Linear Algebra Report

Introduction to Probabilistic Principal Component Analysis

Tran Trung Viet, Huynh Huu Nhat, Le Hoai Nam &
Huynh Hoang Trung Nghia

May 2020

1 Introduction

Principal component analysis (PCA) (Jolliffe 1986) is an extremely popular technique to reduce dimension in multivariate analysis. Its many application areas include data compression, image analysis, visualization, pattern recognition, regression and time series prediction.

The most common definition of PCA, introduced by Hotelling (1933), for a set of observed d-dimensional data vectors $\{\mathbf{t}_n\}, n \in \{1...N\}$, the q principal axes $\mathbf{w}_j, j \in \{1...q\}$, are those orthonormal axes onto which the retained variance under projection is maximal. It can be shown that the vectors $\mathbf{w}_j$ are given by the q dominant eigenvectors (i.e. those with the largest associated eigenvalues) of the sample covariance matrix $\mathbf{S} = \sum_n (\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^T / N$ such that $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$ and where $\bar{\mathbf{t}}$ is the sample mean. The vector $\mathbf{x}_n = \mathbf{W}^T(\mathbf{t}_n - \bar{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_q)$, is thus a q-dimensional reduced representation of the observed vector $\mathbf{t}_n$.

A complementary property of PCA is that the projection onto the principal subspace minimizes the squared reconstruction error $\sum \|\mathbf{t}_n - \hat{\mathbf{t}}_n\|$. The optimal linear reconstruction of $\mathbf{t}_n$ is given by $\hat{\mathbf{t}}_n = \mathbf{W}\mathbf{x}_n + \bar{\mathbf{t}}$, where $\mathbf{x}_n = \mathbf{W}^T(\mathbf{t}_n - \bar{\mathbf{t}})$, and the orthogonal columns of $\mathbf{W}$ span the space of the leading q eigenvectors of $\mathbf{S}$.

One limiting disadvantage of these definitions of PCA is the absence of an associated probability density or generative model. Deriving PCA from the perspective of density estimation would offer a number of important advantages including the following:

- The corresponding likelihood would permit comparison with other density-estimation techniques and facilitate statistical testing.
- Bayesian inference methods could be applied by combining the likelihood with a prior.
- In classification, PCA could be used to model class-conditional densities, thereby allowing the posterior probabilities of class membership to be computed.

- The value of the probability density function could be used as a measure of the 'degree of novelty' of a new data point.
- The probability model would offer a methodology for obtaining a principal component projection when data values are missing.
- The single PCA model could be extended to a mixture of such models.

M.E Tipping and C.M Bishop introduces a probability formula of PCA from the Gaussian latent variable model that is closely related to statistical factor analysis and is discussed in Section 2. The PPCA model is described in detail in Section 3, the principal axes emerge as maximum likelihood parameter estimates which may be computed by the usual eigen-decomposition of the sample covariance matrix and subsequently incorporated in the model. Alternatively, the latent variable formulation leads naturally to an iterative, and computationally efficient, expectation–maximization (EM) algorithm for effecting PCA are presented in Section 3.5.

2 Main Concept behind PPCA

2.1 Latent Variable Models

A latent variable model seeks to relate a d-dimensional observed data vector t to a corresponding q-dimensional vector of latent variables x:

$$\mathbf{t} = \mathbf{y}(\mathbf{x};\mathbf{w}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}(\cdot;\cdot)$ is a function of the latent variables $\mathbf{x}$ with parameters $\mathbf{w}$, and $\boldsymbol{\epsilon}$ is an $\mathbf{x}$-independent noise process. Generally, $q < d$ such that the latent variables offer a more parsimonious description of the data. By defining a prior distribution over $\mathbf{x}$, together with the distribution of $\boldsymbol{\epsilon}$, equation (1) induces a corresponding distribution in the data space, and the model parameters may then be determined by maximum-likelihood techniques. Such a model may also be termed 'generative', as data vectors $\mathbf{t}$ may be generated by sampling from the $\mathbf{x}$ and $\boldsymbol{\epsilon}$ distributions and applying (1).

2.2 Factor Analysis

Factor analysis is a useful tool for investigate the relationship between direct variables (observation variables) and indirect variables (latent variables). The key concept of factor analysis is that observation variables have similar patterns behaviors because they are caused by the latent variables. In other words, the factor analysis tried to mapping latent variables dimension to observation dimension. Like GLM, this mapping function have to consider some noise [1].

