BAYESIAN ESTIMATION OF THE NUMBER OF PRINCIPAL COMPONENTS

Abd-Krim Seghouane¹ and Andrzej Cichocki²

¹ National ICT Australia, Canberra research laboratory and, Research School of Information Sciences and Engineering, The Australian National University, Locked Bag 8001, Canberra ACT 2601, Australia
Phone: +61 2 6125 8621, Fax: +61 2 6125 8660, E-mail: Abd-krim.seghouane@nicta.com.au

² Brain Science Institute, RIKEN Laboratory for Advanced, Brain Signal Processing, Wako-shi, Saitama, 351-0198, Japan
Phone: +81-48-467-9668, Fax: +81-48-467-9686, E-mail: cia@brain.riken.jp

ABSTRACT

Recently, the technique of principal component analysis (PCA) has been expressed as the maximum likelihood solution for a generative latent variable model. A central issue in PCA is choosing the number of principal components to retain. This can be considered as a problem of model selection. In this paper, the probabilistic reformulation of PCA is used as a basis for a Bayesian approach of PCA to derive a model selection criterion for determining the true dimensionality of data. The proposed criterion is similar to the Bayesian Information Criterion, BIC, with a particular goodness of fit term and it is consistent. A simulation example that illustrates its performance for the determination of the number of principal components to be retained is presented.

1. INTRODUCTION

Principal component analysis (PCA) [1] is a well established technique for data analysis and processing. It has been successfully applied in a number of areas from which one can quote; image processing, data visualization and pattern recognition. The general motivation for PCA and the shared root of all its application areas is dimension reduction. Indeed, PCA decomposes high dimensional data into a low dimensional subspace component and a noise component. Modelling complexity in data using a linear projection is an attractive paradigm offering both computational and algorithmic advantages along with increased ease of interpretability. However, this technique of dimension reduction can not be completely satisfactory without a procedure for choosing the number of principal components to be retained.

The choice of the number of components to retain is a problem of model selection [2]. Underestimation of this number will discard valuable information and results in biased estimation of the true dimensionality of data. Overestimation results in a large number of spurious components due to underconstrained estimation and a factorization that will overfit the data.

Two main approaches have been investigated to address the problem of model order determination. One can design an hypothesis testing procedure or develop a model selection criterion. Model selection criteria are often preferred due to their simplicity of application. One only has to evaluate two simple terms that trade off data fitting and model’s complexity. However, the development of a model selection criterion for estimating the number of principal components to be retained requires a probabilistic formulation of PCA.

Recently in [3], it has been shown that a specific form of Gaussian latent variable (where the latent variables offer a more parsimonious description of the data) which is closely related to statistical factor analysis has the property that its maximum likelihood solution extracts the principal subspace of the observed data set. This probabilistic reformulation of PCA permits many extensions. For example, this model has been used in nonlinear image modelling [15]. In this paper, we use it as the basis for a Bayesian formulation of PCA. The issue of model complexity can be handled naturally within a Bayesian paradigm [6][4][5]. Therefore, based on the Bayesian formulation of PCA, we develop a model selection criterion for estimating the true dimensionality of the observed data set (or the number of principal components to retain). This reformulation of PCA associated with integration over the Steifel manifold has also been used in [9] to estimate dimensionality.

In the next section we review PCA and probabilistic PCA. In section 3, the Bayesian formulation of PCA is introduced and the criterion derived. Its consistency is also discussed. A simulation example is presented in section 4 and concluding remarks are given in section 5.

