Bayesian Inference on Principal Component Analysis using Reversible Jump Markov Chain Monte Carlo

Zhihua Zhang[1] and **Kap Luk Chan**[2] and **James T. Kwok**[1] and **Dit-Yan Yeung**[1]

[1]Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{zhzhang,jamesk,dyyeung}@cs.ust.hk

[2]School of Electrical and Electronic Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
eklchan@ntu.edu.sg

Abstract

Based on the probabilistic reformulation of principal component analysis (PCA), we consider the problem of determining the number of principal components as a model selection problem. We present a hierarchical model for probabilistic PCA and construct a Bayesian inference method for this model using reversible jump Markov chain Monte Carlo (MCMC). By regarding each principal component as a point in a one-dimensional space and employing only birth-death moves in our reversible jump methodology, our proposed method is simple and capable of automatically determining the number of principal components and estimating the parameters simultaneously under the same disciplined framework. Simulation experiments are performed to demonstrate the effectiveness of our MCMC method.

Introduction

Principal component analysis (PCA) is a powerful tool for data analysis. It has been widely used for such tasks as dimensionality reduction, data compression and visualization. The original derivation of PCA is based on a standardized linear projection that maximizes the variance in the projected space. Recently, Tipping & Bishop (1999) proposed the probabilistic PCA which explores the relationship between PCA and factor analysis of generative latent variable models. This opens the door to various Bayesian treatments of PCA. In particular, Bayesian inference can now be employed to solve the central problem of determining the number of principal components that should be retained. Bishop (1999a; 1999b) addressed this by using automatic relevance determination (ARD) (Neal 1996) and Bayesian variational methods. Minka (2001), on the other hand, adopted a Bayesian method which is based on the Laplace approximation. In this paper, we propose a hierarchical model for Bayesian inference on PCA using the novel reversible jump Markov chain Monte Carlo (MCMC) algorithm of Green (1995).

In brief, reversible jump MCMC is a random-sweep Metropolis-Hastings method for varying-dimension problems. It constructs a dimension matching transform using the reversible jump methodology and estimates the parameters using Gibbs sampling. Richardson & Green (1997), by developing the split-merge and birth-death moves for the reversible jump methodology, performed a fully Bayesian analysis on univariate data generated from a finite Gaussian mixture (GM) with an unknown number of components. This was then further extended to univariate hidden Markov models (HMM) by Robert, Rydén, & Titterington (2000). In general, reversible jump MCMC is attractive in that it can perform parameter estimation and model selection simultaneously within the same framework. In contrast, the other methods mentioned above can only perform model selection separately. In recent years, reversible jump MCMC has also been successfully applied to neural networks (Holmes & Mallick 1998; Andrieu, Djurié, & Doucet 2001) and pattern recognition (Roberts, Holmes, & Denison 2001).

Motivated by these successes, in this paper, we introduce reversible jump MCMC into the probabilistic PCA framework. This provides a disciplined method to perform parameter estimation simultaneously with choosing the number of principal components. In particular, we propose a hierarchical model for probabilistic PCA, together with a Bayesian inference procedure for this model using reversible jump MCMC. Note that PCA is considerably simpler than GMs and HMMs in the following ways. First, PCA has much fewer free parameters than GMs and HMMs. Second, unlike GMs and HMMs, no component in PCA can be empty. Third, using reversible jump MCMC in GMs and HMMs for high-dimensional data is still an open problem, while reversible jump MCMC for PCA is more manageable because, as to be discussed in more detail in later sections, each principal component can be regarded as a point in some one-dimensional space. Because of these, we will only employ birth-death moves for the dimension matching transform in our reversible jump methodology.

The rest of this paper is organized as follows. In the next section, we give a brief overview of probabilistic PCA and the corresponding maximum likelihood estimation problem. A hierarchical Bayesian model and the corresponding reversible jump MCMC procedure are then presented, followed by some experimental results on different data sets. The last section gives some concluding remarks.

