Journal of Statistical and Econometric Methods, vol.1, no.2, 2012, 109-124
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2012

Interpretation of the Probabilistic Principal Components Analysis with Anisotropic Gaussian Distribution of Latent Variables

Adeleh Vosta[1], Farhad Yaghmaei[2] and Manoochehr Babanezhad[3]

Abstract

Principal component analysis (PCA) is a well established technique for data analysis and processing. Recently, it has been shown that the principal axes of a set of observed data vectors might be determined trough maximum likelihood estimation of parameter in a specific form of latent variable model closely related to factor analysis. It is assumed that the latent variables have a unit isotropic Gaussian distribution. In view of this, in this study, we express some interpretation for covariance between PPCs, correlation between PPCs and variables, and covariance matrix between PPCs and PCs in common PCA case. Further, we consider more general case in which the latent variables are independent with different variances. We also investigate properties of the associated likelihood function.

Keywords: Principal component analysis, Latent variable, Maximum likelihood, Dimensionality reduction, Anisotropic distribution

[1] Department of Statistics, Faculty of Sciences, Golestan University, Gorgan, Golestan, Iran, e-mail: adele.vosta@yahoo.com
[2] Department of Statistics, Faculty of Sciences, Golestan University, Gorgan, Golestan, Iran, e-mail: f_yaghmaie@yahoo.com
[3] Department of Statistics, Faculty of Sciences, Golestan University, Gorgan, Golestan, Iran, e-mail: m.babanezhad@gu.ac.ir

Article Info: *Received* : April 5, 2012. *Revised* : May 30, 2012
Published online : July 30, 2012

1 Introduction

It is well known that PCA is a dimensionality reduction technique which is used in many application areas such as data compression, image processing, data visualization, pattern recognition, and so on. Common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space. For a set of d-dimensional observation vectors $\{t_1, ..., t_N\}$, PCA can be obtained by computing the sample variance matrix, $S = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)'$ and by finding the eigenvectors u_i and eigenvalues λ_i $(i = 1, 2, ..., d)$ such that; $Su_i = \lambda_i u_i$, where $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$ is the data sample mean. The q principal axes ($q < d$ for parsimonious representation) $U = (u_1, u_2, ..., u_q)$ is the orthogonal onto which the retained variance under projection is maximized and U is corresponds to the eigenvalues of sample covariance matrix S [2]. A q dimensional reduction representation of the observed vector t_n is thus $Z_n = U'(t_n - \bar{t})$ and the covariance matrix $\frac{1}{N}\sum Z_n Z_n'$ is diagonal with uncorrelated elements $(\lambda_1, \lambda_2, ..., \lambda_q)$. An important property of PCA is that, it corresponds to the linear projection for which the sum of square reconstruction error $\sum \left\|t_n - \hat{t_n}\right\|^2$ is minimized; $\hat{t_n} = UZ_n + \mu$ where μ is mean vector. One limiting disadvantage of common PCA is the absence of a probability density model, where this is solved in a notable paper by Tipping and Bishop (1999). They in fact introduced a probability model in to PCA in which extent to assume the observed data is linear mapping of latent variables, with unit isotropic Gaussian distribution plus Gaussian error [6]. Driving PCA from the perspective of density estimation convert the common PCA into statistical inference problem. Further, Bayesian inference method is also applied into PCA [7].

In view of this, this paper is organized as follows; the latent variable model and probabilistic principal component analysis (PPCA) with unit isotropic Gaussian distribution of latent variable are described in next section. In section 3, we express some interpretation for probabilistic principal components. PPCA with anisotropic Gaussian distribution of latent variables and dimensionality reduction are investigated in section 4. We give an exmple in section 5, and conclusion is summarized in section 6.

2 Latent variable model and PPCA with isotropic Gaussian distribution of latent variables

2.1 Latent Variable Model

The goal of the latent variable model is to express the set of d-dimensional data vectors $\{t_n\}$ in terms of a smaller number of latent variables $X = (X_1, X_2, ..., X_q)$

(where $q < d$), such that,

$$t = y(X; W) + \epsilon$$

where $y(X; W)$ is a function of the latent variables X with parameter W and ϵ is an X-independent noise process. The definition of the latent variable model is completed with determining the distribution of ϵ, the mapping $y(X; W)$ and the prior distribution of latent variable.
One of the simplest latent variable model is Factor analysis, in which the mapping $y(X; W)$ is linear so that:

$$t = WX + \mu + \epsilon \tag{1}$$

where W is $d \times q$ parameter matrix and parameter μ is non-zero mean vector. The distribution of X is defined to be a zero mean unit covariance Gaussian $N(0, I)$. While the noise model for ϵ is also a zero mean Gaussian with a diagonal covariance matrix Ψ [3]. It follows from (2) that the distribution of t is also normal $N(\mu, C)$ where, $C = \Psi + WW'$. Because of WW' term in the covariance C, the likelihood function is invariant with respect to orthogonal post multiplication of W.