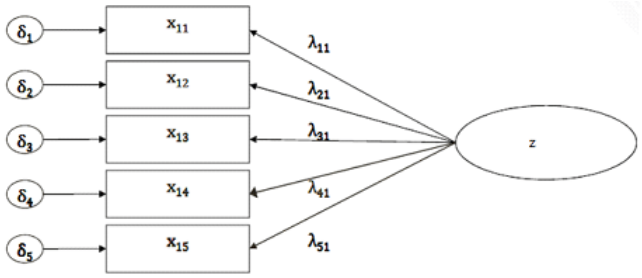

Figure 1: The basic model of Factor analysis

Often the latent have lower dimension than the observation variables. The q-dimensions map into d-dimension with $q < d$. The most common factor analysis model is the linear mapping which is:

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

The W is the weight matrix (i.e. λ_{zi} in figure 5) which transform x from q-dimension (i.e. z in figure 5) to t with d dimension (i.e. x in figure 5), so the W is (q x d) matrix. The $\boldsymbol{\mu}$ is the mean vectors which try to off-scale the t from it center. In the original paper, x is the latent variables with q-dimension vector and follow by the standard normal distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ($\mathbf{I}$ is the identical matrix) with zero mean and unit variance. Also, we assume that the error term have normal distribution as $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$. From this, we can derive the distribution for $\mathbf{t}$ as $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi})$. The $\boldsymbol{\Psi}$ is the variance diagonal matrix where ψ_i is the variance of i-feature/dimension. Then the basis of $\mathbf{W}$ become conditional independent given the latent variables $\mathbf{x}$. We can intuitively understand that when reducing the dimension we have to sacrifice some variance and that variance accounted to $\boldsymbol{\epsilon}$ term. So, the approach of PPCA is a little bit difference from regular PCA is that PPCA use the factor analysis theory to estimate $\mathbf{W}$ and $\boldsymbol{\epsilon}$ matrix, while in PCA we use SVD to calculate for $\mathbf{W}$ and $\boldsymbol{\epsilon}$ is equal $\mathbf{0}$. We can see that from the original PCA we try to find vector $\mathbf{x}$ with lower dimension, and through W matrix from SVD we find $\mathbf{x}$, so the equation of PCA is:

$$\mathbf{x} = \mathbf{W}^{-1}(\mathbf{t} - \boldsymbol{\mu}). \tag{3}$$

In below section (**Section 3**), we will provide the analytical result of PPCA with Maximization Likelihood Estimation (MLE) in case of there is no missing value. However, if there are missing value we can not use that result for calculating $\mathbf{W}$, instead we use EM (expectation maximization) algorithm for estimating the $\mathbf{W}$ and $\boldsymbol{\epsilon}$ matrix.

2.3 Links from Factor Analysis to PCA

In factor analysis the subspace defined by the columns of $\mathbf{W}$ will generally not correspond to the principal subspace of the data. Nevertheless, certain links

between the two methods, it has been observed that the factor loadings and the principal axes are quite similar in situations where the estimates of the elements of $\mathbf{\Psi}$ turn out to be approximately equal (e.g. Rao 1955). Indeed, this is an implied result of the fact that if $\mathbf{\Psi} = \sigma^2\mathbf{I}$ and an isotropic, rather than diagonal, noise model is assumed, then PCA emerges if the $d-q$ smallest eigenvalues of the sample covariance matrix $\mathbf{S}$ are exactly equal. Given this restriction, the factor loadings $\mathbf{W}$ and noise variance σ^2 are identifiable (assuming correct choice of q) and can be determined analytically through eigen-decomposition of $\mathbf{S}$, without resort to iteration (Anderson 1963).
However, this established link with PCA requires that the $d-q$ minor eigenvalues of the sample covariance matrix be equal (or more trivially, be negligible) and thus implies that the covariance model must be exact. Not only is this assumption rarely justified in practice, but when exploiting PCA for dimensionality reduction, we do not require an exact characterisation of the covariance structure in the minor subspace, as this information is effectively 'discarded'. In truth, what is of real interest is the form of the maximum-likelihood solution when the model covariance is not identical to its data sample counterpart.
Importantly, the authors show that PCA does still emerge in the case of an approximate model. The link between factor analysis and PCA was partially discovered by Lawley (1953) and Anderson and Rubin (1956). It was expanded and clearly demonstrated by EM and called PPCA. Michael E. Tipping and Christopher M. Bishop extend this earlier work that we covered again in Section 2.4. They give a full characterisation of the properties of the model we term "probabilistic PCA". Specifically, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, the columns of the maximum-likelihood estimator $\mathbf{W}_{ML}$ are shown to span the principal subspace of the data even when $\mathbf{C} \neq \mathbf{S}$.

