Journal of Applied Mathematics & Bioinformatics, vol.2, no.2, 2012, 35-49
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2012

Bayesian Investigation of Principal Components Using Mixture Priors

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

Abstract

Bayesian approach for principal component analysis (PCA) is a novel method to determine the number of dimensionality through using different prior probabilities. In common strategy one often equally selects variances for the columns of mapping matrix by using the mixture priors. In this article, we generalize this approach by using the mixture priors with different variances in multivariate normal distribution. Further, we employ an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. Estimations of principal component and number of effective dimensionality are performed via Markov Chain Monte Carlo (MCMC) algorithm.

Mathematics Subject Classification: 62G99

Keywords: Principal component analysis, Latent variable model, Dimensionality reduction, Mixture prior, Hierarchical 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. *Corresponding author, e-mail: m.babanezhad@gu.ac.ir

Article Info: *Received* : July 21, 2012. *Revised* : August 27, 2012
Published online : September 15, 2012

1 Introduction

Principal Component Analysis (PCA) is a dimensionality reduction modeling technique that transforms a set of process variables by rotating their axes of representation [1, 2, 3, 4]. It has been successfully applied in many fields such as, data compression, image processing, pattern recognition, data visualization and so on [1, 2, 5, 6]. Probabilistic interpretation of PCA is proposed in [1, 3, 4]. In probabilistic PCA the observed data is assumed a linear mapping of the latent variable plus Gaussian error [2, 7, 8].

In PCA no external information about the data is utilized but with considering probabilistic PCA we can define a Bayesian model and put prior information to it [7, 8, 9, 10]. So the problems in common PCA can be solved with statistical inferential methods. The maximum likelihood method was proposed in [1, 2, 3], which is not applicable for determining number of principal components. Many investigations in determination of the number of appropriate principal component have been done. [1] proposed a Bayesian approach with using a hierarchical prior distribution over the mapping matrix for which each column is assumed with zero mean normal distribution (for more details see, [3, 4]). In this method after estimating the parameters, columns with small variances are ignored, and the number of remaining columns is chosen as the dimension of components. However it is unclear how small the variance can be to ignore the corresponding columns of mapping matrix.

To get around this issue [2] has proposed the mixture priors for mapping matrix. In their method the prior distribution of each column is mixture of zero mean normal distribution and a discrete distribution that assign probability one to point zero. In this paper, we define more general case of this mixture prior in which the continuous part is a normal distribution that governed by a q-dimensional vector of hyper parameters $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_q\}$. Indeed we consider normal distributions with different variances for significant columns of mapping matrix. Also posterior inferences of parameters will be obtain via MCMC algorithm.

In the next Section, we review PCA and probabilistic PCA. In Section 3, the Bayesian model of PCA is introduced. A simulation example is presented in Section 4 and conclusion is given in Section 5.

2 Review of PCA and Probabilistic PCA

2.1 Interpretation of PCA

Consider a data set T of observed d-dimensional data vectors $T = \{t_n\}$ where $n = \{1, 2, 3, ..., N\}$. Common PCA is determined by first computing the sample covariance matrix given by

$$S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \bar{t}) ({}_n - \bar{t})',$$

where $\bar{t} = N^{-1} \sum_{n=1}^{N} t_n$ is the sample mean. Next the eigenvectors u_i and eigenvalues λ_i of S are found such that $Su_i = \lambda_i u_i \quad i = 1, 2, 3, ..., d$. The eigenvectors corresponding to the q largest eigenvalues (where $q < d$, for parsimonious representation) are retained and a reduced dimensionality representation of the data set is defined by $Z_n = U_q' (t_n - \bar{t})$, where $U_q = \{u_1, u_2, ..., u_q\}$. It can be easily shown that PCA corresponds to the linear projection of a data set for which the retained variance is maximum and the sum of squares reconstruction cost is minimized [5, 8].

A significant limitation of common PCA is that, it does not define any probability model for the observed data.

2.2 Interpretation of Probabilistic PCA

Following [6], PCA can be formulated as the maximum likelihood solution of a specific latent variable model. This model relates a d-dimensional vector t_n to a corresponding q-dimensional vector of latent variable

$$t_n = W x_n + \mu + \epsilon, \tag{1}$$

where W is a $d \times q$ matrix that relates the two sets of variables, μ is a d-dimensional mean vector, the latent variables $\{x_n\}$ are defined independent and Gaussian with identity covariance matrix. The noise ϵ is zero mean Gaussian with covariance matrix $\sigma^2 I_d$.