In Factor analysis because of the diagonal noise model Ψ the factor loadings W will in general differ from the principal axes. It is considered that, principal component emerge when the data is assumed to comprise a systematic component plus an independent error term for each variable with common variance σ^2 [8]. Thus the similarity between the factor loadings and the principal axes can be observed when the diagonal element of Ψ be equal.

2.2 PPCA with isotropic Gaussian distribution of latent variables

By considering the model (1) with an isotropic noise structure such that, $\Psi = \sigma^2 I_d$ and isotropic Gaussian distribution of latent variable $X \sim N(0, I)$, it is shown that the columns of maximum likelihood estimation W_{ML} are the scaled and rotated eigenvectors of sample covariance matrix S. In PPCA has been assumed that, the observation t_n is a linear transformation of X_n (where X_n is normally distributed. Here we show this by X $N(0, I)$), with additive Gaussian noise ϵ which is normally distributed ($\epsilon \sim N(0, \Psi)$). It is also follows from (1) that;

$$\begin{aligned}
&t|X \sim N(WX + \mu, \sigma^2 I),\\
&t \sim N(\mu, C), \quad \text{where} \quad C = WW' + \sigma^2 I_d,\\
&X|t \sim N\left(M^{-1}W'(t - \mu), \sigma^2 M^{-1}\right), \quad \text{where} \quad M = WW' + \sigma^2 I_q.
\end{aligned}$$

The parameters of model can be also estimated by maximizing the log likelihood of observed data as follows;

$$\ell = \sum lnp(t_n) = -\frac{Nd}{2}ln(2\pi) - \frac{N}{2}ln\,|C| - \frac{N}{2}tr\left[C^{-1}S\right],$$

where

$$S = \frac{1}{N}\sum_{n=1}^{N}(t_n - \mu)(t_n - \mu)',$$
$$W_{ML} = U_q(\Lambda_q - \sigma^2 I_q)^{\frac{1}{2}}R,$$
$$\sigma^2_{ML} = \frac{1}{d-q}\sum_{j=1}^{d-q}\lambda_j,$$

where U_q is matrix whose columns are eigenvectors of S and Λ_q is diagonal matrix with corresponding q eigenvalues of S, and R is an arbitrary orthogonal rotation matrix.

Also a dimensionality reduction representation for observed data is computed as (Tipping and Bishop, 1999):

$$\langle X_n \rangle = M^{-1}W'(t_n - \mu).$$

We use model (1) with isotropic noise model ϵ with normal distribution $N(0, \Psi)$ and anisotropic Gaussian distribution of latent variable X with normal distribution $N(0, V)$, where V is diagonal matrix with different elements. Further, we investigate the properties of the maximum likelihood estimator for this model under the latter assumptions.

3 Interpretation of probabilistic principal components(PPCs)

3.1 Covariance matrix of probabilistic principal components

According to the maximum likelihood estimator of matrix W, W_{ML}, and with consideration $R = I$, we can calculate the covariance matrix between PPCs as follows;

$$\begin{aligned} M &= W'_{ML}W_{ML} + \sigma^2 I_q \\ &= \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} U'_q U q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} \\ &= \Lambda_q \end{aligned}$$

Therefore

$$\begin{aligned} Cov(\langle X\rangle) &= Cov(M^{-1}W't) \\ &= M^{-1}W'SWM^{-1} \\ &= (\Lambda_q)^{-1}\left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} U_q' S U_q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} (\Lambda_q)^{-1}. \end{aligned} \quad (2)$$

With substituting spectral decomposition of sample covariance matrix, S into (2),

$$Cov(\langle X\rangle) = (\lambda_q)^{-1}(\Lambda_q - \sigma^2 I_q). \quad (3)$$

Note that for simplicity, the mean vector of d-dimensional vector t, is assumed to be zero. So, probabilistic principal components are independent, and variance of each component is given by:

$$Var(\langle X\rangle_k) = \frac{\lambda_k - \sigma^2}{\lambda_k}$$

3.2 Correlations between variables and probabilistic principal components

For interpretation of a probabilistic principal components we can use the correlation between variables and the components.

