Introduction

Principal Component Analysis (PCA) is a well established method that is concerned with multivariate data and presenting the data using fewer dimensions while the most important structure is still preserved. We consider a linear transformation that projects the d-dimensional data to a k-dimensional space where $k < d$ such that each components of transformed data is mutually uncorrelated and explains the maximum variability of the data.

It is desired to present the data with a few number of dimension, however some information must be lost. Therefore, the choice of the number of PC vector's dimension is often critical and controversial. Some rules are available such as inspecting the plot of the variances of the PC vector or choosing the first k PC components such that 90% of the variance is explained. However, those methods are ad hoc. If distributional assumption is available, hypothesis testing on the k is available.

The probabilistic set up for the PCA involves using the normal latent variable which leads to a constrained factor model. Tipping and Bishop [1] show that the analytical solution for the maximum likelihood estimator in term of PCA solution. That suggest the PC scores can be obtained by a likelihood approach. Beside the decision of the dimension of PC scores, PCA can be applied to incomplete data by performing expectation maximisation (EM) algorithm. We will demonstrate the framework for PPCA using two dataset: a simulated dataset and abalone dataset in Section 4.

1 Principal Component Analysis (PCA)

Let $\mathbf{X} \sim (\mu, \Sigma)$ be a d-dimensional random vector with the expected value μ and the $d \times d$ covariance matrix Σ and let $\mathbb{X} \sim \text{sam}(\bar{\mathbf{X}}, S)$ be the samples from the distribution of the random vector $\mathbf{X}$ with sample mean $\bar{\mathbf{X}}$ and sample covariance matrix S.

To understand the variability of the data, the covariance matrix Σ is a natural place to start. Consider the population case. We perform spectrum decomposition on the covariance matrix:

$$\Sigma = \Gamma \Lambda \Gamma^\top, \tag{1}$$

$$\Gamma = [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_d] \qquad \text{such that} \qquad \Sigma \boldsymbol{\eta}_k = \lambda_k \boldsymbol{\eta}_k \tag{2}$$

where column vectors of Γ are the eigenvectors of the covariance matrix Σ and the matrix Λ contains the corresponding eigenvalues. Without loss of generality, the eigenvalues are order ascendingly.

We take the linear transform of the data using matrix Γ_k that contains the first k column vector of the orthogonal matrix Γ. Principal component vector is given by

$$\mathcal{W}^{(k)} = \Gamma_k^\top (\mathbf{X} - \mu), \tag{3}$$

where Λ_k is a sub-diagonal matrix of Λ containing the first k eigenvalues of Σ.

It follows that $\mathcal{W}^{(k)} \sim (\mathbf{0}, \Lambda_k)$, all the k principal components are mutually uncorrelated. The PC vector captures the maximum variance under the constraints of uncorrelatedness. Thus, the PC vector represents the most important variation with k dimensional vector. However, the linear transform Γ_k is not invertible and that implies we discard the $(d-k)$ uncorrelated components in the data and some information must be lost.

In the sample case, we replace the covariance Σ with the sample covariance matrix S. The PC data, the sample case of the PC vector, are given by

$$\mathbb{W}^{(k)} = \hat{\Gamma}_k^\top (\mathbb{X} - \bar{\mathbf{X}}) \tag{4}$$

where the sample covariance matrix has spectrum decomposition

$$S = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^\top.$$

The goal of probabilistic principal component analysis (PPCA) model is to derive the linear transform $\Gamma_k^\top$ through in a probabilistic way, that is, using the likelihood. It makes sense that the probabilistic formulation of the PCA makes use of the standard normal latent variables, denoted $\mathbf{F}$. We employ the model expression, $\mathcal{W}^{(d)} = \Lambda^{-\frac{1}{2}}\mathbf{F} = \Gamma^\top(\mathbf{X} - \mu)$, or a similar argument, which is closing related to Factor Analysis (FA).

2 Factor Analysis (FA)

The construction of the probabilistic PCA model relies on a factor model formulation. There are also important links between PCA and FA. The PCA gives rise to one of the factor model solution. In this section, some background about FA will be provided.

While the PCA attempt to find out the k most important uncorrelated principal components through a linear transformation, with FA, we are interested in identifying the uncorrelated sources of data and how the data are associated with the sources. The goals of two analyses are similar.

