Application of Probabilistic PCA

RnD Project Report

Submitted in partial fulfillment of requirements for the degree of
Bachelor of Technology (Honors)

by
Aditya Kumar Akash
Roll No : 120050046

under the guidance of
Prof. Suyash Awate

Department of Computer Science and Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India

April, 2016

Contents

Abstract

Principal Component analysis (PCA) is a classical data analysis technique that finds linear transformations of data that retain the maximal amount of variance. The classical version is not based on a probability model. Researchers have proposed a probabilistic model of PCA which is closely related to factor analysis. In this work, we understand the Probabilistic PCA (PPCA) and analyse how it handles missing data situations in which we cannot apply standard PCA. We also try to obtain a comparision in performance of PPCA with a variant of PCA.

Acknowledgements

I am sincerely indebted to my advisor, Prof. Suyash Awate, IIT Bombay, for his constant support and guidance throughout the course of this project. His experience and insight in the fields of machine learning, image processing and varied aspects of computer science in general, was valuable in boosting my interest on this topic.

I finally and especially would like to thank my parents and my whole family for their support and trust in all of my endeavours. The morals they have imparted stayed and would stay close to me always. I also thank my friends for giving a helping hand during hard times.

Chapter 1

Introduction

Principal Component analysis (PCA) is a ubiquitous tool for dimensionality reduction. It has many applications data compression, visualization, image processing, exploratory data analysis, pattern recognition and time series prediction.
There are a number of optimization criteria to derive PCA. The most important of these is in terms of stardardized linear projection which maximizes the variance in the projected space [6]. Assume that $\{y_i\}, i \in \{1, 2, ..., n\}$ is a set of d dimensional data vectors. Then the k principal axes $\{w_j\}, j \in \{1, 2, ..., k\}$ are those orthogonal axes onto which the retained variance under projection is maximal. It can be shown that w_j are given by k dominant eigen vectors (those with largest eigen values, λ_j) of the sample covariance matrix

$$\mathbf{S} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})^{\mathbf{T}} \tag{1.1}$$

where $\bar{\mathbf{y}}$ is the sample mean, such that

$$\mathbf{S w_j} = \lambda_\mathbf{j} \mathbf{w_j} \tag{1.2}$$

The k principal components of the observed vector $\mathbf{y_i}$ are given by the vector

$$\mathbf{x_i} = \mathbf{W^T}(\mathbf{y_i} - \bar{\mathbf{y}}) \tag{1.3}$$

where $\mathbf{W} = (\mathbf{w_1}; \mathbf{w_2}; ...; \mathbf{w_k})$. The variables $\mathbf{x_i}$ are uncorrelated such that the covariance matrix $\sum_{i=1}^{n} \frac{\mathbf{x_i x_i^T}}{n}$ is diagonal with elements $lambda_i$.
A complementary property of PCA is that of all the orthogonal linear projections, the principal component minimizes the squared reconstruction error. However, PCA does not provide a probabilistic model of data. This gives motivation for PPCA.

1.1 Motivation

A probabilistic formulation of PCA from a Gaussian latent variable model is obtained. This is closely related to factor analysis. PCA could be viewed a limiting case of such a Gaussian model. In such a formulation, the principal axes emerge as maximum likelihood parameter estimates. Such a probabilistic formulation is intuitively applealing, as the definition of a likelihood measure enables comparison with other probabilistic techniques, while facilitating statistical testing and permitting the application of Bayesian methods. Further motivation behind a probabilistic PCA is that it conveys additional practical advantage as :

- The probability model offers the potential to extend the scope of conventional PCA, such as using probabilistic mixtures and PCA projections in missing data case.
- PPCA can be utilized as a general Gaussian density model. This allows the maximum likelihood estimates for the parameters associated with the covariance matrix to be efficiently computed from the data principal components.

Chapter 2

Background

2.1 Factor Analysis Model and Links to PCA

The factor analysis model is a *latent variable model* which related a d-dimensional observation vector y to k-dimensional vector of latent variable x. Following equation expresses the relationship