3 Probabilistic principal component analysis

3.1 The probability model

3.1.1 The conditional distribution of the observation t given the latent variables x

The use of the isotropic Gaussian noise model $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ for $\boldsymbol{\epsilon}$ in conjunction with (2) implies that the $\mathbf{x}$ conditional probability distribution over $\mathbf{t}$-space is given by

$$\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I}). \tag{4}$$

3.1.2 The conditional distribution of the observation t

With the marginal distribution over the latent variables also Gaussian and conventionally defined by $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the marginal distribution for the observed data t is readily obtained by integrating out the latent variables and is likewise Gaussian:

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \tag{5}$$

where the observation covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$. This result can be derived from (2):

$$\mathbb{E}[\mathbf{t}] = \mathbb{E}[\mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu} \tag{6}$$

$$\begin{aligned} \mathrm{cov}[\mathbf{t}] &= \mathbb{E}[(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})^{\mathrm{T}}] \\ &= \mathbb{E}[\mathbf{W}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I} \end{aligned} \tag{7}$$

3.1.3 Likelihood function and the maximum likelihood estimators

The density function of distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$:

$$f(\mathbf{t}, \boldsymbol{\mu}, \mathbf{C}) = \frac{1}{(2\pi)^{d/2}|(\mathbf{C})|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{t} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{t} - \boldsymbol{\mu})\} \tag{8}$$

The corresponding log-likelihood is then

$$\mathcal{L} = \sum_{k=1}^{N} \ln p(\mathbf{t}_n|\boldsymbol{\mu}, \mathbf{C}) = -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{C}| - \frac{1}{2}\sum_{k=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{t}_n - \boldsymbol{\mu}) \tag{9}$$

Therefore,

$$\mathcal{L} = -\frac{N}{2}\{d\ \ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})\} \tag{10}$$

where

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

The maximum likelihood estimator for $\boldsymbol{\mu}$ is given by the mean of the data, in which case $\mathbf{S}$ is the sample covariance matrix of the observations $\{\mathbf{t}_n\}$. Estimates for $\mathbf{W}$ and σ^2 may be obtained by iterative maximization of $\mathcal{L}$, e.g. by using the **EM** algorithm given in **Section 3.5**, which is based on the algorithm for standard factor analysis of Rubin and Thayer (1982). However, in contrast with factor analysis, maximum likelihood estimators for $\mathbf{W}$ and σ^2 may be obtained explicitly, as we see shortly.

3.1.4 The conditional distribution of the latent variables x given the observed t

The formulation about Marginal and Conditional Gaussians:
Given a marginal Gaussian distribution for x and a conditional Gaussian distribution for y given x in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{12}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{13}$$

the marginal distribution of y and the conditional distribution of x given y are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{14}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{15}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1} \tag{16}$$

We use these formulations to determine the conditional distribution of the latent variables $\mathbf{x}$ given the observed $\mathbf{t}$:

$$\mathbf{x}|\mathbf{t} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}), \tag{17}$$

where we have defined $\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}$. Note that $\mathbf{M}$ is of size q × q whereas $\mathbf{C}$ is d × d.

3.2 Properties of the maximum likelihood estimators

The gradient of the log-likelihood (10) with respect to W may be obtained from standard matrix differentiation results (e.g. see Krzanowski and Marriott 1994, p. 133):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}) \tag{18}$$

At the stationary points:

$$\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W} \tag{19}$$

assuming that $\sigma^2 > 0$, and thus that $\mathbf{C}^{-1}$ exists. This is a necessary and sufficient condition for the density model to remain non-singular, and we will restrict ourselves to such cases. It will be seen shortly that $\sigma^2 > 0$ if $q < rank(\mathbf{S})$, so this assumption implies no loss of practicality.
There are three possible classes of solutions to equation (19):

1. $\mathbf{W} = \mathbf{0}$. This is shown later to be a minimum of the log-likelihood.

2. $\mathbf{C} = \mathbf{S}$, where the covariance model is exact, such as is discussed by Basilevsky (1994, pp 361–363). In this unrealistic case of an exact covariance model, where the $d - q$ smallest eigenvalues of $\mathbf{S}$ are identical and equal to σ^2, $\mathbf{W}$ is identifiable since

$$\begin{aligned} \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T &= \mathbf{S}, \\ \Rightarrow \mathbf{W} &= \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \end{aligned} \tag{20}$$

   where $\mathbf{U}$ is a square matrix whose columns are the eigenvectors of $\mathbf{S}$, with $\mathbf{\Lambda}$ the corresponding diagonal matrix of eigenvalues, and $\mathbf{R}$ is an arbitrary orthogonal (i.e. rotation) matrix.

3. $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$, with $\mathbf{W} = \mathbf{0}$ and $\mathbf{C} \neq \mathbf{S}$.

