Probabilistic Principal Component Analysis for Dimensionality Reduction and Missing Data Reconstruction Using Expectation Maximization

Group: Cohen, Edholm, Mulongo, Zarrinkoub

January 2019

Abstract

Probabilistic principal component analysis (PPCA) is an extension of the well-known principal component analysis (PCA) technique. Two advantages of PPCA is that it allows for treatment of missing data and for more complex projections using mixtures of multiple PPCA models. In this paper, PPCA models are used for reconstructing missing data and performing dimensionality reduction using different expectation maximization algorithms.

1 Introduction

The central idea in principal component analysis (PCA) is to maintain as much variance in the data as possible while reducing the dimensionality in a data set. This is achieved by projecting the data onto the *principal subspace*, a linear space of lower dimension, such that the projected data has maximal variance. PCA has limitations, such as inability to model the data probabilistically, no clear handling of missing values and outliers, and expensive computation time [Bishop, 2006].

Probabilistic PCA (PPCA) tries to remedy the problems stated above by allowing a probabilistic treatment and explanation of the data. Furthermore the data can be explained with a Gaussian density model which enables the extension of the framework from simple PPCA to mixtures of PPCA, in order to represent complex relations in the data. The probabilistic of PCA also allows for a Bayesian treatment [Bishop, 2006].

The Bayesian PCA makes it possible to avoid overfitting when learning the principal components W and the latent space representation of the data $\mathbf{X}$ by regularisation.[Ilin and Raiko, 2010] The Bayesian framework, however, is not investigated here.

The goal of this project is to reproduce a subset of the results presented in [Tipping and Bishop, 1999]. First, a comparison between the maximal variance projection for standard PCA is compared to the projection with PPCA on the same dataset but with missing data. Furthermore, the PPCA mixture model is implemented. Results are discussed for both experiments in terms of their deviation from those in [Tipping and Bishop, 1999]. Lastly, the PPCA techniques described are applied to a real-world data set.

2 Methods

2.1 Datasets

In this project, the *Tobamovirus* data set from [B.D. Ripley, 1996] with 38 data points and 18 dimensions is used. In addition, PCA and PPCA mixtures was performed on data from the World Happiness Report [United Nations,] (which in turn is based on the World Gallup Poll) which measures 9 features for 115 countries. Examples of the features are GDP per capita, life expectancy and absence of corruption.

2.2 PCA

Principal component analysis can be described as a data representation method, although it is usually described in the literature as a dimensionality reduction method. PCA lets us find a representation of the data in a lower dimensional space.

If we consider a set of observations $\{\mathbf{t}_n\}$, with $n = 1, 2, ..., N$ and dimensionality d, PCA finds a projection onto a subspace of dimensionality $q < d$ that maximizes the variance of the data after projection.

We express the relationship between the data space and the lower dimensional space as

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} \tag{1}$$

where $\mathbf{t}$ is a data vector of length d and $\mathbf{x}$ is the q-dimensional representation vector. It can be shown that maximal variance in the principal subspace is obtained when selecting the q eigenvectors of the covariance matrix $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$ with the largest corresponding

eigenvalues and using these as columns of $\mathbf{W}$[Bishop, 2006].

2.3 PPCA

Probabilistic principal component analysis (PPCA) builds on the assumption that the data observed was generated by a latent variable model [Tipping and Bishop, 1999]

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{2}$$

where $\mathbf{t}$ is our d-dimensional data variable and $\mathbf{x}$ is a random M-dimensional variable in latent space. The $\mathbf{W}$ matrix maps from latent to data space and the parameter vector $\boldsymbol{\mu}$ acts as a bias, allowing the model to have a mean other than zero. $\boldsymbol{\epsilon}$ is a zero-mean Gaussian noise variable with independent features and variance σ^2. From this, we can express the probability distribution of $\mathbf{t}$ conditioned on $\mathbf{x}$ as

$$\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \tag{3}$$

In order to obtain the marginal distribution, we need to define a prior over the latent variables. We choose $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which through integrating out the latent variables leads to the marginal

$$p(\mathbf{t}) = (2\pi)^{-d/2}|\mathbf{C}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{t}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{t}-\boldsymbol{\mu})\} \tag{4}$$

where $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T$. From this, the likelihood can be formulated by assuming independently drawn data points and multiplying all $p(\mathbf{t}_n)$:

$$\mathcal{L} = -\frac{N}{2}\{d\ln 2\pi + \ln|\mathbf{C}| + \mathrm{Tr}[\mathbf{C}^{-1}\mathbf{S}]\}. \tag{5}$$

The posterior, used for projection onto the principal subspace, can be found by using Bayes' Rule and is written

$$\mathbf{x}|\mathbf{t} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{t}-\boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}) \tag{6}$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$.