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mu + \epsilon \tag{2.1}$$

where the columns of $\mathbf{W}$, a $d \times k$ matrix, are the factors which relate the two vectors, μ allows a non-zero mean and ϵ is the noise/error.
When $k < d$, the latent variables offer a more parisimonious explanation of the dependencies between the observations.
The underlying assumptions of $x \sim \mathcal{N}(0, \mathbf{I})$ and $\epsilon \sim \mathcal{N}(0, \mathbf{\Psi})$ induces a corresponding Gaussian distribution on observation $\mathbf{y} \sim \mathcal{N}(\mu, \mathcal{W}\mathcal{W}^{\mathcal{T}} + \Psi)$.
The key assumption for the factor analysis model is that, by contraining the error covariance $\mathbf{\Psi}$ to be diagonal, whose elements Ψ_i are estimated from the data, the observed variables t_i are conditionally independent given the values of latent variables $\mathbf{x}$. Thus the latent variables are intended to capture the correlation between the observed variables while the error term ϵ_i represents the variability unique to particular t_i. *This is where PCA differs from factor analysis, as it treats covariance and variance identically.* This distinction in variance and covariance in factor analysis model cause the maximum-likelihood estimates of columns of $\mathbf{W}$ to not correspond the the principal subspace of the observed data. However, the two methods are linked if we consider a special case of isptropic error model, where residual variances $\Psi_i = \sigma^2$ are constrained to be equal [5].

Chapter 3

Probabilistic PCA

3.1 The Probability model

The model bears similarity to the factor analysis model,

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mu + \epsilon \tag{3.1}$$

with the assumption of isotropic gaussian noise model $\mathcal{N}(0, \sigma^2\mathbf{I})$. This gives as x-conditional distribution over y-space as

$$\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \mu, \sigma^{\mathbf{2}}\mathbf{I}) \tag{3.2}$$

With $x \sim \mathcal{N}(0, \mathbf{I})$, the marginal distribution for y is given by

$$\mathbf{t} \sim \mathcal{N}(\mu, \mathbf{C}) \tag{3.3}$$

where oberservation covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I}$.
The log-likelihood is then

$$\mathcal{L} = -\frac{N}{2}d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-\mathbf{1}}\mathbf{S}) \tag{3.4}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y_i} - \mu)(\mathbf{y_i} - \mu)^{\mathbf{T}} \tag{3.5}$$

The maximum-likelihood estimates for μ is given by the mean of the data, in which ase $\mathbf{S}$ is the sample covariance. Estimates of $\mathbf{W}$ and σ^2 is obtained by EM algorithm.

3.2 EM method for PPCA

In the *EM* approach to maximize likelihood for PPCA, we consider latent variables $\mathbf{x_i}$ to be 'missing' data and the 'complete' data to comprise the observations together with latent variables. Corresponding complete log-likelihood is then :

$$\mathcal{L}_\mathcal{C} = \sum_{i=1}^{N} \ln\{p(\mathbf{y_i}, \mathbf{x_i})\} \tag{3.6}$$

, where, in PPCA, we get

$$p(\mathbf{y_i}, \mathbf{x_i}) = (2\pi\sigma^2)^{-d/2}\exp\Big\{ - \frac{||\mathbf{y_i} - \mathbf{W}\mathbf{x_i} - \mu||}{2\sigma^2}\Big\}(2\pi)^{-k/2}\exp\Big\{ - \frac{||\mathbf{x_i}||}{2}\Big\} \tag{3.7}$$

The posterior is given by

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{M^{-1}W^T}(\mathbf{y} - \mu), \sigma^\mathbf{2}\mathbf{M-1}) \tag{3.8}$$

where $\mathbf{M} = W^TW + \sigma^2 I$. From the appendix B of [1] we obtain following

E-Step :

$$\langle \mathbf{x_i} \rangle = \mathbf{M^{-1}W^T}(\mathbf{y_i} - \mu) \tag{3.9}$$

$$\langle \mathbf{x_i}\mathbf{x_i^T} \rangle = \sigma^\mathbf{2}\mathbf{M^{-1}} + \langle \mathbf{x_i} \rangle \langle \mathbf{x_i} \rangle^\mathbf{T} \tag{3.10}$$

M-Step :

$$\tilde{\mathbf{W}} = \big[\sum_\mathbf{i} (\mathbf{y_i} - \mu)\langle \mathbf{x_i} \rangle \big] \big[\sum_\mathbf{i} \langle \mathbf{x_i}\mathbf{x_i^T} \rangle \big]^{-\mathbf{1}} \tag{3.11}$$

$$\sigma^\mathbf{2} = \frac{\mathbf{1}}{\mathbf{Nd}} \sum_\mathbf{i} \big\{ ||\mathbf{y_i} - \mu||^\mathbf{2} - \mathbf{2}\langle \mathbf{x_i} \rangle^\mathbf{T} \tilde{\mathbf{W}}^\mathbf{T}(\mathbf{y_i} - \mu) + \mathrm{tr}(\langle \mathbf{x_i}\mathbf{x_i^T} \rangle \tilde{\mathbf{W}}^\mathbf{T}\tilde{\mathbf{W}}) \big\} \tag{3.12}$$