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Probabilistic PCA

Probabilistic PCA was proposed by Tipping & Bishop (1999). In this model, a high-dimensional random vector $\mathbf{x}$ is expressed as a linear combination of basis vectors ($\mathbf{h}_j$'s) plus noise ($\boldsymbol{\epsilon}$):

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{q} \mathbf{h}_j w_j + \mathbf{m} + \boldsymbol{\epsilon} \\ &= \mathbf{H}\mathbf{w} + \mathbf{m} + \boldsymbol{\epsilon}, \qquad (1) \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{V}), \qquad (2) \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{w} = (w_1, \ldots, w_q)^T \in \mathbb{R}^q$, $q < d$, and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_q]$ is a $d \times q$ matrix that relates the two sets of variables $\mathbf{x}$ and $\mathbf{w}$. The vector $\mathbf{m}$ allows the model to have non-zero mean. In PCA, the noise variance matrix $\mathbf{V}$ is hyperspherical, i.e.,

$$\mathbf{V} = \sigma^2 \mathbf{I}_d, \qquad (3)$$

and the latent variables $w_1, \ldots, w_q$ are independent Gaussians with zero mean and unit variance, i.e.,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q).$$

Note that the probabilistic PCA is closely related to factor analysis, with the only difference being that the noise variance matrix $\mathbf{V}$ in factor analysis is a general diagonal matrix.

Given an observed data set $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x}_N\}$, the goal of PCA is to estimate the matrix $\mathbf{H}$ in (1) and the noise variance σ^2 in (3). From (1) and (2), we can obtain the conditional probability of the observation vector $\mathbf{x}$ as

$$\mathbf{x} | \mathbf{w}, \mathbf{H}, \mathbf{m}, \sigma^2 \sim \mathcal{N}(\mathbf{H}\mathbf{w} + \mathbf{m}, \sigma^2 \mathbf{I}),$$

and so, by integrating out $\mathbf{w}$, we have

$$p(\mathbf{x}|\mathbf{H}, \mathbf{m}, \sigma^2) = \int p(\mathbf{x}|\mathbf{w}, \mathbf{H}, \mathbf{m}, \sigma^2) p(\mathbf{w}) d\mathbf{w}.$$

Note that (Tipping & Bishop 1999)

$$\mathbf{x}|\mathbf{H}, \mathbf{m}, \sigma^2 \sim \mathcal{N}(\mathbf{m}, \mathbf{C}),$$

with $\mathbf{C} = \mathbf{H}\mathbf{H}^\mathbf{T} + \sigma^2 \mathbf{I}$. The corresponding likelihood is therefore

$$\begin{aligned} p(\mathcal{D}|\mathbf{H}, \mathbf{m}, \sigma^2) &= \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{H}, \mathbf{m}, \sigma^2) \\ &= (2\pi)^{-Nd/2} |\mathbf{H}\mathbf{H}^T + \sigma^2 \mathbf{I}|^{-N/2} \\ &\quad \times e^{-\frac{N}{2} \mathrm{tr}((\mathbf{H}\mathbf{H}^T + \sigma^2 \mathbf{I})^{-1} \mathbf{S})}, \qquad (4) \end{aligned}$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T. \qquad (5)$$

From (Tipping & Bishop 1999), the maximum likelihood estimates of $\mathbf{m}$ and $\mathbf{H}$ are

$$\begin{aligned} \widehat{\mathbf{m}} &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i, \qquad (6) \\ \widehat{\mathbf{H}} &= \mathbf{U}_q (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I}_q)^{1/2} \mathbf{R}, \end{aligned}$$

respectively, where $\mathbf{U}_q$ is a $d \times q$ orthogonal matrix in which the q column vectors are the principal eigenvectors of $\mathbf{S}$, $\boldsymbol{\Lambda}_q$ is a $q \times q$ diagonal matrix containing the corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal matrix. For $\mathbf{H} = \widehat{\mathbf{H}}$, the maximum likelihood estimate of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j,$$

which implies that the maximum likelihood noise variance is equal to the average of the left-out eigenvalues.