2.4 PPCA with missing data

To replicate the results of [Tipping and Bishop, 1999] with regards to missing data, we employed a technique used by [Vries, 2019]. Missing values are ignored, creating shortened feature vectors $\mathbf{t}_{ns}$. Each vector $\mathbf{t}_{ns}$ is a copy of $\mathbf{t}_n$ where the elements at indices in $idx(n)$ have been removed, where $idx(n)$ is the set of indices of missing values in $\mathbf{t}_n$. The mean $\boldsymbol{\mu}$ was calculated as equation 7,

$$\boldsymbol{\mu}(i) = \frac{1}{N_{obs,i}}\sum_{n=1}^{N}\mathbf{t}_n(i) \tag{7}$$

where $N_{obs,i}$ is the number of observed values for feature i among all data points. With this, the missing values of feature vectors $\mathbf{t}_n$ are replaced by their respective mean in $\boldsymbol{\mu}$. Individualised mean vectors $\boldsymbol{\mu}_n$ were also created. Each vector $\boldsymbol{\mu}_n$ is a copy of $\boldsymbol{\mu}$, where the elements at indices in $idx(n)$ have been removed. The EM algorithm was implemented, using equation 8 to calculate expected mean and covariance of latent variables.

$$\begin{aligned}\langle\mathbf{x}_n\rangle &= \mathbf{M}^{-1}\mathbf{W}_n^\top(\mathbf{t}_{ns} - \boldsymbol{\mu}_n)\\ \langle\mathbf{x}_n\mathbf{x}_n^\top\rangle &= \sigma^2\mathbf{M}^{-1} + \langle\mathbf{x}_n\rangle\langle\mathbf{x}_n\rangle^\top\end{aligned} \tag{8}$$

Where each $\mathbf{W}_n$ is a copy of $\mathbf{W}$, where the rows at indices in $idx(n)$ have been removed. Equations 9 and 10 are then used to update $\mathbf{W}$ and σ^2,

$$\mathbf{W}_{new} = \mathbf{SW}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\mathbf{SW})^{-1}) \tag{9}$$

$$\sigma^2_{new} = \frac{1}{d}\operatorname{tr}\left(\mathbf{S} - \mathbf{SWM}^{-1}\mathbf{W}^\top_{new}\right) \tag{10}$$

where $\mathbf{S}$ is calculated from 11.

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^\top \tag{11}$$

To check convergence, the expected complete log-likelihood L_C is calculated through equation 12.

$$\begin{aligned}\langle L_C\rangle = -\sum_{n=1}^{N}\Bigg\{&\frac{d}{2}\ln\sigma^2 + \frac{1}{2}\operatorname{tr}\left(\langle\mathbf{x}_n\mathbf{x}_n^\top\rangle\right) + \frac{1}{2\sigma^2}(\mathbf{t}_n - \boldsymbol{\mu})^\top(\mathbf{t}_n - \boldsymbol{\mu}) \\ &-\frac{1}{\sigma^2}\langle\mathbf{x}_n\rangle^\top\mathbf{W}_n^\top(\mathbf{t}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\operatorname{tr}\left(\mathbf{W}^\top\mathbf{W}\langle\mathbf{x}_n\mathbf{x}_n^\top\rangle\right)\Bigg\}\end{aligned} \tag{12}$$

The matrix $\mathbf{W}$ is initialised as a matrix of evenly distributed numbers between -1 and 1.

2.5 Mixture PCA

In the mixture PCA model, it is assumed that several different PCA models are combined with mixture components to generate each data point:

$$p(\mathbf{t}) = \sum_{i=1}^{M}\pi_i p(\mathbf{t}|i) \tag{13}$$

where π_i is the mixture component for each model and $p(\mathbf{t}|i)$ is an individual PCA model. The log-likelihood then takes the form

$$\mathcal{L} = \sum_{n=1}^{N}\ln\left\{\sum_{i=1}^{M}\pi_i p(\mathbf{t}|i)\right\}. \tag{14}$$