We are interested in case 3 where $\mathbf{C} - \mathbf{S}$ and the model covariance need not be equal to the sample covariance. First, we express the weight matrix $\mathbf{W}$ in terms of its singular value decomposition:

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^T, \tag{21}$$

where $\mathbf{U}$ is a $d \times q$ matrix of orthonormal column vectors, $\mathbf{L} = diag(l_1, l_2, ..., l_q)$ is the $q \times q$ diagonal matrix of singular values, and $\mathbf{V}$ is a $q \times q$ orthogonal matrix. Now,

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{W} &= (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}, \\ &= \mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^T\mathbf{W})^{-1}, \\ &= \mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^T. \end{aligned} \tag{22}$$

Then at the stationary points, $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$ implies that

$$\begin{aligned} \mathbf{S}\mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^T &= \mathbf{U}\mathbf{L}\mathbf{V}^T, \\ \Rightarrow \mathbf{S}\mathbf{U}\mathbf{L} &= \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}. \end{aligned} \tag{23}$$

For $l_j \neq 0$, equation (39) implies that if $\mathbf{U} = (u_1, u_2, ..., u_q)$, then the corresponding column vector $\mathbf{u}_j$ must be an eigenvector of $\mathbf{S}$, with eigenvalue λ_j such that $\sigma^2 + l_j^2 = \lambda_j$, and so

$$l_j = (\lambda_j - \sigma^2)^{\frac{1}{2}} \tag{24}$$

For $l_j = 0$, $\mathbf{u}_j$ is arbitrary (and if all l_j are zero, then we recover case 1). All potential solutions for $\mathbf{W}$ may thus be written as

$$\mathbf{W} = \mathbf{W}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{25}$$

where $\mathbf{U}_q$ is a $d \times q$ matrix comprising q column eigenvectors of $\mathbf{S}$, and $\mathbf{K}_q$ is a $q \times q$ diagonal matrix with elements: $k_j = \lambda_j$ corresponding eigenvalue to $\mathbf{u}_j$ or

$k_j = \sigma^2$ if $l_j = 0$. $\mathbf{R}$ is an arbitrary orthogonal matrix, equivalent to a rotation in the principal subspace.

Other combinations of eigenvectors (i.e. non-principal ones) correspond to saddlepoints of the likelihood function. Thus, from equation (**??**), the latent variable model defined by (2) effects a mapping from the latent space into the *principal subspace* of the observed data.

The matrix $\mathbf{U}_q$ may contain any of the eigenvectors of $\mathbf{S}$, so to identify those which maximize the likelihood, the expression for W in (25) is substituted into the log-likelihood function (10) to give

$$\mathcal{L} = -\frac{N}{2}\Big\{dln(2\pi) + \sum_{j=1}^{q'} ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=q'+1}^{d}\lambda_j + (d-q')ln(\sigma^2) + q'\Big\} \qquad (26)$$

where q' is the number of non-zero $l_j, \{\lambda_1, ..., \lambda_{q'}\}$ are the eigenvalues corresponding to those 'retained' in $\mathbf{W}$, and $\{\lambda_{q'+1}, ..., \lambda_d\}$ are those 'discarded'. Maximizing (26) with respect to σ^2 gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j, \qquad (27)$$

and so

$$\mathcal{L} = -\frac{N}{2}\Big\{\sum_{j=1}^{q'} ln(\lambda_j) + (d-q')ln\Big(\frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j\Big) + dln(2\pi) + d\} \qquad (28)$$

Note that (27) implies that $\sigma^2 > 0$ if $\mathrm{rank}(\mathbf{S}) > q$ as stated earlier. We wish to flnd the maximum of (28) with respect to the choice of eigenvectors/eigenvalues to retain in $\mathbf{W}, j \in \{1, ..., q'$ and those to discard, $j \in \{q'+1, ..., d$. By exploiting the constancy of the sum of all eigenvalues with respect to this choice, the condition for maximization of the likelihood can be expressed equivalently as minimization of the quantity

$$E = ln\Big(\frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j\Big) - \frac{1}{d-q'}\sum_{j=q'+1}^{d}(\lambda_j) \qquad (29)$$

which depends only on the discarded values and is non-negative (Jensen's inequality).
We consider minimization of E by first assuming that $d-q'$ discarded eigenvalues have been chosen arbitrarily, and, by differentiation, consider how a single such value λ_k affects the value of E:

$$\frac{\partial E}{\partial \lambda_k} = \frac{1}{\sum_{j=q'+1}^{d}\lambda_j} - \frac{1}{(d-q')\lambda_k} \qquad (30)$$