The paper [1] shows the combination of both of the above steps rewritten as

$$\tilde{\mathbf{W}} = \mathbf{SW}(\sigma^\mathbf{2}\mathbf{I} + \mathbf{M^{-1}W^TSW})^{-\mathbf{1}} \tag{3.13}$$

$$\sigma^\mathbf{2} = \mathrm{tr}(\mathbf{S} - \mathbf{SWM^{-1}}\tilde{\mathbf{W}}^\mathbf{T}) \tag{3.14}$$

$\mathbf{S}$ is the sample covariance.
Analysis of thes equations show that in normal PCA calculation we require calculation of $\mathbf{S}$ which takes $\mathcal{O}(\mathcal{N}d^2)$ operations. But in case of above EM formulation, we only need to compute $\mathbf{SW}$ as $\sum_\mathbf{i} \mathbf{x_i}(\mathbf{x_i^TW})$ which takes $\mathcal{O}(\mathcal{N}dk)$ operations. Thus when $k \ll d$, considerable computational savings would be obtained. This is one of the benefits of using the EM version of PPCA.

3.3 Properties of MLEs

In paper [1] it is shown that with $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathbf{T}} + \sigma^{\mathbf{2}}\mathbf{I}$, likelihood 3.6 is maximized when

$$\mathbf{W_{ML}} = \mathbf{U_k}(\mathbf{\Lambda_k} - \sigma^{\mathbf{2}}\mathbf{I})^{\mathbf{1/2}}\mathbf{R} \tag{3.15}$$

where k column vectors in $\mathbf{U_k}$ are the principal eigenvectors of $\mathbf{S}$, with corresponding eigenvalues in Λ_k, and $\mathbf{R}$ is arbitary orthogonal rotation matrix.
When $\mathbf{W} = \mathbf{W_{ML}}$, MLE for σ^2 is

$$\sigma^2_{ML} = \frac{1}{d-k}\sum_{j=k+1}^{d}\lambda_j \tag{3.16}$$

which has **interpretation of variance lost in projection, averaged over the lost dimension**. Using this we can see how this lost variance is subtracted from the eigen vectors in the estimation of $\mathbf{W_{ML}}$.

Chapter 4

Missing Data

4.1 PPCA with Missing Data

One of the motivation of using PPCA is that it provides interpretation to the data that is missing from the observation variable. Such *missing data variables are assumed to be 'parameters' in the model* and a generic EM algorithm is designed to handle the case.
An example of missing data case would be in computer vision field, when we model a dodecahedran from a sequence of segmented images. One sample of data would contain only information (in form of normals) for only 6 of the faces, while rest is missing data.

In these cases the **E-step** of EM algorithm is generalized to following :
Generalized E-step [2]

- If $\mathbf{y}$ is incomplete, then we find a unique pair of points $\mathbf{x}^*, \mathbf{y}^*$ (such that $\mathbf{x}^*$ lies in the current principal subspace and $\mathbf{y}^*$ lies in the subspace defined by the known information about $\mathbf{y}$) which minimize the norm $||\mathbf{W}\mathbf{x}^* - \mathbf{y}^* + \mu||^2$. Now we set the corresponding expectation of $\mathbf{x}$ to $\mathbf{x}^*$ and correspoinding observed variable $\mathbf{y}$ to $\mathbf{y}^*$. The solution is obtained by finding solution to a particular constrained matrix.

- If $\mathbf{y}$ is complete, then $\langle \mathbf{x} \rangle$ is found as before.

The above steps emerge as a result of treating the missing values as parameters to the model. The optimization problem results from maximizing the likelihood of the complete data.

4.2 PCA with Missing Data

In this work we also try to compare following PCA modification for missing data.

- **PCA with reference to Factor analysis** : In this approach we try to estimate the missing values using the minimization of $||\mathbf{W}\mathbf{x}^* - \mathbf{y}^* + \mu||^2$. $\mathbf{x}$ is the latent variable. In each iteration, first $\mathbf{y}$ is estimated using this minimization, then $\mathbf{W}$ is estimated by finding k principal component of covariance of $\mathbf{Y}$. The optimization problem emerges as a result of assuming the data generation model for $\mathbf{y}$ based on the factor model.

- **Standard PCA with missing values filled by mean** : This case is based on direct estimation of missing values by assuming Gassian model for the data. We assume observation having mean μ and covariance $\mathbf{S}$. If there was no missing value, the estimation of mean and covariance would be sample mean and covariance. Now with missing data being there, we first estimate mean to be sample mean of data which is present. The missing data then obtained by taking derivative comes out to be the components from mean. Thus in the next step, taking the mean over entire data does not change it. *Effectively in this method we replace the missing data with the mean of the non-missing data.*

Chapter 5

Experiments

We give examples to show how PPCA can be exploited for practical examples. The experiments focus on the application of PPCA to the dataset with missing values.