To obtain an expression for $\rho(t_j, \langle X\rangle_k)$, the correlation between jth variable in observed vector t and kth probabilistic principal component, we begin with the vector of sample covariance between the variables in t and the kth component $\langle X\rangle_k$ that compute as follows;

$$\begin{aligned} \langle X\rangle &= (\Lambda_q)^{-1}W't \\ &= \begin{pmatrix} \frac{1}{\lambda_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & 0 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & \frac{1}{\lambda_{q-1}} & 0 \\ 0 & \cdots & 0 & 0 & \frac{1}{\lambda_q} \end{pmatrix} \begin{pmatrix} W_{11} & W_{12} & W_{13} & \cdots & W_{1q} \\ W_{21} & W_{22} & W_{13} & \cdots & W_{2q} \\ W_{31} & W_{32} & W_{13} & \cdots & W_{3q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ W_{d1} & W_{d2} & W_{32} & \cdots & W_{dq} \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ \vdots \\ t_d \end{pmatrix} \\ &= \begin{pmatrix} (\frac{1}{\lambda_1}, 0, \cdots, 0)W't \\ (0, \frac{1}{\lambda_2}, \cdots, 0)W't \\ \vdots \\ (0, 0, \cdots, \frac{1}{\lambda_q})W't \end{pmatrix}. \end{aligned}$$

Therefore

$$\langle X\rangle_k = \left(0, 0, \cdots, \frac{1}{\lambda_k}, \cdots, 0\right) W' t. \tag{4}$$

Now, we compute $Cov(t_j, \langle X\rangle_k)$;

$$\begin{aligned} Cov(t_j, \langle X\rangle_k) &= Cov(I_j' t, \langle X\rangle_k) \\ &= Cov\left(I_j' t, \left(0, 0, \cdots, \frac{1}{\lambda_k}, \cdots, 0\right) W' t\right), \end{aligned}$$

where I_j is a $q \times 1$ vector in which its jth element is one and others are zero. Then

$$\begin{aligned} Cov(t_j, \langle X\rangle_k) &= I_j' S W \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\lambda_k} \\ \vdots \\ 0 \end{pmatrix} \\ &= I_j' U \Lambda U' U_q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\lambda_k} \\ \vdots \\ 0 \end{pmatrix} \\ &= I_j' U_q \Lambda_q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{1}{\lambda_k} \\ \vdots \\ 0 \end{pmatrix} \\ &= I_j' \begin{pmatrix} U_{11} & U_{12} & U_{13} & \cdots & U_{1q} \\ U_{21} & U_{22} & U_{13} & \cdots & U_{2q} \\ U_{31} & U_{32} & U_{13} & \cdots & U_{3q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ U_{d1} & U_{d2} & U_{32} & \cdots & U_{dq} \end{pmatrix} \begin{pmatrix} (\lambda_1 - \sigma^2)^{\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & (\lambda_2 - \sigma^2)^{\frac{1}{2}} & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & 0 & (\lambda_q - \sigma^2)^{\frac{1}{2}} \end{pmatrix} \\ &= (0, 0, \cdots, 1, \cdots, 0) \begin{pmatrix} U_{1k}(\lambda_k - \sigma^2)^{\frac{1}{2}} \\ U_{2k}(\lambda_k - \sigma^2)^{\frac{1}{2}} \\ \vdots \\ U_{jk}(\lambda_k - \sigma^2)^{\frac{1}{2}} \\ \vdots \\ U_{dk}(\lambda_k - \sigma^2)^{\frac{1}{2}} \end{pmatrix} \\ &= U_{jk}(\Lambda_k - \sigma^2 I_q)^{\frac{1}{2}}. \end{aligned} \tag{5}$$

The standard variation of t_j is $\sqrt{S_{jj}}$, the square root of jth diagonal element of S, and the standard deviation of $\langle X\rangle_k$ is $\left(\frac{\lambda_k-\sigma^2}{\lambda_k}\right)^{\frac{1}{2}}$.

Hence the correlation between the jth variable and kth component, $\langle X\rangle_k$ is given by:

$$\begin{aligned}\rho(t_j,\langle X\rangle_k) &= \frac{Cov\left(t_j,\langle X\rangle_k\right)}{(Var(t_j))^{\frac{1}{2}}(Var(\langle X\rangle_k))^{\frac{1}{2}}}\\ &= \frac{U_{jk}\left(\Lambda_q-\sigma^2 I_q\right)^{\frac{1}{2}}}{(S_{jj})^{\frac{1}{2}}\left(\frac{\lambda_q-\sigma^2}{\lambda_q}\right)^{\frac{1}{2}}}\\ &= \frac{U_{jk}\sqrt{\lambda_k}}{\sqrt{S_{jj}}}.\end{aligned} \tag{6}$$

As we can see, it is the same as correlation between jth variable and kth component in common principal component analysis and is proportional to U_{jk}.

