Journal of the Korean Statistical Society 39 (2010) 387–396

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Bayesian principal component analysis with mixture priors[☆]

Hyun Sook Oh[a], Dai-Gyoung Kim[b,*]

[a] Department of Applied Statistics, Kyungwon University, Sujung-gu, Sungnam 461-701, Republic of Korea

[b] Department of Applied Mathematics, Hanyang University, Sangnok-gu, Ansan 426-791, Republic of Korea

ARTICLE INFO

Article history:
Received 30 November 2009
Accepted 1 April 2010
Available online 5 May 2010

AMS 2000 subject classifications:
62

Keywords:
Probabilistic principal component analysis
Dimension reduction
Probabilistic latent variable model

ABSTRACT

A central issue in principal component analysis (PCA) is that of choosing the appropriate number of principal components to be retained. Bishop (1999a) suggested a Bayesian approach for PCA for determining the effective dimensionality automatically on the basis of the probabilistic latent variable model. This paper extends this approach by using mixture priors, in that the choice dimensionality and estimation of principal components are done simultaneously via MCMC algorithm. Also, the proposed method provides a probabilistic measure of uncertainty on PCA, yielding posterior probabilities of all possible cases of principal components.

1. Introduction

Principal component analysis (PCA) is a dimension reduction technique which is used in many application areas such as data compression, image compressing, data visualization, pattern recognition, etc. Notable recent progress in PCA has been made by Tipping and Bishop (1999a) who introduced a probability model into PCA. The probabilistic PCA (PPCA) of Tipping and Bishop (1999a) introduced a latent variable of a reduced dimension and assumed that the observed data is a linear mapping of the latent variable plus Gaussian error. PPCA converts the traditional PCA problems into statistical inference problems, hence existing statistical inferential methods such as statistical testing, Bayesian inference and use of mixture models can be utilized for PCA.

In Tipping and Bishop (1999a), a maximum likelihood method is used to find principal components. However, the maximum likelihood method is not easily applicable for determining an appropriate dimension for the principal components. A possible approach would be to compare all possible cases by using cross validation, but obviously it is computationally demanding when data is high-dimensional.

There are various works on the determination of appropriate dimensionality in PCA. Bishop (1999a) proposed a Bayesian approach for PCA, considering a hierarchical prior distribution over the matrix which conducts a linear mapping of the latent variable to data. A zero mean normal distribution is used as a prior distribution of each column of the linear mapping matrix. Posterior estimates of parameters are obtained and columns with small variance estimates are deleted and the number of retaining columns is the effective dimension of the latent space (corresponding to the number of retained principal components). In Bishop (1999a) the posterior estimation requires analytically intractable integrations and an approximation scheme based on a local Gaussian representation of the posterior distributions is employed, which can be computationally complicated. To get around this problem, Bishop (1999b) proposed a variational PCA which provides a computationally efficient approximation of the posterior distributions, based on a factorial representation of the posterior distributions.

[☆] This work was supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract UD070039AD and was supported by the Kyungwon University Research Fund in 2010.

[*] Corresponding author. Tel.: +82 31 400 5468; fax: +82 31 416 5460.
E-mail addresses: hoh@kyungwon.ac.kr (H.S. Oh), dgkim@hanyang.ac.kr (D.-G. Kim).

1226-3192/$ – see front matter © 2010 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jkss.2010.04.001

In this paper, we propose a simple and efficient Bayesian PCA procedure. We employ a prior in each column of the transformation matrix that is a mixture of continuous distribution and a discrete distribution assigning probability 1 to point zero. We assume independence among the priors. Thus, the prior of the transformation matrix is a product of the mixture prior for columns, i.e. a mixture of different dimensional distributions reflecting all possible combinations of principal components. Posterior inference on the parameters is done by using a simple Markov chain Monte Carlo method. From the posterior probabilities of all possible combinations we determine the columns retained in the principal components.

The proposed method has several advantages over existing Bayesian PCA. First, in the proposed method the choice of dimensionality and estimation of principal components are done simultaneously. Second, the method employs a simple MCMC algorithm by using a conjugate prior for the continuous part of the mixture prior so that a non-expert can easily apply the method. Third, unlike other methods which choose the columns of the transformation matrix with "large" variance estimates as principal components, the proposed method provides a probabilistic measure of uncertainty on PCA, yielding posterior probabilities of all possible cases of principal components. The uncertainty measure is practically important since in some cases it may not be easy to choose a threshold of the variance estimates or there may be several choices of principal components which yield a similar fit for the given data.

This paper is organized as follows. The probabilistic PCA model is described in Section 2. The mixture prior and posterior are introduced in Section 3 and the conditional posterior distributions which are necessary for MCMC are given in Section 4. In Section 5, we compare the proposed method with Bishop (1999a)'s Bayesian PCA using an example data set given in Bishop (1999a). We also apply the method to a data set representing measurements taken from a pipeline containing a mixture of oil, water, and gas (Bishop & James, 1993) in Section 6. Summary and conclusions are given in Section 7.

2. The probabilistic PCA model

Suppose we observe $D = \{\mathbf{t}_i\}$, $i = 1, \ldots, n$, where the $\mathbf{t}_i$'s are d-dimensional data vectors. For $q < d$, q-dimensional latent variables $\{\mathbf{x}_i\}$, $i = 1, \ldots, n$, are considered, where $\mathbf{x}_i$'s are independent q-dimensional normal, $N_q(\mathbf{0}, \mathbf{I}_q)$, random variables.