Bayesian Formalism for PCA

Hierarchical Model and Priors

In a fully Bayesian framework, both the number of principal components (q) and the model parameters ($\boldsymbol{\theta} = \{\mathbf{H}, \mathbf{m}, \sigma^2\}$) are considered to be drawn from appropriate prior distributions. We assume that the joint density of all these variables takes the form

$$p(q, \boldsymbol{\theta}, \mathcal{D}) = p(q) p(\boldsymbol{\theta}\,|q)\, p(\mathcal{D}\,|\boldsymbol{\theta}, q)\,. \qquad (7)$$

Following (Minka 2001), we decompose the matrix $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{U}_q (\mathbf{L}_q - \sigma^2 \mathbf{I}_q)^{1/2} \mathbf{R},$$

where $\mathbf{U}_q^T \mathbf{U}_q = \mathbf{I}_q, \mathbf{R}^T \mathbf{R} = \mathbf{I}_q$, and $\mathbf{L}_q = \mathrm{diag}(l_1, \ldots, l_q)$. It is easy to extend the $d \times q$ matrix $\mathbf{U}_q$ to a $d \times d$ orthogonal matrix $\mathbf{U}$ such that $\mathbf{U} = (\mathbf{U}_q, \mathbf{U}_{d-q})$ and $\mathbf{U}_{d-q}^T \mathbf{U}_{d-q} = \mathbf{I}_{d-q}$. Letting

$$\mathbf{L} = \begin{array}{c c} & \begin{array}{cc} q & d-q \end{array} \\ \begin{array}{c} q \\ d-q \end{array} & \left(\begin{array}{cc} \mathbf{L}_q - \sigma^2 \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right) \end{array},$$

we have

$$\begin{aligned} \mathbf{U}\mathbf{L}\mathbf{U}^T &= (\mathbf{U}_q, \mathbf{U}_{d-q}) \begin{pmatrix} \mathbf{L}_q - \sigma^2 \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U}_q^T \\ \mathbf{U}_{d-q}^T \end{pmatrix} \\ &= \mathbf{U}_q (\mathbf{L}_q - \sigma^2 \mathbf{I}) \mathbf{U}_q^T \\ &= \mathbf{H}\mathbf{H}^T. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{H}\mathbf{H}^T + \sigma^2 \mathbf{I}_d &= \mathbf{U}\mathbf{L}\mathbf{U}^T + \sigma^2 \mathbf{I}_d \\ &= \mathbf{U} \begin{pmatrix} \mathbf{L}_q & \mathbf{0} \\ \mathbf{0} & \sigma^2 \mathbf{I}_{d-q} \end{pmatrix} \mathbf{U}^T. \qquad (8) \end{aligned}$$

Since the matrix $\mathbf{S}$ in (5) is symmetric and positive definite, we can decompose it as $\mathbf{S} = \mathbf{A}\mathbf{G}\mathbf{A}^T$, where $\mathbf{A}$ is an orthogonal matrix consisting of the eigenvectors of $\mathbf{S}$, and $\mathbf{G}$ is diagonal with diagonal elements g_i's being the eigenvalues of $\mathbf{S}$.

For simplicity, we set $\mathbf{R} = \mathbf{I}_q$ and use the maximum likelihood estimators for $\mathbf{m}$ and $\mathbf{U}$. In other words, we obtain $\mathbf{m}$ from (6) and set $\mathbf{U}$ to be the eigenvector matrix $\mathbf{A}$ of $\mathbf{S}$.

Combining with (4) and (8), we can rewrite the likelihood as:

$$
\begin{aligned}
& p(\mathcal{D}|l_1^{-1},\dots,l_q^{-1},\sigma^{-2}) \\
&= \prod_{i=1}^{N} p(\mathbf{x}_i|l_1^{-1},\dots,l_q^{-1},\sigma^{-2}) \\
&= (2\pi)^{-\frac{Nd}{2}} \prod_{j=1}^{q} l_j^{-N/2} \sigma^{-N(d-q)} \times e^{-\frac{N}{2}\sum_{j=1}^{q} l_j^{-1} g_j} \\
&\quad \times e^{-\frac{N\sigma^{-2}}{2}\sum_{j=q+1}^{d} g_j}.
\end{aligned}
$$

Now, our goal is to estimate the parameters (l_j's and σ^2) and the number of principal components (q) via reversible jump MCMC. First of all, we have to choose a proper prior distribution for each parameter. A common choice for q is the Poisson distribution with hyperparameter λ. Here, for convenience of presentation and interpretation, we assume that q follows a uniform prior on $\{1, 2, \cdots, d-1\}$.

To ensure identifiability, we impose the following ordering constraint on l_j's and σ^2:

$$l_1 > l_2 > \cdots > l_q > \sigma^2.$$

The prior joint density for these parameters is then given by:

$$
\begin{aligned}
p(l_1,\dots,l_q,\sigma^2|q) &= (q+1)!\, p(l_1,\dots,l_q,\sigma^2) \\
&\quad \times \mathbf{I}_{l_1>l_2>\cdots>l_q>\sigma^2}(l,\sigma^2),
\end{aligned}
$$

where **I** denotes the indicator function, and the $(q+1)!$ term arises from the ordering constraint.