5.1 Dataset

We use following dataset :

1. **Tobamovirus** dataset : 38 virus , each with 18 features
2. **MNIST** dataset : The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image of 28×28 pixels.
3. **USPS** dataset : Handwritten Digits, 8-bit grayscale images of "0" through "9"; 1100 examples of each class.
4. **Binary Alphadigits** : Binary 20x16 digits of "0" through "9" and capital "A" through "Z". 39 examples of each class.

5.2 Experiment Design

- For the **Tobamovirus** dataset, the data is projected into 2 dimensions for the purpose of visualization of dimension reduction by PCA and PPCA. The dataset is claimed to have three sub-groups. The missing data is simulated by randomly removing each value in the dataset with probability 20%. The aim is to find how much of the sub-groups is being preserved.
- For the handwritten digits datasets, the data was randomly divided intp 7:3 ratio for training and testing, in case the two sets are not present.

We train a *classifier based on mahalanobis distance*. For each digit a factor matrix ($\mathbf{W}$) is obtained using PCA/PPCA on the training data. Based on the factor matrix, we find the projections of all the training data points. Mahanalobis distance of each test data sample is calculated in latent dimension from the training set of each digit. The digit which gives least distance is predicted as the label.

For missing data case, the factor matrix and latent variables are learnt from training data having missing values. The missing values are simulated by randomly removing the data with a given probability. The prediction is done using these learnt values. The algorithms are analysed for different amount of missing values.

With this experiment, we try to find the behaviour (prediction accuracy) of each of the algorithms with different amount of training data.

Chapter 6

Results and Observations

6.1 Tobamovirus Data

For the Tobamovirus data we can see that the projection 6.1 of the complete data obtained using standard PCA gives three sub-groups. Then we have 6.2 which is the projection obtained using PPCA closed form given in []. Figure 6.3 is the projection obtained by the PPCA run with EM algorithm. The projection is the same except for being rotated about some point. The roation is dependent on the initialization of the algorithm.

For the missing data case (20% missing), it is clear that both figure 6.4 and 6.5 is able to obtain three sub-grounps. The salient features of the projection is clear, even when all data points have suffered from at least one missing value. Both the algorithms seems to perform equally good in this dataset.