The probabilistic PCA model assumed that the observation $\mathbf{t}_i$ is a linear transformation of $\mathbf{x}_i$ with additive Gaussian noise such that

$$\mathbf{t}_i = \mathbf{W}\mathbf{x}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{W}$ is a $d \times q$ matrix, $\boldsymbol{\mu}$ is a mean vector, and $\boldsymbol{\epsilon}_i$ is a zero mean Gaussian-distributed vector with covariance $\sigma^2\mathbf{I}_d$. Thus, given $\mathbf{W}$, $\boldsymbol{\mu}$, and σ^2,

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

Note that elements of $\mathbf{t}_i$ are conditionally independent given the latent variable $\mathbf{x}_i$. Now, since $\mathbf{x}_i \sim N_q(\mathbf{0}, \mathbf{I}_q)$, the posterior distribution of $\mathbf{x}_i$ given $\mathbf{W}$, $\boldsymbol{\mu}$, and σ^2 is

$$\mathbf{x}_i|\mathbf{t}_i \sim N(\mathbf{M}^{-1}\mathbf{W}'(\mathbf{t}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}), \tag{2}$$

where $\mathbf{M} = \mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I}_q$.

Then $\mathbf{t}_i$ can be expressed in the latent space as the posterior mean of $\mathbf{x}_i$ in (2),

$$\mathbf{x}_i^* = \mathbf{M}^{-1}\mathbf{W}'(\mathbf{t}_i - \boldsymbol{\mu}), \quad i = 1, \ldots, n. \tag{3}$$

Since $\mathbf{x}_i^*$ is determined for given q, determination of q is a very important issue here. Also, note that $\mathbf{t}_i$ can be reconstructed from $\mathbf{x}_i^*$ by using

$$\hat{\mathbf{t}}_i = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{t}_i, \tag{4}$$

which is the optimal least-squares linear reconstruction of the data from $\mathbf{x}_i^*$ (Tipping & Bishop, 1999b).

In Bishop (1999a), the dimensionality of the latent space is initially set to its maximum possible value $q = d - 1$ and a hierarchical prior $P(\mathbf{W}|\boldsymbol{\alpha})$ over the matrix $\mathbf{W}$ is considered. For $P(\mathbf{W}|\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_q\}$ is a q-dimensional vector of hyper parameters in which each α_j controls $\mathbf{w}_j$ through a conditional Gaussian distribution of the form

$$P(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{j=1}^{q} \left(\frac{\alpha_j}{2\pi}\right)^{d/2} \exp\left\{-\frac{1}{2}\alpha_j\|\mathbf{w}_j\|^2\right\}, \tag{5}$$

where $\mathbf{w}_j$ is the jth column vector of $\mathbf{W}$, $j = 1, \ldots, q$.

To estimate $\mathbf{W}$ and $\boldsymbol{\alpha}$, Bishop (1999a) considered the mode of marginal posterior distribution of $\mathbf{W}$ and type-II maximum likelihood estimator using a local Gaussian approximation to the posterior distribution, respectively. From the estimated value of $\boldsymbol{\alpha}$, $\mathbf{w}_j$ is set to be $\mathbf{0}$ when α_j is very large since α_j is the inverse of the variance of $\mathbf{w}_j$, $j = 1, \ldots, q$. Then the dimensionality of the latent space is determined as the number of nonzero columns of $\mathbf{W}$.

However, it can be unclear how large α_j should be in order to ignore the corresponding column of $\mathbf{W}$, $\mathbf{w}_j$. Also, it can be computationally complicated to use an approximation scheme based on a local Gaussian representation of the posterior distributions. To get around this problem, Bishop (1999b) proposed a variational PCA which provides a computationally efficient approximation of the posterior distributions, based on factorial representation of the posterior distributions.

In this article, we propose a fully Bayesian PCA procedure using a mixture prior on each column of the matrix $\mathbf{W}$ and suggest a simple Markov chain Monte Carlo (MCMC) algorithm for posterior inference on the parameters.

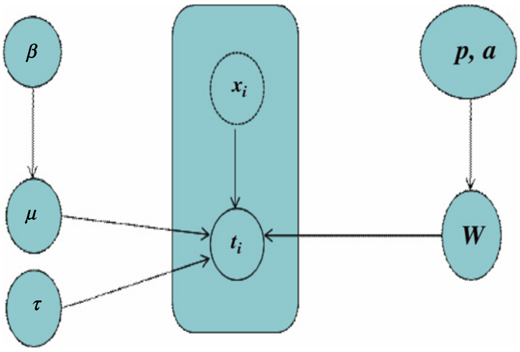


Fig. 1. Graphical representation of the model.