The k-factors model expresses the data as a transform of the set of k common latent variables. Each component of the data varies according to an uncorrelated noise. We write the data as

$$\begin{aligned} \mathbf{X} &= \mathbf{A}\mathbf{F} + \mu + \epsilon; \\ \mathbf{F} &\sim (\mathbf{0}, \mathbf{I}_{k\times k}); \\ \epsilon &\sim (\mathbf{0}, \Psi_{d\times d}); \\ \text{cov}(\mathbf{F}, \epsilon) &= \mathbf{0}_{k\times d} \end{aligned} \tag{5}$$

where $\mathbf{F}$ is a k-dimension random vector, called the common factors, latent variables or hidden variables, representing the uncorrelated *sources*; $\mathbf{A}$ is a linear transform relates the latent variable to the data, calling the factor loadings; μ is the baseline mean of the random vector $\mathbf{X}$, and ϵ is called specific factor, representing the uncorrelated *noise* for each component. The specific factor has zero mean and a diagonal covariance matrix $\mathbf{\Psi}$. The common factor and the specific factor are assumed to be uncorrelated.

Under the model assumption (5), we can express the covariance matrix as

$$\Sigma = \mathbf{A}\mathbf{A}^\top + \mathbf{\Psi}.$$

Finding the uncorrelated sources and the transform $\mathbf{A}$ are our primary interests. Without distributional assumption, the solution of $\mathbf{A}$ can be estimated using PCA. Fix $k < d$, we put the covariance matrix $\Sigma = \Gamma\Lambda\Gamma^\top$. Then the linear transform matrix $\mathbf{A}$ is estimated by

$$\hat{\mathbf{A}} = \Gamma_k \Lambda_k^{1/2} \tag{6}$$

To ensure our estimate for $\mathbf{F}$ must have an identity covariance matrix, the common factor is then estimated by

$$\hat{\mathbf{F}} = \Lambda_k^{-1/2} \Gamma_k^\top (\mathbf{X} - \mu). \tag{7}$$

The diagonal variance matrix for the specific factors is estimated by

$$\hat{\Psi} = \Sigma_{diag} - (\Gamma_k \Lambda_k \Gamma_k^\top)_{diag} \tag{8}$$

where we use the notation A_{diag} to represent to a matrix only contains its diagonal elements of A and the rest of entries are zeros.

Note that if we calculate the covariance matrix using those estimates, combining (6) and (8) we have

$$\begin{aligned}\hat{\mathbf{A}}\hat{\mathbf{A}}^\top + \hat{\Psi} &= \Gamma_k \Lambda_k^{1/2} \Lambda_k^{1/2} \Gamma_k^\top + \Sigma_{diag} - (\Gamma_k \Lambda_k \Gamma_k^\top)_{diag} \\ &= \Gamma_k \Lambda_k \Gamma_k^\top + \Sigma_{diag} - (\Gamma_k \Lambda_k \Gamma_k^\top)_{diag} \neq \Sigma. \end{aligned} \tag{9}$$

We retain the most important eigenvalues and discard the rest of variance. It means when we express our data using a k-factor model. Some information must be lost like PCA. The number of common factors k is unknown. From the expression (9), it is observed that increasing k yields a better approximation of the variance but our model becomes more complicated to interpret. Both PCA and FA face this type of trade-off. Choosing an appropriate number for the factor model is of the primary interest.

The estimates can also be obtained using a likelihood-based approach. The PCA solution in (6) does not require any distributional assumption. When data are symmetric and continuous, there is no harm to assume normality for the common factor and the specific factor. Then, the estimates of $\hat{\mathbf{A}}$ can be obtained through the normal likelihood. The model (5) has new constraints:

$$\mathbf{F} \sim N(\mathbf{0}, \mathbf{I}) \qquad \text{and} \qquad \epsilon \sim N(\mathbf{0}, \Psi).$$