Remark 3.1. Covariance matrix between probabilistic principal components and components of common PCA can be obtained as follows;

$$\begin{aligned}Cov(\langle X\rangle, Z) &= Cov(M^{-1}W't, U_q't)\\ &= M^{-1}W'SU_q\\ &= (\Lambda_q)^{-1}\left(\Lambda_q-\sigma^2 I_q\right)^{\frac{1}{2}}U_q'U\Lambda U'U_q\\ &= (\Lambda_q)^{-1}\left(\Lambda_q-\sigma^2 I_q\right)^{\frac{1}{2}}U_q'U_q\Lambda_q\\ &= \left(\Lambda_q-\sigma^2 I_q\right)^{\frac{1}{2}}.\end{aligned} \tag{7}$$

Therefore probabilistic principal components and components of common PCA are uncorrelated.

4 PPCA with anisotropic Gaussian distribution of latent variables

By considering an anisotropic Gaussian distribution for latent variables $X\sim N(0,V)$, where V is diagonal matrix with elements $v_1,v_2,...,v_q$, in the

latent variable model presented in (1), the probability distribution over t-space for given X is in the form;

$$p\left(t|x\right)=\left(2\pi\sigma^{2}\right)^{-\frac{d}{2}}exp\left\{-\frac{1}{2}\left\|t-WX-\mu\right\|^{2}\right\}\Rightarrow t|X\sim N(WX+\mu,\sigma^{2}I_{d}).$$

The Gaussian prior over the latent variables is defined as follows;

$$p\left(X\right)=\left(2\pi\right)^{-\frac{q}{2}}\left|V\right|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}X'V^{-1}X\right\}\Rightarrow X\sim N(0,V).$$

The marginal distribution of t can then be obtained as the follows;

$$\begin{aligned}p\left(t\right) &= \int p\left(t|X\right)p\left(X\right)dX,\\ &= \left(2\pi\right)^{-\frac{d}{2}}\left|G\right|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}\left(t-\mu\right)'G^{-1}\left(t-\mu\right)\right\},\end{aligned}$$

where

$$G=\gamma\gamma'+\sigma^{2}I_{d},\qquad\gamma=WV^{\frac{1}{2}}.$$

The posterior distribution of the latent variables given the observed vector t can be obtained by using Baye's rule;

$$\begin{aligned}p\left(X|t\right)=\left(2\pi\right)^{-\frac{q}{2}}\left|\sigma^{-2}H\right|^{\frac{1}{2}}exp\{-\frac{1}{2}\left(X-H^{-1}W'\left(t-\mu\right)\right)\left(\sigma^{-2}H\right)\left(X-H^{-1}W'\left(t-\mu\right)\right)\},\\ \Rightarrow X|t\sim N(H^{-1}W'\left(t-\mu\right),\sigma^{2}H^{-1}).\end{aligned}$$

where $H=W'W+\sigma^{2}V^{-1}$.

Note that H is $q\times q$ while G is $d\times d$ matrix. The log-likelihood of the observed data is given by:

$$\begin{aligned}\ell &= \sum_{n=1}^{N}lnp\left(t_{n}\right),\\ &= -\frac{N}{2}\left\{dln\left(2\pi\right)+ln\left|G\right|+tr\left(G^{-1}S\right)\right\},\end{aligned}\tag{8}$$

where

$$S=\frac{1}{N}\sum_{n=1}^{N}\left(t_{n}-\mu\right)\left(t_{n}-\mu\right)'.$$

4.1 Properties of maximum likelihood estimation

In this section we estimate the parameters of model (1) by maximizing the log-likelihood ℓ. First we consider the derivation of ℓ with respect to γ [9]:

$$\frac{\partial \ell}{\partial \gamma} = N\left(G^{-1}SG^{-1}\gamma - G^{-1}\gamma\right). \tag{9}$$

At the stationary point:

$$SG^{-1}\gamma = \gamma,$$

assume that $\text{rank}(S) > q$ and thus G^{-1} to be existed. This is a necessary and sufficient condition for the density model to be nonsingular.

There are three possible classes of solution:

[i] $\gamma = 0 \rightarrow W = 0$, this will yeild minimum of the likelihood function.