Under model (1) the probability distribution of the observed variable t_n given x_n is $N(Wx_n + \mu, \sigma^2 I_d)$. So, the elements of t_n given the latent variable

x_n are independent. The marginal distribution of the observed variable is given by

$$p\left(t_n\right)=\int p\left(t_n|x_n\right)p\left(x_n\right)dx_n=N\left(\mu,C\right),$$

where the covariance matrix $C=WW'+\sigma^2I_d$. The log probability of the parameters under the observed data set T is

$$\ell\left(\mu,W,\sigma^2\right)=-\frac{N}{2}\left\{dln\left(2\pi\right)+ln|C|+trace\left(C^{-1}S\right)\right\},$$

where S is the sample covariance matrix. The maximum likelihood solution for μ, W and σ^2 is seen to be as follow;

$$\begin{aligned}\mu_{ML}&=\bar{t_n}\\ W_{ML}&=U_q\left(\Lambda_q-\sigma^2I_q\right)^{\frac{1}{2}}\\ \sigma^2_{ML}&=\frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i.\end{aligned}$$

The posterior distribution of x_n is given by using Bayes'rule

$$x_n|t_n\sim N\left(M^{-1}W'\left(t_n-\mu\right),\sigma^2M^{-1}\right),$$

where $M=W'W+\sigma^2I_q$.
Also the dimensionality reduction representation for observed data is considered as

$$\langle x_n\rangle=M^{-1}W'\left(t_n-\mu\right)\qquad n=1,2,...,N.$$

The optimal reconstruction of the observed data is obtained by using posterior mean of latent variable as follows

$$\hat{t_n}=W\left(W'W\right)^{-1}W't_n.$$

[1, 3] introduced a hierarchical prior $p\left(W|\alpha\right)$ over the matrix W that governed by a vector of hyper parameters $\alpha=\{\alpha_1,\alpha_2,\alpha_3,...,\alpha_q\}$. Also the dimensionality of the latent space was assumed to its maximum possible value, $q=d-1$. In $p\left(W|\alpha\right)$, each α_i controls the corresponding column in matrix W trough a conditional Gaussian distribution of the form

$$p\left(W|\alpha\right)=\prod_{i=1}^{q}\left\{\left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}}\exp(-\frac{1}{2}\alpha_i||w_i||^2)\right\},\tag{2}$$

where $\{W_i\}$ are the columns of W. The latent space dimensionality also is obtained as the number of large values in estimated elements of α , but it is not clear that how large α_i is.

In [2] a Bayesian model with mixture priors was applied for which solve the problem [4]. Their proposed mixture prior was as

$$W_i|\alpha, p \sim (1-p)\,\delta_0\left(w_j\right) + P\left(1-\delta_0\left(w_i\right)\right) N\left(0, \frac{1}{\alpha} I_d\right), \qquad i = 1, 2, ..., q,$$

where

$$\delta_0\left(w_i\right) = \begin{cases} 1 & w_i = 0, \\ 0 & w_i \neq 0. \end{cases}$$

In this article we consider a fully Bayesian model and generalize [2] approach by using a mixture prior with specific variance in each column of mapping matrix W. Further a MCMC algorithm is applied for posterior inference of parameters.

3 Bayesian Model of PCA

The main goal in probabilistic PCA is to determine the insignificant columns of W. In the other hand, whether or not w_i is zero, by considering a continuous prior for w_i, the probability of $w_i = 0$ become zero. Because of this problem we use the below prior distribution over the columns of matrix W

$$w_i|\alpha_i, p \sim (1-p)\,\delta_0\left(w_i\right) + p\left(1-\delta_0\left(w_i\right)\right) N\left(0, \frac{1}{\alpha_i} I_d\right), \qquad i = 1, 2, 3, ..., q.$$