3. Bayesian PCA

The key issue in PPCA is whether each column of the transformation matrix $\mathbf{W}$ is either insignificant or significant, i.e. $\mathbf{w}_j = \mathbf{0}$ or $\mathbf{w}_j \neq \mathbf{0}$. If we assume a continuous prior on $\mathbf{w}_j$ then we effectively assign zero probability to $\mathbf{w}_j = \mathbf{0}$. To get around this problem, we assume a mixture prior distribution on $\mathbf{w}_j$ that gives a positive probability to the point $\mathbf{w}_j = \mathbf{0}$. In particular, we assume

$$\mathbf{w}_j | p, \alpha \sim (1-p)\delta_0(\mathbf{w}_j) + p(1-\delta_0(\mathbf{w}_j)) N\left(\mathbf{0}, \frac{1}{\alpha}\mathbf{I}_d\right), \quad j = 1, \ldots, q,$$

where

$$\delta_0(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{0} \\ 0, & \text{otherwise} \end{cases}$$

and p and α are hyper parameters such that p $(0 < p < 1)$ is the proportion of nonzero columns of the matrix $\mathbf{W}$ and $1/\alpha$ $(\alpha > 0)$ is the common variance of variables of $\mathbf{w}_j$. Note that α is a scalar here while $\boldsymbol{\alpha}$ is a vector in (5).

We can complete the specification of the Bayesian model by defining priors over the parameters $\boldsymbol{\mu}$, σ^2, p and α. Conjugate priors are considered for computational convenience. Specifically, we assume

$$\begin{aligned} & p \sim \text{Beta}(c_0, c_1), \qquad \alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \\ & \boldsymbol{\mu} | \beta \sim N\left(\mathbf{0}, \frac{1}{\beta}\mathbf{I}_d\right), \qquad \tau \sim \text{Gamma}(a_\tau, b_\tau), \qquad \beta \sim \text{Gamma}(a_\beta, b_\beta), \end{aligned} \tag{6}$$

where $\tau = 1/\sigma^2$, $\text{Beta}(c_0, c_1)$ and $\text{Gamma}(a, b)$ denote a beta distribution with shape parameters c_0 and c_1 and a gamma distribution with shape a and rate b, respectively.

Fig. 1 diagrams the described model, showing its hierarchical structure. Let Θ be a set of the whole unknown parameters,

$$\Theta = \{\boldsymbol{\mu}, p, \alpha, \tau, \beta, \{\mathbf{w}_j,\ j = 1, \ldots, q\}, \{\mathbf{x}_i,\ i = 1, \ldots, n\}\}.$$

Then the joint posterior density of Θ is given as

$$\begin{aligned} P(\Theta | \{\mathbf{t}_i\}) & \propto f(\{\mathbf{t}_i\} | \Theta) \pi(\Theta) \\ & \propto \prod_{i=1}^{n} \left(\frac{\tau}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau}{2}(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)'(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)\right] \\ & \quad \times \prod_{j=1}^{q} \left[(1-p)\delta_0(\mathbf{w}_j) + p(1-\delta_0(\mathbf{w}_j)) \left(\frac{\alpha}{2\pi}\right)^{d/2} \mathrm{e}^{-\frac{\alpha}{2}\mathbf{w}_j'\mathbf{w}_j}\right] \\ & \quad \times \prod_{i=1}^{n} \left(\frac{1}{2\pi}\right)^{q/2} \mathrm{e}^{-\frac{1}{2}\mathbf{x}_i'\mathbf{x}_i} \times \left(\frac{\beta}{2\pi}\right)^{d/2} \mathrm{e}^{-\frac{\beta}{2}\boldsymbol{\mu}'\boldsymbol{\mu}} p^{c_0-1}(1-p)^{c_1-1} \\ & \quad \times \alpha^{a_\alpha-1}\mathrm{e}^{b_\alpha\alpha} \times \tau^{a_\tau-1}\mathrm{e}^{b_\tau\tau} \times \beta^{a_\beta-1}\mathrm{e}^{b_\beta\beta}. \end{aligned} \tag{7}$$

If we get the marginal posterior distribution of $\mathbf{w}_j$ from the above joint posterior distribution, then the posterior probability of $\{\mathbf{w}_j = \mathbf{0}\}$ for $j = 1, \ldots, q$ is derived, based on which the significance of $\mathbf{w}_j$ is determined. However, it is not possible to derive analytical expression for the marginal posterior distribution of $\mathbf{w}_j$. Thus we need to rely on a numerical approximation of $P(\mathbf{w}_j = 0 | \{\mathbf{t}_i\})$.

4. Conditional posterior distribution

From the joint posterior distribution given in (7), we can easily derive the full conditional posterior distribution of each parameter given all the other parameters as follows.

$$\begin{aligned}
\boldsymbol{\mu}|\text{others} &\sim N\left(\frac{\tau}{n\tau+\beta}\sum_{i=1}^{n}(\mathbf{t}_i-\mathbf{W}\mathbf{x}_i),\ \frac{1}{n\tau+\beta}\mathbf{I}_d\right)\\
\tau|\text{others} &\sim \text{Gamma}\left(\frac{nd}{2}+a_\tau,\ b_\tau+\frac{1}{2}\sum_{i=1}^{n}(\mathbf{t}_i-\boldsymbol{\mu}-\mathbf{W}\mathbf{x}_i)'(\mathbf{t}_i-\boldsymbol{\mu}-\mathbf{W}\mathbf{x}_i)\right)\\
p|\text{others} &\sim \text{Beta}\left(c_0+\sum_{j=1}^{q}\gamma_j,\ c_1+q-\sum_{j=1}^{q}\gamma_j\right),\\
\alpha|\text{others} &\sim \text{Gamma}\left(\frac{d}{2}\sum_{j=1}^{q}\gamma_j+a_\alpha,\ b_\alpha+\frac{1}{2}\sum_{j;\gamma_j=1}\mathbf{w}_j'\mathbf{w}_j\right)\\
\beta|\text{others} &\sim \text{Gamma}\left(\frac{d}{2}+a_\beta,\ b_\beta+\frac{1}{2}\boldsymbol{\mu}'\boldsymbol{\mu}\right)\\
\mathbf{x}_i|\text{others} &\sim N\left(\mathbf{M}^{-1}\mathbf{W}'(\mathbf{t}_i-\boldsymbol{\mu}),\ \frac{1}{\tau}\mathbf{M}^{-1}\right),\quad i=1,\ldots,n\\
\mathbf{w}_j|\text{others} &\sim p_{j0}^{*}\delta_0(\mathbf{w}_j)+(1-p_{j0}^{*})N\left(\boldsymbol{\xi}_j,\ \frac{1}{\eta_j}\mathbf{I}_d\right)
\end{aligned}\tag{8}$$