[ii] $G = S$, in this case the covariance model is exact and factor loadings are identical from the eigen-decomposition of S [1]:

$$\gamma = U_q\left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} \rightarrow W_{ML} = U_q\left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} V^{-\frac{1}{2}}. \tag{10}$$

[iii] In this case we assume that

$$SG^{-1}\gamma = \gamma, \quad \gamma \neq 0, \quad S \neq G$$

where

$$\gamma = ULZ'$$

and $U = (u_1, u_2, ..., u_q)$ is a $d \times q$ matrix whose columns are orthonormal and eigenvector of $\gamma\gamma'$, $L = diag\,(l_1, l_2, ..., l_q)$ is diagonal matrix of singular values, Z is $q \times q$ orthogonal matrix whose columns are q eigenvectors of $\gamma\gamma'$, and

$$\begin{aligned} G^{-1}\gamma &= \left(\sigma^2 I_q + \gamma\gamma'\right), \\ &= \gamma\left(\sigma^2 I_q + \gamma'\gamma\right)^{-1}, \\ &= ULZ'\left(\sigma^2 I_q + RL^2R'\right)^{-1}, \\ &= ULZ'Z\left(L^2 + \sigma^2 I_q\right)^{-1}Z', \\ &= UL\left(L^2 + \sigma^2 I_q\right)^{-1}Z'. \end{aligned}$$

Then, substituting the latter in stationary point;

$$SG^{-1}\gamma = \gamma \quad \Rightarrow \quad SUL\left(L^2 + \sigma^2 I_q\right)^{-1} Z' = ULZ',$$
$$SUL = U\left(\sigma^2 I_q + L^2\right) L.$$

Thus $Su_j = (\sigma^2 + {l_j}^2)\, u_j$ for $l_j \neq 0$ and each columns of U must be an eigenvector of S with corresponding eigenvalue $\lambda_j = \sigma^2 + l_j^2$, therefore

$$l_j = \left(\lambda_j - \sigma^2\right)^{\frac{1}{2}}.$$

Then all potential solutions for γ may be obtained as:

$$\gamma = U_q \left(K_q - \sigma^2 I_q\right)^{\frac{1}{2}} R$$

where U_q is $d \times q$ matrix whose columns are eigenvector of S, R is an arbitrary orthogonal matrix, K_q is a diagonal matrix as follows:

$$k_j = \begin{cases} 1 & l_j \neq 0, \\ 0 & l_j = 0. \end{cases}$$

So for $l_j \neq 0 \quad (j = 1, 2, ..., q)\,, \;\; \gamma = U_q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} R,$

$$\Rightarrow W_{ML} = U_q \left(\Lambda_q - \sigma^2 I_q\right)^{\frac{1}{2}} R V^{-\frac{1}{2}} \tag{11}$$

where the columns in matrix U_q are the principal eigenvectors of S. Elements of diagonal matrixΛ_q are eigenvalues for S, and R is an arbitrary $q \times q$ orthogonal rotation matrix where for simplicity we would effectively ignore R (*i.e* choose $R = I$).

By substituting W_{ML} into log-likelihood (8), we obtain,

$$\ell = -\frac{N}{2}\left\{ dln2\pi + \sum_{j=1}^{q} \lambda_j + \frac{1}{\sigma^2} \sum_{j=q+1}^{d} \lambda_j + (d-q)\, ln\sigma^2 + q \right\}. \tag{12}$$

The maximum likelihood estimator of σ^2 is given by:

$$\sigma^2_{ML} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j. \tag{13}$$

The latter expresses the variance lost in the projection that averaged over lost dimension.

By substituting σ^2_{ML} and W_{ML} in ℓ, it can be seen easily that the matrix U which maximizes the likelihood function must be corresponds to q largest eigenvalue of sample covariance matrix S. Thus with using the anisotropic Gaussian distribution of latent variable ($X \sim N(0, V)$) the columns of W are correspond to principal axes.

Remark 4.1. The columns of matrix W_{ML} are not orthogonal, because

$$W'_{ML} W_{ML} = V^{-\frac{1}{2}} R' \left(\Lambda_q - \sigma^2 I_q\right) R V^{-\frac{1}{2}}.$$