An iterative EM algorithm can be developed for optimizing the parameters of this model according to [Tipping and Bishop, 2006]. The responsibility of each model over each data point is written

$$R_{ni} = \frac{p(\mathbf{t}|i)\pi_i}{p(\mathbf{t})}. \tag{15}$$

From this, it can be shown that the following parameter updates can be obtained:

$$\pi_i = \frac{1}{N}\sum_{n=1}^{N}R_{ni} \tag{16}$$

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{n=1}^{N}R_{ni}\mathbf{t}_n}{\sum_{n=1}^{N}R_{ni}} \tag{17}$$

$$\tilde{\mathbf{W}}_i = \mathbf{S}_i \mathbf{W}_i (\sigma_i^2 \mathbf{I} + \mathbf{M}_i^{-1} \mathbf{W}_i^T \mathbf{S}_i \mathbf{W}_i)^{-1} \tag{18}$$

$$\tilde{\sigma}_i^2 = \frac{1}{d} \mathrm{Tr}[\mathbf{S}_i - \mathbf{S}_i \mathbf{W}_i \mathbf{M}_i^{-1} \tilde{\mathbf{W}}_i^T] \tag{19}$$

where

$$\mathbf{S}_i = \frac{1}{\tilde{\pi}_i N} \sum_{n=1}^{N} R_{ni} (\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \tilde{\boldsymbol{\mu}}_i)^T. \tag{20}$$

and d is the dimensionality of the data. Iteration of these updates, in this order, are guaranteed to find the local maximum of the likelihood.

These updates were made until the likelihood converged. Then, for each of the data points, the mapping from data space onto latent space was made for the model with the highest responsibility for the given data point. The mapping into latent space can be obtained by using the mean of the posterior in (6).

3 Results

3.1 PPCA with missing data

A visual representation of these results is given by Figure 1, where three distinct groupings appear in the standard PCA and complete-data PPCA cases. The numbers represent individual data points, so that data point 1 is denoted 10, data point 2 is denoted 11, etc. From the left and middle figure it is clear that standard PCA and PPCA yield similar results. It is noteworthy that the missing data instance of PPCA does not have the same conspicuous grouping characteristic as standard PCA or complete-data PPCA. This differs from the results in [Tipping and Bishop, 1999], where the grouping tendency remains intact when running the algorithm on the incomplete data. One possible reason for this deviation is the further treatment of the mean values, where in the original paper these were averaged over a conditional distribution of the missing, given observed, values. However in this experiment the means were simply the arithmetic mean (equation 7) for each missing data vector.

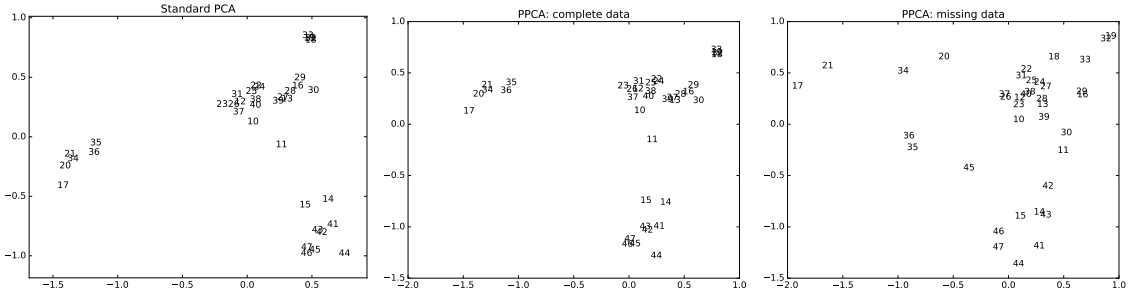

Figure 1: Projections of the 18-dimensional *Tobamovirus* dataset onto a 2-dimensional space according to different methods. Left: standard PCA projection. Middle: PPCA projection with complete data. Right: PPCA projection with 20 percent of data randomly removed.

3.2 Mixture of PPCA

Using the EM algorithm for mixtures of PPCA described in the method, we plotted each data set observation for the model in the mixture that had the highest responsibility for it. The

responsibility describes the level of confidence that the observation was generated by a particular mixture. The figure below shows the result of the projection of the Tobamovirus data for 3 mixtures in 2-dimensional latent space. The deviations from [Tipping and Bishop, 1999] can have several explanations. First, it is stated in [Tipping and Bishop, 1999] that the **W** matrix that maximizes the likelihood can be subject to an arbitrary rotation, with unchanged likelihood. Rotating figure 2 and 3 greatly increases the similarity to the images in [Tipping and Bishop, 1999]. Further differences can be explained by the fact that the EM algorithm implemented only guarantees convergence to a local minimum. It is thus possible that the minimum reached in figures 2 to 4 is not the same minimum as in [Tipping and Bishop, 1999].