We assume that $l_1,\dots,l_q$ and σ^2 are distributed *a priori* as independent variables conditioned on some hyperparameter. In this case, we consider the prior distributions for the parameters $l_1,\dots,l_q$ and σ^2 as conjugate priors:

$$
\begin{aligned}
l_j^{-1} &\sim \Gamma(r,\tau), \quad j = 1, 2, \dots, q, \\
\sigma^{-2} &\sim \Gamma(r,\tau),
\end{aligned}
$$

where $\Gamma(\cdot,\cdot)$ denotes the Gamma distribution.[1] Since the hyperparameter $r > 0$ represents the shape of the Gamma distribution, it is appropriate to pre-specify it. In this paper, r is held fixed while $\tau > 0$ is also given a Gamma prior:

$$\tau \sim \Gamma(\alpha,\eta),$$

where $\alpha > 0$ and $\eta > 0$. We have thus obtained a complete hierarchical model (Figure 1), which can be represented graphically in the form of a directed acyclic graph (DAG).

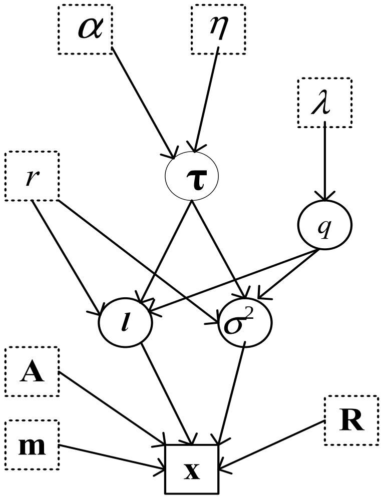


Figure 1: DAG for the proposed probabilistic PCA model.

Reversible Jump MCMC Methodology

For the hierarchical model proposed above, the goal of Bayesian inference is to generate realizations from the conditional joint density $p(q,\boldsymbol{\theta}|\mathcal{D})$ derived from (7). The reversible jump MCMC algorithm in (Green 1995) allows us to handle this problem even when the number of principal components is unknown. The algorithm proceeds by augmenting the usual proposal step of a random-sweep Metropolis-Hastings method with variable dimensions. It constructs the dimension matching transform with the split-merge or birth-death moves by adding or removing a latent variable. As discussed in the introduction section, we only use the birth-death moves in this paper. Consequently, each sweep of our reversible jump MCMC procedure consists of three types of moves:

(a) update the parameters l_j's and σ;
(b) update the hyperparameter τ;
(c) the birth or death of a component.

Move types (a) and (b) are used for parameter estimation via Gibbs sampling. Since move type (c) involves changing the number of components q by 1, it constitutes the reversible jump and is used for model selection via the Metropolis-Hastings algorithm. Assume that a move of type m, which moves from the current state $\mathbf{s}$ to a new state $\mathbf{s}'$ in a higher-dimensional space, is proposed. This is often implemented by drawing a vector $\mathbf{v}$ of continuous random variables, independent of $\mathbf{s}$, and denoting $\mathbf{s}'$ by using an invertible deterministic function $f(\mathbf{s},\mathbf{v})$. Then, the acceptance probabilities from $\mathbf{s}$ to $\mathbf{s}'$ and from $\mathbf{s}'$ to $\mathbf{s}$ are $\min(1,R)$ and $\min(1,R^{-1})$, respectively, where

$$R = \frac{p(\mathbf{s}'|\mathbf{x})\, r_m(\mathbf{s}')}{p(\mathbf{s}|\mathbf{x})\, r_m(\mathbf{s}) p(\mathbf{v})} \left| \frac{\partial \mathbf{s}'}{\partial(\mathbf{s},\mathbf{v})} \right|, \qquad (9)$$

$r_m(\mathbf{s})$ is the probability of choosing move type m in state $\mathbf{s}$, $p(\mathbf{v})$ is the density function of $\mathbf{v}$, and $\left|\frac{\partial \mathbf{s}'}{\partial(\mathbf{s},\mathbf{v})}\right|$ is the Jacobian arising from the change of variables from $(\mathbf{s},\mathbf{v})$ to $\mathbf{s}'$. Using (9) for our probabilistic PCA model, the acceptance probability for the birth move from $\{q, l_1^{-1},\dots,l_q^{-1}\}$