In this distribution $\alpha = (\alpha_1, \alpha_2, \alpha_3, ..., \alpha_q)$ and p are hyper parameters, where $\frac{1}{\alpha_i}$ $(\alpha_i > 0)$ is the specific variance of column w_i and p $(0 < p < 1)$ is the proportion of insignificant columns of matrix W. In what follows we adopt a fully Bayesian model with assuming prior distribution for μ, σ^2, p, α :

$$\begin{aligned} \alpha_i &\sim Gamma\left(a_{\alpha_i}, b_\alpha\right), \qquad i = 1, 2, ..., q \\ p &\sim Beta\left(c_p, d_p\right) \\ \mu|\beta &\sim N\left(0, \frac{1}{\beta} I_d\right) \\ \beta &\sim Gamma\left(a_\beta, b_\beta\right) \\ \tau &\sim Gamma\left(a_\tau, b_\tau\right), \qquad where \quad \tau = \frac{1}{\sigma^2}. \end{aligned} \tag{3}$$

Note that we assume independence among the priors. If we assume all unknown parameters of model as

$$\Theta = \{\mu, \beta, \tau, p, \{w_i; \quad i = 1, 2, ..., q\}, \{\alpha_i; \quad i = 1, 2, ..., q\}, \{x_n; \; n = 1, 2, ..., N\}\},$$

then the joint posterior distribution over the parameters is given by

$$\begin{aligned} f(\Theta|\{t_n\}) &= \frac{f(\{t_n\}|\Theta) f(\Theta)}{f(\{t_n\})} \\ &\propto f(\{t_n\}|\Theta) f(\Theta) \\ &\propto \prod_{n=1}^{N} \left\{ \left(\frac{\tau}{2\pi}\right)^{\frac{d}{2}} \exp\left[-\frac{\tau}{2}(t_n - \mu - Wx_n)'(t_n - \mu - Wx_n)\right]\right\} \\ &\quad \times \prod_{i=1}^{q} \left\{(1-p)\delta_0(w_i) + p(1-\delta_0(w_i))\left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}} \exp\left[-\frac{\alpha_i}{2} w_i' w_i\right]\right\} \\ &\quad \times \prod_{n=1}^{N} \left\{ \left(\frac{1}{2\pi}\right)^{\frac{q}{2}} \exp -\frac{x_n' x_n}{2}\right\} \times \left(\frac{\beta}{2\pi}\right)^{\frac{d}{2}} \exp\left[-\frac{\beta}{2}\mu'\mu\right] \\ &\quad \times p^{c_p - 1}(1-p)^{d_p - 1} \times \prod_{i=1}^{q} \left\{\alpha_i^{a_{\alpha_i} - 1} \exp(-b_\alpha \alpha_i)\right\} \\ &\quad \times \tau^{a_\tau - 1} \exp(-b_\tau) \times \beta^{a_\beta - 1} \exp(-b_\beta \beta). \end{aligned} \quad (4)$$

In order to determine insignificancy of the columns of matrix W, we must marginalized this model over w_i which is analytically intractable a numeric approximation algorithm MCMC. One of the attractive methods for setting up an MCMC algorithm is Gibbs sampling. The Gibbs sampler does this by successively and repeatedly simulating from the conditional distributions of each component given the other components. Also this procedure is particularly useful where we have conditional conjugacy, so that the resulting conditional distributions are from standard distributions.

3.1 Conditional Posterior Distributions of Gibbs Sampling Procedure

The conditional posterior distribution of each parameter given all the other parameters given as follow

$$\begin{aligned}
\mu|others &\sim N\left(\frac{\tau}{n\tau+\beta}\sum_{n=1}^{N}\left(t_n - Wx_n\right), \frac{1}{n\tau+\beta}I_d\right)\\
\tau|others &\sim Gamma\left(\frac{nd}{2}+a_\tau, b_\tau+\frac{1}{2}\sum_{n=1}^{N}\left(t_n-\mu-Wx_n\right)'\left(t_n-\mu-Wx_n\right)\right)\\
x_n|others &\sim N\left(M^{-1}W'\left(t_n-\mu\right), \frac{1}{\tau}M^{-1}\right)\\
\beta &\sim Gamma\left(\frac{d}{2}+a_\beta, b_\beta+\frac{1}{2}\mu'\mu\right).
\end{aligned}$$

For the other parameters, corresponding conditional posterior distributions are computed by considering two case $w_i = 0$, $w_i \neq 0$. So the conditional distribution of p can be derived as follow

