14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

BAYESIAN ESTIMATION OF THE NUMBER OF PRINCIPAL COMPONENTS

Abd-Krim Seghouane[1] *and Andrzej Cichocki*[2]

[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,
Brian 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 Bayasian 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 $D = (\mathbf{t}_1, ..., \mathbf{t}_N)$. Derivation of PCA is obtained by first computing the sample covariance matrix $S = N^{-1}\sum_{i=1}^{N}(\mathbf{t}_i - \bar{\mathbf{t}})(\mathbf{t}_i - \bar{\mathbf{t}})^T$ where $\bar{\mathbf{t}} = N^{-1}\sum_{i=1}^{N}\mathbf{t}_i$ is the data sample mean, and second by finding the eigenvectors u_i and eigenvalues λ_i such that

National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence Program.

$Su_i = \lambda_i u_i$. The q principal axes (where $q < d$ for parsimonious representation) $U = (u_1, ..., u_q)$ are those 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 $\mathbf{x}_i = U^T(\mathbf{t}_i - \bar{\mathbf{t}})$ is thus a q-dimensional reduced representation of the observed vector $\mathbf{t}_i$ and the covariance matrix $\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T / N$ 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 $\sum_{i=1}^N (\mathbf{t}_i - \widehat{\mathbf{t}}_i)^T (\mathbf{t}_i - \widehat{\mathbf{t}}_i)$ is minimized; $\widehat{\mathbf{t}}_i = U\mathbf{x}_i + \bar{\mathbf{t}}$ define the linear optimal reconstruction of $\mathbf{t}_i$.
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 $\mathbf{t}$ to a corresponding q-dimensional vector of latent variable $\mathbf{x}$. For parsimony purposes, $q < d$. This model is related to factor analysis and written as

$$\mathbf{t} = W\mathbf{x} + \mu + \varepsilon, \tag{1}$$

where W is a $d \times q$ matrix that relates the two sets of variables, μ is a d-dimensional vector, the latent variables are defined to be independent and Gaussian with unit variance, so $p(\mathbf{x}) = N(0, I_q)$ and the noise ε is zero mean Gaussian with covariance matrix $\sigma_d^2 I_d$. The difference with factor analysis is the covariance matrix of ε, 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 σ_d^2 that maximize the likelihood of the observed data vectors $D = (\mathbf{t}_1, ..., \mathbf{t}_N)$.
Under model (1), the probability distribution of the observed variable $\mathbf{t}$ given $\mathbf{x}$ is $N(W\mathbf{x} + \mu, \sigma_d^2 I_d)$. The marginal distribution of the observed variable is then given by

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = N(\mu, C), \tag{2}$$

where the observation covariance matrix $C = WW^T + \sigma_d^2 I_d$.
Under this model, the probability of the observed data set is

$$p(D|W, \mu, \sigma^2) = (2\pi)^{-Nd/2} |C|^{-N/2} \exp\left\{ -\frac{1}{2} tr(C^{-1}S) \right\}, \tag{3}$$

where

$$S = \frac{1}{N} \sum_{i=1}^N (\mathbf{t}_i - \hat{\mu})(\mathbf{t}_i - \hat{\mu})^T = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i \mathbf{y}_i^T$$

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

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{t}_i.$$

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

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

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 Λ_q and R is an arbitrary $q \times q$ orthogonal rotation matrix. The maximum likelihood estimator of σ_d^2 is given by

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

where λ_i is the ith eigenvalue of S. This represents the average variance lost over the discarded dimension.
For this choice of W and σ_d^2, the covariance matrix C reduces to a diagonal matrix $C = diag(\sigma_1^2, ..., \sigma_d^2)$ with $\sigma_{q+1}^2 = ... = \sigma_d^2$. The maximized likelihood can therefore be rewritten as,

$$p(D|\hat{W}, \hat{\mu}, \hat{\sigma}^2) = (2\pi)^{-Nd/2} \left(\prod_{i=1}^q \sigma_i^2 \right)^{-N/2} (\sigma_d^2)^{-N(d-q)/2} \exp\left\{ -\frac{N}{2} \sum_{i=1}^q \frac{v_i}{\sigma_i^2} \right\} \exp\left\{ -\frac{N}{2\sigma_d^2} \sum_{i=q+1}^d v_i \right\} \tag{4}$$

where

$$v_i = \frac{1}{N} \sum_{j=1}^N y_{ji}^2.$$

This proposed form of the maximized likelihood is more adapted for the derivation of model selection criteria. The v_i's are consistent unbiased estimators of σ_j^2 for $j = 1, ..., q$.
Probabilistic PCA (PPCA) suggests it self 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 θ_q represents the parameter vector in the probability model of order q for the data, then within a Bayesian paradigm

$$p(q|D) = \int p(q, \theta_q|D) d\theta_q \propto \int p(D|\theta_q) p(\theta_q) d\theta_q. \tag{5}$$