where

$$\begin{aligned}
p_{j0}^{*} &= \left(1+\frac{p}{1-p}\left(\frac{\alpha}{\eta_j}\right)^{d/2}\exp\left[\frac{\eta_j}{2}\boldsymbol{\xi}_j'\boldsymbol{\xi}_j\right]\right)^{-1},\\
\gamma_j &= \begin{cases}0, & \text{if } \mathbf{w}_j=\mathbf{0}\\ 1, & \text{otherwise,}\end{cases}\\
\eta_j &= \tau\sum_{i=1}^{n}x_{ij}^2+\alpha,\\
\boldsymbol{\xi}_j &= \frac{\tau}{\eta_j}\sum_{i=1}^{n}x_{ij}(\mathbf{t}_i-\boldsymbol{\mu}-\mathbf{W}_{(-j)}\mathbf{x}_{i(-j)}),\\
\mathbf{W}_{(-j)} &= (\mathbf{w}_1,\ldots,\mathbf{w}_{j-1},\mathbf{w}_{j+1},\ldots,\mathbf{w}_q),
\end{aligned}$$

and $\mathbf{x}_{i(-j)} = (x_{i,1},\ldots,x_{i,j-1},x_{i,j+1},\ldots,x_{i,q})'$. It is easy to verify the full conditional posterior distributions of (8) except $\mathbf{w}_j$|others for which the proof is given in Appendix.

With these full conditional posterior distributions, a MCMC method can be applied to get random samples of Θ from the joint posterior distribution of Θ given in (7). Note that in the MCMC algorithm, γ_j is generated from Binomial distribution with the probability of zero, p_{j0}^{*} and if $\gamma_j = 0$, $\mathbf{w}_j = \mathbf{0}$ and otherwise (i.e. $\gamma_j = 1$), $\mathbf{w}_j$ is a random sample generated from the Gaussian distribution, $N\left(\boldsymbol{\xi}_j, \frac{1}{\eta_j}\mathbf{I}_d\right)$ in (8) for $j = 1,\ldots,d-1$. Thus insignificant columns of the matrix $\mathbf{W}$ automatically become zero through the Gibbs sampler.

5. Example 1

First, we consider a data set given in Bishop (1999a). A data set of 20 points in 10-dimensional space is generated from a Gaussian distribution having standard deviation in 5 directions given by (1.0, 0.8, 0.6, 0.4, 0.2) and standard deviation 0.04 in the remaining 5 directions, and it is assumed $\boldsymbol{\mu} = \mathbf{0}$.

Applying Bishop's Bayesian PCA (Bishop, 1999a) which uses EM algorithms, we get

$$\boldsymbol{\alpha} = \{23, 4175, 16, 17, 1784, 22, 321, 10^{10}, 3687\}.$$

Then the variance of each $\mathbf{w}_j$ is $1/\alpha_j$ for $j = 1,\ldots,9$, which is shown in Fig. 2. From Fig. 2, the appropriate dimensionality for the principal component subspace would be $\hat{q} = 4$.

Fig. 3 shows Hinton diagrams for the matrix $\mathbf{W}$. The plot on the left shows $\mathbf{W}$ before fading out while the one on the right shows $\mathbf{W}$ after fading out columns with small variances. The error variance is estimated as $\hat{\sigma}^2 = 0.0001$, but it is underestimated since it is based on $\mathbf{W}$ before fading out insignificant columns.