4.2 Dimensionality reduction

From a probabilistic perspective, process of dimensionality reduction consider in term of the posterior distribution of latent variable can be summarized by posterior mean of latent variable as;

$$\langle X_n \rangle = H^{-1} W' \left(t_n - \mu \right). \tag{14}$$

If $\sigma^2 \to 0$, then

$$H^{-1} = \left(W'W \right)^{-1} \quad \text{and} \quad W H^{-1} W'$$

represent an orthogonal projection in to data and PCA is recovered:

$$W \langle X_n \rangle = W H^{-1} W' \left(t_n - \mu \right)$$

However because of $\sigma^2 \to 0$ the density model is singular and undefinable. Also with $\sigma^2 > 0$, $W \langle X_n \rangle$ is not an orthogonal projection of t_n, but with

$$W = W_{ML},$$

we can obtain optimal reconstruction of the observed data by using posterior mean of latent variable as follows:

$$\begin{aligned} \hat{t_n} &= W_{ML} \left(W'_{ML} W_{ML} \right)^{-1} H \langle X_n \rangle + \mu, \\ &= W_{ML} \left(W'_{ML} W_{ML} \right) W'_{ML} \left(t_n - \mu \right) + \mu. \end{aligned} \tag{15}$$

Remark 4.2. In the case that latent variables are anisotropic normal distributed the covariance matrix of PPCs can be expressed, as:

$$Cov_V(\langle X \rangle) = Cov_V(H^{-1} W' t)$$

With respect to W_{ML} in (11) and substuting $R = I$;

$$\begin{aligned} H &= W'_{ML} W_{ML} + \sigma^2 I_q \\ &= V^{\frac{-1}{2}} \left(\Lambda_q - \sigma^2 I_q \right)^{\frac{1}{2}} U'_q U_q \left(\Lambda_q - \sigma^2 I_q \right)^{\frac{1}{2}} V^{\frac{-1}{2}} + \sigma^2 I_q \\ &= V^{-1} \left(\Lambda_q - \sigma^2 I_q \right) + \sigma^2 I_q \end{aligned}$$

Therefore

$$\begin{aligned} \langle X \rangle &= H^{-1} W' t \\ &= \left(V^{-1} (\Lambda_q - \sigma^2 I_q) + \sigma^2 I_q \right)^{-1} V^{-\frac{1}{2}} (\Lambda_q - \sigma^2 I_q)^{\frac{1}{2}} U'_q t \end{aligned}$$

$$= \begin{pmatrix} \frac{V_1}{\lambda_1+\sigma^2(V_1-1)} & 0 & \cdots & 0 \\ 0 & \frac{V_2}{\lambda_2+\sigma^2(V_2-1)} & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \frac{V_{q-1}}{\lambda_{q-}+\sigma^2(V_{q-1}-1)} & \vdots \\ 0 & \cdots & 0 & \frac{V_q}{\lambda_q+\sigma^2(V_q-1)} \end{pmatrix} W't$$

$$= \begin{pmatrix} (\frac{V_1}{\lambda_1+\sigma^2(V_1-1)}, 0, \cdots, 0)W't \\ (0, \frac{V_2}{\lambda_2+\sigma^2(V_2-1)}, \cdots, 0)W't \\ \vdots \\ (0, 0, \cdots, \frac{V_q}{\lambda_q+\sigma^2(V_q-1)})W't \end{pmatrix}$$

Therefore kth component can be obtained as follows:

$$\langle X \rangle_k = (0, 0, \cdots, \frac{V_k}{\lambda_k + \sigma^2(V_k - 1)})W't \tag{16}$$

Covariance matrix is given by:

$$\begin{aligned} Cov_V(H^{-1}W't) &= Cov\left(\left(V^{-1}(\Lambda_q - \sigma^2 I_q) + \sigma^2 I_q\right)^{-1} V^{-\frac{1}{2}}(\Lambda_q - \sigma^2 I_q)^{\frac{1}{2}} U_q' t\right) \\ &= V^{-1}(\Lambda_q - \sigma^2 I_q)\left(V^{-1}(\Lambda_q - \sigma^2 I_q) + \sigma^2 I_q\right)^{-2} \end{aligned} \tag{17}$$

So components are independent, as case that latent variables were isotropic normal distributed. Variance of each component is then;

$$Var_V(\langle X_k \rangle) = \frac{\lambda_k V_k(\lambda_k - \sigma^2)}{(\lambda_k + \sigma^2(V_k - 1))^2}$$

Remark 4.3. The correlation between kth probabilistic principal component and jth variable, in this case is given by:

$$\rho_V(t_j, \langle X \rangle_k) = \frac{Cov_V(t_j, \langle X_k \rangle_V)}{(Var(t_j))^{\frac{1}{2}} (Var(\langle X_k \rangle_V))^{\frac{1}{2}}}.$$

where $Cov_V(t_j, \langle X_k \rangle)$ is calculated by:

$$\begin{aligned} Cov_V(t_j, \langle X \rangle_k) &= Cov\left(I_j' t, \left(0, 0, \cdots, \frac{V_k}{\lambda_k + \sigma^2(V_k - 1)}, \cdots, 0\right) W't\right) \\ &= I_j' SW \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{V_k}{\lambda_k+\sigma^2(V_k-1)} \\ \vdots \\ 0 \end{pmatrix} \end{aligned}$$

$$SW = U\Lambda U' U_q(\lambda_q - \sigma^2 I_q)^{\frac{1}{2}} V^{-\frac{1}{2}}.$$