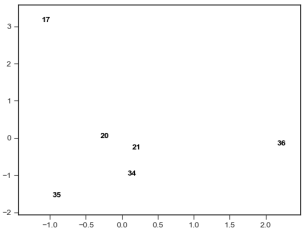

Figure 2: Cluster 1

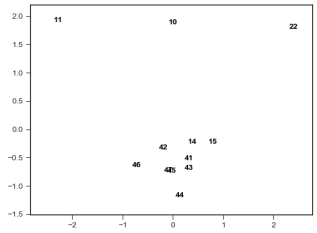

Figure 3: Cluster 2

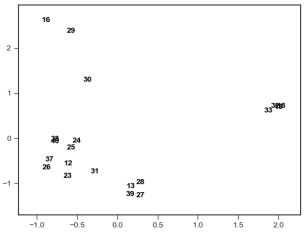

Figure 4: Cluster 3

Projections of the 18-dimensional *Tobamovirus* dataset onto a 2-dimensional latent space with 3 mixtures.

3.3 Application to World Happiness data

3.3.1 PCA

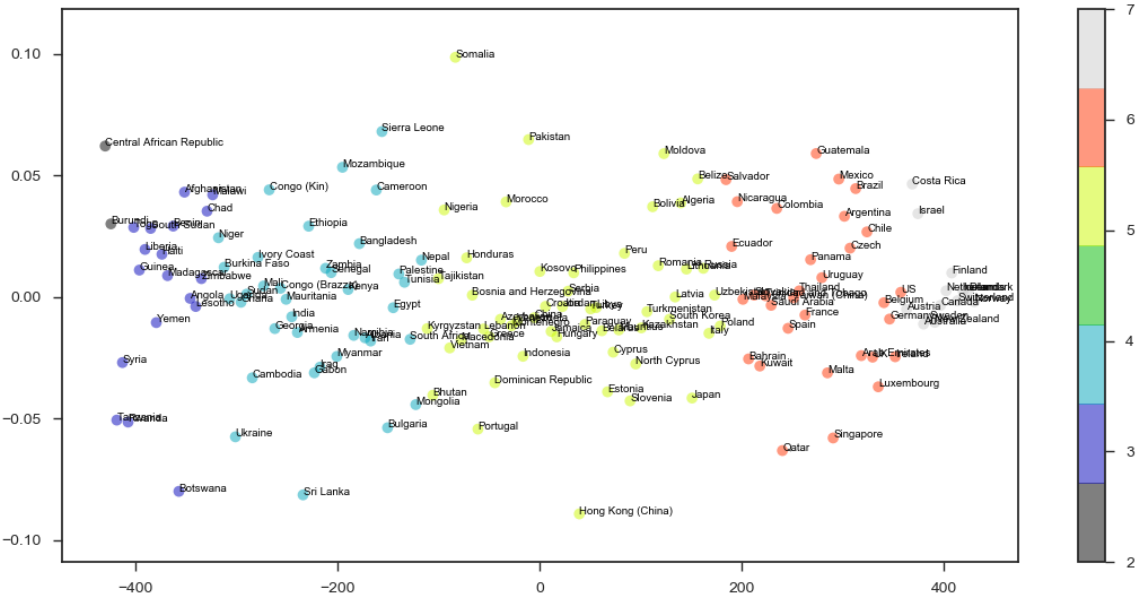

Figure 5: Projection of the World Happiness data to a 2D subspace. The colors represent the happiness score in the data set.

Figure 5 shows the projection of World Happiness data onto a two dimensional space. Looking at the projected countries, we can see that similar countries are close to each other in the projection.