2. REVIEW OF PCA AND PROBABILISTIC PCA

2.1 Description of PCA

Consider a data set of d-dimensional observation vectors \( D = \{t_1, ..., t_N\} \). Derivation of PCA is obtained by first computing the sample covariance matrix \( S = N^{-1} \sum_{i=1}^{N} (t_i - \bar{t})(t_i - \bar{t})^T \) where \( \bar{t} = N^{-1} \sum_{i=1}^{N} t_i \) is the data sample mean, and second by finding the eigenvectors \( u_i \) and eigenvalues \( \lambda_i \) such that
The $q$ principal axes (where $q < d$ for parsimonious representation) $U = (u_1, ..., u_q)$ are orthonormal onto which the retained variance under projection is maximal [7]. It can be shown that they correspond to the eigenvectors associated to the largest eigenvalues. The vector $x_i = U^T (t_i - \bar{t})$ is thus a $q$-dimensional reduced representation of the observed vector $t_i$ and the covariance matrix $\sum_{i=1}^{N} x_i x_i^T / N$ is diagonal with elements $(\lambda_1, ..., \lambda_q)$. An important property is that PCA corresponds to the linear projection for which the sum of squares reconstruction error $\sum_{i=1}^{N} (t_i - \hat{t}_i)^T (t_i - \hat{t}_i)$ is minimized; $\hat{t}_i = U x_i + \bar{t}$ define the linear optimal reconstruction of $t_i$.

A significant limitation of PCA is the absence of an associated probabilistic model for the observed data.

### 2.2 Description of probabilistic PCA

Following [3], PCA can be reformulated as the maximum likelihood solution of a linear latent variable model that relates a $d$-dimensional observation vector $t$ to a corresponding $q$-dimensional vector of latent variable $x$. For parsimony purposes, $q < d$. This model is related to factor analysis and written as

$$t = W x + \mu + \epsilon,$$

where $W$ is a $d \times q$ matrix that relates the two sets of variables, $\mu$ is a $d$-dimensional vector, the latent variables are defined to be independent and Gaussian with unit variance, so $p(x) = N(0, I_q)$ and the noise $\epsilon$ is zero mean Gaussian with covariance matrix $\sigma^2_d I_d$. The difference with factor analysis is the covariance matrix of $\epsilon$, which here, is not a general diagonal matrix.

Based on this model, PCA can be expressed as the estimation of the basis vectors $W$ and the noise variance $\sigma^2_d$ that maximize the likelihood of the observed data vectors $D = (t_1, ..., t_N)$. Under model (1), the probability distribution of the observed variable $t$ given $x$ is $N(Wx + \mu, \sigma^2_d I_d)$. The marginal distribution of the observed variable is then given by

$$p(t) = \int p(t|x)p(x)dx = N(\mu, C),$$

where the observation covariance matrix $C = WW^T + \sigma^2_d I_d$.

Under this model, the probability of the observed data set is

$$p(D|W, \mu, \sigma^2) = (2\pi)^{-Nd/2}|C|^{-N/2} \exp \left\{ -\frac{1}{2} tr(C^{-1}S) \right\},$$

where

$$S = \frac{1}{N} \sum_{i=1}^{N} (t_i - \hat{\mu})(t_i - \hat{\mu})^T = \frac{1}{N} \sum_{i=1}^{N} y_i y_i^T$$

is the sample covariance matrix of the observed data and $\hat{\mu}$ is the maximum likelihood estimate of $\mu$.

The log-likelihood is therefore given by

$$L = -\frac{N}{2} \left\{ d \ln(2\pi) + \ln|C| + tr(C^{-1}S) \right\}. $$

The maximum likelihood estimate of the parameter $\mu$ is the sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} t_i,$$

As shown in [3], the maximum likelihood solution of $W$ is given by

$$\hat{W} = U_q(\Lambda_q - \sigma^2_d I_q)R,$$

where the columns of the $d \times q$ matrix $U_q$ are the eigenvectors of $S$, with corresponding eigenvalues in the $q \times q$ diagonal matrix $\Lambda_q$ and $R$ is an arbitrary $q \times q$ orthogonal rotation matrix. The maximum likelihood estimator of $\sigma^2_d$ is given by

$$\hat{\sigma}^2_d = \frac{1}{d-q} \sum_{i=q+1}^{d} \lambda_i,$$

where $\lambda_i$ is the $i$th eigenvalue of $S$. This represents the average variance lost over the discarded dimension. For this choice of $W$ and $\sigma^2_d$, the covariance matrix $C$ reduces to a diagonal matrix $C = diag(\sigma^2_1, ..., \sigma^2_q)$ with $\sigma^2_1 = ... = \sigma^2_d$. The maximized likelihood can therefore be rewritten as,