As a consequence of the normality, the common factor and the specific factor are now independent. Under the model, the random vector $\mathbf{X}$ is normal distribution

$$\mathbf{X} \sim N(\mu, \mathbf{A}\mathbf{A}^\top + \Psi).$$

We make use of the information about the distribution of the data and one standard way is to employ the maximum likelihood (ML) estimation. The properties of the maximum likelihood such as deriving an asymptotically efficient estimator and performing a hypothesis testing can be enjoyed. Consider $\mathbb{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_n]$ such that $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \mathbf{C})$ where $\mathbf{C} = \mathbf{A}\mathbf{A}^\top + \boldsymbol{\Psi}$. The maximum likelihood estimator $\hat{\mathbf{A}}, \hat{\mu}, \hat{\Psi}$ for $\mathbf{A}, \mu, \Psi$, respectively, are obtained by maximising the likelihood

$$\begin{aligned} L(\mathbf{A}, \mu, \Psi | \mathbb{X}) &= \prod_{i=1}^{n} (2\pi)^{-\frac{d}{2}} \det(\mathbf{C})^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{X}_i - \mu)^\top \mathbf{C}^{-1} (\mathbf{X}_i - \mu) \right) \\ &= (2\pi)^{-\frac{nd}{2}} \det(\mathbf{C})^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(\mathbf{C}^{-1} (\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^\top \right). \end{aligned} \tag{10}$$

However, there exists no analytical solution for the maximum estimator. The solutions must be attained using an iterative scheme such as expectation maximisation (EM) algorithm. The `factoran` function in Matlab also provides framework for such computation.

An analytical solution of the maximum likelihood exists if the variance of each specific factor is constrained to be equal. This assumption leads to the *probabilistic principal component analysis*.

3 Probabilistic Principal Component Analysis

Probabilistic principal component model is a constrained factor analysis model. The major difference is that common factor and specific factor are now assumed to be normal and the variance for each component of the random vector is equal. We write the d-dimensional random vector as

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \boldsymbol{\mu} + \epsilon; \tag{11}$$

$$\mathbf{F} \sim N(\mathbf{0}, \mathbf{I}_{k\times k});$$

$$\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{d\times d});$$

$$\mathbf{F} \text{ and } \epsilon \text{ are independent;}$$

where σ^2 is the variance for the specific factors.

Since $\mathbf{X}$ is linear transformation of $\mathbf{F}$ and ϵ, it follows that $\mathbf{X}$ must be normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}\mathbf{A} + \sigma^2\mathbf{I}$. Hence, we have

$$\mathbf{X} \sim N(\mu, \mathbf{A}\mathbf{A} + \sigma^2\mathbf{I}).$$

The likelihood for observing n d-dimensional data is

$$L(\mathbf{A}, \mu, \sigma^2|\mathbb{X}) = (2\pi)^{-\frac{nd}{2}} \det(\mathbf{C})^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}(\mathbf{C}^{-1}(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^\top\right) \tag{12}$$

where $\mathbf{C} = \mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}$.

This likelihood expression is almost as same as the likelihood expression (10), except we replace the matrix Ψ with $\sigma^2\mathbf{I}$. The maximum likelihood estimator is then attained by the $\hat{\mu}$, $\hat{A}$ and $\hat{\sigma}^2$ that maximise this likelihood expression (12).

3.0.1 Maximum likelihood

As discussed, an analytical solution for the maximum likelihood estimator for this model exists for fixed k. Tipping and Bishop [1] show the maximum likelihood estimator (MLE) with a fixed number of factors k is

$$\hat{\mu} = \bar{\mathbf{X}} \tag{13}$$

$$\hat{\mathbf{A}}_k = \hat{\Gamma_k} \left(\hat{\Lambda}_k - \hat{\sigma}^2 \mathbf{I}_{k \times k} \right)^{1/2} \mathbf{R} \tag{14}$$

$$\hat{\sigma}_k^2 = \frac{1}{(d-k)} \sum_{j>k} \hat{\lambda}_j. \tag{15}$$

where Γ_k and Λ_k are derived from the spectrum decomposition from the sample covariance matrix S. The λ_j are the eigenvalues of S, ordered decreasingly; the matrix $\mathbf{R}$ is any orthogonal matrix so any rotation of $\mathbf{A}$ is also the solution of $\mathbf{A}$.

The solution of PPCA is closely related to the PCA, although it is derived from a k-factors model. Compared to the PCA solution (6), the estimator for $\mathbf{A}$ is adjusted by the variance of specific factor $\hat{\sigma}^2$ on the diagonal of $\hat{\Lambda}_k$.