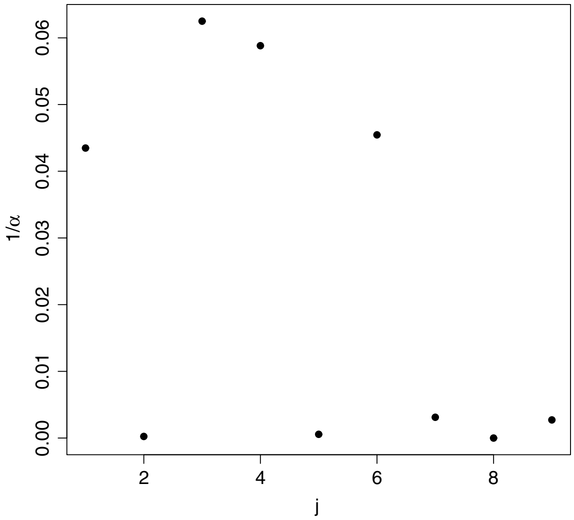
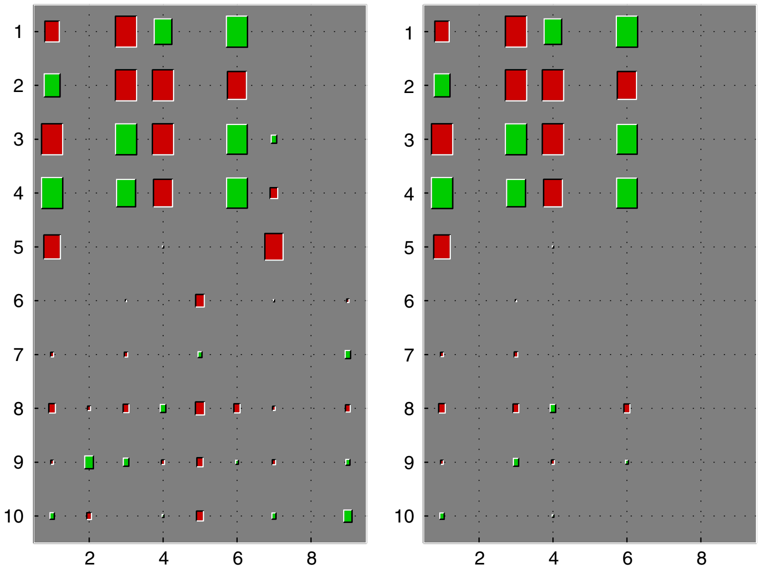


Fig. 2. Variance of $\mathbf{w}_j$ by Bishop's BPCA.



(a) W before fading out insignificant columns.

(b) W after fading out insignificant columns.

Fig. 3. Hinton diagrams of **W** by Bishop's BPCA.

We apply Bayesian PCA with mixture priors proposed in this paper to the same data. The parameters for Beta and Gamma prior distributions defined in (6) are specified as follows; for p, the proportion of nonzero columns of W, uniform prior distribution with $c_0 = c_1 = 1$ is considered as a non-informative prior. Now, since we assumed $\boldsymbol{\mu} = \mathbf{0}$, priors for α and τ are to be concerned. We use $a_\alpha = a_\tau = 3$ for easy convergence of MCMC simulation and then b_α and b_τ can be selected so that prior information is much weaker than the information from data. Since the model covariance of $\mathbf{t}$ is $\sigma^2\mathbf{I} + \mathbf{WW}'$, $\hat{\tau}$ and $\hat{\alpha}$ can be estimated empirically from the covariance matrix of data, and based on that b_α and b_τ are selected. Note that the priors for small values of b_α and b_τ reflect the information from data more.

With specified priors, 30,000 random samples for the conditional posterior distributions in Section 4 are generated by the MCMC method and then the first 35,000 samples are discarded as burn-in. From the remaining 5000 samples, the average of the random samples for each parameter is taken as an estimate of the parameter.

For several choices for parameters, b_α and b_τ, the results are shown in Table 1. It shows that the number of principal components ($\hat{q}$ and $\hat{p}$) is increasing and the variances ($1/\hat{\alpha}$ and $\hat{\sigma}^2$) are decreasing as $b_\alpha (=b_\tau)$ decreases. Actually, it can be proved that $1 - p_{j0}^*$, the posterior probability of $\{\mathbf{w}_j \neq \mathbf{0}\}, j = 1, \ldots, 9$ is decreasing as either α or τ increases from the formula given in Section 4. Hence, if we choose priors representing data well, then the procedure finds principal

Table 1
Posteror estimates for parameters given b_α and b_τ.

b_α $(=b_\tau)$	$\hat{q}$	$\hat{p}$	$\hat{\alpha}$	$\hat{\sigma}^2$
0.1	5	0.54	27.07	0.004
0.5	4	0.45	14.11	0.0142
1	4	0.45	10.05	0.0143
2	3	0.36	5.42	0.0512

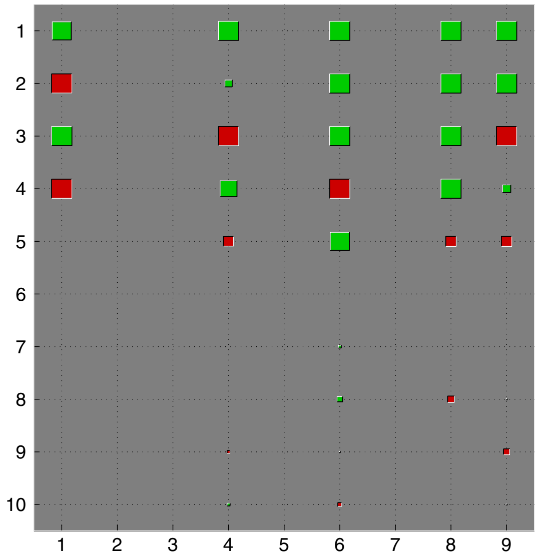


Fig. 4. Hinton diagram of $\mathbf{W}$.

components more precisely. In this example, our data has 5 components with larger variance relatively than the other remaining components, which is detected when b_α and b_τ are small. On the other hand, it has been checked that the results are not sensitive for the choice of parameters, c_0 and c_1.

Let us consider more detail in the case of $b_\alpha = b_\tau = 0.1$. The Hinton diagram of $\mathbf{W}$ is given in Fig. 4. We can see that insignificant columns of $\mathbf{W}$ have been faded out automatically.