From (30), it can be seen that $E(\lambda_k)$ is convex and has a single minimum when λ_k is equal to the mean of the discarded eigenvalues (including itself). The eigenvalue λ_k can only take discrete values, but if we consider exchanging k for some retained eigenvalue $\lambda_j, j \in \{1...q^{'}\}$, then if λ_j lies between λ_k and the current mean retained eigenvalue, swapping λ_j and λ_k must decrease E. If we consider that the eigenvalues of $\mathbf{S}$ are ordered, for any combination of discarded eigenvalues which includes a 'gap' occupied by a retained eigenvalue, there will always be a sequence of adjacent eigenvalues with a lower value of E. It follows then that to minimize E, the discarded eigenvalues $\lambda_{q'+1}, ..., \lambda_d$ must be chosen to be adjacent amongst the ordered eigenvalues of S.
This alone is not sufficient to show that the smallest eigenvalues must be discarded in order to maximize the likelihood. However, a further constraint is available from equation (24), since $l_j = (\lambda_j - \sigma^2)$ implies that there can be no real solution to the stationary equations of the log-likelihood if any retained eigenvalue $\lambda_j < \sigma^2$. Since, from (27), σ^2 is the average of the discarded eigenvalues, this condition would be violated if the smallest eigenvalue were not discarded. Now, combined with the previous result, this indicates that E must be minimized when $\lambda_{q'+1}, ..., \lambda_d$ are the smallest $d - q^{q^{'}}$ eigenvalues and so L is maximized when $\lambda_1, ..., \lambda_1$ are the principal eigenvalues of S. It should also be noted that the log-likelihood $\mathcal{L}$ is maximized, with respect to $q^{'}$, when there are fewest terms in the sum in (29) which occurs when $q^{'} = q$ and therefore no l_j is zero.

It may also be shown that for $\mathbf{W} = \mathbf{W}_{\mathrm{ML}}$, the maximum likelihood estimator for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \tag{31}$$

which has a clear interpretation as the variance 'lost' in the projection, averaged over the lost dimensions.
In practice, to find the most likely model given $\mathbf{S}$, we would first estimate σ^2_{ML} from equation (31), and then $\mathbf{W}_{\mathrm{ML}}$ from equation (25), where for simplicity we would effectively ignore $\mathbf{R}$ (i.e. choose $\mathbf{R} = \mathbf{I}$). Alternatively, we might employ the **EM** algorithm detailed in Section 3.5, where $\mathbf{R}$ at convergence can be considered arbitrary.

3.3 Factor analysis revisited

Although the above estimators result from the application of a simple constraint to the standard factor analysis model, we note that an important distinction resulting from the use of the isotropic noise covariance $\sigma^2\mathbf{I}$ is that PPCA is covariant under rotation of the original data axes, as is standard PCA, whereas factor analysis is covariant under componentwise rescaling. Another point of contrast is that in factor analysis neither of the factors found by a two-factor model is necessarily the same as that found by a single-factor model. In PPCA, we see above that the principal axes may be found incrementally.

3.4 Dimensionality reduction

The general motivation for PCA is to transform the data into some reduced dimensionality representation, and with some minor algebraic manipulation of $\mathbf{W}_{\mathrm{ML}}$ we may indeed obtain the standard projection onto the principal axes if desired. However, it is more natural from a probabilistic perspective to consider the dimensionality reduction process in terms of the distribution of the latent variables, conditioned on the observation. From expression (17), this distribution may be conveniently summarized by its *mean*:

$$\langle \mathbf{x}_n | \mathbf{t}_n \rangle = \mathbf{M}^{-1} \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} (\mathbf{t}_n - \boldsymbol{\mu}). \tag{32}$$

(Note, also from expression (17), that the corresponding conditional covariance is given by $\sigma^2_{\mathrm{ML}} \mathbf{M}^{-1}$ and is thus independent of n.) It can be seen that, when $\sigma^2 \to 0, \mathbf{M}^{-1} \to (\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \mathbf{W}_{\mathrm{ML}})^{-1}$ and equation (32) then represents an orthogonal projection into latent space and so standard PCA is recovered. However, the density model then becomes singular, and thus undefined. In practice, with $\sigma^2 > 0$ as determined by (31), the latent projection becomes skewed towards the origin as a result of the Gaussian marginal distribution for $\mathbf{x}$. Because of this, the reconstruction $\mathbf{W}_{\mathrm{ML}} \langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$ is *not* an orthogonal projection of $\mathbf{t}_n$ and is therefore not optimal (in the squared reconstruction error sense). Nevertheless, optimal reconstruction of the observed data from the conditional latent mean may still be obtained, in the case of $\sigma^2 > 0$, and is given by $\mathbf{W}_{\mathrm{ML}} (\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \mathbf{W}_{\mathrm{ML}})^{-1} \mathbf{M} \langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$.