$$p(D|\hat{W}, \hat{\mu}, \hat{\sigma}^2) = (2\pi)^{-Nd/2} \left( \prod_{i=1}^{q} \sigma^2_i \right)^{-N/2} \exp \left\{ -\frac{N}{2} \sum_{j=q+1}^{d} \frac{y_j^2}{\sigma^2_j} \right\},$$

where

$$v_j = \frac{1}{N} \sum_{i=1}^{N} y_j^2.$$

This proposed form of the maximized likelihood is more adapted for the derivation of model selection criteria. The $v_j$’s are consistent unbiased estimators of $\sigma^2_j$ for $j = 1, ..., q$.

Probabilistic PCA (PPCA) suggests itself as an adapted tool in a number of problems of data compression and visualization. However, as with PCA, PPCA suffers from the absence of a method for determining the value of the latent space dimensionality $q$. The choice of $q$ corresponds to a problem of model selection. The most convenient way to choose $q$ is by the optimization of model selection criterion that trade off data fitting and model complexity.

In what follows we adopt a Bayesian approach to derive an appropriate model selection criterion for the choice of the dimensionality $q$.

### 3. BAYESIAN ESTIMATION OF THE NUMBER OF PRINCIPAL COMPONENTS

A Bayesian choice of the latent space dimensionality $q \in \{1, ..., d\}$ is obtained by maximizing the probability $p(q|D)$. If $\theta_q$ represents the parameter vector in the probability model of order $q$ for the data, then within a Bayesian paradigm

$$p(q|D) = \int p(q, \theta_q|D)d\theta_q \propto \int p(D|\theta_q)p(\theta_q)d\theta_q.$$  

This expression is valid for models with equal uniform prior. Armed with the probabilistic reformulation of PCA defined in the previous section, a Bayesian approach of PCA is obtained by first introducing a prior distribution $p(\mu, W, \sigma^2)$ over the parameters of the model. Based on the fact that the only information we have is the data set $D$, the most convenient prior in this case is the noninformative prior. Based on the model (3) a criterion has been proposed in [9]. In this paper we use the model (4) introduced earlier for the derivation
of a new criterion for the choice of the dimensionality \( q \). In the model (4) the parameter vector is \( \theta = (\sigma_1, \ldots, \sigma_d) \) with \( \sigma_i = \sigma_d \) for \( i > q \).

Since information about \( \sigma_1, \ldots, \sigma_d \) is not available, we will choose noninformative prior distributions for \( \sigma_1, \ldots, \sigma_d \) using Jeffrey’s invariance theory [10]

\[
p(\sigma) \propto \frac{1}{\sigma},
\]

(6)

Substituting (4) and (6) in (5) gives

\[
p(q|D) \propto \int p(\theta) p(\theta_d) d\theta_d
\]

\[
\propto \int (2\pi)^{-Nd/2} \left( \prod_{i=1}^{q} \sigma_i^{2} \right)^{-N/2} \left( \sigma_d^{2} \right)^{-N(d-q)/2} \exp \left\{ -\frac{N}{2} \sum_{i=1}^{q} v_i \left( \prod_{i=1}^{q} \sigma_i \right)^{-1} \right\} \sigma_d^{-1} \exp \left\{ -\frac{N}{2\sigma_d} \sum_{i=q+1}^{d} v_i \right\} d\sigma_1 \ldots d\sigma_d \sigma_d.
\]

(7)

To evaluate this integral, we use the identity

\[
\int_{0}^{\infty} x^{-(a+1)} e^{-bx} dx = \frac{1}{2} b^{-a/2} \Gamma(a/2),
\]

where \( a > 0, b > 0 \). Using this we have,

\[
\int_\sigma \sigma^{-N} e^{-\frac{Nv}{2\sigma}} \sigma^{-1} d\sigma = \frac{1}{2} \left( \frac{Nv}{2} \right)^{-N/2} \Gamma(N/2)
\]

(8)

and

\[
\int_{\sigma_d} \sigma_d^{-N(d-q)} \sigma_d^{-1} \exp \left\{ -\frac{N}{2\sigma_d} \sum_{i=q+1}^{d} v_i \right\} d\sigma_d = \frac{1}{2} \left( \frac{N\sum_{i=q+1}^{d} v_i}{2} \right)^{-N(d-q)/2} \Gamma \left( \frac{N(d-q)}{2} \right)
\]

(9)

Integrals of this form are described in [12] and are related to the Student-t distribution.