[1] The Gamma density of a random variable $x \sim \Gamma(\alpha,\lambda)$ is defined as:

$$p(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},$$

where $\alpha > 0, \lambda > 0$ are the shape and scaling parameters, respectively.

to $\{q+1, l_1^{-1}, \dots, l_q^{-1}, l_{q+1}^{-1}\}$ is $\min(1, R)$, where

$$\begin{aligned} R &= \text{(likelihood ratio)}\frac{p(q+1)}{p(q)}(q+2) \\ &\quad \times \frac{p(l_1^{-1}, \dots, l_{q+1}^{-1}, \sigma^{-2})}{p(l_1^{-1}, \dots, l_q^{-1}, \sigma^{-2})} \frac{d_{q+1}}{b_q} \frac{1}{p(l_{q+1}^{-1})}. \end{aligned} \tag{10}$$

Here, d_k and b_k are the probabilities of attempting death and birth moves, respectively, when the current state has k latent variables. Usually, $b_j = d_j = 0.5$ when $j = 2, \dots, d-2$, $d_1 = b_{d-1} = 0$ and $b_1 = d_{d-1} = 1$. Our death proposal proceeds by choosing the latent variable with the smallest eigenvalue.

The correspondence between (9) and (10) is fairly straightforward. The first two terms of (10) form the ratio $\frac{p(\mathbf{s}'|\mathbf{x})}{p(\mathbf{s}|\mathbf{x})}$, the $(q+2)$-factor is the ratio $\frac{(q+2)!}{(q+1)!}$ from the order statistics densities for the parameters l_j's and σ^2, and the last term is the proposal ratio $\frac{r_m(\mathbf{s}')}{r_m(\mathbf{s})p(\mathbf{v})}$. The Jacobian is equal to unity because we are drawing new principal components independent of the current parameters.

Reversible Jump MCMC Algorithm for Probabilistic PCA

We use Gibbs sampling (Gilks, Richardson, & Spiegelhalter 1996) to simulate the parameters and hyperparameters in our model. The reversible jump MCMC algorithm for the proposed probabilistic PCA method is described as follows:

Reversible Jump MCMC Algorithm

1. Initialization: Sample $(q, \sigma^{-2}, l_1^{-1}, \dots, l_q^{-1}, \tau)$ from their priors.
2. Iteration t:
   - Update the parameters and hyperparameters using Gibbs sampling.
   - Draw a uniform random variable $u \sim \mathcal{U}(0,1)$;
   - If $u \le b_q$, then perform the Birth move.
   - Else if $u \le b_q + d_q$, then perform the Death move.
   - End if
3. Set $t = t+1$ and go back to Step 2 until convergence.

Gibbs Sampler

1. For $j = 1, \cdots, q$, simulate from the full conditionals

$$\begin{aligned} l_j^{-1}|\cdots &\sim \Gamma\Big(\frac{N}{2} + r, \frac{N g_j}{2} + \tau\Big)\mathbf{I}_{\left[l_{j-1}, l_{j+1}\right]}(l_j), \\ \sigma^{-2}|\cdots &\sim \Gamma\Big(\frac{N(d-q)}{2} + r, \frac{N\sum_{j=q+1}^{d} g_j}{2} + \tau\Big) \\ &\quad \times \mathbf{I}_{\left(0, l_q\right]}(\sigma^2). \end{aligned}$$

2. Simulate the hyperparameter τ from its full conditional:

$$\tau|\cdots \sim \Gamma\Big((q+1)r + \alpha, \sum_{j=1}^{q} l_j^{-1} + \sigma^{-2} + \eta\Big).$$

Birth move

1. Draw l_{q+1}^{-1} from its prior $\Gamma(r, \tau)\mathbf{I}_{[l_q, \sigma^2]}(l_{q+1})$.
2. Calculate the acceptance probability $\alpha = \min(1, R)$ of the birth move using (10).
3. Draw a uniform random variable $v \sim \mathcal{U}(0,1)$.
4. If $v < \alpha$, then accept the proposed state; otherwise, set the next state to be the current state.