This tells us that the linear transform Γ_k for PCA can be obtained through maximising the likelihood. The maximum likelihood estimator is invariant under the invertible transformation. One way to acquire the maximum likelihood estimator for Γ_k is to work out the spectrum decomposition of $\hat{\mathbf{A}}_k \hat{\mathbf{A}}_k^\top$.

This model has two main advantages in practice:

1. The number of latent variables k is unknown and it can be chosen using hypothesis testing based on the maximum likelihood estimator. The estimators (14) and (15) are the maximiser of the likelihood conditioned on a fixed k. One of the main goals of factor analysis is to find out a minimal but sufficient number of factors to represent the data. It is equivalent to finding out a "good" number of dimensions of the PC vector. The selection of k involves a series of hypothesis testing which will be demonstrated in the first example of Section 4.

2. Missing data can be dealt with. Under the PPCA model, as discussed, Γ_k can be acquired in two ways: first, spectrum decomposition of S, and second maximising the likelihood and performing some algebraic calculations to give $\hat{\Gamma}_k$. When missing data presents, the calculation of S is not available. In contrast, by treating the missing data as latent variables, the maximum likelihood estimator for Γ_k can still be obtained using Expectation Maximisation

(EM) algorithm. We will demonstrate this feature in the second example of Section 4.

4 Examples

4.1 Simulated data

We now consider two sets of simulated dataset from a normal distribution. The goal of considering the simulated data is to assess the performance of PPCA by comparing the estimates of interest to the true answer.

We simulate in total two sets of data. For the first set, we simulate the 5000 data points from $N(\mathbf{0}, \mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I})$. We assume $\mathbf{A}$ and σ^2 to be:

$$\mathbf{A} = \begin{pmatrix} 1 & 0.5 \\ 0.1 & 2 \\ 1.2 & 0.2 \\ 0.5 & 2 \\ 0.8 & 1 \end{pmatrix} \qquad \sigma^2 = 1.5.$$

For the second set of data, we only change the variance matrix of the specific factor form $\sigma^2\mathbf{I}$ to Ψ where Ψ is taken to be

$$\Psi = \begin{pmatrix} 1.2 & 0 & 0 & 0 & 0 \\ 0 & 3.7 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

The aim of analysing the second dataset using PPCA is to assess the performance of this model when the assumption of the equal variance of specific factor is violated.

PPCA is conducted and $k = 2$ is assumed. Figure 1 shows that estimates for the $\hat{\mathbf{A}}$ for the two datasets using biplots. The three matrix have been rotated for comparison. Observed from the figure, maximum likelihood estimate for $\mathbf{A}$ for equal noise variance has a very similar structure and the vectors in the biplot are pointing slightly upward compared to the true $\mathbf{A}$. Consider the dataset with unequal equal noise variance. The structure is still preserved compared to the $\mathbf{A}$, except for the magnificence of top vector is a bit over-stated. From this show, PPCA gives fairly good estimates even though the equal noise variance assumption