Now, substituting (8) and (9) in (7) gives

\[
p(q|D) \propto \left( \frac{1}{2} \right)^{q+1} \Gamma \left( \frac{N}{2} \right)^{q} \Gamma \left( \frac{N(d-q)}{2} \right) \prod_{i=1}^{q} \left( \frac{Nv_i}{2} \right)^{-\frac{N}{2}} \left( \frac{N\sum_{i=q+1}^{d} v_i}{2} \right)^{-N(d-q)/2}
\]

(10)

Now consider minimizing \(-2 \ln p(q|D)\) as opposed to maximizing \(p(q|D)\). We have

\[
-2 \ln p(q|D) \propto 2(q+1) \ln 2 - 2q \ln \Gamma \left( \frac{N}{2} \right)
\]

\[
+ N \ln \left( \prod_{i=1}^{q} v_i \right) - 2 \ln \left( \Gamma \left( \frac{N(d-q)}{2} \right) \right)
\]

\[
+ N(d-q) \ln \left( \frac{N}{2} \right) + qN \ln \left( \frac{N}{2} \right)
\]

\[
+ N(d-q) \ln \sum_{i=q+1}^{d} v_i.
\]

(11)

To approximate the \( \Gamma \) function, we use the Stirling’s formula

\[
\Gamma(x) = (2\pi)^{1/2} e^{-x} e^{O(N^{-1})}.
\]

Hence,

\[
2 \ln \Gamma \left( \frac{N}{2} \right) = \ln(2\pi) + (N-1) \ln \left( \frac{N}{2} \right) - N + O \left( \frac{1}{N} \right),
\]

and

\[
2 \ln \left( \Gamma \left( \frac{N(d-q)}{2} \right) \right) = \ln(2\pi) + O \left( \frac{1}{N} \right)
\]

\[
+ N(d-q) \left[ \ln \left( \frac{N}{2} \right) \right]
\]

\[
+ \ln \left( \frac{N}{2} \right) + \ln(d-q)
\]

Substituting these two expressions in (11) and removing the \( O(1/N) \) and constant terms gives

\[
-2 \ln p(q|D) / N = \ln \left( \prod_{i=1}^{q} v_i \right) \left( \frac{1}{d-q} \sum_{i=q+1}^{d} v_i \right)^{d-q} + \frac{q}{N} \ln(N).
\]

(12)

The proposed criterion for selecting the number of principal components is identical to the MDL [13] and BIC [4][6] where the first term of the right hand side of equation (12) play the role of the data fitting or goodness of fit term. The form of this term is similar to the likelihood term described in many papers on signal detection [14]. The difference is the presence of the quantities \( v_i \)’s defined above in place of the eigenvalues of the covariance matrix of the observed data.

**Lemma:** The proposed information criterion given by

\[
ICPPA(k) = \ln \left( \prod_{i=1}^{k} v_i \right) \left( \frac{1}{d-k} \sum_{i=k+1}^{d} v_i \right)^{d-k} + \frac{k}{N} \ln(N).
\]

where ICPPA stands for Information Criterion for Principal Component Analysis, is consistent.

**Proof:** Let \( q \) be the correct order. The consistency of the criterion (12) is proved by by showing that in the large sample limit ICPPA(\( k \)) is minimized for \( k = q \).