Death move

1. Remove the qth principal component.
2. Calculate the acceptance probability $\alpha = \min(1, R^{-1})$ of the death move using (10).
3. Draw a uniform random variable $v \sim \mathcal{U}(0,1)$.
4. If $v < \alpha$, then accept the proposed state; otherwise, set the next state to be the current state.

Experiments

In this section, we perform experiments on several data sets to demonstrate the efficacy of the reversible jump MCMC algorithm for probabilistic PCA. We adopt the recommendation of Richardson & Green (1997) on the choice of hyperparameters and set $r > 1 > \alpha$. Also, we set $r = 3.0$, $\alpha = 0.5$, and $\eta = 1.2/V$, where V is the standard deviation of the data. We run our algorithm for 20,000 sweeps in the following two experiments. The first 10,000 sweeps are discarded as burn-in. All our inferences are based on the last 10,000 sweeps.

Experiment 1

In the first experiment, we generate a data set (Set 1) of 1,000 points from a 6-dimensional Gaussian distribution, with variances in the 6 dimensions equal to 10, 7, 5, 3, 1 and 1, respectively. The eigenvalues of the observed covariance matrix on the data so generated are 8.9580, 7.2862, 5.3011, 2.8964, 1.1012 and 0.9876, respectively. Table 1 shows the posterior probabilities for different numbers of principal components (q) and the corresponding estimated values of the parameters (l_j's and σ^2). As we can see, the posterior probability of q is tightly concentrated at $q = 4$, which agrees with our intuition that there are 4 dominant dimensions in this data set. Moreover, the estimated values of the parameters are very close to the true ones. Figure 2 depicts the jumping in the number of principal components on the last 10,000 sweeps.

Table 1: Posterior probabilities for different numbers of principal components (q's) and the estimated values of l_j's and σ^2.

q	$p(q\|\mathcal{D})$	σ^2	l_j's
4	**0.8666**	1.0573	9.0342, 7.3198 5.2214, 2.9420
5	0.1334	1.0263	9.0322, 7.3187 5.2201, 2.9403, 1.0930

Like the variational method in (Bishop 1999b), our proposed MCMC method does not provide one specific value on the number of principal components that should be retained. Instead, it provides posterior probability estimate for each possible dimensionality over the complete range. This leaves room for us to make an appropriate decision. In many applications, however, these probabilities are tightly concentrated at a specific dimensionality or only a few dimensionalities that are close to each other. Notice that the number of principal components q only jumps between 4 and 5 after the burn-in period. Since the noise variance σ^2 is very close to the variance of either of the last two principal components (Table 1), either one may be treated as a noise term and hence q is sometimes estimated to be equal to 5.

Experiment 2

In the second experiment, we use three data sets similar to those used by Minka (2001). The first data set (Set 2 in Table 2) consists of 100 points generated from a 10-dimensional Gaussian distribution, with variances in the first 5 dimensions equal to 10, 8, 6, 4 and 2 respectively, and with variance equal to 1 in the last 5 dimensions. The second data set (Set 3 in Table 2) consists of 100 points generated from a 10-dimensional Gaussian distribution, with variances in the first 5 dimensions equal to 10, 8, 6, 4 and 2 respectively, and with variance 0.1 in the last 5 dimensions. The third data sets (Set 4 in Table 2), consisting of 10,000 points, is generated from a 15-dimensional Gaussian distribution, with variances in the first 5 dimensions equal to 10, 8, 6, 4 and 2 respectively, and with variance equal to 0.1 in the last 10 dimensions.

Table 2 shows the posterior probabilities for different numbers of principal components (q) and Figure 2 depicts the jumping in the number of principal components during the last 10,000 sweeps. As we can see, the posterior probabilities are all tightly concentrated at $q = 5$, which agrees with our intuition that there are 5 dominant dimensions in these data sets.

Concluding Remarks

In this paper, we have proposed a Bayesian inference method for PCA using reversible jump MCMC. This allows simultaneous determination of the number of principal components and estimation of the corresponding parameters. Moreover, since each principal component in this PCA framework is considered as a point in a one-dimensional space, the use of reversible jump MCMC becomes feasible. Also, as the proposed probabilistic PCA framework is only one of the possible generative latent variable models, in the future, it is worthy to explore the reversible jump MCMC methodology for other generative models, such as the mixture of probabilistic PCA, factor analysis and its mixture, and independent component analysis.