3.5 Expectation Maximization of PPCA

In the above section, we have visited the closed form of PPCA with MLE. However, this form only works with complete data and is computational costly. In this section, we will derive an EM algorithms for PPCA from above assumptions. EM algorithm estimates the model parameters from latent variable model. It includes 2 main steps: E-step is the computation of the expectation of complete data log-likelihood and M-step is the maximization of the result from E-step. In the high-level concept of EM, we try to compute $q(\mathbf{x})^*$ with a choosing of θ_{old}, then use $q(\mathbf{x})^*$ to compute θ_{new} until they converge [2]. The author also proved that the current value of $\mathcal{L}(q, \theta)_{new}$ is always larger than the former value of $\mathcal{L}(q, \theta)_{old}$

$$\mathcal{L}(q, \theta) = \sum_z q(z) \ln p(x, z | \theta) \tag{33}$$

As mentioned above, we have assumed that $\mathbf{x}$ is normally distributed with $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$, so we only have to estimate the parameters. In the EM approach, to maximize the likelihood for PPCA, we consider the latent variables $\mathbf{x}_n$ to be 'missing' data and the 'complete' data to comprise the observations together with these latent variables. The corresponding complete-data log-likelihood is then:

$$\mathcal{L}_{\mathrm{C}} = \sum_{n=1}^{N} \ln \left\{ p\left(\mathbf{t}_n, \mathbf{x}_n | \mathbf{W}, \boldsymbol{\epsilon}\right) \right\} \tag{34}$$

Where, in PPCA, section 3:

$$p\left(\mathbf{t}_n, \mathbf{x}_n | \mathbf{W}, \boldsymbol{\epsilon}\right) = \left(2\pi\sigma^2\right)^{\frac{-d}{2}} \exp\left(-\frac{\left\| \mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu} \right\|^2}{2\sigma^2} \right) (2\pi)^{\frac{-q}{2}} \exp\left(-\frac{\left\| \mathbf{x}_n \right\|^2}{2} \right) \tag{35}$$

So, we will begin with the E-step to compute the parameters. We take the expectation of $\mathcal{L}_{\mathrm{C}}$ with respected to $p\left(\mathbf{t}_n, \mathbf{x}_n | \mathbf{W}, \boldsymbol{\epsilon}\right)$:

$$\begin{aligned} \left\langle \mathcal{L}_{\mathrm{C}} \right\rangle = -\sum_{n=1}^{N} \Big\{ & \frac{d}{2} \ln\left(\sigma^2\right) + \frac{1}{2} \operatorname{tr}\left(\left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle \right) + \frac{1}{2\sigma^2} \left(\mathbf{t}_n - \boldsymbol{\mu} \right)^{\mathrm{T}} \left(\mathbf{t}_n - \boldsymbol{\mu} \right) \\ & - \frac{1}{\sigma^2} \left\langle \mathbf{x}_n \right\rangle^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} \left(\mathbf{t}_n - \boldsymbol{\mu} \right) + \frac{1}{2\sigma^2} \operatorname{tr}\left(\mathbf{W}^{\mathrm{T}} \mathbf{W} \left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle \right) \Big\} \end{aligned} \tag{36}$$

Some terms were omitted because they are independent of the parameters, and also:

$$\left\langle \mathbf{x}_n \right\rangle = \mathbf{M}^{-1} \mathbf{W}^{\mathrm{T}} \left(\mathbf{t}_n - \boldsymbol{\mu} \right) \tag{37}$$

$$\left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle = \sigma^2 \mathbf{M}^{-1} + \left\langle \mathbf{x}_n \right\rangle \left\langle \mathbf{x}_n \right\rangle^{\mathrm{T}} \tag{38}$$

With $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$. Reminded that these statistics are the current parameters (fixed or θ_{old}) and follow distribution (17).

In the M-step, we try to maximize $\left\langle \mathcal{L}_{\mathrm{C}} \right\rangle$ according to $\mathbf{W}, \boldsymbol{\epsilon}$, so we have:

$$\tilde{\mathbf{W}} = \left\{ \sum_{n=1}^{N} \left(\mathbf{t}_n - \boldsymbol{\mu} \right) \left\langle \mathbf{x}_n \right\rangle^{\mathrm{T}} \right\} \left(\sum_{n=1}^{N} \left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle \right)^{-1} \tag{39}$$

$$\tilde{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \left\{ \left\| \mathbf{t}_n - \boldsymbol{\mu} \right\|^2 - 2 \left\langle \mathbf{x}_n \right\rangle^{\mathrm{T}} \widetilde{\mathbf{W}}^{\mathrm{T}} \left(\mathbf{t}_n - \boldsymbol{\mu} \right) + \operatorname{tr}\left(\left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle \widetilde{\mathbf{W}}^{\mathrm{T}} \widetilde{\mathbf{W}} \right) \right\} \tag{40}$$