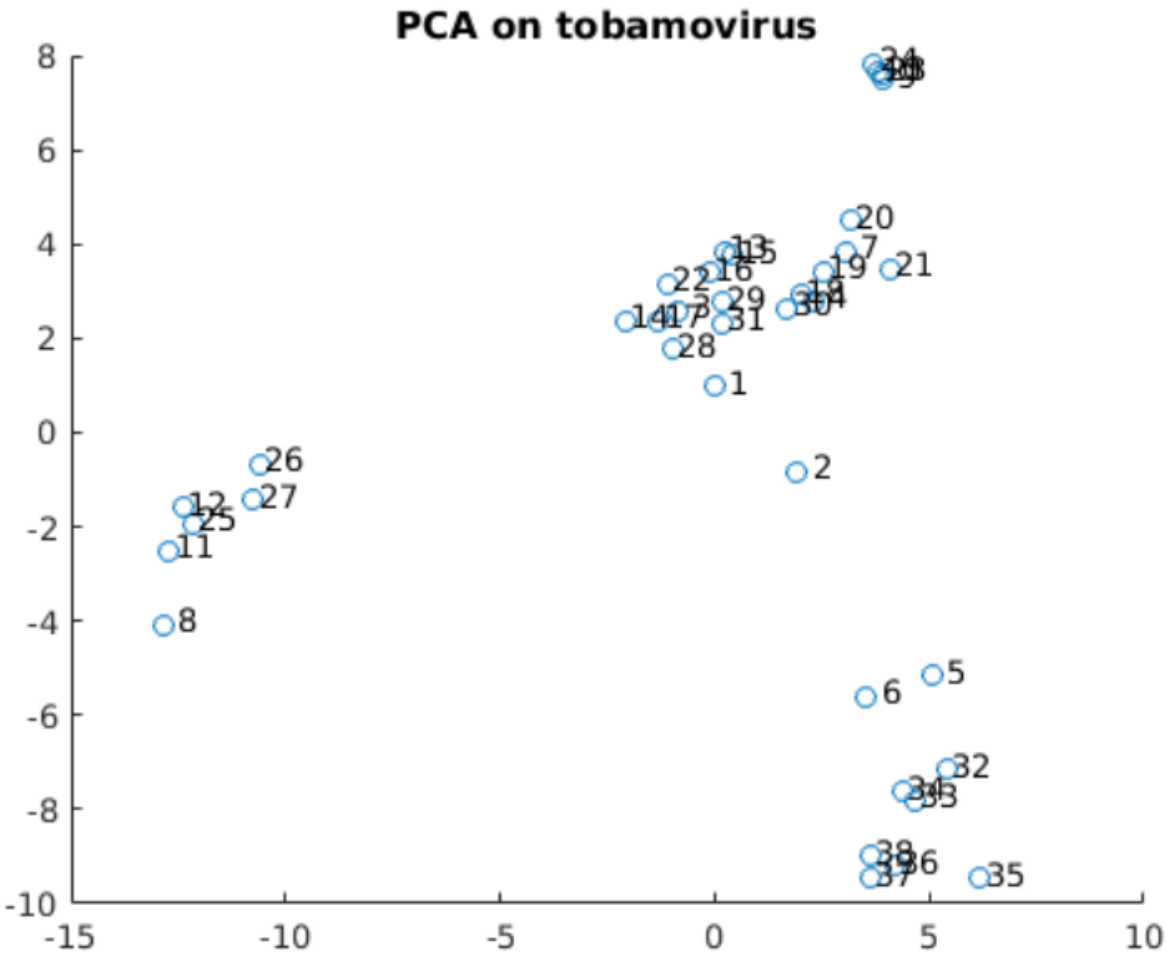

Figure 6.1: Standard PCA

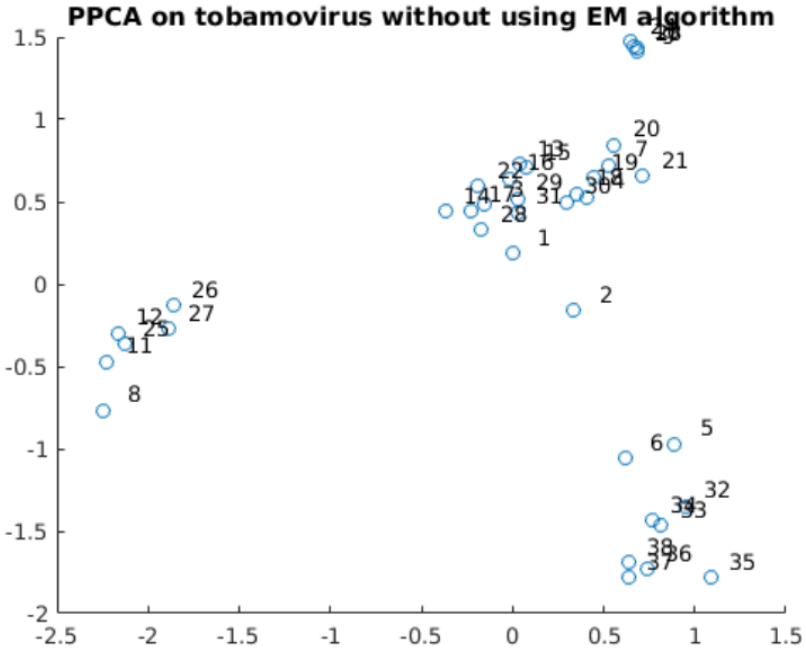

Figure 6.2: PPCA using closed form formula

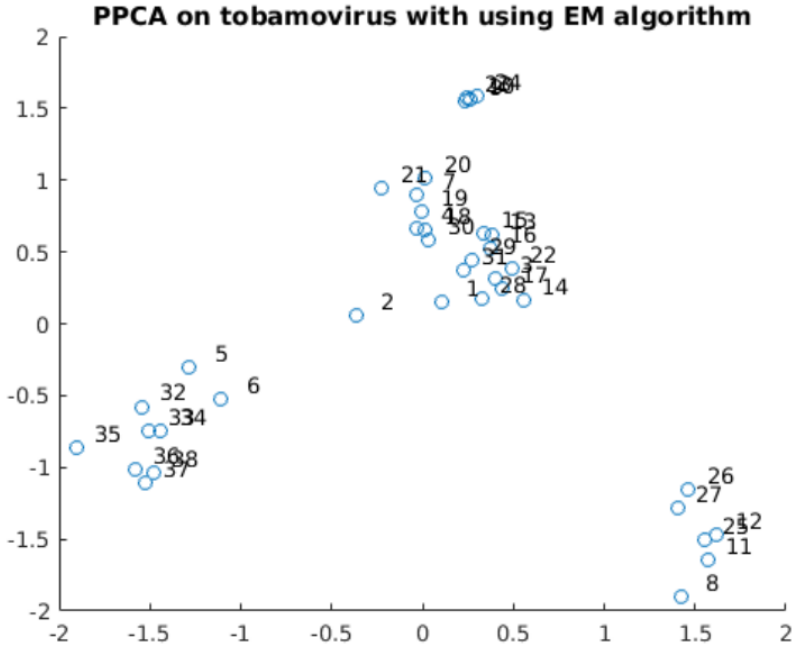

Figure 6.3: PPCA using EM algorithm

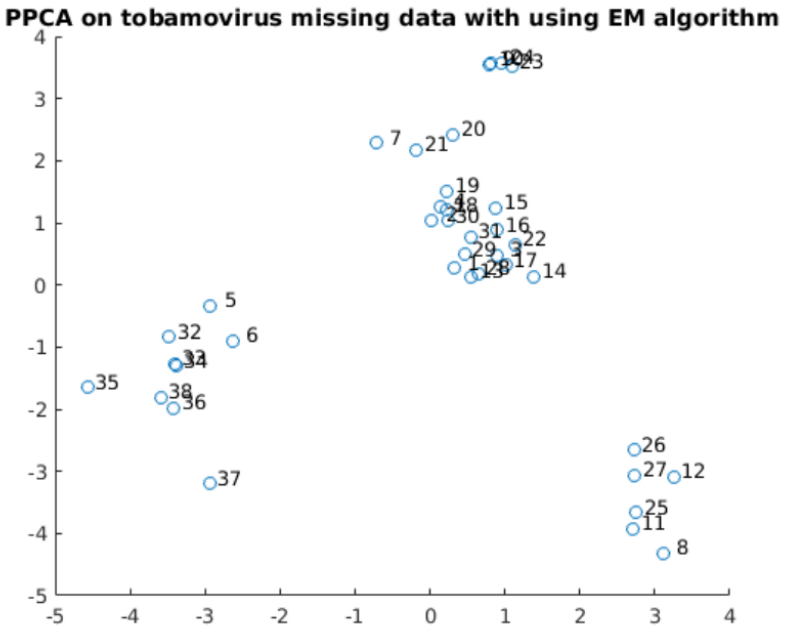

Figure 6.4: PPCA projection with 20% missing data using EM

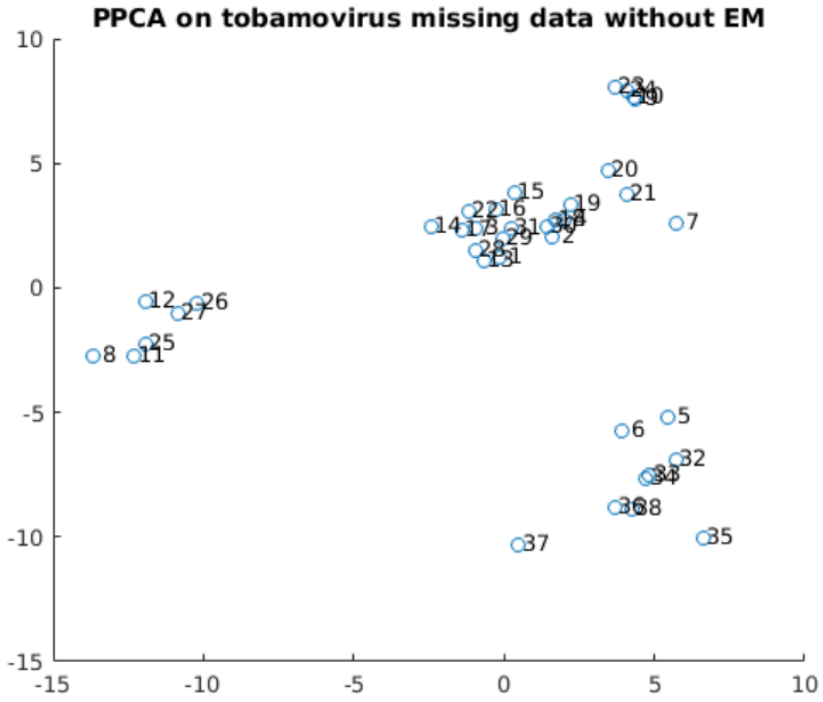

Figure 6.5: PCA projection with 20% missing data

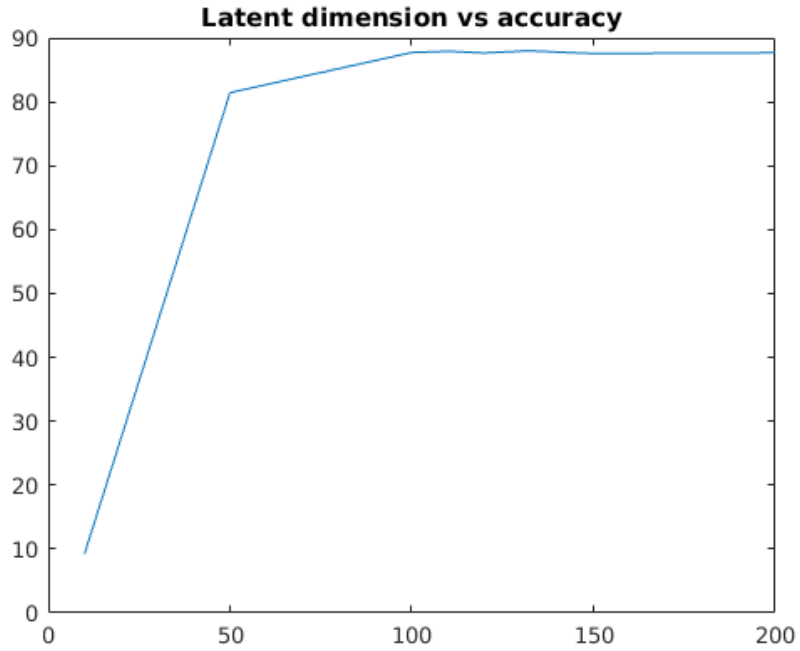

Figure 6.6: Accuracy vs latent dimension for MNIST

6.2 MNIST Data

MNIST data contains 28×28 images. Each images contains a handwritten digit. The task is predition of the digits. The experiment design outlines in the last chapter is followed.

The plot 6.6 shows how the accuracy increases rapidly at start but then becomes static. The latent dimension $k = 133$ is a good choice as it has highest accuracy in the region plotted and also the increase becomes very less after that point.

The experiments that follow have $k = 133$ fixed. Following is the table showing accuracy of PPCA with missing values and comparision with other variants PCA for handling missing data.