With the retained columns of $\mathbf{W}$, 5-dimensional principal components are obtained by (3) in Section 2. Also, the reconstruction of the data from these principal components is possible by using the formula (4) in Section 2. Fig. 5 shows image plots for the original data, the compressed data (principal components) and the reconstructed data, from the left to the right. 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 components of the given data set.

The time sequence plots for the posterior probability of $\{\mathbf{w}_j \neq \mathbf{0}\}, j = 1, \ldots, 9$ are given in Fig. 6. It shows strong evidence that convergence has been achieved.

6. Example 2

Our second example is a synthetically generated data set representing measurements taken from a pipeline containing a mixture of oil, water, and gas (Bishop & James, 1993). Each data point consists of 12 measurements and belongs to one of three different geometrical configurations corresponding to laminar (stratified), homogeneous, and annular flows. Details of the data are described in Bishop (2006) and the data set is available from the web site "http://research.microsoft.com/cmbishop/PRML ".

The data set is divided into training, validation, and test sets. We use the test set here, containing $n = 1000$ samples of $d = 12$ dimensional vectors. The Eigen values obtained from the correlation matrix are

$$5.010, 2.238, 2.171, 0.832, 0.748, 0.366, 0.202, 0.169, 0.131, 0.092, 0.031, 0.011.$$

There is a big decrease between the 1st and the 2nd Eigen values and also between the 3rd and the 4th Eigen values. From this information the effective dimensionality seems to be 1 or 3 based on the conventional PCA.

Now, the MCMC algorithm for the proposed Bayesian PCA with mixture priors is applied to the data set with empirical priors as in Example 1 and we get the posterior probability,

$$P(\mathbf{w}_1 \neq \mathbf{0}, \ldots, \mathbf{w}_9 \neq \mathbf{0}) = (1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1),$$

which clearly indicates $\hat{q} = 3$. The convergence has been achieved well as in the previous example.

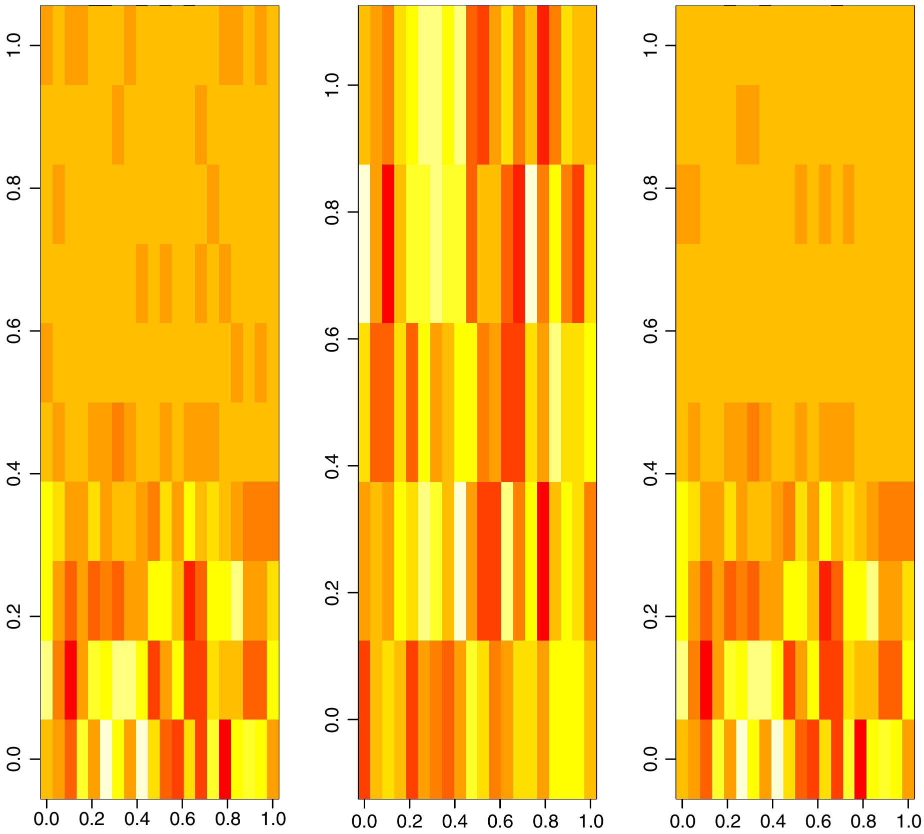


Fig. 5. Image plots for the original data (left), compressed data (middle), and reconstructed data (right).