This is the complete algorithm of EM for PPCA. To maximize the likelihood then, the sufficient statistics of the conditional distributions are calculated from equations (37) and (38), after which revised estimates of the parameters are obtained from equations (38) and (39). These four equations are iterated in sequence until the algorithm is judged to have converged. We may gain considerable insight into the operation of the EM algorithm by substituting for $\left\langle \mathbf{x}_n \right\rangle$ and $\left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle$ from equations (37) and (38) into equations (38) and (39). Some further manipulation leads to both the E-step and the M-step being combined and rewritten as

$$\tilde{\mathbf{W}} = \mathbf{S}\mathbf{W} \left(\sigma^2 \mathbf{I} + \mathbf{M}^{-1} \mathbf{W}^{\mathrm{T}} \mathbf{S} \mathbf{W} \right)^{-1} \tag{41}$$

$$\tilde{\sigma}^2 = \frac{1}{d} \operatorname{tr}\left(\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\right) \tag{42}$$

With $\mathbf{S}$ from equation (11). We can see that EM is the iteration algorithm so the $\tilde{\mathbf{W}}$ is represented for θ_{new} and $\mathbf{W}$ is represented for θ_{old}. From (41) and (42), we know that the data go through the EM algorithm only from $\mathbf{S}$ from equation (11).

3.6 Handling Missing value with EM

In practice, the original data sometimes contains missing value due to error measurement or human errors. So utilizing EM algorithm may help us to estimate the missing value. As we mentioned earlier in Section 3.5, the EM algorithm consists of two recursive steps, i.e. choose θ_{old} to calculate $q(\mathbf{x})^*$, and form $q(\mathbf{x})^*$ to calculate θ_{new}, with x is the latent variables. In case there are missing values in the dataset, we introduce another latent variables representing for the missing data $q(\mathbf{t}_m)$. With this we will have to construct the EM algorithm again. The implementation form "pcaMethods" package of Bioconductor constructed the $\tilde{\sigma}^2$ and $\tilde{\mathbf{W}}$ equation for EM iteration for missing values [3].

$$\tilde{\mathbf{W}} = \left(N\mathbf{S} + \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^{\top}\right)^{-1}\bar{\boldsymbol{X}}\bar{\boldsymbol{T}}^{\top} \tag{43}$$

$$\tilde{\boldsymbol{\sigma}^{\mathbf{2}}} = \frac{1}{Nd}\left(N \operatorname{Tr}\left\{\mathbf{W}^{\top}\boldsymbol{\Sigma}\mathbf{W}\right\} + \sum_{n}\left\|\bar{\mathbf{t}}_n - \mathbf{W}^{\top}\bar{\mathbf{x}}_n\right\| + D_h\boldsymbol{\sigma}^2\right) \tag{44}$$

We have that D_h is the total number of missing values, and N is the total number of observations. For each iteration, we will update our parameters as $\mathbf{W}$, $\tilde{\boldsymbol{\sigma}^{\mathbf{2}}}$ and missing values. We recalculated $\hat{\mathbf{t}}_n$ as [4]

$$\begin{aligned}\hat{\mathbf{t}}_n &= \mathbf{W}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{M}\left\langle\mathbf{x}_n\right\rangle \\ &= \mathbf{W}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n\end{aligned} \tag{45}$$

When we have $\hat{\mathbf{t}}_n$, we only update the missing values of $\mathbf{t}_n$ from $\hat{\mathbf{t}}_n$. Noted that we could replace the "NaN" value in the dataset with "0" or other simple imputation methods (i.e. mean, mode, median) before running the PPCA algorithm.

4 Application

4.1 Running Time Comparation

We test the calculate time between the closed form of PPCA and EM of PPCA. It shows that the EM algorithm used in PPCA is relatively faster than the closed form. In the test, we use the breast cancer dataset with missing value. The original data consists of 9 features and 699 observations. However, this

advantage is not always hold. The difference between running time often depend on the size of the data and the choice of q dimensions. Sometime, choosing a bad q might lead to "forever iteration" because the algorithm is not converged.

```
1  %time
2  PPCA(data, 3)

CPU times: user 3 µs, sys: 1e+03 ns, total: 4 µs
Wall time: 6.91 µs
(array([[-1.89,  0.11,  1.14],
        [-2.53, -0.42,  0.17],
        [-2.5 , -0.3 ,  0.08],
        [-2.12,  0.06, -0.81],
        [-1.55, -0.3 , -0.01],
        [-2.84,  1.39, -0.11],
        [-1.86, -0.06, -0.13],
        [-2.26, -0.76, -0.19],
        [-0.82, -0.23, -0.09]]), 2.24623849594828, array([[4.4248927 ],
        [3.14306152],
        [3.21459227],
        [2.85121602],
        [3.1702432 ],
        [3.67095851],
        [3.4148784 ],
        [2.86409156],
        [1.63948498]]))
```