Case \( k < q \), it follows from (12) that

\[
ICPPA(q) - ICPPA(k) = \ln \left( \frac{\prod v_{i=k+1}^{q} v_i}{\frac{1}{q-k} \sum_{i=k+1}^{q} v_i (q-k)} \right)
\]

\[
+ \ln \left( \frac{1}{q-k} \sum_{i=k+1}^{q} v_i (q-k) \frac{1}{d-k} \sum_{i=q+1}^{d} v_i (d-q) \right) \left( \frac{1}{d-k} \sum_{i=q+1}^{d} v_i (d-k) \right) \]

\[
+ \frac{q-k}{N} \ln(N).
\]

(13)
Using the arithmetic mean, geometrical mean inequality it follows
\[
\frac{1}{q-k} \sum_{i=k+1}^{q} v_i > \prod_{i=k+1}^{q} v_i^{1/(q-k)}.
\]
This implies that the first term of (13) is negative. If we define
\[
A_1 = \frac{1}{q-k} \sum_{i=k+1}^{q} v_i \quad A_2 = \frac{1}{d-q} \sum_{i=q+1}^{d} v_i
\]
\[
\alpha_1 = \frac{q-k}{d-k} \quad \alpha_2 = \frac{d-q}{d-k}
\]
then the second term of (13) can be rewritten as
\[
B = (d-k) \ln \left( \frac{A_1^{\alpha_1} A_2^{\alpha_2}}{\alpha_1 A_1 + \alpha_2 A_2} \right).
\]
By the generalized arithmetic mean geometric mean inequality, we have \(\alpha_1 A_1 + \alpha_2 A_2 \geq A_1^{\alpha_1} A_2^{\alpha_2}\) which implies \(\ln(B) < 0\).

Now, since the last term in (13) goes to zero as the sample size increases it follows that the difference \(\text{ICPPA}(q) - \text{ICPPA}(k)\) is negative and then
\[
\text{ICPPA}(q) < \text{ICPPA}(k) \quad \text{a.s.} \quad N \to \infty.
\]

Taking now \(k > q\), it follows from [14] (Lemma 3.2) and a Taylor expansion of the logarithm that
\[
L_k = \ln \left( \prod_{i=1}^{k} v_i \right) \times \left( \frac{1}{d-k} \sum_{i=k+1}^{d} v_i \right)^{(d-k)}
\]
\[
= O \left( \frac{\ln \ln N}{N} \right) \text{a.s.} \quad (14)
\]

substituting (14) into (12), recalling that as \(N \to \infty\)
\[
\ln N / \ln \ln N \to \infty.
\]
It follows that for \(k > q\)
\[
\text{ICPPA}(k) > \text{ICPPA}(q) \quad \text{a.s.} \quad N \to \infty,
\]
This completes the proof.

4. SIMULATION EXAMPLE

In order to illustrate the performance of the proposed criterion in estimating the number of principal components to retain, a computer simulation of 1000 trials was performed with \(d = 20\), \(N = 15\) and \(N = 100\), exact order \(q_0 = 10\) with \(\sigma_j = 10^j \sigma_d\) for \(j \leq q\). This yielded the following percentages of correct detection in terms of the final SNR, \(FSNR = -10 \log_{10} \sigma_d^2\).

Figure 1 illustrate the performance in term of sample size for FSNR=30 dB. On Table 1, it is seen that the proposed criterion provides good results for small sample data sets independently of the level of the FSNR. These results are better when the sample size increases. On Figure 1, we can observe the influence of the sample size on the performance of the criterion. The performance exceeds 90% of correct selection when the sample size \(\geq 50\).

5. CONCLUSION

The main objective of this paper was to show that a consistent criterion for estimating the number principal components to retain in PCA can be obtained by a blend of Bayesian arguments and PPCA. The final criterion which is obtained by approximating the posterior probability distribution \(p(q|D)\) is more suited for large sample applications. This is due to the number of \(O(1/N)\) terms that have been removed for the derivation of this simplified version which has a less accurate penalty term. For the derivation of this criterion, a more simplified version of the maximized likelihood in comparison to the one used in [9] has been used. The obtained criterion is identical to \(BIC\) if we consider the first term of the right hand side of (12) as the goodness of fit term. Its good performance in estimating the number of principal component to retain has been shown in a simulation example.

REFERENCES