3.3.2 Mixtures of PPCA

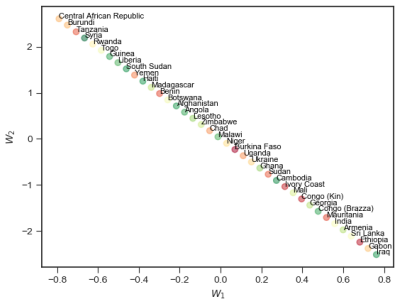

Figure 6: Cluster 1

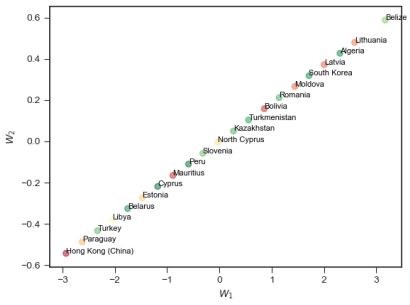

Figure 7: Cluster 2

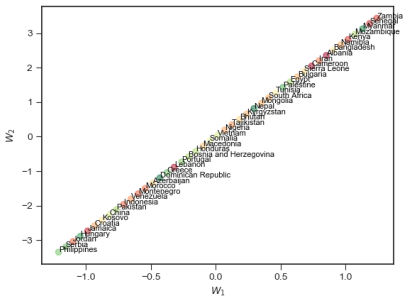

Figure 8: Cluster 3

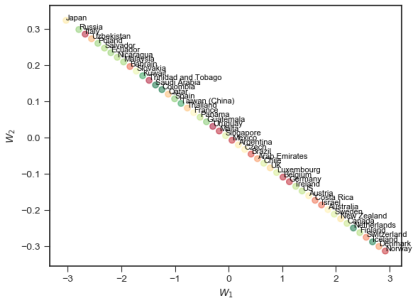

Figure 9: Cluster 4

In figures 6 to 9 we project the data onto a 2D latent space with 4 different mixtures. By inspecting each image we see that similar countries are assigned to the same cluster. For example, cluster 1 in figure 6 seems to mostly contain countries with weak economies whereas cluster 4 in figure 9 mostly contains countries with well developed economies.

4 Discussion

In the missing data experiment, the amount of missing data (20 percent) was quite small. More recent research suggests that PPCA should be able to handle even sparser data vectors, especially for smaller problems and when $d \ll n$.[Ilin and Raiko, 2010]

When using mixture models, we found that the results differed slightly from those obtained in [Tipping and Bishop, 1999]. As pointed out, this difference can be due to convergence to a local maximum of the likelihood. Thus, it is likely that the random initialization of the parameters plays a role in whether the global minimum is found or not. This is supported by [Blömer and Bujna, 2013] and [Baudry and Celeux, 2015].

PPCA in its original form has the advantage that it provides closed-form solutions for the maximum likelihood estimators. But it is, however, limited in its capability to scale, since the computation of the co-variance matrix is very expensive for big amount of data with high dimensionality. Furthermore, the linearity of PPCA limits its capabilities.

Mixtures of PPCA models can be used for representing complex data. However, it is difficult to know beforehand the numbers of components or mixtures components appropriate when using PCA or PPCA. This difficulty can be circumvented by using a variational Bayesian treatment of PPCA [Bishop, 2000].

In order to allow for non-linearity and increase scalability, autoencoders (a type of neural network) can be used. Recent developments in the field [Plaut, 2018] have shown that autoencoders can be used for dimensionality reduction with results equivalent to those of standard PCA. Autoencoders have the advantage of being able to handle large data sets and data with high dimensionality, such as images. For large data sets with high dimensionality, even the EM approach presented in this report can become computationally expensive due to the required computation of the co-variance matrix $\mathbf{S}$. Furthermore, autoencoders allow for sequential (online) learning which makes them preferable to PPCA in some applications.

References

[Baudry and Celeux, 2015] Baudry, J.-P. and Celeux, G. (2015). Em for mixtures. *Statistics and Computing*, 25(4):713–726.

[B.D. Ripley, 1996] B.D. Ripley (1996). Weblet Importer.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

[Bishop, 2000] Bishop, C. M. (2000). Variational principal components.

[Blömer and Bujna, 2013] Blömer, J. and Bujna, K. (2013). Simple methods for initializing the em algorithm for gaussian mixture models.

[Ilin and Raiko, 2010] Ilin, A. and Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *978-951-22-9481-7*, 11.

[Plaut, 2018] Plaut, E. (2018). From principal subspaces to principal components with linear autoencoders. *arXiv preprint arXiv:1804.10253*.

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 61(3):611–622.

[Tipping and Bishop, 2006] Tipping, M. E. and Bishop, C. M. (2006). Probabilistic principal component analysis. *MIT Press*, pages 443–482.

[United Nations,] United Nations. World Happiness Report — Kaggle.

[Vries, 2019] Vries, B. D. (2019). Lecture notes from adaptive information processing 11: Continuous latent variable models - pca and fa.