Figure 2: Running time of MLE closed form

```
[44] 1  %time
     2  PPCA(data, 3, EM=True)

CPU times: user 1e+03 ns, sys: 0 ns, total: 1e+03 ns
Wall time: 4.29 µs
(array([[0.34735852, 0.48085443, 2.12993241],
        [1.04890626, 1.50187771, 1.80768724],
        [1.15143957, 1.40747757, 1.74730125],
        [1.71467381, 1.17493721, 0.91217847],
        [0.6926017 , 0.98142298, 1.02401569],
        [2.35235559, 0.21221763, 2.10668832],
        [1.06528794, 0.96229442, 1.1901868 ],
        [0.9540902 , 1.7593317 , 1.31583803],
        [0.37828725, 0.60309452, 0.46779596]]),
 2.2462471956575O6,
 array([[4.4248927 ],
        [3.14306152],
        [3.21459227],
        [2.85121602],
        [3.1702432 ],
        [3.67095851],
        [3.4148784 ],
        [2.86409156],
        [1.63948498]]))
```

Figure 3: Running time of EM

4.2 Missing data Dimension Reduction

The most interesting about PPCA is that it can run with the missing data. By employing the EM, the PPCA algorithm can reconstruct the missing value and produce the important components out of it. The flow is to update the missing values in every iteration of EM. This can be done by considering the missing values as another latent variable's distribution. We run the PPCA on the Breast Cancer dataset with 50 missing values in each feature. We can see that PPCA is effectively projected to lower dimensions in case that there is missing data compared to full data PCA.

4.3 Missing data Imputation

PPCA offers a natural approach to the estimation of the principal axes in cases where some, or indeed all, of the data vectors $\mathbf{t}_n = (t_{n1}, ..., t_{nd})$ exhibit one

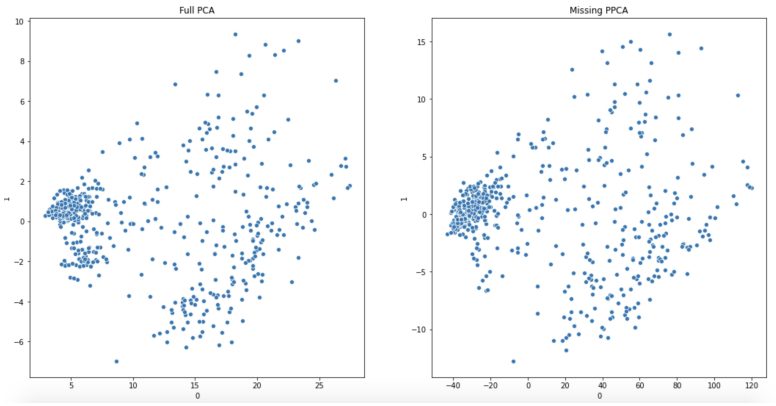

Figure 4: Comparison between Full data PCA vs Missing data PPCA

or more missing (at random) values. Drawing on the standard methodology for maximizing the likelihood of a Gaussian model in the presence of missing values (Little and Rubin, 1987) and the EM algorithm for PPCA is an iterative algorithm for maximum likelihood estimation of the principal axes, where both the latent variables $\mathbf{x}_n$ and the missing observations t_{nj} make up the 'complete' data.

We used William H. Wolberg's Diagnostic Wisconsin Breast Cancer Database dataset to test PPCA's ability to fill in NaN values with a natural method, filling the NaN value with mean. We randomly dropped 50 values for each characteristic column of the data set and revisited their positions. We then filled these missing values with mean values and ran PPCA on this data set and obtained the $\mathbf{W}_{ML}$ weight matrix. We use the reconstruction property of PPCA to rebuild missing values using the W matrix. We perform root mean square error (RMSE) calculation between missing values reconstructed by PPCA and its true value from the original data. We also calculate the RMSE between the values filled with Mean and the true value. PPCA's RMSE are 13% to 54% lower than filled with Mean and 26% lower on average as shown in Figure 2.

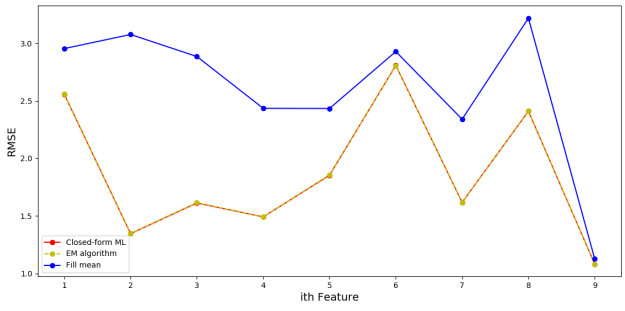

Figure 5: PPCA compared to fixed filled with Mean

References

[1] David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons, 2011.

[2] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[3] Josep Porta, Jakob Verbeek, and Ben Krose. Active appearance-based robot localization using stereo vision. *Autonomous Robots*, 18(1):59–80, 2005.

[4] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analysers. 1998.