Then the sample covariance vector is given by:

$$\begin{aligned} Cov_V(t_j, \langle X_k \rangle) &= I_j' \begin{pmatrix} U_{11} & U_{12} & U_{13} & \cdots & U_{1q} \\ U_{21} & U_{22} & U_{13} & \cdots & U_{2q} \\ U_{31} & U_{32} & U_{13} & \cdots & U_{3q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ U_{d1} & U_{d2} & U_{32} & \cdots & U_{dq} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)} \\ \vdots \\ 0 \end{pmatrix} \\ &= (0, 0, \cdots, 1, \cdots, 0) \begin{pmatrix} U_{1k} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)} \\ U_{2k} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)} \\ \vdots \\ U_{jk} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)} \\ \vdots \\ U_{dk} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)} \end{pmatrix} \\ &= U_{jk} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \rho_V(t_j, \langle X_k \rangle) &= \frac{U_{jk} \frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)}}{(S_{jj})^{\frac{1}{2}} (\frac{V_k^{\frac{1}{2}} \lambda_k (\lambda_k - \sigma^2)^{\frac{1}{2}}}{\lambda_k + \sigma^2 (V_k - 1)})^{\frac{1}{2}}} \\ &= \frac{U_{jk} \sqrt{\lambda_k}}{\sqrt{S_{jj}}}. \end{aligned} \tag{18}$$

As it is yielded, the correlation between jth variable and kth component in the case that latent variables are anisotropic normal distributed is same as the correlation of component with jth variable, where latent variables are isotropic normal ditributed and both of them are same as corresponding correlation in common PCA.

Remark 4.4. Probabilistic principal components and principal components of common PCA are uncorrelated, because :

$$\begin{aligned} Cov_V(\langle X\rangle, Z) &= Cov(H^{-1}W't, U_q't) \\ &= H^{-1}W'WSU_q \\ &= H^{-1}W'U_q\Lambda_q \\ &= \left(V^{-1}(\Lambda_q-\sigma^2I_q)+\sigma^2I_q\right)^{-1}V^{\frac{-1}{2}}(\Lambda_q-\sigma^2I_q)^{\frac{1}{2}}U_q'U_q\Lambda_q \\ &= V^{-\frac{1}{2}}\Lambda_q(\Lambda_q-\sigma^2I_q)^{\frac{1}{2}}\left(V^{-1}(\Lambda_q-\sigma^2I_q)+\sigma^2I_q\right)^{-1} \qquad (19) \end{aligned}$$

So with respect to the diagonal matrix that resulted above, in this case, PPCs and PCs are also uncorrelated.

5 Simulation Study

We consider now a data set of 20 points in 10-dimensional space that generated from a Gaussian distribution that have standard deviation in first 5 dimension as $(1.0, 0.8, 0.6, 0.4, 0.2)$ and standard deviation 0.04 in the remaining 5 directions, and it is assumed, $\mu = 0$ and latent variables are from the Gaussian distribution whit mean $= 0$ and covariance matrix,

$$V = diag(0.1, 0.2, 0.3, 0.4, 0.5) \quad (X \sim N(0, V)).$$

Applying PPCA with anisotropic Gaussian distribution for latent variable, the effective dimensionality for principal component that is correspond to $q = 5$ largest eigenvalues of ariance matrix S. We obtained 5-dimensional principal component by (14) in section 4, also the reconstruction of the data from these principal component can be obtained by using (15) in section 4.

Figure1 shows the image plots for original data, the compressed data (probabilistic principal components) and the reconstructed data.
It can be seen that the compressed data represents the original data appropriately and the reconstruction from the compressed data also recovered the original data well with exact recovery up to the first 5 component of the data.