Missing Data %	Accuracy for PPCA with EM	Accuracy for PCA based on Factor Model	Accuracy for PCA based on μ for missing value
0	88.01	87.97	87.97
1	88.25	88.27	88.40
5	89.62	91.19	90.30
20	92.74	93.08	93.01
40	92.44	71.71	93.05
60	83.49	2.81	91.44
80	-	-	86.31
90	-	-	78.08
99	-	-	42.94

Table 6.1: Accuracy for Missing data

From table 6.1, we can find following observations

1. The accuracy increase with increase in missing data till a certain fraction. After a threshold the accuracy drops.

2. The accuracy of PCA handling missing data based on factor model drops significantly (Column 2). The possible explanation for this is that the number of unknowns in the minimization problem $||\mathbf{y} - \mathbf{W}\mathbf{x} - \mu||$ becomes large as x is latent and y also has missing data. So the system of equations gives poor estimates of the missing data.

3. PCA based on missing data filled by μ components show the best acuracy for missing data. A possible explanation is that a large part of data is a background image and digits occupy only lesser fraction. So larger number of missing data are estimated correclty using mean which would be towards background pixel side.

6.3 USPS Dataset

USPS also contains handwritten digit images. The experiment outlined for MNIST is performd again for USPS. For each experiment data is randomly divided into $7:3$ ratio for training and testing. From figure 6.7, it can be seen tha $k = 100$ is an ideal choice for the latent dimension. Further experiments keep $k = 100$. We can also see similar behaviour of this data set as the MNIST. Table 6.2 also contain the accuracy on conplete data as last column as the data partitioning is random.

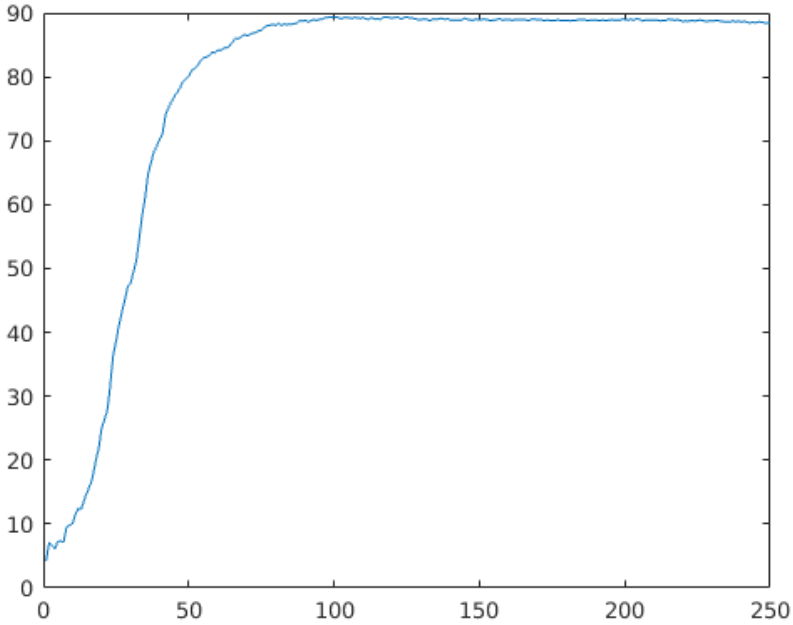

Figure 6.7: Accuracy vs latent dimension for USPS

Missing Data %	Accuracy for PPCA with EM	Accuracy for PCA based on Factor Model	Accuracy for PCA based on μ for missing value	Accuracy Standard PCA on complete data
0.5	89.96	89.57	90.12	89.42
1	89.93	89.72	89.84	88.84
5	90.51	90.54	91.90	88.36
10	91.87	91.60	93.36	89.39
20	92.30	92.48	93.87	90.39
40	83.69	84.63	91.51	88.30
60	55.21	52.30	88.75	88.60