In our reversible jump method, the dimension matching transform only employs birth-death moves, but not both split-merge and birth-death moves as in the reversible jump method for Gaussian mixtures (Richardson & Green 1997). Note that the birth-death moves of (Richardson & Green 1997) are developed for the empty components, which is only a supplement of the split-merge moves in order to enhance the robustness of the reversible jump method. Apparently, it is possible to use split-merge moves instead of birth-death moves to develop a reversible jump method for PCA. Recently, Stephens (2000) used the birth-death process instead of the reversible jump methodology and described an alternative of the reversible jump MCMC, called the birth-death MCMC. This birth-death MCMC differs from our method in that ours still follows the setting of standard reversible jump MCMC. Moreover, the computational cost of the birth-death MCMC is far higher than that of reversible jump MCMC. The relationship between these two has also been recently studied in (Cappé, Robert, & Rydén 2003).

Appendix

We assume independence between $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ given all model parameters, and between $l_1^{-1}, \ldots, l_q^{-1}$ and σ^{-2} given the hyperparameter τ. For convenience, we denote $\mathbf{l^{-1}} = \{l_1^{-1}, \ldots, l_q^{-1}\}$ and $\boldsymbol{\theta}$ =$\{\mathbf{l}^{-1}$, τ, $\sigma^{-2}\}$. The joint distribution of the data and parameters is:

$$\begin{aligned}
p(\mathcal{D}, \boldsymbol{\theta}) &= \Big\{ \prod_{i=1}^{N} p(\mathbf{x}_i | \mathbf{l}^{-1}, \sigma^{-2}) \Big\} \\
&\quad \times \Big\{ \prod_{j=1}^{q} p(l_j^{-1} | \tau) \Big\} \times p(\sigma^{-2}|\tau) p(\tau) \\
&= (2\pi)^{-\frac{Nd}{2}} \prod_{j=1}^{q} l_j^{-\frac{N}{2}} \sigma^{-N(d-q)} \\
&\quad \times e^{-\frac{N}{2} \left(\sum_{j=1}^{q} l_j^{-1} g_j + \sigma^{-2} \sum_{j=q+1}^{d} g_j \right)} \\
&\quad \times \Big\{ \prod_{j=1}^{q} p(l_j^{-1} | \tau) \Big\} \times p(\sigma^{-2}|\tau) p(\tau).
\end{aligned}$$

Then, the full conditionals for l_j^{-1}, σ^{-2} and τ are

$$\begin{aligned}
l_j^{-1} | \cdots &\sim l_j^{-\frac{N}{2}} e^{-\frac{N}{2} l_j^{-1} g_j} p(l_j^{-1}|\tau) \\
&\sim \Gamma\Big(\frac{N}{2} + r, \frac{N g_j}{2} + \tau \Big), \\
\sigma^{-2} | \cdots &\sim \sigma^{-N(d-q)} e^{-\frac{N\sigma^{-2}}{2} \sum_{j=q+1}^{d} g_j} p(\sigma^{-2}|\tau) \\
&\sim \Gamma\Big(\frac{N(d-q)}{2} + r, \frac{N \sum_{j=q+1}^{d} g_j}{2} + \tau \Big), \\
\tau | \cdots &\sim \Big\{ \prod_{j=1}^{q} p(l_j^{-1}|\tau) \Big\} \times p(\sigma^{-2}|\tau) p(\tau) \\
&\sim \Gamma\Big((q+1) r + \alpha, \sum_{j=1}^{q} l_j^{-1} + \sigma^{-2} + \eta \Big),
\end{aligned}$$