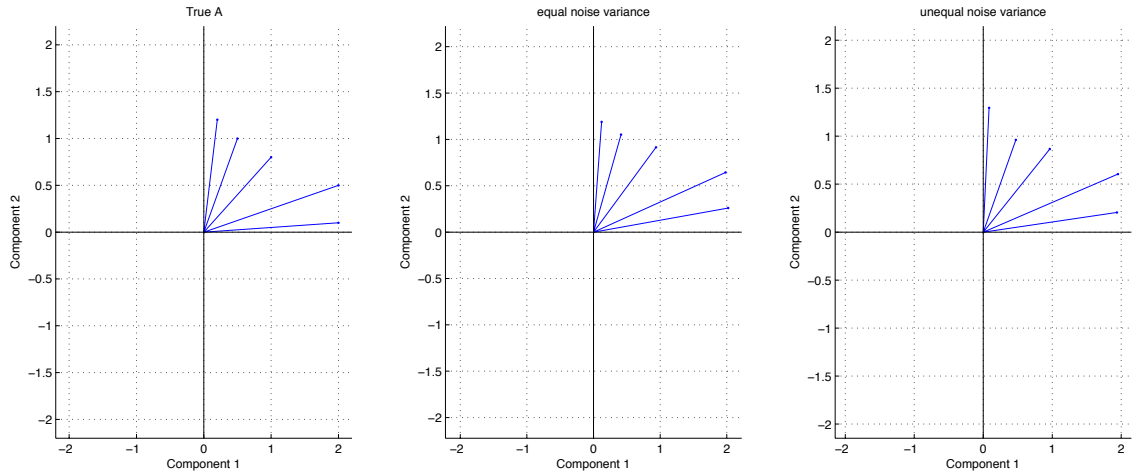

Figure 1: Biplots for the true linear transform $\mathbf{A}$ (left), the estimate of $\hat{\mathbf{A}}$ for the both datasets: equal (middle) and unequal (right) variances for specific factors.

is violated.

In the above analysis, $k = 2$ is assumed because the data are simulated from the two-factors model. However, in practice, the number of latent variance k is unknown. Without the distributional assumption, the number of latent variable is chosen by inspecting the plots or making use of some prior knowledge. In contrast, with the distributional assumption, this number can be systematically chosen using a hypothesis test. We will show the framework for deciding the number.

We are interested in choosing the minimal but sufficient number k. Under the factor model and PPCA model, the covariance matrix does not equal to the true covariance matrix. The sample covariance matrix under the two models is not equal to the sample covariance matrix. It is simply because we discard some information and capture the most important variance. The hypothesis test is relied on assessing whether the variance is sufficiently explained. Hence, we are test the covariance matrix under the factor model.

We test the covariance matrix under each k. Since minimal k is desired, we start with $k = 1$. If $k = 1$ is shown to be not sufficient, discarding too much variance, then we will need to consider the one more common factor. When $k = 1$ is sufficient, we will not have to test other choice of k. It leads us to the following scheme of hypothesis test.

Start with $k = 1$, we test the null hypothesis $H_{0,k} : \Sigma = \mathbf{A}_k\mathbf{A}_k^\top + \sigma_k^2\mathbf{I}$ (The construction of the test statistic will be discussed in the next part).

- If the null hypothesis is retained, we conclude this k-factors model is sufficient to explain the data. Testing on further k is not necessary.
- If the null hypothesis is rejected, this k-factors model is insufficient to explain the data. We need to consider the $(k+1)$ factors model. Therefore, we set $k = k+$ and perform the hypothesis test again.

We now discuss the test statistic for the null hypothesis

$$H_{0,k} : \Sigma = \mathbf{A}_k\mathbf{A}_k^\top + \sigma_k^2\mathbf{I}$$

against the alternative hypothesis

$$H_{1,k} : \Sigma \neq \mathbf{A}_k\mathbf{A}_k^\top + \sigma_k^2\mathbf{I}$$

where We use the likelihood ratio test for a fixed k

$$\Lambda_{H_0,k}(\mathbb{X}) = \frac{\sup_{H_{0,k}} L(\Sigma|\mathbb{X})}{\sup_{H_{1,k}} L(\Sigma|\mathbb{X})} \tag{16}$$

$$= \frac{L(\mathbf{A}_k\mathbf{A}_k^\top + \sigma_k^2\mathbf{I}|\mathbb{X})}{L(\hat{\Sigma}|\mathbb{X})}. \tag{17}$$

where $\hat{\Sigma}$ is the maximum likelihood estimator for the variance matrix Σ; we use the notation $\sup_{H_{0,k}}$ to represent the supremum of the likelihood function under the constraints of the PPCA model.

Koch [2] shows the test statistic for the model of unequal variance for specific factor. Restricting to equal specific factor variance yields

$$\Lambda_{H_0,k}(\mathbb{X}) = \left(\frac{\det(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top + \hat{\sigma}^2\mathbf{I})}{\det(\hat{\Sigma})}\right)^{-\frac{n}{2}} \exp\left(\frac{nd}{2} - \frac{(n-1)}{2}\mathrm{tr}\left[(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top + \hat{\sigma}^2\mathbf{I})^{-1}S\right]\right). \tag{18}$$