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 $\theta = (\sigma_1, ..., \sigma_d)$ with $\sigma_i = \sigma_d$ for $i > q$.
Since information about $\sigma_1, ..., \sigma_d$ is not available, we will choose noninformative prior distributions for $\sigma_1, ..., \sigma_d$ using Jeffrey's invariance theory [10]

$$p(\sigma_i) \propto \frac{1}{\sigma_i}. \tag{6}$$

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

$$\begin{aligned} p(q|D) &\propto \int p(D|\theta_q)p(\theta_q)d\theta_q \\ &\propto \int (2\pi)^{-Nd/2}\left(\prod_{i=1}^{q}\sigma_i^2\right)^{-N/2}(\sigma_d^2)^{-N(d-q)/2} \\ &\quad \exp\left\{-\frac{N}{2}\sum_{i=1}^{q}\frac{v_i}{\sigma_i^2}\right\}\left(\prod_{i=1}^{q}\sigma_i^{-1}\right)\sigma_d^{-1} \\ &\quad \exp\left\{-\frac{N}{2\sigma_d^2}\sum_{i=q+1}^{d}v_i\right\}d\sigma_1...d\sigma_q d\sigma_d. \end{aligned} \tag{7}$$

To evaluate this integral, we use the identity

$$\int_0^{+\infty} x^{-(a+1)}e^{-bx^{-2}}dx = \frac{1}{2}b^{-a/2}\Gamma(a/2),$$

where $a > 0$, $b > 0$. Using this we have,

$$\int_\sigma \sigma^{-N}e^{-\frac{Nv}{2\sigma^2}}\sigma^{-1}d\sigma = \frac{1}{2}\left(\frac{Nv}{2}\right)^{-N/2}\Gamma(N/2) \tag{8}$$

and

$$\begin{aligned} &\int_{\sigma_d} \sigma^{-N(d-q)}\sigma_d^{-1}\exp\left\{-\frac{N}{2\sigma_d^2}\sum_{i=q+1}^{d}v_i\right\}d\sigma_d = \\ &\frac{1}{2}\left(\frac{N\sum_{i=q+1}^{d}v_i}{2}\right)^{-N(d-q)/2}\Gamma\left(\frac{N(d-q)}{2}\right) \end{aligned} \tag{9}$$

Integrals of this form are described in [12] and are related to the Student-t distribution.
Now, substituting (8) and (9) in (7) gives

$$\begin{aligned} p(q|D) &\propto \left(\frac{1}{2}\right)^{q+1}\Gamma\left(\frac{N}{2}\right)^{q}\Gamma\left(\frac{N(d-q)}{2}\right) \\ &\quad \prod_{i=1}^{q}\left(\frac{Nv_i}{2}\right)^{-\frac{N}{2}}\left(\frac{N\sum_{i=q+1}^{d}v_i}{2}\right)^{-N(d-q)/2} \end{aligned} \tag{10}$$

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

$$\begin{aligned} -2\ln p(q|D) &\propto 2(q+1)\ln 2 - 2q\ln\Gamma\left(\frac{N}{2}\right) \\ &+ N\ln\left(\prod_{i=1}^{q}v_i\right) - 2\ln\left(\Gamma\left(\frac{N(d-q)}{2}\right)\right) \\ &+ N(d-q)\ln\left(\frac{N}{2}\right) + qN\ln\left(\frac{N}{2}\right) \\ &+ N(d-q)\ln\sum_{i=q+1}^{d}v_i. \end{aligned} \tag{11}$$

To approximate the Γ function, we use the Stirling's formula [11]

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

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

$$\begin{aligned} 2\ln\left(\Gamma\left(\frac{N(d-q)}{2}\right)\right) &= \ln(2\pi) + O\left(\frac{1}{N}\right) \\ &+ N(d-q)\left[\ln\left(\frac{N}{2}\right)\right. \\ &+ \left.\ln(d-q) - 1\right] \\ &- \left[\ln\left(\frac{N}{2}\right) + \ln(d-q)\right] \end{aligned}$$

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

$$\begin{aligned} \frac{-2\ln p(q|D)}{N} &\propto \ln\left(\left(\prod_{i=1}^{q}v_i\right)\times\left(\frac{1}{d-q}\sum_{i=q+1}^{d}v_i\right)^{(d-q)}\right) \\ &+ \frac{q}{N}\ln(N). \end{aligned} \tag{12}$$

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

$$\begin{aligned} ICPPA(k) &= \ln\left(\left(\prod_{i=1}^{k}v_i\right)\times\left(\frac{1}{d-k}\sum_{i=k+1}^{d}v_i\right)^{(d-k)}\right) \\ &+ \frac{k}{N}\ln(N). \end{aligned}$$