- $w_i = 0 \;\Rightarrow\; p|others \sim Beta\left(c_p, d_p+q\right).$

- $w_i \neq 0 \;\Rightarrow\; p|others \sim Beta\left(c_p+q, d_p\right).$

By mixing these two cases the conditional posterior distribution is given as

$$p|others \sim Beta\left(c_p+\sum_{i=1}^{q}\gamma_i, d_p+q-\sum_{i=1}^{q}\gamma_i\right),$$

where $\gamma_i = \begin{cases} 1 & w_i \neq 0, \\ 0 & w_i = 0. \end{cases}$

Similarly the corresponding distribution for elements of hyper parameters vector α is achieved as follow

- $w_i = 0 \;\Rightarrow\; \alpha_i \sim Gamma\left(a_{\alpha_i}, b_\alpha\right).$

- $w_i \neq 0 \;\Rightarrow\; \alpha_i \sim Gamma\left(a_{\alpha_i}+\frac{||w_i||^2}{2}, b_\alpha+\frac{d}{2}\right).$

Therefore

$$\alpha_i|others \sim Gamma\left(a_{\alpha_i}+\frac{\gamma_i||w_i||^2}{2}, b_\alpha+\frac{\gamma_i d}{2}\right), \quad i=1,2,...,q.$$

The conditional posterior distribution of w_j also can be derived in two following steps;

- In case where $w_i = 0$

$$p\left(w_i|others\right) = \frac{p\left(t|\alpha_i,\beta,\tau,\mu,X,W\left(-i\right),w_i=0\right)\times p\left(w_i=0\right)}{\int p\left(t|\alpha_i,\beta,\tau,\mu,X,W\right)dw_i}. \tag{5}$$

$$p\left(t|\alpha_i,\beta,\tau,\mu,X,W\left(-i\right),w_i=0\right)\times p\left(w_i=0\right) = \left(1-p\right)C_{1i}, \tag{6}$$

where C_{1i} is :

$$\left(\frac{\tau}{2\pi}\right)^{\frac{nd}{2}}\exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}\left(t_n - W\left(-i\right)x_{n(-i)}-\mu\right)'\left(t_n - W\left(-i\right)x_{n(-i)}-\mu\right)\right\}$$

also

$$W\left(-i\right) = \left(W\left(1\right),W\left(2\right),...,W\left(i+1\right),...,W\left(q\right)\right)$$

and

$$x_{n(-i)} = \left(x_{n,1},x_{n,2},...,x_{n,i+1},...,x_{n,q}\right)'.$$

The marginal distribution of w_i from the joint distribution of observations $f\left(\{t_n\}|\Theta\right)$ is given by

$$\int p\left(t|\alpha_i,\beta,\tau,\mu,X,W\right)dw_i = \left(1-p\right)C_{1i}+pC_{2i}, \tag{7}$$

where C_{2i} is

$$\int_{w_i\neq 0}\left(\frac{\tau}{2\pi}\right)^{\frac{nd}{2}}\exp\left\{-\frac{\tau}{2}\left(t_n - Wx_n-\mu\right)'\left(t_n - Wx_n-\mu\right)\left(\frac{\alpha_i}{2\pi}\right)^{\frac{d}{2}}\exp(-\frac{w_i'w_i}{2})\right\}.$$

Substituting (6) and (7) in (5)

$$p\left(w_i|others\right) = \frac{C_{1i}\left(1-p\right)}{C_{1i}\left(1-p\right)+C_{2i}p}. \tag{8}$$

- In case where $w_i \neq 0$

$$\begin{aligned} p\left(w_i|others\right) &= \frac{p\left(t|\alpha_i,\beta,\tau,\mu,X,W\right)\times p\left(w_i\right)}{\int p\left(t|\alpha_i,\beta,\tau,\mu,X,W\right)dw_i} \\ &\propto \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}\left(t_n - Wx_n-\mu\right)'\left(t_n - Wx_n-\mu\right)\right\} \\ &\quad\times\exp\left\{-\frac{\alpha_i}{2}w_i'w_i\right\}. \end{aligned} \tag{9}$$