Under the null hypothesis, the asymptotic result is

$$-2\log\Lambda_{H_0,k}(\mathbb{X}) \to \chi^2_\nu$$

k-factor model	$-2\log\Lambda_{H_0,k}$	df	p-values	Decision
$k=1$	"∞"	9	0.0000	Reject
$k=2$	5.4094	5	0.3680	Retain

Table 1: Scheme of choosing k using hypothesis test.

k-factor model	Retained hypotheses %
$k=1$	0%
$k=2$	96.1%
$k\geq 3$	3.9%

Table 2: Retained hypotheses of k-factor model for each simulated dataset.

with degree of freedom

$$\nu = \frac{1}{2}d(d+1) - (dk + 1 - \frac{1}{2}k(k-1)). \tag{19}$$

We now demonstrate the scheme of choosing k for the first dataset. The significance level for the test is set to be 0.05. We first set the one-factor model. The log likelihood ratio does not converge to a bound number that the computer can handle. Thus, with p-value equaling zero, we reject the null hypothesis and conclude that the data should have more than one source. We then consider the two-factors model for the data. The test statistic for the two-factors model is 5.4094 with degrees of freedom 9. Hence, the p-value is 0.3680 and we retain the hypothesis. Two-factors model is sufficient for the data and further testing $k > 2$ is not necessary. Therefore, it is concluded that the data belong to two major sources which agree with the true model.

The next step is to assess the robustness of the choice of k using the scheme of hypothesis testing. We simulated 1000 sets of data using the same setup as the first dataset and select a k using the scheme. The result of the 1000 tests is shown in Table 2. No dataset is suited by an one-factor model. Over 96.1% of dataset are two-factors model and the rest of 3.9% are 3-factors model or more than 3 factors. Testing all potential k is not feasible because the degree of freedom (19) will become a non-positive number as k increases. From this, under-estimation of the number of the latent variables is not likely. The performance of this scheme is quite accurate considering that the misclassification is 3.9% which is better the pre-determined type-I error, 0.05, of our test.

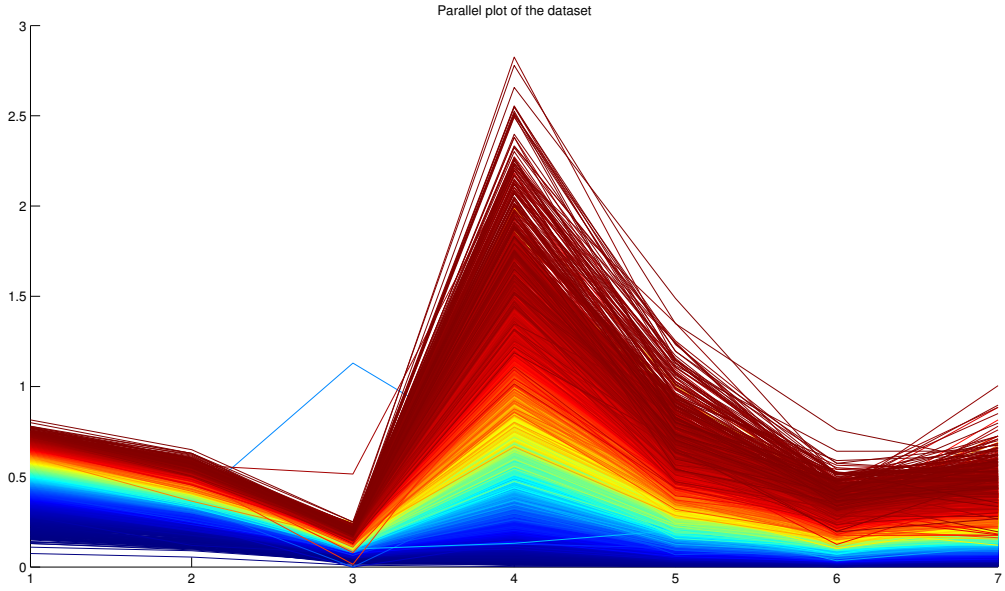

Figure 2: Parallel plot of the abalone dataset; the colours are assigned according to the order of the first variable, red for greater values and blue for low values.

