = \begin{cases} \lambda_j & \text{eigenvalue corresponding to } u_j \\ \sigma^2 & o/w \end{cases} \tag{27}$$

R is any rotation matrix.

$$C = WW^T + \sigma^2 I \tag{28}$$

$$= ULV^T VLU^T + \sigma^2 I \tag{29}$$

$$= UL^2U^T + \sigma^2 I \tag{30}$$

$$|C| = |UL^2U^T + \sigma^2 I| \tag{31}$$

Identity: $|I + AB| = |I + BA|$

$$\therefore |C| = |\sigma^2 I + L^2| \tag{32}$$

Using (19) and (32)

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ln(2\pi) + \frac{1}{\sigma^2}\sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')ln(\sigma^2) + q' \right\} \tag{33}$$

Minimizing wrt σ

$$\sigma^2 = \frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j \tag{34}$$

If $\sigma^2 > 0$, then the $Rank(S) > n$

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ln(2\pi) + \sum_{j=1}^{q'} \ln\lambda_j + (d-q')ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j\right) + d \right\} \tag{35}$$

Interestingly, the minimization of E only leads to the requirement of λ_j to be adjacent in the spectrum of eigenvalues. Since the diagonal entries in K have to be at least as big as σ^2, the biggest q eigenvalues are considered in the K matrix.

4 Observations

4.1 Equivalence with PCA

It can be seen that, when $\sigma^2 \to 0$, $M^{-1} \to (W^T W)^{-1}$.

The maximum likelihood reconstruction could be written as

$$\tilde{x} = WM^{-1}W^T x \tag{36}$$
$$= U\Lambda^{1/2}(\Lambda^{1/2}U^T U\Lambda^{1/2})^{-1}\Lambda^{1/2}U^T x \tag{37}$$
$$= UU^T x \tag{38}$$

This is same as PCA (in the reconstruction sense)

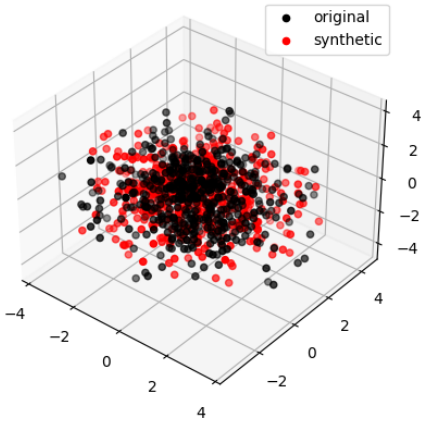

Figure 1: Synthetic data generated using PPCA

4.2 reconstruction

When $\sigma^2 > 0$ then the latent projection becomes skewed.

$$\langle z|x\rangle = M^{-1}W^T x$$

The above equation doesn't represent an orthogonal projection of z and is therefore not optimal in squared reconstruction loss sense. However, the optimal reconstruction can still be obtained from the conditional latent mean and is given by

$$\tilde{x} = W(W^T W)^{-1} M\langle z|x\rangle \tag{39}$$

The reconstruction obtained here would be the same as PCA. An example for this has been included in the sample notebooks.

4.3 Data Generation

We have already obtained distribution of $x|z$ in section 3. Using this we can generate more data. One exmaple where this would be useful is when one has limited samples of a defective class and more data is required to train a classifier, this technique could be used to generate more samples. In figure 1, we can see that the augmented data is similar to the actual data.

4.4 Data Imputation

We aren't just limited to generating entirely new data, we can also fill in the missing values for features in a given data point. This can be achieved using

the conditional property of gaussians. An example for this has been included in the sample notebooks.

5 Conclusion and Further Reading

In this article we delved into the theory of proabilistic PCA. We saw how principal component analysis can be viewed as a maximum-likelihood procedure based on a probability density model of the observed data. We derived the closed form solution to obtain W. We saw that PPCA is as good as PCA in terms of reconstruction but also has the ability to generate more data and also it can be used as an imputation technique.

5.1 Further Reading

1. Heteroscedasticity: The noise doesn't follow homoscedasticity
2. Bayesian PCA: Find the number of components for latent space.
3. Outlier Detection