By using equation $(t_n - \mu - Wx_n) = \left(t_n - \mu - W(-i)\, x_{n(-i)}\right) - w_i x_{ni}$,

$$C_{2i} = C_{1i} \left(\frac{\alpha_i}{\eta_i}\right)^{\frac{d}{2}} \exp\{\frac{\eta_i}{2}\xi_i'\xi_i\},$$

and (9) derives as

$$\begin{aligned} p\,(w_i|others) &\propto C_{1i} \times \left(\frac{\alpha_i}{\eta_i}\right)^{\frac{d}{2}} \exp\left\{\frac{\eta_i}{2}\xi_i'\xi_i\right\} \times N\left(\xi_i, \frac{1}{\eta_i}I_d\right) \\ &\propto N\left(\xi_i, \frac{1}{\eta_i}I_d\right), \end{aligned}$$

where η_i and ξ_i, are the forms

$$\begin{aligned} \eta_i &= \tau \sum_{n=1}^{N} {x_{ni}}^2 + \alpha_i, \\ \xi_i &= \frac{\tau}{\eta_i} \sum_{n=1}^{N} x_{ni} \left(t_n - \mu - W(-i)\, x_{n(-i)}\right). \end{aligned}$$

So the conditional posterior distribution of w_i can be derived as

$$w_i|others \sim \delta_0(w_i)\, p_{i.}^* + (1 - \delta_0(w_i))\, N\left(\xi_i, \frac{1}{\eta_i}I_d\right),$$

where the posterior probability of $w_i = 0$, $p_{i.}^*$ is

$$\begin{aligned} p_{i.}^* &= p\,(w_i = 0|others) \\ &= \frac{C_{1i}(1-p)}{C_{1i}(1-p) + C_{2i}p} \\ &= \left(\frac{C_{1i}(1-p) + pC_{1i}\left(\frac{\alpha_i}{\eta_i}\right)^{\frac{d}{2}} \exp(\frac{\eta_i}{2}\xi_i'\xi_i)}{C_{1i}(1-p)}\right)^{-1} \\ &= \left(1 + \frac{p}{1-p}\left(\frac{\alpha_i}{\eta_i}\right)^{\frac{d}{2}} \exp(\frac{\eta_i}{2}\xi_i'\xi_i)\right)^{-1}, \quad i = 1, 2, 3, ..., q. \end{aligned}$$

Now by using a Gibbs sampling algorithm, random samples of Θ from (4) can be driven.

4 Simulation Study

We generate now a data set of 20 points in 10-dimensional space with $\mu = 0$; where the standard deviation is taking the values $1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.01, 0.02, 0.03, 0.04$. By applying the mixture priors which are proposed in this article, we run a Gibbs sampling algorithm in 40000 iterates, and discard 35000 samples. Then the average of remaining samples for each parameter have been considered as it's estimation. Note that, the dimensionality of principal components is initially set to $d - 1$.
In this example the prior distribution of β and γ in (3) are defined as follows. Since the results are not sensitive for different choice of c_p and d_p, a non informative prior is used to p ($c_p = d_p = 1$). By considering $a_\tau = 0.5$ and $a_\alpha = (0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.2, 1.5, 2.0)$; the simulations are carried out by taking several values of $b_\alpha = b_\tau = b$. As it shown in Figure 4 and Table 1; increasing b leads to decreasing of the variances of columns of matrix W and increases the error variance.

Table 1: Estimation of posterior parameters given different values for b.

b	$\hat{p}$	$\hat{q}$	$\widehat{\sigma^2}$
0.10	0.56	5.00	1.25×10^{-4}
0.50	0.54	5.00	1.43×10^{-4}
1.00	0.51	5.00	1.75×10^{-4}
1.50	0.48	4.00	2.33×10^{-4}
2.00	0.45	4.00	3.02×10^{-4}
3.00	0.37	3.00	3.15×10^{-4}

Our data in this study has 5 components with large variance; this value of the variance is achieved when the values of b are decreased. In addition, we obtain the reconstruction of the original data and 5-dimensional principal components. The image plots of original data, reconstruction data and principal components are shown in Figure 1. The variability of error variance for several sample size is checked in Figure 2 and Figure 3.

Posterior inferences about p, σ^2 and variance of each column of matrix W are investigated for different choice of parameters c_p and d_p. As it can be seen in Figures 4 and 5; the results are not sensitive for different values of c_p and d_p.