4.2 Abalone data

The abalone dataset contains measurements of 4177 abalones that were collected from the Marine Research Laboratories in Taroona in Tasmania [3]. Eight measurements are taken for each abalone. The seven variables are *Length, Diameter and Height, Whole Weight, Shucked Weight, Viscera Weight* and *Dried-Shell Weight*. The last variable is *numbers of rings*, representing the age of an abalone. In this analysis, the last variable is left out because *numbers of rings* is a discrete random variable.

The Figure 2 shows the parallel co-ordinate plot of the abalone data. Each line in the plot represent one observation (values against the number of the variable). Red colours are assigned according to high values of the first variables and the blue colours, in contrast, are assigned to the low values. From this plot, we observe that the the abalone which had greater length (variable 1) often resulted in greater values in other variables. This indicates that variables are highly correlated and the variability of data may be attributed to a few number of sources

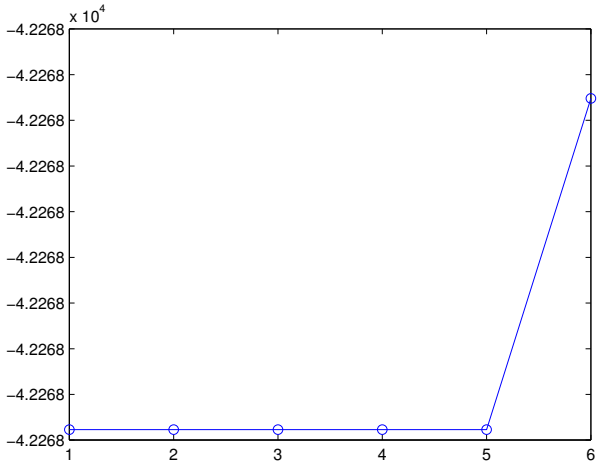

Figure 3: The log likelihood conditioned a fixed $k = 1, \cdots, 6$.

4.2.1 Decision of the number of latent variable through EM

The first step of the analysis is to select an appropriate number of the latent variables. It is observed that the determinant of the sample covariance is very close to zero, that is $\det(S) < 10^{-20}$. The hypothesis testing scheme would very unstable since the test statistic (18) depends on the reciprocal of the $\det(S)$. Instead of work out the test-statistic, which potentially causes numerical problem, we directly maximise the log-likelihood using EM algorithm.

The Figure 3 shows the log-likelihood of the k-factors model for $k = 1, \cdots, 6$. Although the maximum likelihood is obtained at $k = 6$, the log-likelihood for each k is very close. Considering more factor does not improve the log-likelihood very significantly. Hence, we take $k = 1$ and conclude the data essentially belong to one source.

Fixed $k = 1$, we then estimate $\mathbf{A}$. Since we only consider one factor, $\mathbf{A}$ is a $d \times 1$ vector. Table 3 gives the estimates corresponding to the variables. All the estimates are positive. The magnitude of the variable *Whole Weight* is much greater than the rest of the variable, a double of the second greatest. Those estimates are quite close to the standard deviations of each component and hence they also reflect the variation in the data.

Variable	Std. err.	Estimates
Length	0.1201	0.1121
Diameter	0.0992	0.0926
Height	0.0418	0.0344
Whole Weight	0.4904	0.4889
Shucked Weight	0.2200	0.2158
Viscera Weight	0.1096	0.1057
Dried-Shell Weight	0.1392	0.1325

Table 3: Table of the estimate of $\hat{\mathbf{A}}$ corresponding to the variables. In the middle column, it gives the standard deviation of each variable.

4.2.2 PC scores in the presence of missing data

As discussed, PPCA can deal with incomplete data. Making use of the likelihood, missing values are also treated as latent variable when performing EM algorithm and reconstruction of the PC scores (4) is available. The linear transform Γ_k is given by the spectrum decomposition of the matrix

$$\hat{\mathbf{A}}\hat{\mathbf{A}}^\top = \hat{\Gamma}_k(\hat{\Lambda}_k - \hat{\sigma}^2\mathbf{I})\hat{\Gamma}_k^\top$$

where the matrix $\hat{\mathbf{A}}$ and noise variance $\hat{\sigma}^2$ are attained from EM algorithm.

The PC score are reconstructed by

$$\mathbb{W}^{(k)} = \hat{\Gamma}_k^\top(\mathbb{X}_{imp} - \bar{\mathbf{X}}_{imp})$$

where $\mathbb{X}_{imp}$ is the imputed data matrix using the likelihood.