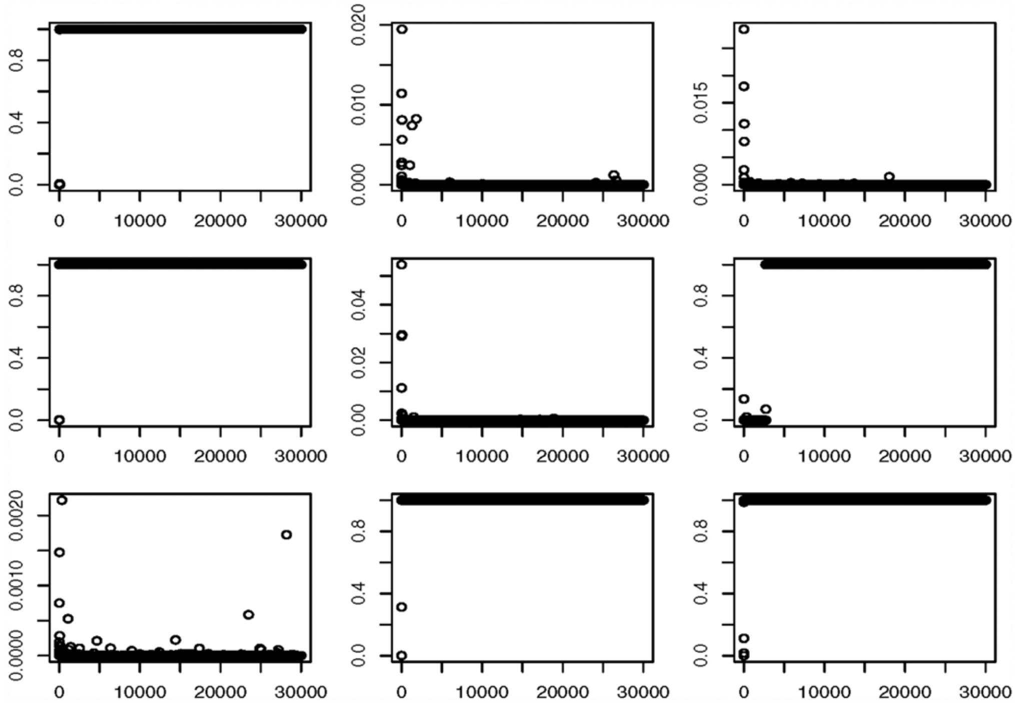


Fig. 6. Plots of the posterior probability of $\{\mathbf{w}_j \neq \mathbf{0}\}, j = 1, \ldots, 9$.

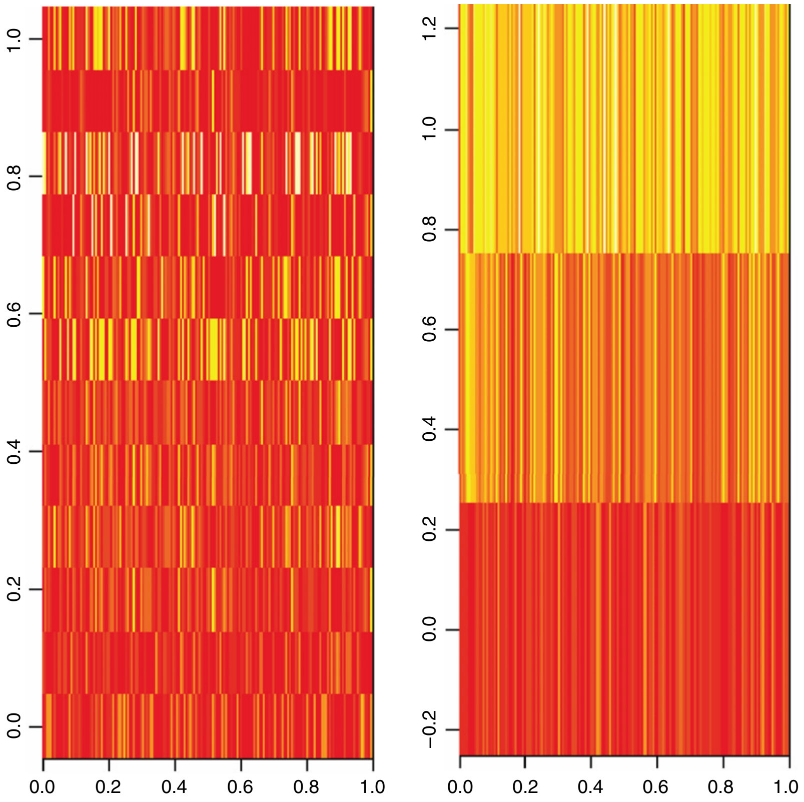


Fig. 7. Image plots for oil flow data in the original 12-dimensional data space (left) and in 3-dimensional latent space (right).

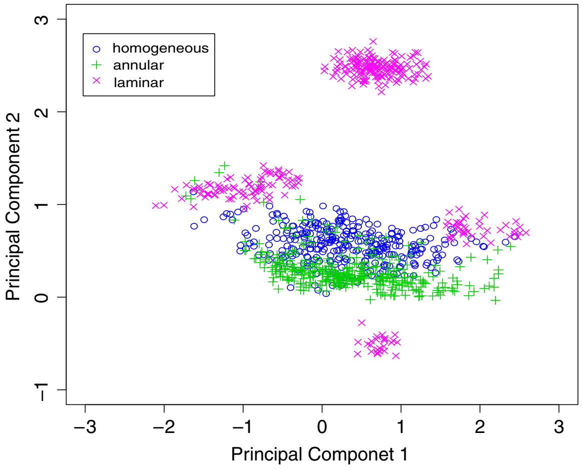


Fig. 8. Visualization of oil flow data in 2 dimensional space.

Fig. 7 shows image plots for the data in 12-dimensional and 3-dimensional space, respectively. The image plot in 3-dimensional latent space shows 3 different geometrical configurations in the data more evidently than the one in the original 12-dimensional data space.

For data visualization, we take the latent variable space to be 2-dimensional by fixing $q = 2$ in model (1) and let $p = 2/11$ for the proportion of nonzero columns of the matrix $\mathbf{W}$. Fig. 8 shows the oil flow data visualized in 2-dimensional space in which each point is symbolized according to its geometrical configuration. Note that it is very simple and efficient to get the principal components in any dimensional space by fixing q and the prior probability p for the proportion of nonzero columns of the matrix $\mathbf{W}$.