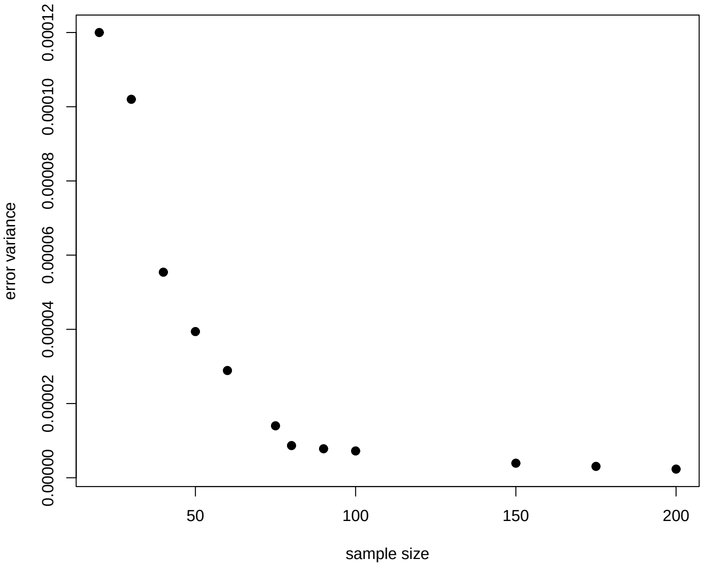

Figure 1: Variability of error variance

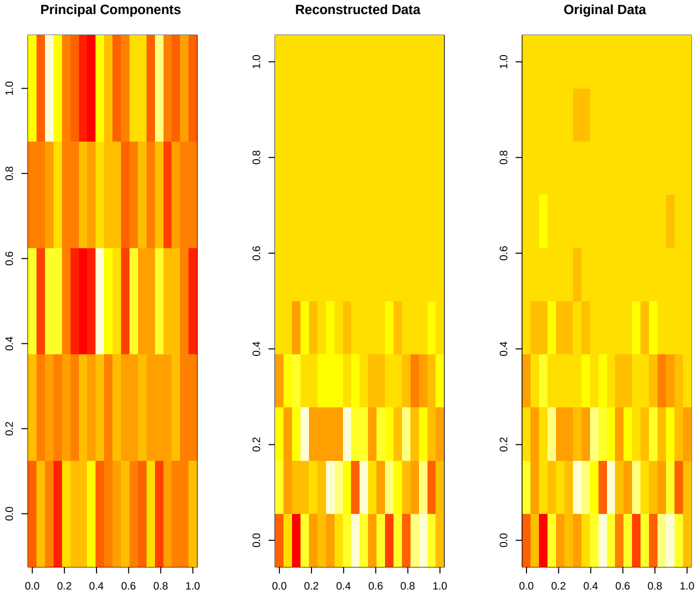

Figure 2: Image plots

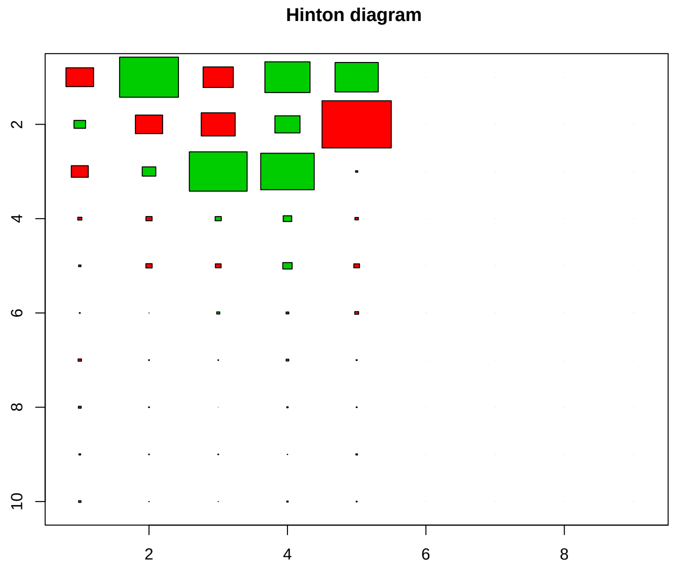

Figure 3: Hinton diagram of the mapping matrix

5 Conclusion

The main issue in PCA is to determine the number of effective dimension. In common strategy, PCA does not apply any external information about the data, since it is not based on the probability model [1, 3]. With regards to the probabilistic for PCA, we can consider a Bayesian approach and utilize prior information about the parameters of model. The Bayesian modeling framework basically proves to be very flexible, allowing simultaneous estimation of model parameters, in particular in PCA [1, 3, 5]. Noting that the aforementioned abilities of the Bayesian modeling framework in PCA, the first Bayesian method is introduced by [1, 3], with defining a continuous hierarchical prior distribution over the mapping matrix.