The function `ppca` in Matlab also provides a user-friendly framework for performing PPCA on incomplete data set.

We demonstrate the performance of the PPCA on the abalone in which 20% missing values are randomly simulated. Figure 4 shows the PC scores plot using PCA and PC scores plot of incomplete data using PPCA. On the right plot, we observe a pointy tail in the scatters, suggesting the abalone dataset is not normal. However, with 20% missing data, the PC scores plot we reconstructed using PPCA also gives a very similar structure. A notable difference is found at the tails. Since PPCA assumes normality, the imputed data are more evenly spread. In short, the performance of the PC scores reconstruction using PPCA is accurate although the generally normality assumption is violated.

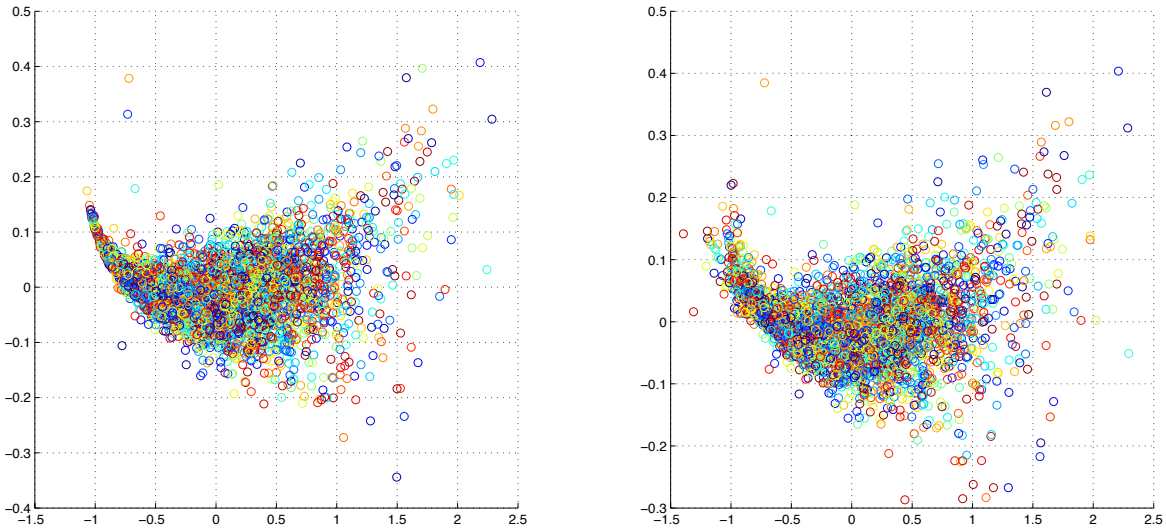

Figure 4: The PC1 vs PC2 scores plot of the original data using PCA (left) and the PC1 vs PC2 scores plot of the abalone data with 20% missing values using PPCA.

Conclusion

The goal of this project is to understand probabilistic principal component analysis (PPCA). It is to perform the principal component analysis on the data in the probabilistic way. Principal component analysis attempts to reduce the dimensionality of the data by discarding some variance.

The formulation of the principal component model involves the standard normal latent variables which leads to a constrained factor analysis model. Bishop and Tipping (1999) shows that an analytical for the maximum likelihood estimator exists for the probabilistic PC model. This suggest that, beside the spectrum decomposition of the sample covariance matrix, the PC scores can also be acquired through the likelihood.

Thus, PPCA has two practical advantages. Firstly, the number of latent variable or dimension of PC vector can be decided using hypothesis test. The scheme of the selecting such a number requires a series of hypothesis tests on a number of candidates. This feature is demonstrated using simulated dataset in Section 4. Secondly, the PC scores can be re-constructed for incomplete data using EM algorithm which is demonstrated using the abalone dataset in Section 4.

References

[1] M. E. Tipping and C. M. Bishop. Probabilitic principal component analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1999.

[2] I. Koch. *Analysis of Multivariate and High-Dimensional data*. Cambrigde Univsersity Press, 2013.

[3] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B Ford. The population biology of abalone in tasmania. i. blacklip abalone from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 1994.