Table 6.2: Accuracy for Missing data for USPS

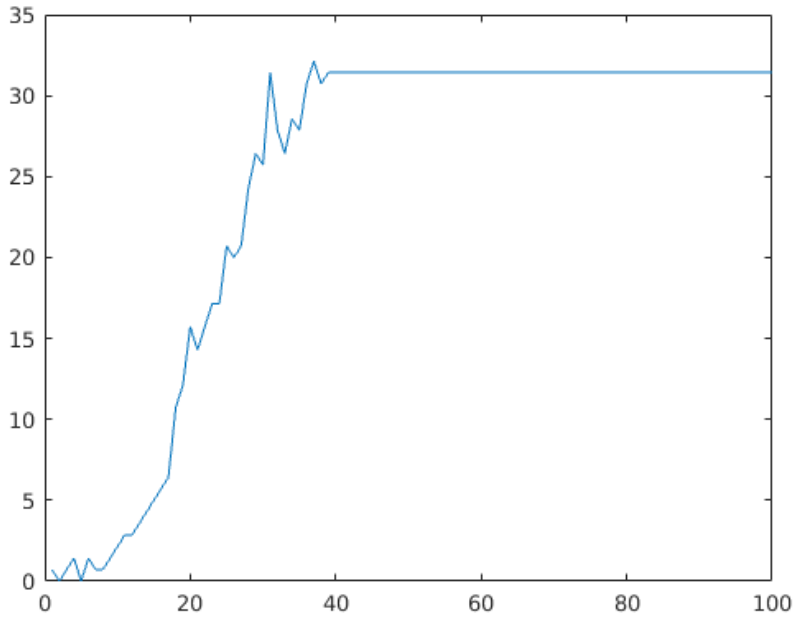

Figure 6.8: Accuracy vs latent dimension for Alpha digits

6.4 Binary AlphaDigits Dataset

The data is present in form of 39 samples for each 36 classes. Each sample is a 20×16 image. Since the amount of data present is low picking up a large latent dimension is not feasible for calculating the mahalanobis distance.

From the plot 6.8, it is clear that the accuracy increases till the number of latent dimension k reaches the number of training sample ($k \sim 40$). After that the accuracy does not increase. This makes this dataset difficult to analyse. We donot any further prediction analysis for this dataset.

Chapter 7

Conclusion

In this work, we see how principal component analysis model can be viewed as a maximum likelihood procedure based on a probability density model of the observed data. We also see links to PPCA to factor analysis and subtle difference between the two, as well as the underlying assumption for PPCA. An EM algorithm was discussed for PPCA which iteratively maximized the likelihood of the data.

The main aim of this work was understanding PPCA and its application on real world dataset. We apply PPCA to the case of missing data and compare its performance with different version of PCA made to handle missing data. In the *Results* chapter we see how PPCA is capable of handling missing data along with providing a reasonable model for interpretation of the results. But we found out that for the datasets which we covered PCA with missing data replaced by mean of non-missing data gives best performance. This leads us to conclude that for cases where data is not generated using mixtures of gaussian, the sophistication of PPCA is only to provide a probabilistic touch to the classical version of PCA.
The importance for EM algorithm of PPCA is more when the data has an inherent mixture model distribution. In such cases, PCA cannot trivially handle it. But PPCA could easily incorporate it into the model and handle such cases.

Bibliography

[1] Tipping, Michael E., and Christopher M. Bishop. "Probabilistic principal component analysis." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61.3 (1999): 611-622.

[2] Roweis, Sam. "EM algorithms for PCA and SPCA." Advances in neural information processing systems (1998): 626-632.

[3] Chen, Haifeng. "Principal component analysis with missing data and outliers." `http://www.cmlab.csie.ntu.edu.tw/~cyy/learning/papers/PCA_Tutorial.pdf` (2002).

[4] LeCun, Yann, Corinna Cortes, and Christopher JC Burges. "The MNIST database of handwritten digits." (1998).

[5] Whittle, Peter. "On principal components and least square methods of factor analysis." Scandinavian Actuarial Journal 1952.3-4 (1952): 223-239.

[6] Hotelling, Harold. "Analysis of a complex of statistical variables into principal components." Journal of educational psychology 24.6 (1933): 417.