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

$$\begin{aligned} ICPPA(q) - ICPPA(k) &= \ln\left(\frac{\left(\prod_{i=k+1}^{q}v_i\right)}{\left(\frac{1}{q-k}\sum_{i=k+1}^{q}v_i\right)^{(q-k)}}\right) \\ &+ \ln\left(\frac{\left(\frac{1}{q-k}\sum_{i=k+1}^{q}v_i\right)^{(q-k)}\left(\frac{1}{d-q}\sum_{i=q+1}^{d}v_i\right)^{(d-q)}}{\left(\frac{1}{d-k}\sum_{i=k+1}^{d}v_i\right)^{(d-k)}}\right) \\ &+ \frac{q-k}{N}\ln N. \end{aligned} \tag{13}$$

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

$$A_1 = \frac{1}{q-k}\sum_{i=k+1}^{q} v_i \qquad A_2 = \frac{1}{d-q}\sum_{i=q+1}^{d} v_i$$
$$\alpha_1 = \frac{q-k}{d-k} \qquad \alpha_2 = \frac{d-q}{d-k}$$

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 $ICPPA(q) - ICPPA(k)$ is negative and then

$$ICPPA(q) < ICPPA(k) \quad a.s. \quad N \to \infty.$$

Taking now $k > q$, it follows from [14] (Lemma 3.2) and a Taylor expansion of the logarithm that

$$\begin{aligned} L_k &= \ln\left(\left(\prod_{i=1}^{k} v_i\right) \times \left(\frac{1}{d-k}\sum_{i=k+1}^{d} v_i\right)^{(d-k)}\right) \\ &= O\left(\frac{\ln\ln N}{N}\right) a.s. \end{aligned} \tag{14}$$

substituting (14) into (12), recalling that as $N \to \infty$ $\ln N / \ln\ln N \to \infty$.
It follows that for $k > q$

$$ICPPA(k) > ICPPA(q) \quad a.s. \quad N \to \infty.$$

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 $\sigma_j = 10\sigma_d$ 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

Table 1: Percentage of correct selection by the criterion for 1000 realizations.

FSNR dB	10	20	30	40	50	60	70
N=15	82	82	80	81	80	82	82
N=100	96	94	96	95	96	96	97

term of sample size for FSNR=30 dB. On Table 1, it is seen

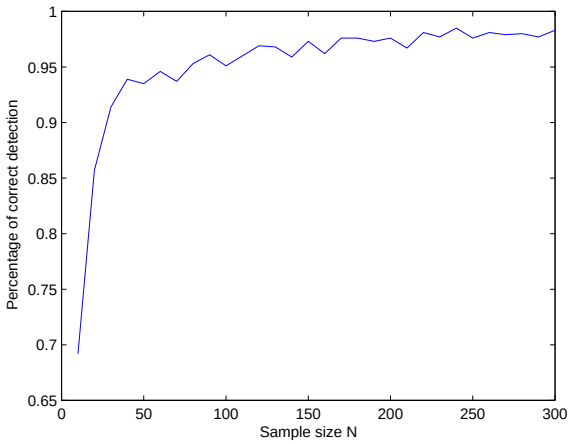

Figure 1: Percentage of correct selection for different sample sizes, FSNR=30 dB.

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 ≥ 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

[1] I. T. Jolliffe, *Principal Component Analysis*, New York, Springer-Verlag, 1986.

[2] H. Linhart and W. Zucchini, *Model Selection*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, 1986.

[3] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Jour. Roy. Stat. Soc., S. B*, Vol. 63, pp. 611-622, 1999.

[4] H. Akaike, "On entropy maximization principle," *Proc. Symp. on Applications of Statistics*, 1977.

[5] L. Wasserman, " Bayesian Model Selection and Model Averaging," *Technical Report, No.666*, Carnegie Mellon

University Department of Statistics, 1997. http://www.stat.cmu.edu/cmu-stats/tr/tr666/tr666.html

[6] G. Schwartz, "Estimating the dimension of a model," *Annals of Statictics*, Vol. 6, pp. 461-464, 1978.

[7] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.* Vol. 24, pp. 417-441, 1933.

[8] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Lond. Edinb. Dub. Phil. Mag. J. Sci.*, 6th ser., 2, pp. 559-572, 1901.

[9] T. Minka, "Automatic Choice of Dimensionality for PCA," *In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.)*, Advances in Neural Information Processing systems, Vol. 13, pp. 598-604, 2001. http://research.microsoft.com/minka/papers/ and http://research.microsoft.com/minka/statlearn/demo/.

[10] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, Reading, MA:Addison-Wesley, 1973.

[11] L. V. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1979.

[12] D. S. Sivia, *Data Analysis: A Bayesian Tutorial* , Oxford University Press, USA, 1996.

[13] J. Rissanen, "Modeling by shortest data description," *Automatica*, Vol. 14, pp. 465-471, 1978.

[14] L. C. Zhao, P. R. Krishnaiah and Z. D. Bai, "On detection of the number of signals in presence of white noise," *Jour. Mutli. Anal.*, Vol. 20, pp. 1-25, 1986.

[15] C. M. Bishop and J. Winn, "Non-linear Bayesian image modelling," *In Proceedings Sixth European Conference on Computer Vision, Dublin*, Vol. 1, pp. 3-17, 2000.