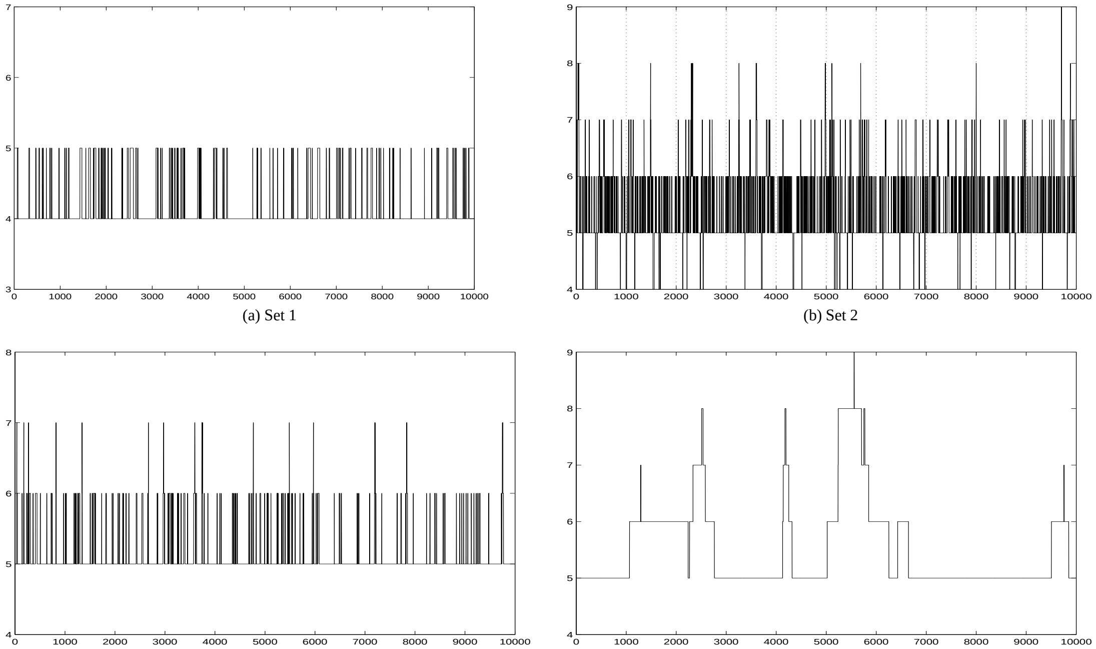
respectively.

References

Andrieu, C.; Djurié, P. M.; and Doucet, A. 2001. Model selection by MCMC computation. *Signal Processing* 81:19–37.

Table 2: Posterior probabilities for different numbers of principal components (q's).

Data set	Number of points	Dimensionality	$p_i \equiv p(q = i \mid \mathcal{D})$
Set 2	100	10	$p_4 = 0.0231$, $\mathbf{p_5 = 0.6830}$, $p_6 = 0.2329$, $p_7 = 0.0566$, $p_8 = 0.0043$, $p_9 = 0.0001$
Set 3	100	10	$\mathbf{p_5 = 0.8907}$, $p_6 = 0.1023$, $p_7 = 0.0070$
Set 4	10000	15	$\mathbf{p_5 = 0.6346}$, $p_6 = 0.2681$, $p_7 = 0.0436$, $p_8 = 0.0532$, $p_8 = 0.0005$



(a) Set 1

(b) Set 2

(c) Set 3

(d) Set 4

Figure 2: Number of components vs. number of sweeps after the burn-in period of 10,000 sweeps.

Bishop, C. M. 1999a. Bayesian PCA. In *Advances in Neural Information Processing Systems 11*, volume 11, 382–388.

Bishop, C. M. 1999b. Variational principal components. In *Proceedings of the International Conference on Artificial Neural Networks*, volume 1, 509–514.

Cappé, O.; Robert, C. P.; and Rydén, T. 2003. Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society, B* 65:679–700.

Gilks, W. R.; Richardson, S.; and Spiegelhalter, D. J. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732.

Holmes, C. C., and Mallick, B. K. 1998. Bayesian radial basis functions of variable dimension. *Neural Computation* 10:1217–1233.

Minka, T. P. 2001. Automatic choice of dimensionality for PCA. In *Advances in Neural Information Processing Systems 13*.

Neal, R. M. 1996. *Bayesian Learning for Neural Networks*. New York: Springer-Verlag.

Richardson, S., and Green, P. J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B* 59:731–792.

Robert, C. P.; Rydén, T.; and Titterington, D. M. 2000. Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society Series B* 62:57–75.

Roberts, S. J.; Holmes, C.; and Denison, D. 2001. Minimum entropy data partitioning using reversible jump Markov chain Monte Carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(8):909–915.

Stephens, M. 2000. Bayesian analysis of mixtures with an unknown number of components — an alternative to reversible jump methods. *Annals of Statistics* 28:40–74.

Tipping, M. E., and Bishop, C. M. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B* 61(3):611–622.