In this view, we use the mixture priors to investigate the Bayesian PCA in this paper. Specifically, we extend this method by introducing a mixture prior for mapping matrix which the continuous part of it, for significant columns, is Bishop's hierarchical prior. Also the posterior inferences on the parameters of model have been done via MCMC algorithm. In the procedure of the MCMC algorithm, we realized that the error variance decreases by increasing the sam-

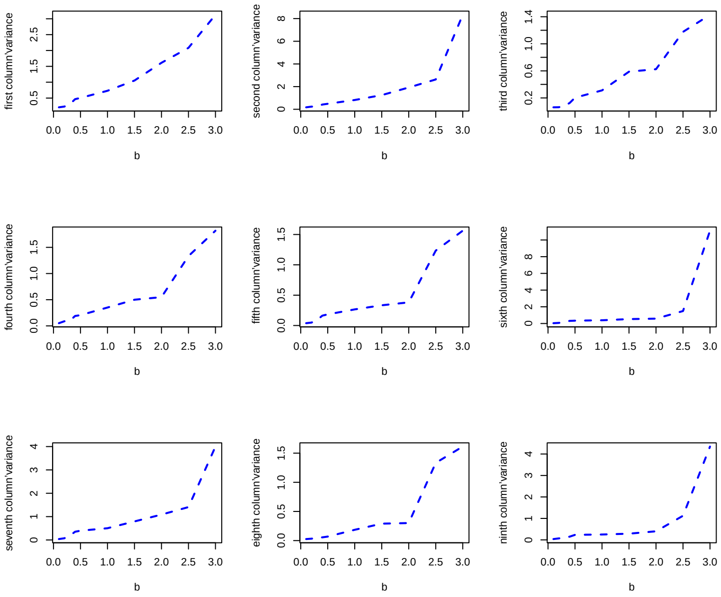

Figure 4: Variability of p by using the different choice of c_p, d_p

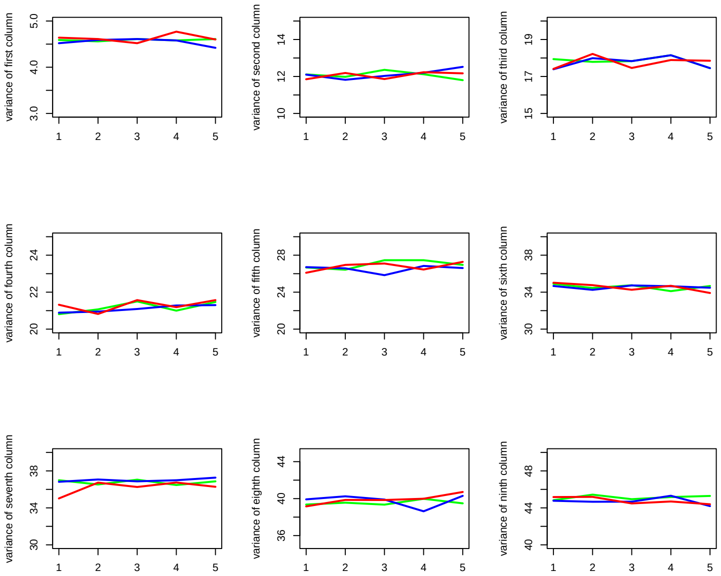

Figure 5: Variances of mapping matrix's columns for different choice of c_p, d_p

ple size N. We obtained the posterior probability for insignificant columns of mapping matrix. More important, our findings have shown that the results are not sensitive for different values of c_p and d_p. This can be specifically proven by the sensitivity analysis in further research. This might also be due to mixture priors that we employed. In addition, the variability of posterior inferences has been checked with several choices for parameters of defined priors.

Acknowledgement

This research was granted by Golestan University. We would also like to thank the referees for their very valuable comments.

References

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

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

[3] C.M. Bishop, Bayesian PCA, *Advances in neural information processing systems*, **11**, (1999) ,382-388. MIT Press.

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

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

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

[7] R. Gottardo and A.E. Raftery, Markov chain Monte Carlo with mixtures of mutually singular distributions, *Journal of Computational and Graphical Statistics*, **17**, (2008), 949-975.

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

[9] H. Hung, P.S. Wu, I.P. Tu and S.Y. Huang, On multilinear principal component analysis of order-two tensors, (2011), manuscript.

[10] T.G. Kolda and B.W. Bader, Tensor decompositions and applications, *SIAM Review*, **51**, (2009), 455-500.