7. Conclusions

There have been various works for PCA based on the PPCA model since its introduction by Tipping and Bishop (1999a). The key issue in PPCA is determination of the appropriate dimensions in PCA. Bishop (1999a) proposed a Bayesian PCA for determining the effective dimensionality automatically, considering a hierarchical prior distribution over the matrix which conducts a linear mapping of the latent variable to data.

In this paper, we provide a fully Bayesian procedure, extending Bishop's approach by using mixture priors on the transformation matrix. The mixture priors are given reflecting all possible combinations of principal components. This provides a probabilistic measure of uncertainty on PCA, yielding posterior probabilities of all possible case of principal components. This is practically important since in some cases there may be several choices of principal components which yield a similar fit for the given data. Also, the choice of dimensionality and estimation of principal components are done simultaneously and a simple MCMC algorithm is employed, so that a non-expert can easily apply the method.

The proposed method can be applied for very high-dimensional data, however, it takes a lot of time to get the result since all possible cases of principal components are considered. To avoid this problem, a small value of q can be given initially so that the range for possible dimensionality can be restricted priorly. This will be considered in detail in our future work with applications in the area of pattern recognition.

Appendix

Proof of the formula for $\mathbf{w}_j|$others *in* (8); Let $C_{1j} = f(\mathbf{t}|\boldsymbol{\mu}, \mathbf{w}_j = \mathbf{0}, \mathbf{W}_{(-j)}, \alpha, \mathbf{x}, \tau, \beta)$ and

$$C_{2j} = \int_{\mathbf{w}_j \neq \mathbf{0}} f(\mathbf{t}|\boldsymbol{\mu}, \mathbf{w}_j \neq \mathbf{0}, \mathbf{W}_{(-j)}, \alpha, \mathbf{x}, \tau, \beta) \left(\frac{\alpha}{2\pi}\right)^{d/2} \mathrm{e}^{-\frac{\alpha}{2}\mathbf{w}_j'\mathbf{w}_j} d\mathbf{w}_j.$$

Then (Gottardo & Raftery, 2008)

$$P(\mathbf{w}_j = \mathbf{0}|\mathbf{t}, \boldsymbol{\mu}, \mathbf{W}_{(-j)}, \alpha, \mathbf{x}, \tau, \beta) = \frac{C_{1j}(1-p)}{C_{1j}(1-p) + C_{2j}p}. \tag{A.1}$$

Now,

$$C_{1j} = \left(\frac{\tau}{2\pi}\right)^{nd/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}_{(-j)}\mathbf{x}_{i(-j)})'(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}_{(-j)}\mathbf{x}_{i(-j)})\right],$$

and

$$C_{2j} = \int_{\mathbf{w}_j \neq \mathbf{0}} \left(\frac{\tau}{2\pi}\right)^{nd/2} \left(\frac{\alpha}{2\pi}\right)^{d/2} \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)'(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)\right] \exp\left[-\frac{\alpha}{2}\mathbf{w}_j'\mathbf{w}_j\right] d\mathbf{w}_j.$$

After straightforward calculation of C_{2j},

$$C_{2j} = C_{1j}\left(\frac{\alpha}{\eta_j}\right)^{d/2} \exp\left[\frac{\eta_j}{2}\boldsymbol{\xi}_j'\boldsymbol{\xi}_j\right].$$

Hence

$$\text{(A.1)} = \left(1 + \frac{p}{1-p}\frac{C_{2j}}{C_1j}\right)^{-1} = \left(1 + \frac{p}{1-p}\left(\frac{\alpha}{\eta_j}\right)^{d/2} \exp\left[\frac{\eta_j}{2}\boldsymbol{\xi}_j'\boldsymbol{\xi}_j\right]\right)^{-1},$$

which is p_{j0}^*.

Next, from the posterior probability distribution of Θ given in (7) the conditional posterior probability distribution of $\mathbf{w}_j$ when $\mathbf{w}_j \neq \mathbf{0}$ is

$$P(\mathbf{w}_j|\mathbf{w}_j \neq \mathbf{0}, \mathbf{t}, \boldsymbol{\mu}, \mathbf{W}_{(-j)}, \alpha, \mathbf{x}, \tau, \beta) \propto \exp\left[-\frac{\tau}{2}\sum_{i=1}^{n}(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)'(\mathbf{t}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{x}_i)\right] \exp\left[-\frac{\alpha}{2}\mathbf{w}_j'\mathbf{w}_j\right].$$

Then it is easy to verify the formula $\mathbf{w}_j|$others from the above formula.

References

Bishop, C. M. (1999a). Bayesian PCA. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems*: *Vol. 11* (pp. 382–388). MIT Press.

Bishop, C. M. (1999b). Variational principal components. In *Proceedings ninth international conference on artificial neural networks, vol. 1* (pp. 509–514). IEE.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer Science + Business Media, LLC.

Bishop, C. M., & James, G. D. (1993). Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research, Section A*, *327*, 580–593.

Gottardo, R., & Raftery, A. E. (2008). Markov chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics*, *17*(4), 949–975.

Tipping, M. E., & Bishop, C. M. (1999a). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, *21*(3), 611–622.

Tipping, M. E., & Bishop, C. M. (1999b). Mixtures of probabilistic principal component analyzers. *Neural Computation*, *11*(2), 443–482.