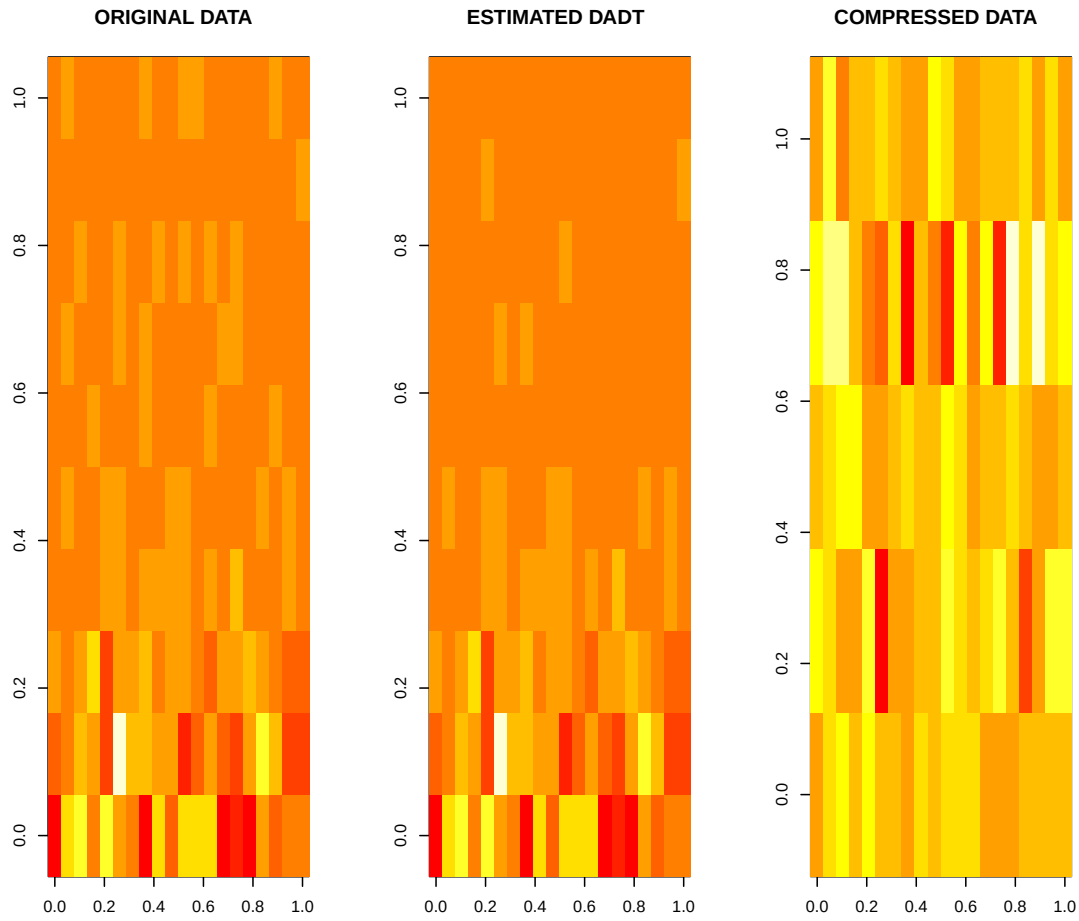

Figure 1: Image plots for the original data (left), reconstructed data(middle), and compressed data (right).

6 Conclusion

There have been various works for PCA based on the PPCA model since its introduction by Tipping and Bishop (1999). In all of these works, an isotropic Gaussian distribution for latent variables has been used.
In this paper, we provide some interpretation for PPCs and extended Bishop and Tipping's approach by using anisotropic Gaussian distribution for latent variables. Furthermore it is resulted that, common PCs and PPCs are uncorrelated. The latent variable model with anisotropic Gaussian distribution of latent variable may be used in Bayesian PCA [4]. This will be considered in detail in our further work.

References

[1] A. Basilevsky, *Statistical Factor Analysis and Related Methods*, Wiley, New York, 1994.

[2] Alvin C. Rencher, *Multivariate Statistical Inference and applications*, Wiley, New York, 1998.

[3] C.M. Bishop and M.E. Tipping , A hierarchical latent variable model for data visualization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(3), (1998), 281-293.

[4] H.S. Oh and D.G. Kim, Bayesian principal component analysis whit mixture priors, *Jurnul of the korean statistical society*, **39**, (2010), 387-396.

[5] I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 2002.

[6] M.E. Tipping and C.M. Bishop, Probabilistic principal component analysis, *Journal of the Royal Statistical Society*, **21**(3), (1999), 611-622.

[7] Seghouane, Abd-krim and Cichocki, Andrzej, Bayesian estimation of the number of principal components, *Signal Processing*, **87**, (2007), 562-568.

[8] T.W. Anderson, Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, **34**, (1963), 122-148.

[9] W.J. Krzanowski and F.H.C. Marriott, *Multivariate Analysis Part I: Distributions, Ordination and Inference*, Edward Arnold, London, 1994.